# Analysis of Variance, Design, and Regression: Applied Statistical Methods

Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico

**To Mark, Karl, and John**

It's been great fun.

# Contents

# Preface

This book examines the application of basic statistical methods: primarily analysis of variance and regression but with some discussion of count data. It is directed primarily towards Masters degree students in statistics studying analysis of variance, design of experiments, and regression analysis. I have found that the Masters level regression course is often popular with students outside of statistics. These students are often weaker mathematically and the book caters to that fact while continuing to give a complete matrix formulation of regression.

The book is complete enough to be used as a second course for upper division and beginning graduate students in statistics and for graduate students in other disciplines. To do this, one must be selective in the material covered, but the more theoretical material appropriate only for Statistics Masters students is generally isolated in separate subsections and, less often, in separate sections.

For a Masters level course in analysis of variance and design, I have the students review Chapter 2, I present Chapter 3 while simultaneously presenting the examples of Section 4.2, I present Chapters 5 and 6, very briefly review the first five sections of Chapter 7, present Sections 7.11 and 7.12 in detail and then I cover Chapters 9, 10, 11, 12, and 17. Depending on time constraints, I will delete material or add material from Chapter 16.

For a Masters level course in regression analysis, I again have the students review Chapter 2 and I review Chapter 3 with examples from Section 4.2. I then present Chapters 7, 13, and 14, Appendix A, Chapter 15, Sections 16.1.2, 16.3, 16.5 (along with analysis of covariance), Section 8.7 and finally Chapter 18 . All of this is done in complete detail. If any time remains I like to supplement the course with discussion of response surface methods.

As a second course for upper division and beginning graduate students in statistics and graduate students in other disciplines, I cover the first eight chapters with omission of the more technical material. A follow up course covers the less technical aspects of Chapters 9 through 15 and Appendix A.

I think the book is reasonably encyclopedic. It really contains everything I would like my students to know about applied statistics prior to them taking courses in linear model theory or log-linear models.

I believe that beginning students (even Statistics Masters students) often find statistical procedures to be a morass of vaguely related special techniques. As a result, this book focuses on four connecting themes.

1. Most inferential procedures are based on identifying a (scalar) parameter of interest, estimating that parameter, obtaining the standard error of the estimate, and identifying the appropriate reference distribution. Given these items, the inferential procedures are identical for various parameters.
2. Balanced one-way analysis of variance has a simple, intuitive interpretation in terms of comparing the sample variance of the group means with the mean of the sample variances for each group. All balanced analysis of variance problems are considered in terms of computing sample variances for various group means.
3. Comparing different models provides a structure for examining both balanced and unbalanced analysis of variance problems and for examining regression problems. In some problems the most reasonable analysis is simply to find a succinct model that fits the data well.
4. Checking assumptions is a crucial part of every statistical analysis.

The object of statistical data analysis is to reveal useful structure within the data. In a model-based setting, I know of two ways to do this. One way is to find a *succinct* model for the data. In such a case, the structure revealed is simply the model. The model selection approach is particularly appropriate when the ultimate goal of the analysis is making predictions. This book uses the model selection approach for multiple regression and for general unbalanced multifactor analysis of variance. The other approach to revealing structure is to start with a general model, identify interesting one-dimensional parameters, and perform statistical inferences on these parameters. This parametric approach requires that the general model involve parameters that are easily interpretable. We use the parametric approach for one-way analysis of variance, balanced multifactor analysis of variance, and simple linear regression. In particular, the parametric approach to analysis of variance presented here involves a strong emphasis on examining contrasts, including interaction contrasts. In analyzing two-way tables of counts, we use a partitioning method that is analogous to looking at contrasts.

All statistical models involve assumptions. Checking the validity of these assumptions is crucial because *the models we use are never correct. We hope that our models are good approximations to the true condition of the data and experience indicates that our models often work very well.* Nonetheless, to have faith in our analyses, we need to check the modeling assumptions as best we can. Some assumptions are very difficult to evaluate, e.g., the assumption that observations are statistically independent. For checking other assumptions, a variety of standard tools has been developed. Using these tools is as integral to a proper statistical analysis as is performing an appropriate confidence interval or test. For the most part, using model-checking tools without the aid of a computer is more trouble than most people are willing to tolerate.

My experience indicates that students gain a great deal of insight into balanced analysis of variance by actually doing the computations. The computation of the mean square for treatments in a balanced one-way analysis of variance is trivial on any hand calculator with a variance or standard deviation key. More importantly, the calculation reinforces the fundamental and intuitive idea behind the balanced analysis of variance test, i.e., that a mean square for treatments is just a multiple of the sample variance of the corresponding treatment means. I believe that as long as students find the balanced analysis of variance computations challenging, they should continue to do them by hand (calculator). I think that automated computation should be motivated by boredom rather than bafflement.

In addition to the four primary themes discussed above, there are several other characteristics that I have tried to incorporate into this book.

I have tried to use examples to motivate theory rather than to illustrate theory. Most chapters begin with data and an initial analysis of that data. After illustrating results for the particular data, we go back and examine general models and procedures. I have done this to make the book more palatable to two groups of people: those who only care about theory after seeing that it is useful and those unfortunates who can never bring themselves to care about theory. (The older I get, the more I identify with the first group. As for the other group, I find myself agreeing with W. Edwards Deming that experience without theory teaches nothing.) As mentioned earlier, the theoretical material is generally confined to separate subsections or, less often, separate sections, so it is easy to ignore.

I believe that the *ultimate* goal of all statistical analysis is prediction of observable quantities. I have incorporated predictive inferential procedures where they seemed natural.

The object of most statistics books is to illustrate techniques rather than to analyze data; this book is no exception. Nonetheless, I think we do students a disservice by not showing them a substantial portion of the work necessary to analyze even 'nice' data. To this end, I have tried to consistently examine residual plots, to present alternative analyses using different transformations and case deletions, and to give some final answers in plain English. I have also tried to introduce such material as early as possible. I have included reasonably detailed examinations of a three-factor analysis of variance and of a split plot design with four factors. I have included some examples in which, like real life, the final answers are not 'neat.' While I have tried to introduce statistical ideas as soon as possible, I have tried to keep the mathematics as simple as possible for as long as possible.

For example, matrix formulations are postponed to the last chapter on multiple regression and the last section on unbalanced analysis of variance.

I never use side conditions or normal equations in analysis of variance.

In multiple comparison methods, (weakly) controlling the experimentwise error rate is discussed in terms of first performing an omnibus test for no treatment effects and then choosing a criterion for evaluating individual hypotheses. Most methods considered divide into those that use the omnibus $F$ test, those that use the Studentized range test, and the Bonferroni method, which does not use any omnibus test.

I have tried to be very clear about the fact that experimental designs are set up for arbitrary groups of treatments and that factorial treatment structures are simply an efficient way of defining the treatments in some problems. Thus, the nature of a randomized complete block design does not depend on how the treatments happen to be defined. The analysis always begins with a breakdown of the sum of squares into treatments, blocks, and error. Further analysis of the treatments then focuses on whatever structure happens to be present.

The analysis of covariance chapter includes an extensive discussion of how the covariates must be chosen to maintain a valid experiment. Tukey's one degree of freedom test for nonadditivity is presented as an analysis of covariance test for the need to perform a power transformation rather than as a test for a particular type of interaction.

The chapter on confounding and fractional replication has more discussion of analyzing such data than many other books contain.

Minitab commands are presented for most analyses. Minitab was chosen because I find it the easiest of the common packages to use. However, the real point of including computer commands is to illustrate the kinds of things that one needs to specify for any computer program and the various auxiliary computations that may be necessary for the analysis. The other statistical packages used in creating the book were BMDP, GLIM, and MSUSTAT.

*Acknowledgements*

Many people provided comments that helped in writing this book. My colleagues Ed Bedrick, Aparna Huzurbazar, Wes Johnson, Bert Koopmans, Frank Martin, Tim O'Brien, and Cliff Qualls helped a lot. I got numerous valuable comments from my students at the University of New Mexico. Marjorie Bond, Matt Cooney, Jeff S. Davis, Barbara Evans, Mike Fugate, Jan Mines, and Jim Shields stand out in this regard. The book had several anonymous reviewers, some of whom made excellent suggestions.

I would like to thank Martin Gilchrist and Springer-Verlag for permission to reproduce Example 7.6.1 from *Plane Answers to Complex Questions: The Theory of Linear Models*. I also thank the Biometrika Trustees for permission to use the tables in Appendix B.5. Professor John Deely and the University of Canterbury in New Zealand were kind enough to support completion of the book during my sabbatical there.

Now my only question is what to do with the chapters on quality control, $p^n$ factorials, and response surfaces that ended up on the cutting room floor.

Ronald Christensen
Albuquerque, New Mexico
February 1996

BMDP Statistical Software is located at 1440 Sepulveda Boulevard, Los Angeles, CA 90025, telephone: (213) 479-7799

MINITAB is a registered trademark of Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801, telephone: (814) 238-3280, telex: 881612.

MSUSTAT is marketed by the Research and Development Institute Inc., Montana State University, Bozeman, MT 59717-0002, Attn: R.E. Lund.

Chapter 1

# Introduction

In this chapter we introduce basic ideas of probability and some related mathematical concepts that are used in statistics. Values to be analyzed statistically are generally thought of as random variables; these are numbers that result from random events. The mean (average) value of a population is defined in terms of the expected value of a random variable. The variance is introduced as a measure of the variability in a random variable (population). We also introduce some special distributions (populations) that are useful in modeling statistical data. The purpose of this chapter is to introduce these ideas, so they can be used in analyzing data and in discussing statistical models.

In writing statistical models, we often use symbols from the Greek alphabet. A table of these symbols is provided in Appendix B.6.

Rumor has it that there are some students studying statistics who have an aversion to mathematics. Such people might be wise to focus on the concepts of this chapter and not let themselves get bogged down in the details. The details are given to provide a more complete introduction for those students who are not math averse.

## 1.1 Probability

Probabilities are numbers between zero and one that are used to explain random phenomena. We are all familiar with simple probability models. Flip a standard coin; the probability of heads is 1/2. Roll a die; the probability of getting a three is 1/6. Select a card from a well-shuffled deck; the probability of getting the queen of spades is 1/52 (assuming there are no jokers). One way to view probability models that many people find intuitive is in terms of random sampling from a fixed population. For example, the 52 cards form a fixed population and picking a card from a well-shuffled deck is a means of randomly selecting one element of the population. While we will exploit this idea of sampling from fixed populations, we should also note its limitations. For example, blood pressure is a very useful medical indicator, but even with a fixed population of people it would be very difficult to define a useful population of blood pressures. Blood pressure depends on the time of day, recent diet, current emotional state, the technique of the person taking the reading, and many other factors. Thinking about populations is very useful, but the concept can be very limiting both practically and mathematically. For measurements such as blood pressures and heights, there are difficulties in even specifying populations mathematically.

For mathematical reasons, probabilities are defined not on particular outcomes but on sets of outcomes (events). This is done so that continuous measurements can be dealt with. It seems much more natural to define probabilities on outcomes as we did in the previous paragraph, but consider some of the problems with doing that. For example, consider the problem of measuring the height of a corpse being kept in a morgue under controlled conditions. The only reason for getting morbid here is to have some hope of defining what the height is. Living people, to some extent, stretch and contract, so a height is a nebulous thing. But even given that someone has a fixed height, we can never know what it is. When someone's height is measured as 177.8 centimeters (5 feet 10 inches), their height is not really 177.8 centimeters, but (hopefully) somewhere between 177.75 and 177.85 centimeters. There is really no chance that anyone's height is *exactly* 177.8 cm, or exactly 177.8001 cm,

or exactly 177.800000001 cm, or exactly $56.5955\pi$ cm, or exactly $(76\sqrt{5}+4.5\sqrt{3})$ cm. In any neighborhood of 177.8, there are more numerical values than one could even imagine counting. The height should be somewhere in the neighborhood, but it won't be the particular value 177.8. The point is simply that trying to specify all the possible heights and their probabilities is a hopeless exercise. It simply cannot be done.

Even though individual heights cannot be measured exactly, when looking at a population of heights they follow certain patterns. There are not too many people over 8 feet (244 cm) tall. There are lots of males between 175.3 cm and 177.8 cm ($5'9''$ and $5'10''$). With continuous values, each possible outcome has no chance of occurring, but outcomes do occur and occur with regularity. If probabilities are defined for sets instead of outcomes, these regularities can be reproduced mathematically. Nonetheless, initially the best way to learn about probabilities is to think about outcomes and their probabilities.

There are five key facts about probabilities:

1. Probabilities are between 0 and 1.
2. Something that happens with probability 1 is a sure thing.
3. If something has no chance of occurring, it has probability 0.
4. If something occurs with probability, say, .25, the probability that it will not occur is $1-.25=.75$.
5. If two events are mutually exclusive, i.e., if they cannot possibly happen at the same time, then the probability that either of them occurs is just the sum of their individual probabilities.

Individual outcomes are always mutually exclusive, e.g., you cannot flip a coin and get both heads and tails, so probabilities for outcomes can always be added together. Just to be totally correct, I should mention one other point. It may sound silly, but we need to assume that *something* occurring is always a sure thing. If we flip a coin, we must get either heads or tails with probability 1. We could even allow for the coin landing on its edge as long as the probabilities for all the outcomes add up to 1.

EXAMPLE 1.1.1. Consider the nine outcomes that are all combinations of three heights, tall (T), medium (M), short (S) and three eye colors, blue (Bl), brown (Br) and green (G). The combinations are displayed below.

Height–eye color combinations

| | | Eye color | | |
|---|---|---|---|---|
| | | Blue | Brown | Green |
| Height | Tall | T, Bl | T, Br | T, G |
| | Medium | M, Bl | M, Br | M, G |
| | Short | S, Bl | S, Br | S, G |

The set of all outcomes is

$$\{(\mathrm{T},\mathrm{Bl}),(\mathrm{T},\mathrm{Br}),(\mathrm{T},\mathrm{G}),(\mathrm{M},\mathrm{Bl}),(\mathrm{M},\mathrm{Br}),(\mathrm{M},\mathrm{G}),(\mathrm{S},\mathrm{Bl}),(\mathrm{S},\mathrm{Br}),(\mathrm{S},\mathrm{G})\}.$$

The event that someone is tall consists of the three pairs in the first row of the table, i.e.,

$$\{\mathrm{T}\}=\{(\mathrm{T},\mathrm{Bl}),(\mathrm{T},\mathrm{Br}),(\mathrm{T},\mathrm{G})\}.$$

This is the union of the three outcomes $(\mathrm{T},\mathrm{Bl})$, $(\mathrm{T},\mathrm{Br})$, and $(\mathrm{T},\mathrm{G})$. Similarly, the set of people with blue eyes is obtained from the first column of the table; it is the union of $(\mathrm{T},\mathrm{Bl})$, $(\mathrm{M},\mathrm{Bl})$, and $(\mathrm{S},\mathrm{Bl})$ and can be written

$$\{\mathrm{Bl}\}=\{(\mathrm{T},\mathrm{Bl}),(\mathrm{M},\mathrm{Bl}),(\mathrm{S},\mathrm{Bl})\}.$$

If we know that $\{\mathrm{T}\}$ *and* $\{\mathrm{Bl}\}$ both occur, there is only one possible outcome, (T, Bl).

Table 1.1: *Height–eye color probabilities*

| | | Eye color | | |
|---|---|---|---|---|
| | | Blue | Brown | Green |
| Height | Tall | .12 | .15 | .03 |
| | Medium | .22 | .34 | .04 |
| | Short | .06 | .01 | .03 |

The event that $\{\mathrm{T}\}$ *or* $\{\mathrm{Bl}\}$ occurs consists of all outcomes in either the first row or the first column of the table, i.e.,

$$\{(\mathrm{T},\mathrm{Bl}),(\mathrm{T},\mathrm{Br}),(\mathrm{T},\mathrm{G}),(\mathrm{M},\mathrm{Bl}),(\mathrm{S},\mathrm{Bl})\}. \qquad \square$$

EXAMPLE 1.1.2. Table 1.1 contains probabilities for the nine outcomes that are combinations of height and eye color from Example 1.1.1.

Note that each of the nine numbers is between 0 and 1 and that the sum of all nine equals 1. The probability of blue eyes is

$$\begin{aligned}\Pr(\mathrm{Bl}) &= \Pr[(\mathrm{T},\mathrm{Bl}),(\mathrm{M},\mathrm{Bl}),(\mathrm{S},\mathrm{Bl})]\\ &= \Pr(\mathrm{T},\mathrm{Bl})+\Pr(\mathrm{M},\mathrm{Bl})+\Pr(\mathrm{S},\mathrm{Bl})\\ &= .12+.22+.06\\ &= .4.\end{aligned}$$

Similarly, $\Pr(\mathrm{Br}) = .5$ and $\Pr(\mathrm{G}) = .1$. The probability of not having blue eyes is

$$\begin{aligned}\Pr(\text{not Bl}) &= 1-\Pr(\mathrm{Bl})\\ &= 1-.4\\ &= .6.\end{aligned}$$

Note also that $\Pr(\text{not Bl}) = \Pr(\mathrm{Br})+\Pr(\mathrm{G})$.

The (*marginal*) probabilities for the various heights are:

$$\Pr(\mathrm{T}) = .3, \quad \Pr(\mathrm{M}) = .6, \quad \Pr(\mathrm{S}) = .1. \qquad \square$$

Even if there are a countable (but infinite) number of possible outcomes, one can still define a probability by defining the probabilities for each outcome. It is only for measurement data that one really needs to define probabilities on sets.

Two random events are said to be independent if knowing that one of them occurs provides no information about the probability that the other event will occur. Formally, two events $A$ and $B$ are *independent* if

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B).$$

Thus the probability that *both* events $A$ and $B$ occur is just the product of the individual probabilities that $A$ occurs and that $B$ occurs. As we will begin to see in the next section, independence plays an important role is statistics.

EXAMPLE 1.1.3. Using the probabilities of Table 1.1 and the computations of Example 1.1.2, the events tall and brown eyes are independent because

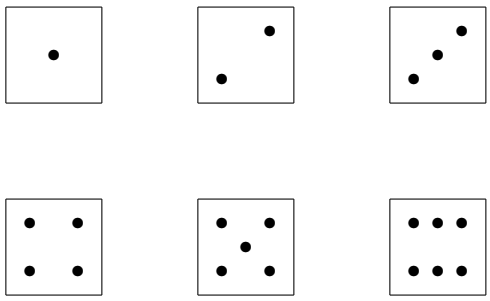
$$\Pr(\text{tall and brown}) = \Pr(\mathrm{T},\mathrm{Br}) = .15 = (.3)(.5) = \Pr(\mathrm{T})\times\Pr(\mathrm{Br}).$$

On the other hand, medium height and blue eyes are *not* independent because

$$\Pr(\text{medium and blue}) = \Pr(\mathrm{M},\mathrm{Bl}) = .22 \neq (.6)(.4) = \Pr(\mathrm{M})\times\Pr(\mathrm{Bl}). \qquad \square$$

## 1.2 Random variables and expectations

A *random variable* is simply a function that relates outcomes with numbers. The key point is that any probability associated with the outcomes induces a probability on the numbers. The numbers and their associated probabilities can then be manipulated mathematically. Perhaps the most common and intuitive example of a random variable is rolling a die. The outcome is that a face of the die with a certain number of spots ends up on top. These can be pictured as



Without even thinking about it, we define a random variable that transforms these six faces into the numbers 1, 2, 3, 4, 5, 6.

In statistics we think of observations as random variables. These are often some number associated with a randomly selected member of a population. For example, one random variable is the height of a person who is to be randomly selected from among University of New Mexico students. (A random selection gives the same probability to every individual in the population. This random variable presumes that we have well-defined methods for measuring height and defining UNM students.) Rather than measuring height, we could define a different random variable by giving the person a score of 1 if that person is female and 0 if the person is male. We can also perform mathematical operations on random variables to yield new random variables. Suppose we plan to select a random sample of 10 students, then we would have 10 random variables with female and male scores. The sum of these random variables is another random variable that tells us the (random) number of females in the sample. Similarly, we would have 10 random variables for heights and we can define a new random variable consisting of the average of the 10 individual height random variables. Some random variables are related in obvious ways. In our example we measure both a height and a sex score on each person. If the sex score variable is a 1 (telling us that the person is female), it suggests that the height may be smaller than we would otherwise suspect. Obviously some female students are taller than some male students, but knowing a person's sex definitely changes our knowledge about their probable height.

We do similar things in tossing a coin.

EXAMPLE 1.2.1. Consider tossing a coin twice. The four outcomes are ordered pairs of heads ($H$) and tails ($T$). The outcomes can be denoted as

$$(H,H) \qquad (H,T) \qquad (T,H) \qquad (T,T)$$

where the outcome of the first toss is the first element of the ordered pair.

The standard probability model has the four outcomes equally probable, i.e., $1/4 = \Pr(H,H) = \Pr(H,T) = \Pr(T,H) = \Pr(T,T)$. Equivalently

| | | Second toss | | |
|---|---|---|---|---|
| | | Heads | Tails | Total |
| First toss | Heads | 1/4 | 1/4 | 1/2 |
| | Tails | 1/4 | 1/4 | 1/2 |
| | Total | 1/2 | 1/2 | 1 |

The probability of heads on each toss is 1/2. The probability of tails is 1/2. We will define two random variables:

$$y_1(r,s) = \begin{cases} 1 & \text{if } r = H \\ 0 & \text{if } r = T \end{cases}$$

$$y_2(r,s) = \begin{cases} 1 & \text{if } s = H \\ 0 & \text{if } s = T \end{cases} .$$

Thus, $y_1$ is 1 if the first toss is heads and 0 otherwise. Similarly, $y_2$ is 1 if the second toss is heads and 0 otherwise.

The event $y_1 = 1$ occurs if and only if we get heads on the first toss. We get heads on the first toss by getting either of the outcome pairs $(H,H)$ or $(H,T)$. In other words, the event $y_1 = 1$ is equivalent to the event $\{(H,H),(H,T)\}$. The probability of $y_1 = 1$ is just the sum of the probabilities of the outcomes in $\{(H,H),(H,T)\}$.

$$\begin{aligned} \Pr(y_1 = 1) &= \Pr(H,H) + \Pr(H,T) \\ &= 1/4 + 1/4 = 1/2. \end{aligned}$$

Similarly,

$$\begin{aligned} \Pr(y_1 = 0) &= \Pr(T,H) + \Pr(T,T) \\ &= 1/2 \\ \Pr(y_2 = 1) &= 1/2 \\ \Pr(y_2 = 0) &= 1/2 . \end{aligned}$$

Now define another random variable,

$$W(r,s) = y_1(r,s) + y_2(r,s) .$$

The random variable $W$ is the total number of heads in two tosses:

$$\begin{aligned} W(H,H) &= 2 \\ W(H,T) &= W(T,H) = 1 \\ W(T,T) &= 0 . \end{aligned}$$

Moreover,

$$\begin{aligned} \Pr(W = 2) &= \Pr(H,H) = 1/4 \\ \Pr(W = 1) &= \Pr(H,T) + \Pr(T,H) = 1/2 \\ \Pr(W = 0) &= \Pr(T,T) = 1/4 . \end{aligned}$$

These three equalities define a probability on the outcomes 0, 1, 2. In working with $W$, we can ignore the original outcomes of head-tail pairs and work only with the new outcomes 0, 1, 2 and their associated probabilities. We can do the same thing for $y_1$ and $y_2$. The probability table given earlier can be rewritten in terms of $y_1$ and $y_2$.

| | | $y_2$ | | |
|---|---|---|---|---|
| | | 1 | 0 | $y_1$ totals |
| $y_1$ | 1 | 1/4 | 1/4 | 1/2 |
| | 0 | 1/4 | 1/4 | 1/2 |
| | $y_2$ totals | 1/2 | 1/2 | 1 |

Note that, for example, $\Pr[(y_1,y_2)=(1,0)] = 1/4$ and $\Pr(y_1=1) = 1/2$. This table shows the *distribution* of the probabilities for $y_1$ and $y_2$ both separately (marginally) and jointly.

□

For any random variable, *a statement of the possible outcomes and their associated probabilities is referred to as the (marginal)* probability distribution *of the random variable. For two or more random variables, a table or other statement of the possible joint outcomes and their associated probabilities is referred to as the* joint probability distribution *of the random variables.*

All of the entries in the center of the distribution table given above for $y_1$ and $y_2$ are independent. For example,

$$\Pr[(y_1,y_2)=(1,0)] \equiv \Pr(y_1=1 \text{ and } y_2=0) = \Pr(y_1=1)\Pr(y_2=0).$$

We therefore say that $y_1$ and $y_2$ are independent. In general, *two random variables $y_1$ and $y_2$ are independent if any event involving only $y_1$ is independent of any event involving only $y_2$.*

Independence is an extremely important concept in statistics. Observations to be analyzed are commonly assumed to be independent. This means that *the random aspect of one observation contains no information about the random aspect of any other observation.* (However, every observation tells us about fixed aspects of the underlying population such as the population center.) *For most purposes in applied statistics, just this intuitive understanding of independence is sufficient.*

### *1.2.1 Expected values and variances*

The *expected value (population mean)* of a random variable is a number characterizing the middle of the distribution. For a random variable $y$ with a discrete distribution (i.e., one having a finite or countable number of outcomes), the expected value is

$$\mathrm{E}(y) \equiv \sum_{\text{all } r} r\Pr(y=r).$$

EXAMPLE 1.2.2. Let $y$ be the result of picking one of the numbers 2, 4, 6, 8 at random. Because the numbers are chosen at random,

$$1/4 = \Pr(y=2) = \Pr(y=4) = \Pr(y=6) = \Pr(y=8).$$

The expected value in this simple example is just the mean (average) of the four possible outcomes.

$$\begin{aligned}\mathrm{E}(y) &= 2\left(\frac{1}{4}\right)+4\left(\frac{1}{4}\right)+6\left(\frac{1}{4}\right)+8\left(\frac{1}{4}\right)\\ &= (2+4+6+8)/4\\ &= 5.\end{aligned}$$

□

EXAMPLE 1.2.3. Five pieces of paper are placed in a hat. The papers have the numbers 2, 4, 6, 6, and 8 written on them. A piece of paper is picked at random. The expected value of the number drawn is the mean of the numbers on the five pieces of paper. Let $y$ be the random variable that

relates a piece of paper to the number on that paper. Each piece of paper has the same probability of being chosen, so, because the number 6 appears twice, the distribution of the random variable $y$ is

$$\begin{aligned} \frac{1}{5} &= \Pr(y=2) = \Pr(y=4) = \Pr(y=8) \\ \\ \frac{2}{5} &= \Pr(y=6)\,. \end{aligned}$$

The expected value is

$$\begin{aligned} \mathrm{E}(y) &= 2\left(\frac{1}{5}\right)+4\left(\frac{1}{5}\right)+6\left(\frac{2}{5}\right)+8\left(\frac{1}{5}\right) \\ &= (2+4+6+6+8)/5 \\ &= 5.2\,. \end{aligned}$$

□

EXAMPLE 1.2.4. Consider the coin tossing random variables $y_1$, $y_2$, and $W$ from Example 1.2.1. Recalling that $y_1$ and $y_2$ have the same distribution,

$$\begin{aligned} \mathrm{E}(y_1) &= 1\left(\frac{1}{2}\right)+0\left(\frac{1}{2}\right) = \frac{1}{2} \\ \mathrm{E}(y_2) &= \frac{1}{2} \\ \mathrm{E}(W) &= 2\left(\frac{1}{4}\right)+1\left(\frac{1}{2}\right)+0\left(\frac{1}{4}\right) = 1\,. \end{aligned}$$

The variable $y_1$ is the number of heads in the first toss of the coin. The two possible values 0 and 1 are equally probable, so the middle of the distribution is $1/2$. $W$ is the number of heads in two tosses; the expected number of heads in two tosses is 1. □

The expected value indicates the middle of a distribution, but does not indicate how spread out (dispersed) a distribution is.

EXAMPLE 1.2.5. Consider three gambles that I will allow you to take. In game $z_1$ you have equal chances of winning 12, 14, 16, or 18 dollars. In game $z_2$ you can again win 12, 14, 16, or 18 dollars, but now the probabilities are .1 that you will win either \$14 or \$16 and .4 that you will win \$12 or \$18. The third game I call $z_3$ and you can win 5, 10, 20, or 25 dollars with equal chances. Being no fool, I require you to pay me \$16 for the privilege of playing any of these games. We can write each game as a random variable.

| | | | | | |
|---|---|---|---|---|---|
| $z_1$ | outcome | 12 | 14 | 16 | 18 |
| | probability | .25 | .25 | .25 | .25 |
| $z_2$ | outcome | 12 | 14 | 16 | 18 |
| | probability | .4 | .1 | .1 | .4 |
| $z_3$ | outcome | 5 | 10 | 20 | 25 |
| | probability | .25 | .25 | .25 | .25 |

I try to be a good casino operator, so none of these games is fair. You have to pay \$16 to play, but you only expect to win \$15. It is easy to see that

$$\mathrm{E}(z_1) = \mathrm{E}(z_2) = \mathrm{E}(z_3) = 15\,.$$

But don't forget that I'm taking a loss on the ice-water I serve to players and I also have to pay for the pictures of my extended family that I've decorated my office with.

Although the games $z_1$, $z_2$, and $z_3$ have the same expected value, the games (random variables) are very different. Game $z_2$ has the same outcomes as $z_1$, but much more of its probability is placed farther from the middle value 15. The extreme observations 12 and 18 are much more probable under $z_2$ than $z_1$. If you currently have \$16, need \$18 for your grandmother's bunion removal, and anything less than \$18 has no value to you, then $z_2$ is obviously a better game for you than $z_1$.

Both $z_1$ and $z_2$ are much more tightly packed around 15 than is $z_3$. If you needed \$25 for the bunion removal, $z_3$ is the game to play because you can win it all in one play with probability .25. In either of the other games you would have to win at least five times to get \$25, a much less likely occurrence. Of course you should realize that the most probable result is that Grandma will have to live with her bunion. You are unlikely to win either \$18 or \$25. While the ethical moral of this example is that a fool and his money are soon parted, the statistical point is that there is more to a random variable than its mean. The variability of random variables is also important. □

The *(population) variance* is a measure of how spread out a distribution is from its expected value. Let $y$ be a random variable having a discrete distribution with $\mathrm{E}(y) = \mu$, then the variance of $y$ is

$$\mathrm{Var}(y) \equiv \sum_{\text{all } r} (r-\mu)^2 \Pr(y = r).$$

This is the average squared distance of the outcomes from the center of the population. More technically, it is the expected squared distance between the outcomes and the mean of the distribution.

EXAMPLE 1.2.6. Using the random variables of Example 1.2.5,

$$\begin{aligned}
\mathrm{Var}(z_1) &= (12-15)^2(.25)+(14-15)^2(.25) \\
&\quad +(16-15)^2(.25)+(18-15)^2(.25) \\
&= 5 \\
\mathrm{Var}(z_2) &= (12-15)^2(.4)+(14-15)^2(.1) \\
&\quad +(16-15)^2(.1)+(18-15)^2(.4) \\
&= 7.4 \\
\mathrm{Var}(z_3) &= (5-15)^2(.25)+(10-15)^2(.25) \\
&\quad +(20-15)^2(.25)+(25-15)^2(.25) \\
&= 62.5
\end{aligned}$$

The increasing variances from $z_1$ through $z_3$ indicate that the random variables are increasingly spread out. However, the value $\mathrm{Var}(z_3) = 62.5$ seems too large to measure the relative variabilities of the three random variables. More on this later. □

EXAMPLE 1.2.7. Consider the coin tossing random variables of Examples 1.2.1 and 1.2.4.

$$\begin{aligned}
\mathrm{Var}(y_1) &= \left(1-\frac{1}{2}\right)^2\frac{1}{2}+\left(0-\frac{1}{2}\right)^2\frac{1}{2} = \frac{1}{4} \\
\mathrm{Var}(y_2) &= \frac{1}{4} \\
\mathrm{Var}(W) &= (2-1)^2\left(\frac{1}{4}\right)+(1-1)^2\left(\frac{1}{2}\right)+(0-1)^2\left(\frac{1}{4}\right) = \frac{1}{2}.
\end{aligned}$$

□

A problem with the variance is that it is measured on the wrong scale. If $y$ is measured in meters,

$\mathrm{Var}(y)$ involves the terms $(r-\mu)^2$; hence it is measured in meters squared. To get things back on the original scale, we consider the *standard deviation* of $y$

$$\text{Std. dev. } (y) \equiv \sqrt{\mathrm{Var}(y)}\,.$$

EXAMPLE 1.2.8. Consider the random variables of Examples 1.2.5 and 1.2.6.

$$\begin{array}{lcll} \text{Std. dev. } (z_1) & = & \sqrt{5} & \doteq 2.236 \\ \text{Std. dev. } (z_2) & = & \sqrt{7.4} & \doteq 2.720 \\ \text{Std. dev. } (z_3) & \equiv & \sqrt{62.5} & \doteq 7.906 \end{array}$$

The standard deviation of $z_3$ is 3 to 4 times larger than the others. From examining the distributions, the standard deviations seem to be more intuitive measures of relative variability than the variances. The variance of $z_3$ is 8.5 to 12.5 times larger than the other variances; these values seem unreasonably inflated. □

Standard deviations and variances are useful as measures of the relative dispersions of different random variables. The actual numbers themselves do not mean much. Moreover, there are other equally good measures of dispersion that can give results that are somewhat inconsistent with these. One reason standard deviations and variances are so widely used is because they are convenient mathematically. In addition, normal (Gaussian) distributions are widely used in applied statistics and are completely characterized by their expected values (means) and variances (or standard deviations). Knowing these two numbers, the mean and variance, one knows everything about a normal distribution.

### *1.2.2 Chebyshev's inequality*

Another place in which the numerical values of standard deviations are useful is in applications of Chebyshev's inequality. Chebyshev's inequality gives a lower bound on the probability that a random variable is within an interval. Chebyshev's inequality is important in quality control work (control charts) and in evaluating prediction intervals.

Let $y$ be a random variable with $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \sigma^2$. Chebyshev's inequality states that for any number $k > 1$,

$$\Pr[\mu - k\sigma < y < \mu + k\sigma] \geq 1 - \frac{1}{k^2}.$$

Thus the probability that $y$ will fall within $k$ standard deviations of $\mu$ is at least $1-(1/k^2)$.

The beauty of Chebyshev's inequality is that it holds for absolutely any random variable $y$. Thus we can always make some statement about the probability that $y$ is in a symmetric interval about $\mu$. In many cases, for particular choices of $y$, the probability of being in the interval can be much greater than $1-k^{-2}$. For example, if $k=3$ and $y$ has a normal distribution as discussed in the next section, the probability of being in the interval is actually .997, whereas Chebyshev's inequality only assures us that the probability is no less than $1-3^{-2} = .889$. However, we know the lower bound of .889 applies regardless of whether $y$ has a normal distribution.

### *1.2.3 Covariances and correlations*

Often we take two (or more) observations on the same member of a population. We might observe the height and weight of a person. We might observe the IQs of a wife and husband. (Here the population consists of married couples.) In such cases we may want a numerical measure of the relationship between the pairs of observations. Data analysis related to these concepts is known as regression analysis and is discussed in Chapters 7, 13, 14, and 15. These ideas are also briefly used for testing normality in Section 2.4.

The *covariance* is a measure of the linear relationship between two random variables. Suppose $y_1$ and $y_2$ are discrete random variables. Let $\mathrm{E}(y_1) = \mu_1$ and $\mathrm{E}(y_2) = \mu_2$. The covariance between $y_1$ and $y_2$ is

$$\mathrm{Cov}(y_1, y_2) \equiv \sum_{\text{all } (r,s)} (r-\mu_1)(s-\mu_2)\mathrm{Pr}(y_1 = r, y_2 = s).$$

Positive covariances arise when relatively large values of $y_1$ tend to occur with relatively large values $y_2$ and small values of $y_1$ tend to occur with small values of $y_2$. On the other hand, negative covariances arise when relatively large values of $y_1$ tend to occur with relatively small values $y_2$ and small values of $y_1$ tend to occur with large values of $y_2$. It is simple to see from the definition that, for example,

$$\mathrm{Var}(y_1) = \mathrm{Cov}(y_1, y_1).$$

In an attempt to get a handle on what the numerical value of the covariance means, it is often rescaled into a *correlation coefficient*.

$$\mathrm{Corr}(y_1, y_2) \equiv \mathrm{Cov}(y_1, y_2) \big/ \sqrt{\mathrm{Var}(y_1)\mathrm{Var}(y_2)}.$$

Positive values of the correlation have the same qualitative meaning as positive values of the covariance, but now a *perfect* increasing linear relationship is indicated by a correlation of 1. Similarly, negative correlations and covariances mean similar things, but a perfect decreasing linear relationship gives a correlation of $-1$. The absence of any linear relationship is indicated by a value of 0.

A perfect linear relationship between $y_1$ and $y_2$ means that an increase of one unit in, say, $y_1$ dictates an exactly proportional change in $y_2$. For example, if we make a series of very accurate temperature measurements on something and simultaneously use one device calibrated in Fahrenheit and one calibrated in Celsius, the pairs of numbers should have an essentially perfect linear relationship.

EXAMPLE 1.2.9. Let $z_1$ and $z_2$ be two random variables defined by the following probability table:

| | | $z_2$ | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | $z_1$ totals |
| $z_1$ | 6 | 0 | 1/3 | 0 | 1/3 |
| | 4 | 1/3 | 0 | 0 | 1/3 |
| | 2 | 0 | 0 | 1/3 | 1/3 |
| | $z_2$ totals | 1/3 | 1/3 | 1/3 | 1 |

Then

$$\begin{aligned} E(z_1) &= 6\left(\frac{1}{3}\right) + 4\left(\frac{1}{3}\right) + 2\left(\frac{1}{3}\right) = 4, \\ E(z_2) &= 0\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right) + 2\left(\frac{1}{3}\right) = 1, \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(z_1) &= (2-4)^2\left(\frac{1}{3}\right) + (4-4)^2\left(\frac{1}{3}\right) + (6-4)^2\left(\frac{1}{3}\right) \\ &= 8/3, \\ \mathrm{Var}(z_2) &= (0-1)^2\left(\frac{1}{3}\right) + (1-1)^2\left(\frac{1}{3}\right) + (2-1)^2\left(\frac{1}{3}\right) \\ &= 2/3, \end{aligned}$$

$$\begin{aligned}\text{Cov}(z_1,z_2) &= (2-4)(0-1)(0)+(2-4)(1-1)(0)+(2-4)(2-1)\left(\frac{1}{3}\right)\\ &\quad +(4-4)(0-1)\left(\frac{1}{3}\right)+(4-4)(1-1)(0)+(4-4)(2-1)(0)\\ &\quad +(6-4)(0-1)(0)+(6-4)(1-1)\left(\frac{1}{3}\right)+(6-4)(2-1)(0)\\ &= -2/3,\end{aligned}$$

$$\begin{aligned}\text{Corr}(z_1,z_2) &= (-2/3)\Big/\sqrt{(8/3)(2/3)}\\ &= -1/2.\end{aligned}$$

This correlation indicates that relatively large $z_1$ values tend to occur with relatively small $z_2$ values. However, the correlation is considerably greater than $-1$, so the linear relationship is less than perfect. Moreover, the correlation measures the linear relationship and *fails to identify the perfect nonlinear relationship* between $z_1$ and $z_2$. If $z_1 = 2$, then $z_2 = 2$. If $z_1 = 4$, then $z_2 = 0$. If $z_1 = 6$, then $z_2 = 1$. If you know one random variable, you know the other, but because the relationship is nonlinear, the correlation is not $\pm 1$. □

EXAMPLE 1.2.10. Consider the coin toss random variables $y_1$ and $y_2$ from Example 1.2.1. We earlier observed that these two random variables are independent. If so, there should be no relationship between them (linear or otherwise). We now show that their covariance is 0.

$$\begin{aligned}\text{Cov}(y_1,y_2) &= \left(0-\frac{1}{2}\right)\left(0-\frac{1}{2}\right)\frac{1}{4}+\left(0-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\frac{1}{4}\\ &\quad +\left(1-\frac{1}{2}\right)\left(0-\frac{1}{2}\right)\frac{1}{4}+\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\frac{1}{4}\\ &= \frac{1}{16}-\frac{1}{16}-\frac{1}{16}+\frac{1}{16}=0.\end{aligned}$$ □

*In general, whenever two random variables are independent, their covariance (and thus their correlation) is 0. However, just because two random variables have 0 covariance* does not *imply that they are independent. Independence has to do with not having any kind of relationship; covariance examines only linear relationships. Random variables with nonlinear relationships can have zero covariance but not be independent.*

### *1.2.4 Rules for expected values and variances*

We now present some extremely useful results that allow us to show that statistical estimates are reasonable and to establish the variability associated with statistical estimates. These results relate to the expected values, variances, and covariances of linear combinations of random variables. A linear combination of random variables is something that only involves multiplying random variables by fixed constants, adding such terms together, and adding a constant.

**Proposition 1.2.11.** Let $y_1$, $y_2$, $y_3$, and $y_4$ be random variables and let $a_1$, $a_2$, $a_3$, and $a_4$ be real numbers.

1. $\text{E}(a_1y_1 + a_2y_2 + a_3) = a_1\text{E}(y_1) + a_2\text{E}(y_2) + a_3$.
2. If $y_1$ and $y_2$ are independent, $\text{Var}(a_1y_1 + a_2y_2 + a_3) = a_1^2\text{Var}(y_1) + a_2^2\text{Var}(y_2)$.
3. $\text{Var}(a_1y_1 + a_2y_2 + a_3) = a_1^2\text{Var}(y_1) + 2a_1a_2\text{Cov}(y_1,y_2) + a_2^2\text{Var}(y_2)$.

4. $\mathrm{Cov}(a_1y_1 + a_2y_2, a_3y_3 + a_4y_4) = a_1a_3\mathrm{Cov}(y_1,y_3) + a_1a_4\mathrm{Cov}(y_1,y_4) + a_2a_3\mathrm{Cov}(y_2,y_3) + a_2a_4\mathrm{Cov}(y_2,y_4)$.

*All of these results generalize to linear combinations involving more than two random variables.*

EXAMPLE 1.2.12. Recall that when independently tossing a coin twice, the total number of heads, $W$, is the sum of $y_1$ and $y_2$, the number of heads on the first and second tosses respectively. We have already seen that $\mathrm{E}(y_1) = \mathrm{E}(y_2) = .5$ and that $\mathrm{E}(W) = 1$. We now illustrate item 1 of the proposition by finding $\mathrm{E}(W)$ again. Since $W = y_1 + y_2$,

$$\mathrm{E}(W) = \mathrm{E}(y_1 + y_2) = \mathrm{E}(y_1) + \mathrm{E}(y_2) = .5 + .5 = 1.$$

We have also seen that $\mathrm{Var}(y_1) = \mathrm{Var}(y_2) = .25$ and that $\mathrm{Var}(W) = .5$. Since the coin tosses are independent, item 2 above gives

$$\mathrm{Var}(W) = \mathrm{Var}(y_1 + y_2) = \mathrm{Var}(y_1) + \mathrm{Var}(y_2) = .25 + .25 = .5\,.$$

The key point is that this is an easier way of finding the expected value and variance of $W$ than using the original definitions. □

We now illustrate the generalizations referred to in Proposition 1.2.11. We begin by looking at the problem of estimating the mean of a population.

EXAMPLE 1.2.13. Let $y_1, y_2, y_3$, and $y_4$ be four random variables each with the same (population) mean $\mu$, i.e., $\mathrm{E}(y_i) = \mu$ for $i = 1,2,3,4$. We can compute the *sample mean* (average) of these, defining

$$\begin{aligned} \bar{y}_\cdot &\equiv \frac{y_1 + y_2 + y_3 + y_4}{4} \\ &= \frac{1}{4}y_1 + \frac{1}{4}y_2 + \frac{1}{4}y_3 + \frac{1}{4}y_4. \end{aligned}$$

The $\cdot$ in the subscript of $\bar{y}_\cdot$ indicates that the sample mean is obtained by summing over the subscripts of the $y_i$s. The $\cdot$ notation is not necessary for this problem but becomes useful in dealing with the analysis of variance problems treated later in the book.

Using item 1 of Proposition 1.2.11 we find that

$$\begin{aligned} \mathrm{E}(\bar{y}_\cdot) &= \mathrm{E}\left(\frac{1}{4}y_1 + \frac{1}{4}y_2 + \frac{1}{4}y_3 + \frac{1}{4}y_4\right) \\ &= \frac{1}{4}\mathrm{E}(y_1) + \frac{1}{4}\mathrm{E}(y_2) + \frac{1}{4}\mathrm{E}(y_3) + \frac{1}{4}\mathrm{E}(y_4) \\ &= \frac{1}{4}\mu + \frac{1}{4}\mu + \frac{1}{4}\mu + \frac{1}{4}\mu \\ &= \mu. \end{aligned}$$

Thus one observation on $\bar{y}_\cdot$ would make a reasonable estimate of $\mu$.

If we also assume that the $y_i$s are independent with the same variance, say, $\sigma^2$, then from item 2 of Proposition 1.2.11

$$\begin{aligned} \mathrm{Var}(\bar{y}_\cdot) &= \mathrm{Var}\left(\frac{1}{4}y_1 + \frac{1}{4}y_2 + \frac{1}{4}y_3 + \frac{1}{4}y_4\right) \\ &= \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_1) + \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_2) \\ &\quad + \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_3) + \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_4) \end{aligned}$$

$$= \left(\frac{1}{4}\right)^2\sigma^2+\left(\frac{1}{4}\right)^2\sigma^2+\left(\frac{1}{4}\right)^2\sigma^2+\left(\frac{1}{4}\right)^2\sigma^2$$
$$= \frac{\sigma^2}{4}.$$

The variance of $\bar{y}_\cdot$ is only one fourth of the variance of an individual observation. Thus the $\bar{y}_\cdot$ observations are more tightly packed around their mean $\mu$ than the $y_i$s are. This indicates that one observation on $\bar{y}_\cdot$ is more likely to be close to $\mu$ than an individual $y_i$. □

These results for $\bar{y}_\cdot$ hold quite generally; they are not restricted to the average of four random variables. *If* $\bar{y}_\cdot = (1/n)(y_1+\cdots+y_n) = \sum_{i=1}^n y_i/n$ *is the sample mean of n independent random variables all with the same population mean* $\mu$ *and population variance* $\sigma^2$,

$$\mathrm{E}(\bar{y}_\cdot) = \mu$$

and

$$\mathrm{Var}(\bar{y}_\cdot) = \frac{\sigma^2}{n}.$$

In fact, proving these general results uses exactly the same ideas as the proofs for a sample of size 4.

As with a sample of size 4, the general results on $\bar{y}_\cdot$ are very important in statistical inference. If we are interested in determining the population mean $\mu$ from future data, the obvious estimate is the average of the individual observations, $\bar{y}_\cdot$. The observations are random, so the estimate $\bar{y}_\cdot$ is also a random variable and the middle of its distribution is $\mathrm{E}(\bar{y}_\cdot) = \mu$, the original population mean. Thus $\bar{y}_\cdot$ is a reasonable estimate of $\mu$. Moreover, $\bar{y}_\cdot$ is a better estimate than any particular observation $y_i$ because $\bar{y}_\cdot$ has a smaller variance, $\sigma^2/n$ as opposed to $\sigma^2$ for $y_i$. With less variability in the estimate, any one observation of $\bar{y}_\cdot$ is more likely to be near its mean $\mu$ than a single observation $y_i$. In practice, we obtain data and compute a sample mean. This constitutes one observation on the random variable $\bar{y}_\cdot$. If our sample mean is to be a good estimate of $\mu$, our one look at $\bar{y}_\cdot$ had better have a good chance of being close to $\mu$. This occurs when the variance of $\bar{y}_\cdot$ is small. Note that the larger the sample size $n$, the smaller is $\sigma^2/n$, the variance of $\bar{y}_\cdot$. We will return to these ideas later.

Generally, we will use item 1 of Proposition 1.2.11 to show that estimates are *unbiased*. In other words, we will show that the expected value of an estimate is what we are trying to estimate. In estimating $\mu$, we have $\mathrm{E}(\bar{y}_\cdot) = \mu$, so $\bar{y}_\cdot$ is an unbiased estimate of $\mu$. All this really does is show that $\bar{y}_\cdot$ is a reasonable estimate of $\mu$. More important than showing unbiasedness is using item 2 to find variances of estimates. Statistical inference depends crucially on having some idea of the variability of an estimate. Item 2 is the primary tool in finding the appropriate variance for different estimates.

## 1.3 Continuous distributions

As discussed in Section 1.1, many things that we would like to measure are, in the strictest sense, not measurable. We cannot find a building's exact height even though we can approximate it *extremely* accurately. This theoretical inability to measure things exactly has little impact on our practical world, but it has a substantial impact on the theory of statistics.

The data in most statistical applications can be viewed as counts of how often some event has occurred or as measurements. Probabilities associated with count data are easy to describe. We discuss some probability models for count data in Sections 1.4 and 1.5. With measurement data, we can never obtain an exact value, so we don't even try. With measurement data, we assign probabilities to intervals. Thus we do not discuss the probability that a person has the height 177.8 cm or 177.8001 cm or $56.5955\pi$ cm, but we do discuss the probability that someone has a height *between* 177.75 cm and 177.85 cm. Typically, we think of doing this in terms of pictures. We associate probabilities with areas under curves. (Mathematically, this involves integral calculus and is discussed in a brief

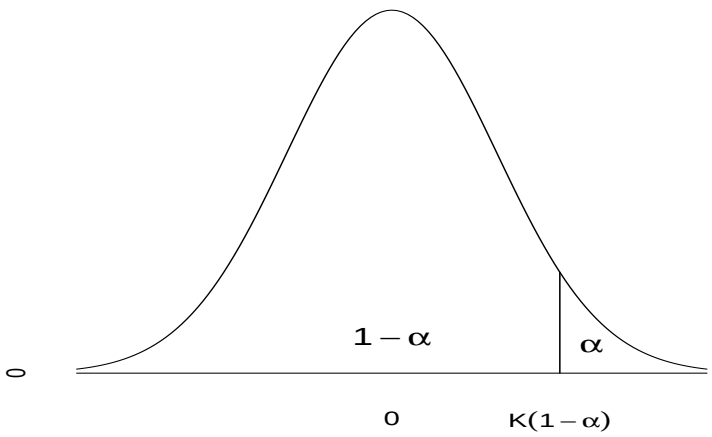



Figure 1.1: *A continuous probability density.*

appendix at the end of the chapter.) Figure 1.1 contains a picture of a continuous probability distribution (*a density*). Probabilities must be between 0 and 1, so the curve must always be nonnegative (to make all areas nonnegative) and the area under the entire curve must be 1.

Figure 1.1 also shows a point $K(1-\alpha)$. This point divides the area under the curve into two parts. The probability of obtaining a number less than $K(1-\alpha)$ is $1-\alpha$, i.e., the area under the curve to the left of $K(1-\alpha)$ is $1-\alpha$ . The probability of obtaining a number greater than $K(1-\alpha)$ is $\alpha$, i.e., the area under the curve to the right of $K(1-\alpha)$. $K(1-\alpha)$ is a particular number, so the probability is 0 that $K(1-\alpha)$ will actually occur. There is no area under a curve associated with any particular point.

Pictures such as Figure 1.1 are often used as models for populations of measurements. With a fixed population of measurements, it is natural to form a histogram, i.e., a bar chart that plots intervals for the measurement against the proportion of individuals that fall into a particular interval. Pictures such as Figure 1.1 can be viewed as approximations to such histograms. The probabilities described by pictures such as Figure 1.1 are those associated with randomly picking an individual from the population. Thus, randomly picking an individual from the population modeled by Figure 1.1 yields a measurement less than $K(1-\alpha)$ with probability $1-\alpha$.

Ideas similar to those discussed in Section 1.2 can be used to define expected values, variances, and covariances for continuous distributions. These extensions involve integral calculus and are discussed in the appendix. In any case, Proposition 1.2.11 continues to apply.

The most commonly used distributional model for measurement data is the *normal* distribution (also called the *Gaussian* distribution). The bell shaped curve in Figure 1.1 is referred to as the standard normal curve. The formula for writing the curve is not too ugly, it is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Here $e$ is the base of natural logarithms. Unfortunately, even with calculus it is very difficult to compute areas under this curve. Finding standard normal probabilities requires a table.

By itself, the standard normal curve has little value in modeling measurements. For one thing, the curve is centered about 0. I don't take many measurements where I think the central value should be 0. To make the normal distribution a useful model, we need to expand the standard normal into

a family of distributions with different centers (expected values) $\mu$ and different spreads (standard deviations) $\sigma$. By appropriate recentering and rescaling of the plot, all of these curves will have the same shape as Figure 1.1.

The standard normal distribution is the special case of a normal with $\mu = 0$ and $\sigma = 1$. The standard normal plays an important role because it is the only normal distribution that we need tabled. (Obviously, we could not table normal distributions for every possible value of $\mu$ and $\sigma$.) Suppose a measurement $y$ has a normal distribution with mean $\mu$, standard deviation $\sigma$, and variance $\sigma^2$. We write this as

$$y \sim N(\mu, \sigma^2).$$

Normal distributions have the property that

$$\frac{y-\mu}{\sigma} \sim N(0,1),$$

cf. Exercise 1.6.2. This standardization process allows us to get by with only the standard normal table for finding probabilities for all normal distributions.

The standard normal distribution is sometimes used in constructing statistical inferences but more often a similar distribution is used. When data are normally distributed, statistical inferences often require something called Student's $t$ distribution. (Student was the pen name of W. S. Gosset.) The $t$ distribution is a family of distributions all of which look roughly like Figure 1.1. They are all symmetric about 0, but they have slightly different amounts of dispersion (spread). The amount of variability in each distribution is determined by a positive integer parameter called the *degrees of freedom*. With only 1 degree of freedom, the mathematical properties of a $t$ distribution are fairly bizarre. (This special case is called a Cauchy distribution.) As the number of degrees of freedom get larger, the $t$ distributions get better behaved and have less variability. As the degrees of freedom gets arbitrarily large, the $t$ distribution approximates the standard normal distribution.

Two other distributions that come up later are the chi-squared distribution ($\chi^2$) and the $F$ distribution. These arise naturally when drawing conclusions about the population variance from data that are normally distributed. Both distributions differ from those just discussed in that both are asymmetric and both are restricted to positive numbers. However, the basic idea of probabilities being areas under curves remains unchanged.

In Section 1.2, we introduced Chebyshev's inequality. Shewhart (1931, p. 177) discusses work by Camp and Meidell that allows us to improve on Chebyshev's inequality for continuous distributions. Once again let $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \sigma^2$. If the density, i.e., the function that defines the curve, is symmetric, unimodal (has only one peak), and always decreases as one moves farther away from the mode, then the inequality can be sharpened to

$$\Pr[\mu - k\sigma < y < \mu + k\sigma] \geq 1 - \frac{1}{(2.25)k^2}.$$

As discussed in the previous section, with $y$ normal and $k = 3$, the true probability is .997, Chebyshev's inequality gives a lower bound of .889, and the new improved Chebyshev inequality gives a lower bound of .951. By making some relatively innocuous assumptions, we get a substantial improvement in the lower bound.

## 1.4 The binomial distribution

There are a few distributions that are used in the vast majority of statistical applications. The reason for this is that they tend to occur naturally. The normal distribution is one. As discussed in the next chapter, the normal distribution occurs in practice because a result called The central limit theorem dictates that many distributions can be approximated by the normal. Two other distributions, the binomial and the multinomial, occur in practice because they are very simple. In this section we

discuss the binomial. The next section introduces the multinomial distribution. The results of this section are only used in Chapter 8 and in discussions of transformations.

If you have independent identical random trials and count how often something (anything) occurs, the appropriate distribution is the binomial. What could be simpler?

EXAMPLE 1.4.1. Being somewhat lonely in my misspent youth, I decided to go to a dating service. The service was to provide me with five dates. Being a very open-minded soul, I convinced myself that the results of one date would not influence my opinion about other dates. From my limited experience with the opposite sex, I have found that I enjoy about 40% of such brief encounters. I decided that my money would be well spent if I enjoyed two or more of the five dates. Unfortunately, my loan shark repossessed my 1954 Studebaker before I could indulge in this taste of nirvana. Back in those days, we chauvinists believed: no wheels – no women. Nevertheless, let us compute the probability that I would have been satisfied with the dating service. Let $W$ be the number of dates I would have enjoyed. The simplest way to find the probability of satisfaction is

$$\begin{aligned}\Pr(W \geq 2) &= 1 - \Pr(W < 2) \\ &= 1 - \Pr(W = 0) - \Pr(W = 1),\end{aligned}$$

but that is much too easy. Let's compute

$$\Pr(W \geq 2) = \Pr(W = 2) + \Pr(W = 3) + \Pr(W = 4) + \Pr(W = 5).$$

In particular, we compute each term on the right-hand side.

Write the outcome of the five dates as an ordered collection of Ls and Ds. For example, (L, D, L, D, D) indicates that I like the first and third dates, but dislike the second, fourth, and fifth.

To like five dates, I must like everyone of them.

$$\Pr(W = 5) = \Pr(L, L, L, L, L).$$

Remember, I assumed that the dates were independent and that the probability of my liking any one is .4. Thus,

$$\begin{aligned}\Pr(W = 5) &= \Pr(L)\Pr(L)\Pr(L)\Pr(L)\Pr(L) \\ &= (.4)^5.\end{aligned}$$

The probability of liking four dates is a bit more complicated. I could only dislike one date, but there are five different choices for the date that I could dislike. It could be the fifth, the fourth, the third, the second, or the first. Any pattern of 4 Ls and a D excludes the other patterns from occurring, e.g., if the only date I dislike is the fourth, then the only date I dislike cannot be the second. Since the patterns are mutually exclusive (disjoint), the probability of disliking one date is the sum of the probabilities of the individual patterns.

$$\begin{aligned}\Pr(W = 4) &= \Pr(L, L, L, L, D) \\ &\quad + \Pr(L, L, L, D, L) \\ &\quad + \Pr(L, L, D, L, L) \\ &\quad + \Pr(L, D, L, L, L) \\ &\quad + \Pr(D, L, L, L, L).\end{aligned} \tag{1.4.1}$$

By assumption $\Pr(L) = .4$, so $\Pr(D) = 1 - \Pr(L) = 1 - .4 = .6$. The dates are independent, so

$$\begin{aligned}\Pr(L, L, L, L, D) &= \Pr(L)\Pr(L)\Pr(L)\Pr(L)\Pr(D) \\ &= (.4)^4.6.\end{aligned}$$

Similarly,

$$\begin{aligned}\Pr(L, L, L, D, L) &= \Pr(L, L, D, L, L) \\ &= \Pr(L, D, L, L, L) \\ &= \Pr(D, L, L, L, L) \\ &= (.4)^4.6.\end{aligned}$$

Summing up the values in equation (1.4.1),

$$\Pr(W = 4) = 5(.4)^4(.6).$$

Computing the probability of liking three dates is even worse.

$$\begin{aligned}\Pr(W=3) &= \Pr(L,L,L,D,D)\\ &\quad+\Pr(L,L,D,L,D)\\ &\quad+\Pr(L,D,L,L,D)\\ &\quad+\Pr(D,L,L,L,D)\\ &\quad+\Pr(L,L,D,D,L)\\ &\quad+\Pr(L,D,L,D,L)\\ &\quad+\Pr(D,L,L,D,L)\\ &\quad+\Pr(L,D,D,L,L)\\ &\quad+\Pr(D,L,D,L,L)\\ &\quad+\Pr(D,D,L,L,L)\end{aligned}$$

Again all of these patterns have exactly the same probability. For example, using independence

$$\Pr(D,L,D,L,L) = (.4)^3(.6)^2\,.$$

Adding up all of the patterns

$$\Pr(W=3) = 10(.4)^3(.6)^2\,.$$

By now it should be clear that

$$\Pr(W=2) = (\text{no. of patterns with 2 Ls and 3 Ds})(.4)^2(.6)^3\,.$$

The number of patterns can be computed as

$$\binom{5}{2} \equiv \frac{5!}{2!(5-2)!} \equiv \frac{5\cdot4\cdot3\cdot2\cdot1}{(2\cdot1)(3\cdot2\cdot1)} = 10\,.$$

The probability that I would be satisfied with the dating service is

$$\begin{aligned}\Pr(W\geq 2) &= 10(.4)^2(.6)^3+10(.4)^3(.6)^2+5(.4)^4.6+(.4)^5\\ &= .663\,.\end{aligned}$$

□

Binomial random variables can also be generated by sampling from a fixed population. If we were going to make 20 random selections from the UNM student body, the number of females would have a binomial distribution. Given a set of procedures for defining and sampling the student body, there would be some fixed number of students of which a given number would be females. Under random sampling, the probability of selecting a female on any of the 20 trials would be simply the proportion of females in the population. Although it is very unlikely to occur in this example, the sampling scheme must allow the possibility of students being selected more than once in the sample. If people were not allowed to be chosen more than once, each successive selection would change the proportion of females available for the subsequent selection. Of course, when making 20 selections out of a population of over 20,000 UNM students, even if you did not allow people to be reselected, the changes in the proportions of females are insubstantial and the binomial distribution makes a good approximation to the true distribution. On the other hand, if the entire student population was 40 rather than 20,000+, it might not be wise to use the binomial approximation when people are not allowed to be reselected.

Typically, the outcome of interest in a binomial is referred to as a success. If the probability of a success is $p$ for each of $N$ independent identical trials, then the number of successes $y$ has a binomial distribution with parameters $N$ and $p$. Write

$$y \sim \text{Bin}(N,p)\,.$$

The distribution of $y$ is

$$\Pr(y=r)=\binom{N}{r}p^r(1-p)^{N-r}$$

for $r=0,1,\ldots,N$. Here

$$\binom{N}{r}\equiv\frac{N!}{r!(N-r)!}$$

where for any positive integer $m$, $m!\equiv m(m-1)(m-2)\cdots(2)(1)$ and $0!\equiv 1$. The notation $\binom{N}{r}$ is read "$N$ choose $r$" because it is the number of distinct ways of choosing $r$ individuals out of a collection containing $N$ individuals.

EXAMPLE 1.4.2. The random variables in Example 1.2.1 were $y_1$, the number of heads on the first toss of a coin, $y_2$, the number of heads on the second toss of a coin, and $W$, the combined number of heads from the two tosses. These have the following distributions:

$$\begin{aligned} y_1 &\sim \text{Bin}\left(1,\frac{1}{2}\right)\\ y_2 &\sim \text{Bin}\left(1,\frac{1}{2}\right)\\ W &\sim \text{Bin}\left(2,\frac{1}{2}\right). \end{aligned}$$

Note that $W$, the $\text{Bin}\left(2,\frac{1}{2}\right)$, was obtained by adding together the two independent $\text{Bin}\left(1,\frac{1}{2}\right)$ random variables $y_1$ and $y_2$. This result is quite general. Any $\text{Bin}(N,p)$ random variable can be written as the sum of $N$ independent $\text{Bin}(1,p)$ random variables. □

Given the probability distribution of a binomial, we can find the mean (expected value) and variance. By definition, if $y\sim\text{Bin}(N,p)$, the mean is

$$\text{E}(y)=\sum_{r=0}^{N}r\binom{N}{r}p^r(1-p)^{N-r}.$$

This is difficult to evaluate directly, but by writing $y$ as the sum of $N$ independent $\text{Bin}(1,p)$ random variables and using Exercise 1.6.1 and Proposition 1.2.11, it is easily seen that

$$\text{E}(y)=Np.$$

Similarly, the variance of $y$ is

$$\text{Var}(y)=\sum_{r=0}^{N}(r-Np)^2\binom{N}{r}p^r(1-p)^{N-r}$$

but by again writing $y$ as the sum of $N$ independent $\text{Bin}(1,p)$ random variables and using Exercise 1.6.1 and Proposition 1.2.11, it is easily seen that

$$\text{Var}(y)=Np(1-p).$$

Exercise 1.6.8 consists of proving these mean and variance formulae.

On occasion we will need to look at both the number of successes from a group of $N$ trials and the number of failures at the same time. If the number of successes is $y_1$ and the number of failures is $y_2$, then

$$\begin{aligned} y_2 &= N-y_1\\ y_1 &\sim \text{Bin}(N,p) \end{aligned}$$

and

$$y_2 \sim \text{Bin}(N, 1-p)\,.$$

The last result holds because, with independent identical trials, the number of outcomes that we call failures must also have a binomial distribution. If $p$ is the probability of success, the probability of failure is $1-p$. Of course,

$$\begin{aligned} E(y_2) &= N(1-p) \\ \text{Var}(y_2) &= N(1-p)p\,. \end{aligned}$$

Note that $\text{Var}(y_1) = \text{Var}(y_2)$ regardless of the value of $p$. Finally,

$$\text{Cov}(y_1, y_2) = -Np(1-p)$$

and

$$\text{Corr}(y_1, y_2) = -1\,.$$

There is a perfect linear relationship between $y_1$ and $y_2$. If $y_1$ goes up one count, $y_2$ goes down one count. When we look at both successes and failures write

$$(y_1, y_2) \sim \text{Bin}\big(N, p, (1-p)\big)\,.$$

This is the simplest case of the multinomial distribution discussed in the next section.

## 1.5 The multinomial distribution

The multinomial distribution is a generalization of the binomial allowing more than two categories. The results in this section are only used in Chapter 8.

EXAMPLE 1.5.1. Consider the probabilities for the nine height and eye color categories given in Example 1.1.2. The probabilities are repeated below.

Height–eye color probabilities

| | | Eye color | | |
|---|---|---|---|---|
| | | Blue | Brown | Green |
| Height | Tall | .12 | .15 | .03 |
| | Medium | .22 | .34 | .04 |
| | Short | .06 | .01 | .03 |

Suppose a random sample of 50 individuals was obtained with these probabilities. For example, one might have a population of 100 people in which 12 were tall with blue eyes, 15 were tall with brown eyes, 3 were short with green eyes, etc. We could randomly select one of the 100 people as the first individual in the sample. Then, returning that individual to the population, take another random selection from the 100 to be the second individual. We are to proceed in this way until 50 people are selected. Note that with a population of 100 and a sample of 50 there is a substantial chance that some people would be selected more than once. The numbers of selections falling into each of the nine categories has a multinomial distribution with $N = 50$ and these probabilities.

It is unlikely that one would actually perform sampling from a population of 100 people as described above. Typically, one would not allow the same person to be chosen more than once. However, if we had a population of 10,000 people where 1200 were tall with blue eyes, 1500 were tall with brown eyes, 300 were short with green eyes, etc., with a sample size of 50 we might be willing to allow the possibility of selecting the same person more than once simply because it is extremely unlikely to happen. Technically, to obtain the multinomial distribution with $N = 50$ and these probabilities, when sampling from a fixed population we need to allow individuals to appear more than once. However, when taking a small sample from a large population, it does not matter

much whether or not you allow people to be chosen more than once, so the multinomial often provides a good approximation even when individuals are excluded from reappearing in the sample. □

Consider a group of $N$ independent identical trials in which each trial results in the occurrence of one of $q$ events. Let $y_i$, $i = 1,\ldots,q$ be the number of times that the $i$th event occurs and let $p_i$ be the probability that the $i$th event occurs on any trial. The $p_i$s must satisfy $p_1 + p_2 + \cdots + p_q = 1$. We say that $(y_1,\ldots,y_q)$ has a multinomial distribution with parameters $N, p_1,\ldots,p_q$. Write

$$(y_1,\ldots,y_q) \sim \text{Mult}(N, p_1,\ldots,p_q)\,.$$

The distribution is given by the probabilities

$$\begin{aligned}\Pr(y_1 = r_1,\ldots,y_q = r_q) &= \frac{N!}{r_1!\cdots r_q!}p_1^{r_1}\cdots p_q^{r_q}\\ &= \left(N!\Big/\prod_{i=1}^{q} r_i!\right)\prod_{i=1}^{q} p_i^{r_i}.\end{aligned}$$

Here the $r_i$s are allowed to be any whole numbers with each $r_i \geq 0$ and $r_1 + \cdots + r_q = N$. Note that if $q = 2$, this is just a binomial distribution. In general, each individual component $y_i$ of a multinomial consists of $N$ trials in which category $i$ either occurs or does not occur, so individual components have the marginal distributions

$$y_i \sim \text{Bin}(N, p_i).$$

It follows that

$$\text{E}(y_i) = Np_i$$

and

$$\text{Var}(y_i) = Np_i(1-p_i)\,.$$

It can also be shown that

$$\text{Cov}(y_i, y_j) = -Np_ip_j \qquad \text{for} \qquad i \neq j\,.$$

EXAMPLE 1.5.2. Suppose that the 50 individuals from Example 1.5.1 fall into the categories as listed below.

Height–eye color observations

| | | Eye color | | |
|---|---|---|---|---|
| | | Blue | Brown | Green |
| Height | Tall | 5 | 8 | 2 |
| | Medium | 10 | 18 | 2 |
| | Short | 3 | 1 | 1 |

The probability of getting this particular table is

$$\frac{50!}{5!8!2!10!18!2!3!1!1!}(.12)^5(.15)^8(.03)^2(.22)^{10}(.34)^{18}(.04)^2(.06)^3(.01)^1(.03)^1.$$

This number is zero to over 5 decimal places. The fact that this is a very small number is not surprising. There are a lot of possible tables, so the probability of getting any particular table is very small. In fact, many of the possible tables are *much* less likely to occur than this table.

Let's return to thinking about the observations as random. The expected number of observations for each category is given by $Np_i$. It is easily seen that the expected counts for the cells are as given below.

Height–eye color expected values

| | | Eye color | | |
|---|---|---|---|---|
| | | Blue | Brown | Green |
| Height | Tall | 6.0 | 7.5 | 1.5 |
| | Medium | 11.0 | 17.0 | 2.0 |
| | Short | 3.0 | 0.5 | 1.5 |

Note that the expected counts need not be integers.

The variance for, say, the number of tall blue-eyed people in this population is $50(.12)(1-.12) = 5.28$. The variance of the number of short green-eyed people is $50(.03)(1-.03) = 1.455$. The covariance between the number of tall blue-eyed people and the number of short green-eyed people is $-50(.12)(.03) = -.18$. The correlation between the numbers of tall blue-eyed people and short green-eyed people is $-.18/\sqrt{(5.28)(1.455)} = -0.065$. □

*Appendix: probability for continuous distributions*

As stated in Section 1.3, probabilities are sometimes defined as areas under a curve. The curve, called a probability density function or just a density, must be defined by some nonnegative function $f(\cdot)$. (Nonnegative to ensure that probabilities are always positive.) Thus the probability that a random observation $y$ is between two numbers, say $a$ and $b$, is the area under the curve measured between $a$ and $b$. Using calculus, this is

$$\Pr[a < y < b] = \int_a^b f(y)\,\mathrm{d}y.$$

Because we are measuring areas under curves, there is no area associated with any one point, so $\Pr[a < y < b] = \Pr[a \le y < b] = \Pr[a < y \le b] = \Pr[a \le y \le b]$. The area under the entire curve must be 1, i.e.,

$$1 = \Pr[-\infty < y < \infty] = \int_{-\infty}^{\infty} f(y)\,\mathrm{d}y.$$

Figure 1.1 indicates that the probability below $K(1-\alpha)$ is $1-\alpha$, i.e.,

$$1-\alpha = \Pr[y < K(1-\alpha)] = \int_{-\infty}^{K(1-\alpha)} f(y)\,\mathrm{d}y$$

and that the probability above $K(1-\alpha)$ is $\alpha$, i.e.,

$$\alpha = \Pr[y > K(1-\alpha)] = \int_{K(1-\alpha)}^{\infty} f(y)\,\mathrm{d}y.$$

The expected value of $y$ is defined as

$$\mathrm{E}(y) = \int_{-\infty}^{\infty} y f(y)\,\mathrm{d}y.$$

For any function $g(y)$, the expected value is

$$\mathrm{E}[g(y)] = \int_{-\infty}^{\infty} g(y) f(y)\,\mathrm{d}y.$$

In particular, if we let $\mathrm{E}(y) = \mu$ and $g(y) = (y-\mu)^2$, we define the variance as

$$\mathrm{Var}(y) = E[(y-\mu)^2] = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,\mathrm{d}y.$$

To define the covariance between two random variables, say $y_1$ and $y_2$, we need a joint density $f(y_1, y_2)$. We can find the density for $y_1$ alone as

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)\, \mathrm{d}y_2$$

and we can write $\mathrm{E}(y_1)$ in two equivalent ways

$$\mathrm{E}(y_1) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1 f(y_1, y_2)\, \mathrm{d}y_1\, \mathrm{d}y_2 = \int_{-\infty}^{\infty} y_1 f_1(y_1)\, \mathrm{d}y_1.$$

Writing $\mathrm{E}(y_1) = \mu_1$ and $\mathrm{E}(y_2) = \mu_2$ we can now define the covariance between $y_1$ and $y_2$ as

$$\mathrm{Cov}(y_1, y_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_1 - \mu_1)(y_2 - \mu_2) f(y_1, y_2)\, \mathrm{d}y_1\, \mathrm{d}y_2.$$

## 1.6 Exercises

EXERCISE 1.6.1. Use the definitions to find the expected value and variance of a $\mathrm{Bin}(1, p)$ distribution.

EXERCISE 1.6.2. Let $y$ be a random variable with $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \sigma^2$. Show that

$$\mathrm{E}\left(\frac{y-\mu}{\sigma}\right) = 0$$

and

$$\mathrm{Var}\left(\frac{y-\mu}{\sigma}\right) = 1.$$

Let $\bar{y}_\cdot$ be the sample mean of $n$ independent observations $y_i$ with $\mathrm{E}(y_i) = \mu$ and $\mathrm{Var}(y_i) = \sigma^2$. What is the expected value and variance of

$$\frac{\bar{y}_\cdot - \mu}{\sigma/\sqrt{n}}?$$

Hint: For the first part, write

$$\frac{y-\mu}{\sigma} \quad \text{as} \quad \frac{1}{\sigma}y - \frac{\mu}{\sigma}$$

and use Proposition 1.2.11.

EXERCISE 1.6.3. Let $y$ be the random variable consisting of the number of spots that face up upon rolling a die. Give the distribution of $y$. Find the expected value, variance, and standard deviation of $y$.

EXERCISE 1.6.4. Consider your letter grade for this course. Obviously, it is a random phenomenon. Define the 'grade point' random variable: $y(\mathrm{A}) = 4$, $y(\mathrm{B}) = 3$, $y(\mathrm{C}) = 2$, $y(\mathrm{D}) = 1$, $y(\mathrm{F}) = 0$. If you were lucky enough to be taking the course from me, you would find that I am an easy grader. I give 5% As, 10% Bs, 35% Cs, 30% Ds, and 20% Fs. I also assign grades at random, that is to say, my tests generate random scores. Give the distribution of $y$. Find the expected value, variance, and standard deviation of the grade points a student would earn in my class. (Just in case you hadn't noticed, I'm being sarcastic.)

EXERCISE 1.6.5. Referring to Exercise 1.6.4, suppose I have a class of 40 students, what is the joint distribution for the numbers of students who get each of the five grades? Note that we are no

longer looking at how many grade points an individual student might get, we are now counting how many occurrences we observe of various events. What is the distribution for the number of students who get Bs? What is the expected value of the number of students who get Cs? What is the variance and standard deviation of the number of students who get Cs? What is the probability that in a class of 5 students, 1 gets an A, 2 get Cs, 1 gets a D, and 1 fails?

EXERCISE 1.6.6. Graph the function $f(x) = 1$ if $0 < x < 1$ and $f(x) = 0$ otherwise. This is known as the uniform density on $(0,1)$. If we use this curve to define a probability function, what is the probability of getting an observation larger than $1/4$? Smaller than $2/3$? Between $1/3$ and $7/9$?

EXERCISE 1.6.7. Arthritic ex-football players prefer their laudanum made with Old Pain-Killer Scotch by two to one. If we take a random sample of 5 arthritic ex-football players, what is the distribution of the number who will prefer Old Pain-Killer? What is the probability that only 2 of the ex-players will prefer Old Pain-Killer? What is the expected number who will prefer Old Pain-Killer? What are the variance and standard deviation of the number who will prefer Old Pain-Killer?

EXERCISE 1.6.8. Let $W \sim \text{Bin}(N,p)$ and for $i = 1,\ldots,N$ take independent $y_i$s that are $\text{Bin}(1,p)$. Argue that $W$ has the same distribution as $y_1 + \cdots + y_N$. Use this fact, along with Exercise 1.6.1 and Proposition 1.2.11, to find $\text{E}(W)$ and $\text{Var}(W)$.

EXERCISE 1.6.9. Appendix B.1 gives probabilities for a family of distributions that all look roughly like Figure 1.1. All members of the family are symmetric about zero and the members are distinguished by having different numbers of degrees of freedom ($df$). They are called $t$ distributions. For $0 \le \alpha \le 1$, the $\alpha$ percentile of a $t$ distribution with $df$ degrees of freedom is the point $x$ such that $\Pr[t(df) \le x] = \alpha$. For example, from Table B.1 the row corresponding to $df = 10$ and the column for the .90 percentile tells us that $\Pr[t(10) \le 1.372] = .90$.

(a) Find the .99 percentile of a $t(7)$ distribution.

(b) Find the .975 percentile of a $t(50)$ distribution.

(c) Find the probability that a $t(25)$ is less than or equal to 3.450.

(d) Find the probability that a $t(100)$ is less than or equal to 2.626.

(e) Find the probability that a $t(16)$ is greater than 2.92.

(f) Find the probability that a $t(40)$ is greater than 1.684.

(g) Recalling that $t$ distributions are symmetric about zero, what is the probability that a $t(40)$ distribution is less than $-1.684$?

(h) What is the probability that a $t(40)$ distribution is between $-1.684$ and 1.684?

(i) What is the probability that a $t(25)$ distribution is less than $-3.450$?

(j) What is the probability that a $t(25)$ distribution is between $-3.450$ and 3.450?

EXERCISE 1.6.10. Consider a random variable that takes on the values 25, 30, 45, and 50 with probabilities .15, .25, .35, and .25, respectively. Find the expected value, variance, and standard deviation of this random variable.

EXERCISE 1.6.11. Consider three independent random variables $X$, $Y$, and $Z$. Suppose $\text{E}(X) = 25$, $\text{E}(Y) = 40$, and $\text{E}(Z) = 55$ with $\text{Var}(X) = 4$, $\text{Var}(Y) = 9$, and $\text{Var}(Z) = 25$.

(a) Find $\text{E}(2X + 3Y + 10)$ and $\text{Var}(2X + 3Y + 10)$.

(b) Find $\text{E}(2X + 3Y + Z + 10)$ and $\text{Var}(2X + 3Y + Z + 10)$.

EXERCISE 1.6.12. As of 1994, Duke University had been in the final four of the NCAA's national basketball championship tournament seven times in nine years. Suppose their appearances were independent and that they had a probability of .25 for winning the tournament in each of those years.

(a) What is the probability that Duke would win two national championships in those seven appearances?

(b) What is the probability that Duke would win three national championships in those seven appearances?

(c) What is the expected number of Duke championships in those seven appearances?

(d) What is the variance of the number of Duke championships in those seven appearances?

EXERCISE 1.6.13. Graph the function $f(x) = 2x$ if $0 < x < 1$ and $f(x) = 0$ otherwise. If we use this curve to define a probability function, what is the probability of getting an observation larger than $1/4$? Smaller than $2/3$? Between $1/3$ and $7/9$?

EXERCISE 1.6.14. A pizza parlor makes small, medium, and large pizzas. Over the years they make 20% small pizzas, 35% medium pizzas, and 45% large pizzas. On a given Tuesday night they were asked to make only 10 pizzas. If the orders were independent and representative of the long-term percentages, what is the probability that the orders would be for four small, three medium, and three large pizzas. On such a night, what is the expected number of large pizzas to be ordered and what is the expected number of small pizzas to be ordered? What is the variance of the number of large pizzas to be ordered and what is the variance of the number of medium pizzas to be ordered?

EXERCISE 1.6.15. When I order a limo, 65% of the time the driver is male. Assuming independence, what is the probability that 6 of my next 8 drivers are male? What is the expected number of male drivers among my next eight? What is the variance of the number of male drivers among my next eight?

EXERCISE 1.6.16. When I order a limo, 65% of the time the driver is clearly male, 30% of the time the driver is clearly female, and 5% of the time the gender of the driver is indeterminant. Assuming independence, what is the probability that among my next 8 drivers 5 are clearly male and 3 are clearly female? What is the expected number of indeterminant drivers among my next eight? What is the variance of the number of clearly female drivers among my next eight?

Chapter 2

# One sample

In this chapter we examine the analysis of a single *random* sample consisting of $n$ independent observations from some population.

## 2.1 Example and introduction

EXAMPLE 2.1.1. Consider the dropout rate from a sample of math classes at the University of New Mexico in the 1984–85 school year as reported by Koopmans (1987). The data are

$$5,22,10,12,8,17,2,25,10,10,7,7,40,7,9,17,12,12,1,$$

$$13,10,13,16,3,14,17,10,10,13,59,11,13,5,12,14,3,14,15.$$

This list of $n = 38$ observations is not very illuminating. A graphical display of the numbers is more informative. Figure 2.1 plots the data above a single axis. This is often called a *dot plot*. From Figure 2.1, we see that most of the observations are between 0 and 18. There are two conspicuously large observations. Going back to the original data we identify these as the values 40 and 59. In particular, these two *outlying* values strongly suggest that the data do not follow a bell shaped curve and thus that the data do not follow a normal distribution.

□

Typically, for one sample of data we assume that the $n$ observations are

| Data | | Distribution |
|---|---|---|
| $y_1, y_2, \ldots, y_n$ | independent | $N(\mu, \sigma^2)$ |

The key assumptions are that the observations are independent and have the same distribution. In particular, we assume they have the same (unknown) mean $\mu$ and the same (unknown) variance $\sigma^2$.

These assumptions of independence and a constant distribution should be viewed as only useful approximations to actual conditions. Often the most valuable approach to evaluating these assumptions is simply to think hard about whether they are reasonable. In any case, the conclusions we reach are only as good as the assumptions we have made. The only way to be positive that these assumptions are true is if we arrange for them to be true. If we have a fixed finite population and take a random sample from the population allowing elements of the population to be observed more than

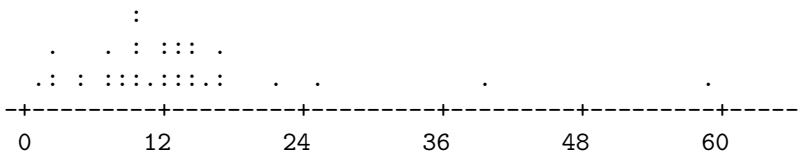



Figure 2.1: *Dot plot for drop rate percentage data.*

once, then the assumptions (other than normality) are true. In Example 2.1.1, if we had the dropout rates for all math classes in the year and randomly selected these 38 while allowing for classes to appear more than once in the sample, the assumptions of independence with the same distribution are satisfied.

The ideal conditions of independent sampling from a fixed population are difficult to achieve. Many populations refuse to hold still while we sample them. For example, the population of students at a large university changes almost continuously (during working hours). To my way of thinking, the populations associated with most interesting data are virtually impossible to define unambiguously. Who really cares about the dropout rates for 1984–85? As such, they can only be used to fix blame. Our real interest is in what the data can tell us about current and future dropout rates. If the data are representative of current or future conditions, the data can be used to fix problems. For example, one might find out whether certain instructors generate huge dropout rates and avoid taking classes from them. It is difficult to decide whether these or any data are representative of current or future conditions because we cannot possibly know the future population and we cannot practically know the current population. As mentioned earlier, often our best hope is to think hard about whether these data approximate independent observations from the population of interest.

Even when sampling from a fixed population, we use approximations. In practice we rarely allow elements of a fixed population to be observed more than once in a sample. This invalidates the assumptions. If the first sampled element is eliminated, the second element is actually being sampled from a different population than the first. (One element has been eliminated.) Fortunately, when the sample contains a small proportion of the fixed population, the standard assumptions make a good approximation. Moreover, the normal distribution is never more than an approximation to a fixed population. The normal distribution has an infinite number of possible outcomes, while fixed populations are finite. Often, the normal distribution makes a good approximation, especially if we do our best to validate it. In addition, the assumption of a normal distribution is only used when drawing conclusions from small samples. For large samples we can get by without the assumption of normality.

Our primary objective is to draw conclusions about the mean $\mu$. We condense the data into summary statistics. These are the sample mean, the sample variance, and the sample standard deviation. The sample mean has the algebraic formula

$$\bar{y}_\cdot \equiv \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\left[y_1 + y_2 + \cdots + y_n\right]$$

where the $\cdot$ in $\bar{y}_\cdot$ indicates that the mean is obtained by averaging the $y_i$s over the subscript $i$. The sample mean $\bar{y}_\cdot$ estimates the population mean $\mu$. The sample variance is an estimate of the population variance $\sigma^2$. The sample variance is *essentially* the average squared distance of the observations from the sample mean,

$$\begin{aligned} s^2 &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_\cdot)^2 \\ &= \frac{1}{n-1}\left[(y_1 - \bar{y}_\cdot)^2 + (y_2 - \bar{y}_\cdot)^2 + \cdots + (y_n - \bar{y}_\cdot)^2\right]. \end{aligned} \tag{2.1.1}$$

The sample standard deviation is just the square root of the sample variance,

$$s \equiv \sqrt{s^2}.$$

EXAMPLE 2.1.2. The sample mean of the dropout rate data is

$$\bar{y}_\cdot = \frac{5 + 22 + 10 + 12 + 8 + \cdots + 3 + 14 + 15}{38} = 13.11.$$

If we think of these data as a sample from the fixed population of math dropout rates in 1984–85, $\bar{y}_{\cdot}$ is obviously an estimate of the simple average of all the dropout rates of all the classes in that academic year. Equivalently, $\bar{y}_{\cdot}$ is an estimate of the expected value for the random variable defined as the dropout rate obtained when we randomly select one class from the fixed population. Alternatively, we may interpret $\bar{y}_{\cdot}$ as an estimate of the mean of some population that is more interesting but less well defined than the fixed population of math dropout rates for 1984–85.

The sample variance is

$$\begin{aligned} s^2 &= \frac{\left[(5-13.11)^2+(22-13.11)^2+\cdots+(14-13.11)^2+(15-13.11)^2\right]}{38-1} \\ &= 106.5. \end{aligned}$$

This estimates the variance of the random variable obtained when randomly selecting one class from the fixed population. The sample standard deviation is

$$s = \sqrt{106.5} = 10.32\,. \qquad \square$$

The only reason $s^2$ is *not* the average squared distance of the observations from the sample mean is that the denominator in (2.1.1) is $n-1$ instead of $n$. If $\mu$ were known, a better estimate of the population variance $\sigma^2$ would be $\hat{\sigma}^2 \equiv \sum_{i=1}^{n}(y_i-\mu)^2/n$. In $s^2$, we have used $\bar{y}_{\cdot}$ to estimate $\mu$. Not knowing $\mu$, we know less about the population, so $s^2$ cannot be as good an estimate as $\hat{\sigma}^2$. The quality of a variance estimate can be measured by the number of observations on which it is based; $\hat{\sigma}^2$ makes full use of all $n$ observations for estimating $\sigma^2$. In using $s^2$, we lose the functional equivalent of one observation for having estimated the parameter $\mu$. Thus $s^2$ has $n-1$ in the denominator of (2.1.1) and is said to have $n-1$ *degrees of freedom*. In nearly all problems that we will discuss, there is one degree of freedom available for every observation. The degrees of freedom are assigned to various estimates and we will need to keep track of them.

The statistics $\bar{y}_{\cdot}$ and $s^2$ are estimates of $\mu$ and $\sigma^2$ respectively. The *law of large numbers* is a mathematical result implying that for large sample sizes $n$, $\bar{y}_{\cdot}$ gets arbitrarily close to $\mu$ and $s^2$ gets arbitrarily close to $\sigma^2$.

Both $\bar{y}_{\cdot}$ and $s^2$ are computed from the random observations $y_i$. The summary statistics are functions of random variables, so they must also be random. Each has a distribution and to draw conclusions about the unknown parameters $\mu$ and $\sigma^2$ we need to know the distributions. In particular, if the original data are normally distributed, the sample mean has the distribution

$$\bar{y}_{\cdot} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

or equivalently,

$$\frac{\bar{y}_{\cdot}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)\,, \tag{2.1.2}$$

see Exercise 1.6.2. In Subsection 1.2.4 we established that $\mathrm{E}(\bar{y}_{\cdot})=\mu$ and $\mathrm{Var}(\bar{y}_{\cdot})=\sigma^2/n$, so the only new claim made here is that the sample mean computed from *independent, identically distributed (iid)* normal random variables is again normally distributed. Moreover, the *central limit theorem* is a mathematical result stating that these distributions are approximately true for 'large' samples $n$, regardless of whether the original data are normally distributed.

As we will see below, the distributions given above are only useful in drawing conclusions about data when $\sigma^2$ is known. Generally, we will need to estimate $\sigma^2$ with $s^2$ and proceed as best we can. By the law of large numbers, $s^2$ becomes arbitrarily close to $\sigma^2$, so for large samples we can substitute $s^2$ for $\sigma^2$ in the distributions above. In other words, for large samples the *approximation*

$$\frac{\bar{y}_{\cdot}-\mu}{\sqrt{s^2/n}} \sim N(0,1) \tag{2.1.3}$$

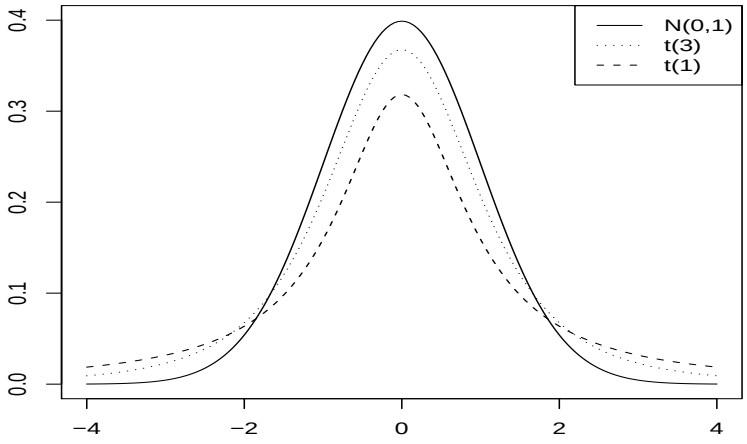



Figure 2.2: *Three distributions: solid, $N(0,1)$; long dashes, $t(1)$; short dashes, $t(3)$.*

holds regardless of whether the data were originally normal.

For small samples we cannot rely on $s^2$ being close to $\sigma^2$, so we fall back on the assumption that the original data are normally distributed. For normally distributed data, the appropriate distribution is called a $t$ distribution with $n-1$ degrees of freedom. In particular,

$$\frac{\bar{y}_\cdot - \mu}{\sqrt{s^2/n}} \sim t(n-1). \tag{2.1.4}$$

The $t$ distribution is similar to the standard normal but more spread out, see Figure 2.2. It only makes sense that if we need to estimate $\sigma^2$ rather than knowing it, our conclusions will be less exact. This is reflected in the fact that the $t$ distribution is more spread out than the $N(0,1)$. In the previous paragraph we argued that for large $n$ the appropriate distribution is

$$\frac{\bar{y}_\cdot - \mu}{\sqrt{s^2/n}} \sim N(0,1).$$

We are now arguing that for normal data the appropriate distribution is $t(n-1)$. It better be the case (and is) that for large $n$ the $N(0,1)$ distribution is approximately the same as the $t(n-1)$ distribution. In fact, we define $t(\infty)$ to be a $N(0,1)$ distribution where $\infty$ indicates an infinitely large number.

*Formal distribution theory*

By definition, the $t$ distribution is obtained as the ratio of two things related to the sample mean and variance. We now present this general definition.

First, for normally distributed data, the sample variance $s^2$ has a known distribution that depends on $\sigma^2$. It is related to a distribution called the *chi-squared* ($\chi^2$) distribution with $n-1$ degrees of freedom. In particular,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1). \tag{2.1.5}$$

Moreover, for normal data, $\bar{y}_\cdot$ and $s^2$ are independent.

**Definition 2.1.3.** A $t$ distribution is the distribution obtained when a random variable with a $N(0,1)$ distribution is divided by an independent random variable that is the square root of a $\chi^2$ random variable over its degrees of freedom. The $t$ distribution has the same degrees of freedom as the chi-square.

In particular, $[\bar{y}_\cdot - \mu]/\sqrt{\sigma^2/n}$ is $N(0,1)$, $\sqrt{[(n-1)s^2/\sigma^2]/(n-1)}$ is the square root of a chi-squared random variable over its degrees of freedom, and the two are independent because $\bar{y}_\cdot$ and $s^2$ are independent, so

$$\frac{\bar{y}_\cdot - \mu}{\sqrt{s^2/n}} = \frac{[\bar{y}_\cdot - \mu]/\sqrt{\sigma^2/n}}{\sqrt{[(n-1)s^2/\sigma^2]/(n-1)}} \sim t(n-1).$$

The $t$ distribution has the same degrees of freedom as the estimate of $\sigma^2$; this is typically the case in other applications.

## 2.2 Inference about $\mu$

Most statistical tests and confidence intervals are applications of a single theory. (Tests and confidence intervals for variances are exceptions.) To use this theory we need to know four things. In the one-sample problem the four things are

1. the parameter of interest, $\mu$,
2. the estimate of the parameter, $\bar{y}_\cdot$,
3. the standard error of the estimate, $\mathrm{SE}(\bar{y}_\cdot) \equiv \sqrt{s^2/n} = s/\sqrt{n}$, and
4. the appropriate distribution for $[\bar{y}_\cdot - \mu]/\sqrt{s^2/n}$ .

Specifically, we need a known (tabled) distribution for $[\bar{y}_\cdot - \mu]/\sqrt{s^2/n}$ that is symmetric about zero and continuous. The standard error, $\mathrm{SE}(\bar{y}_\cdot)$, is the estimated standard deviation of $\bar{y}_\cdot$. Recall that the variance of $\bar{y}_\cdot$ is $\sigma^2/n$, so its standard deviation is $\sqrt{\sigma^2/n}$ and estimating $\sigma^2$ by $s^2$ gives the standard error $\sqrt{s^2/n}$.

The appropriate distribution for $[\bar{y}_\cdot - \mu]/\sqrt{s^2/n}$ when the data are normally distributed is the $t(n-1)$ as in (2.1.4). For large samples, the appropriate distribution is the $N(0,1)$ as in (2.1.3). Recall that for large samples from a normal population, it is irrelevant whether we use the standard normal or the $t$ distribution because they are essentially the same. In the unrealistic case where $\sigma^2$ is known we do not need to estimate it, so we use $\sqrt{\sigma^2/n}$ instead of $\sqrt{s^2/n}$ for the standard error. In this case, the appropriate distribution is (2.1.2) if either the original data are normal or the sample size is large.

We need notation for the percentage points of the known distribution and we need a name for the point that cuts off the top $\alpha$ of the distribution. Typically, we need to find points that cut off the top 5%, 2.5%, 1%, or 0.5% of the distribution, so $\alpha$ is .05, .025, .01, or .005. As discussed in the previous paragraph, the appropriate distribution depends on various circumstances of the problem, so we begin by discussing percentage points with a generic notation. We use the notation $K(1-\alpha)$ for the point that cuts off the top $\alpha$ of the distribution. Figure 2.3 displays this idea graphically for a value of $\alpha$ between 0 and .5. The distribution is described by the curve, which is symmetric about 0. $K(1-\alpha)$ is indicated along with the fact that the area under the curve to the right of $K(1-\alpha)$ is $\alpha$. Formally the point that cuts off the top $\alpha$ of the distribution is $K(1-\alpha)$ where

$$\Pr\left[\frac{\bar{y}_\cdot - \mu}{\mathrm{SE}(\bar{y}_\cdot)} > K(1-\alpha)\right] = \alpha.$$

Note that the same point $K(1-\alpha)$ also cuts off the bottom $1-\alpha$ of the distribution, i.e.,

$$\Pr\left[\frac{\bar{y}_\cdot - \mu}{\mathrm{SE}(\bar{y}_\cdot)} < K(1-\alpha)\right] = 1-\alpha.$$

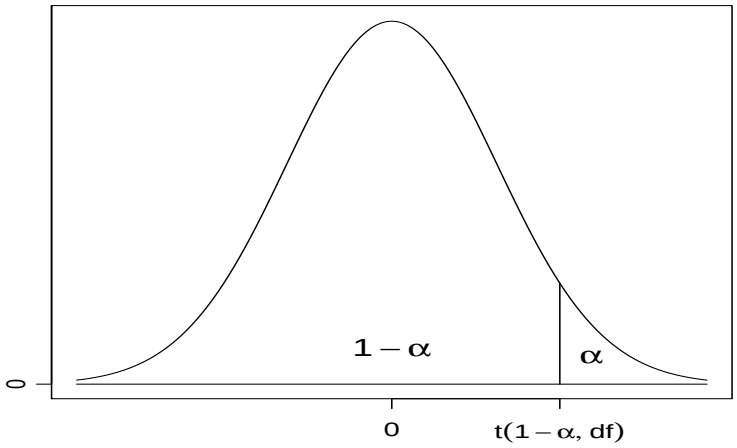



Figure 2.3: $1-\alpha$ *percentile of the distribution of* $[\bar{y}_\cdot - \mu]/\mathrm{SE}(\bar{y}_\cdot)$.

This is illustrated in Figure 2.3 by the fact that the area under the curve to the left of $K(1-\alpha)$ is $1-\alpha$. The reason the point is labeled $K(1-\alpha)$ is because it cuts off the bottom $1-\alpha$ of the distribution. The labeling depends on the percentage to the left even though our interest is in the percentage to the right.

There are at least three different ways to label these percentage points; I have simply used the one I feel is most consistent with general usage in probability and statistics. The key point however is to be familiar with Figure 2.3. We need to find points that cut off a fixed percentage of the area under the curve. As long as we can find such points, what we call them is irrelevant. Ultimately, anyone doing statistics will need to be familiar with all three methods of labeling. One method of labeling is in terms of the area to the left of the point; this is the one we will use. A second method is labeling in terms of the area to the right of the point; thus the point we call $K(1-\alpha)$ could be labeled, say, $Q(\alpha)$. The third method is to call this number, say, $W(2\alpha)$, where the area to the right of the point is doubled in the label. For example, if the distribution is a $N(0,1)$, the point that cuts off the bottom 97.5% of the distribution is 1.96. This point also cuts off the top 2.5% of the area. It makes no difference if we refer to 1.96 as the number that cuts off the bottom 97.5%, $K(.975)$, or as the number that cuts off the top 2.5%, $Q(.025)$, or as the number $W(.05)$ where the label involves $2\times.025$; the important point is being able to identify 1.96 as the appropriate number. Henceforth, we will always refer to points in terms of $K(1-\alpha)$, the point that cuts off the bottom $1-\alpha$ of the distributions. No further reference to the alternative labelings will be made but all three labels are used in Appendix B.1. There $K(1-\alpha)$s are labeled as percentiles and, for reasons related to statistical tests, $Q(\alpha)$s and $W(2\alpha)$s are labeled as *one-sided* and *two-sided* $\alpha$ levels respectively.

A fundamental assumption in inference about $\mu$ is that the distribution of $[\bar{y}_\cdot - \mu]/\mathrm{SE}(\bar{y}_\cdot)$ is symmetric about 0. By the symmetry around zero, if $K(1-\alpha)$ cuts off the top $\alpha$ of the distribution, $-K(1-\alpha)$ must cut off the bottom $\alpha$ of the distribution. Thus for distributions that are symmetric about 0 we have $K(\alpha)$, the point that cuts off the bottom $\alpha$ of the distribution, equal to $-K(1-\alpha)$. This fact is illustrated in Figure 2.4. Algebraically, we write

$$\Pr\left[\frac{\bar{y}_\cdot - \mu}{\mathrm{SE}(\bar{y}_\cdot)} < -K(1-\alpha)\right] = \Pr\left[\frac{\bar{y}_\cdot - \mu}{\mathrm{SE}(\bar{y}_\cdot)} < K(\alpha)\right] = \alpha.$$

Frequently, we want to create a central interval that contains a specified probability, say $1-\alpha$.

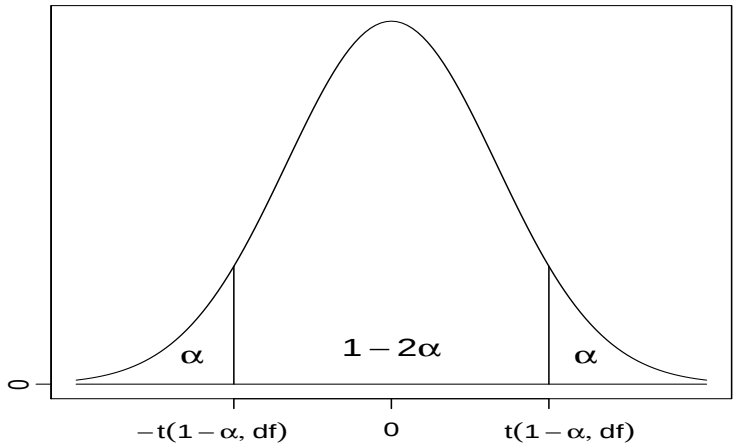



Figure 2.4: *Symmetry about 0 in the distribution of* $[\bar{y}_\cdot - \mu]/\mathrm{SE}(\bar{y}_\cdot)$.

Figure 2.5 illustrates the construction of such an interval. Algebraically, the middle interval with probability $1-\alpha$ is obtained by

$$\Pr\left[-K\Big(1-\frac{\alpha}{2}\Big) < \frac{\bar{y}_\cdot - \mu}{\mathrm{SE}(\bar{y}_\cdot)} < K\Big(1-\frac{\alpha}{2}\Big)\right] = 1-\alpha.$$

The probability of getting something outside of this interval is

$$\alpha = \frac{\alpha}{2} + \frac{\alpha}{2} = \Pr\left[\frac{\bar{y}_\cdot - \mu}{\mathrm{SE}(\bar{y}_\cdot)} < -K\Big(1-\frac{\alpha}{2}\Big)\right] + \Pr\left[\frac{\bar{y}_\cdot - \mu}{\mathrm{SE}(\bar{y}_\cdot)} > K\Big(1-\frac{\alpha}{2}\Big)\right].$$

In practice, the values $K(1-\alpha)$ are found from either a normal table or a $t$ table. For normal percentage points, we use the notation

$$z(1-\alpha) = K(1-\alpha).$$

For percentage points of a $t$ with $df$ degrees of freedom, use

$$t(1-\alpha, df) = K(1-\alpha).$$

Recall that as $df$ gets large, the $t(df)$ distribution converges to a $N(0,1)$, so

$$z(1-\alpha) = t(1-\alpha, \infty).$$

Percentiles of the $t$ distribution are given in Appendix B.1 with the $\infty$ row giving percentiles of the $N(0,1)$ distribution.

### *2.2.1 Confidence intervals*

A confidence interval is an interval of possible $\mu$ values in which we are 'confident' that the true value of $\mu$ lies. Moreover, a numerical level of confidence is specified for the interval. Confidence

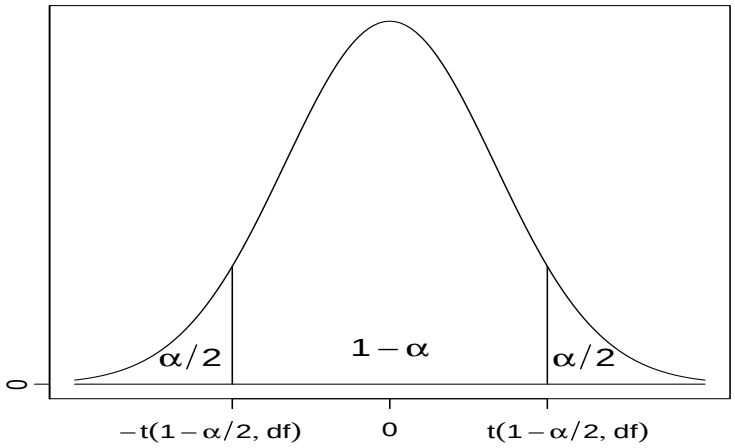



Figure 2.5: $1-\alpha$ *central interval for the distribution of* $[\bar{y}_{\cdot}-\mu]/\mathrm{SE}(\bar{y}_{\cdot})$.

intervals are commonly viewed as the most useful single procedure in statistical inference. A 95% confidence interval for $\mu$ is based on the following probability statements:

$$\begin{aligned} .95 &= \Pr\left[-K(.975) < \frac{\bar{y}_{\cdot}-\mu}{\mathrm{SE}(\bar{y}_{\cdot})} < K(.975)\right] \\ &= \Pr[\bar{y}_{\cdot}-K(.975)\,\mathrm{SE}(\bar{y}_{\cdot}) < \mu < \bar{y}_{\cdot}+K(.975)\,\mathrm{SE}(\bar{y}_{\cdot})] \end{aligned}$$

The first equality given above holds simply by the definition of $K(.975)$ and the symmetry of the distribution; it expresses Figure 2.5 algebraically for $\alpha = .05$. The second equality follows from the fact that the statements within the two sets of square brackets can be shown to be algebraically equivalent.

More generally, a $(1-\alpha)100\%$ confidence interval for $\mu$ is based on the following probability statements:

$$\begin{aligned} 1-\alpha &= \Pr\left[-K\left(1-\frac{\alpha}{2}\right) < \frac{\bar{y}_{\cdot}-\mu}{\mathrm{SE}(\bar{y}_{\cdot})} < K\left(1-\frac{\alpha}{2}\right)\right] \\ &= \Pr\left[\bar{y}_{\cdot}-K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(\bar{y}_{\cdot}) < \mu < \bar{y}_{\cdot}+K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(\bar{y}_{\cdot})\right] \end{aligned}$$

The first equality given above holds simply by the definition of $K\left(1-\frac{\alpha}{2}\right)$ and the symmetry of the distribution. Again, it is just an algebraic statement of Figure 2.5. The second equality follows from the fact that the statements within the square brackets are algebraically equivalent. A proof of the equivalence is given in the appendix to the next chapter.

The probability statement

$$1-\alpha = \Pr\left[\bar{y}_{\cdot}-K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(\bar{y}_{\cdot}) < \mu < \bar{y}_{\cdot}+K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(\bar{y}_{\cdot})\right]$$

is the basis of the confidence interval for $\mu$. The $(1-\alpha)100\%$ confidence interval for $\mu$ is simply the interval within the square brackets, i.e., the points between $\bar{y}_{\cdot}-K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(\bar{y}_{\cdot})$ and $\bar{y}_{\cdot}+K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(\bar{y}_{\cdot})$ *with observed values substituted for* $\bar{y}_{\cdot}$ *and* $\mathrm{SE}(\bar{y}_{\cdot})$. The endpoints can be written

$$\bar{y}_{\cdot} \pm K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(\bar{y}_{\cdot}),$$

or, substituting the form of the standard error,

$$\bar{y}_\cdot \pm K\left(1-\frac{\alpha}{2}\right)\frac{s}{\sqrt{n}}.$$

Note that increasing the sample size $n$ decreases the standard error and thus makes the confidence interval narrower. Narrower confidence intervals give more precise information about $\mu$. In fact, by taking $n$ large enough, we can make the confidence interval arbitrarily narrow.

EXAMPLE 2.2.1. For the dropout rate data presented at the beginning of the chapter, the parameter is the mean dropout rate for math classes, the estimate is $\bar{y}_\cdot = 13.11$, and the standard error is $s/\sqrt{n} = 10.32/\sqrt{38} = 1.67$. As seen in the dot plot, the original data are not normally distributed. The plot looks nothing at all like the bell shaped curve in Figure 1.1, which is a picture of a normal distribution. Thus we hope that a sample of size 38 is sufficiently large to justify use of the $N(0,1)$ distribution via the central limit theorem and the law of large numbers. For a 95% confidence interval, $95 = (1-\alpha)100$, $.95 = (1-\alpha)$, $\alpha = 1-.95 = .05$, and $1-\alpha/2 = .975$, so the number we need from the $t$ table is $z(.975) = t(.975,\infty) = 1.96$. The endpoints of the confidence interval are

$$13.11 \pm 1.96(1.67)$$

giving an interval of

$$(9.8, 16.4).$$

Rounding to simple numbers, we are 95% confident that the true dropout rate is between 10% and 16.5% □

The confidence interval has probability $1-\alpha$ that we are *going to get* a confidence interval that covers what we are trying to estimate, i.e., $\mu$. However, once the data are observed and the interval computed, this is no longer true. The particular interval that we get either covers $\mu$ or it does not. There is no probability associated with the coverage; nothing is random, neither $\mu$ nor the endpoints of the interval. For this reason we say that, 'We are $(1-\alpha)100\%$ *confident* that the true value of $\mu$ is in the interval.' I doubt that anybody has a good definition of what the word 'confident' means in that sentence. Having done my duty to explain the correct meaning of confidence intervals, you can (and will) go back to thinking that the probability is $1-\alpha$ that your interval covers $\mu$. It does not do any real harm and it can be justified using arguments from Bayesian statistics. This issue of interpretation is discussed in much more detail in the next chapter.

### *2.2.2 Hypothesis tests*

An hypothesis test is a procedure for checking the validity of a claim. Someone makes a claim which becomes the *null hypothesis*. We wish to test whether or not the claim is true. If relevant data are available, we can test the claim, but we cannot really test whether it is true or false, we can merely test whether the data are consistent or inconsistent with the claim. Data that are inconsistent with the claim suggest that the claim is false. Data that are consistent with the claim are just that, consistent with the claim; they do not imply that the claim is true because other circumstances could equally well have generated the data.

In a one sample problem, for some fixed known number $m$ we may want to test the *null hypothesis*

$$H_0 : \mu = m$$

versus the *alternative hypothesis*

$$H_A : \mu \neq m.$$

*The number m must be known; it is some number that is of interest for the specific data being analyzed. It is not just an unspecified symbol.*

EXAMPLE 2.2.2. For the dropout rate data, we might be interested in the hypothesis that the true dropout rate is 10%. Thus the null hypothesis is $H_0 : \mu = 10$ and the alternative hypothesis is $H_A : \mu \neq 10$. □

The test is based on the assumption that $H_0$ is true and we check to see if the data are inconsistent with that assumption. The idea is much like the idea of a proof by contradiction. We make an assumption $H_0$. If the data contradict that assumption, we can conclude that the assumption $H_0$ is false. If the data do not contradict $H_0$, we can only conclude that the data are consistent with the assumption; *we cannot conclude that the assumption is true*.

Unfortunately, there are two complicating factors in a statistical test. First, data almost never yield an absolute contradiction to the assumption. We need to quantify the extent to which the data are inconsistent with the assumption. Second, while we wish to test a specific assumption $H_0$, there are other assumptions involved in any statistical procedure. A contradiction only invalidates $H_0$ if the other assumptions are valid. These other assumptions were discussed at the beginning of the chapter. They include such things as independence, normality, and all observations having the same mean and variance. While we can never confirm that these other assumptions are absolutely valid, it is a key aspect of modern statistical practice to validate the assumptions as far as is reasonably possible. When we are convinced that the other assumptions are reasonably valid, data that contradict the assumptions can be reasonably interpreted as contradicting the specific assumption $H_0$.

We need to be able to identify data that are inconsistent with the assumption that $\mu = m$. Note that, regardless of any hypotheses, $\bar{y}_\cdot$ is an estimate $\mu$. For example, suppose $m = 10$. If $\bar{y}_\cdot = 10.1$, $\bar{y}_\cdot$ is an estimate of $\mu$, so the data seem to be consistent with the idea that $\mu = 10$. On the other hand, if $\bar{y}_\cdot = 10{,}000$, we expect that $\mu$ will be near $10{,}000$ and the observed $\bar{y}_\cdot$ seems to be inconsistent with $H_0 : \mu = 10$. The trick is in determining which values of $\bar{y}_\cdot$ are far enough away from 10 for us to be reasonably sure that $\mu \neq 10$. As a matter of fact, in the absence of information about the variability of $\bar{y}_\cdot$, we cannot really say that $\bar{y}_\cdot = 10.1$ is consistent with $\mu = 10$ or that $\bar{y}_\cdot = 10{,}000$ is inconsistent with $\mu = 10$. If the variability associated with $\bar{y}_\cdot$ is extremely small, $\bar{y}_\cdot = 10.1$ may be highly inconsistent with $\mu = 10$. On the other hand, if the variability associated with $\bar{y}_\cdot$ is extremely large, $\bar{y}_\cdot = 10{,}000$ may be perfectly consistent with $\mu = 10$. Obviously, the standard error of $\bar{y}_\cdot$, which is our measure of variability, must play a major role in the analysis.

Generally, since $\bar{y}_\cdot$ estimates $\mu$, if $\mu > m$, then $\bar{y}_\cdot$ tends to be greater than $m$ so that $\bar{y}_\cdot - m$ and thus $[\bar{y}_\cdot - m]/\mathrm{SE}(\bar{y}_\cdot)$ tend to be large positive numbers (larger than they would be if $H_0 : \mu = m$ were true). On the other hand, if $\mu < m$, then $\bar{y}_\cdot - m$ and $[\bar{y}_\cdot - m]/\mathrm{SE}(\bar{y}_\cdot)$ will tend to be a large negative numbers. Data that are inconsistent with the null hypothesis $\mu = m$ are large positive and large negative values of the *test statistic* $[\bar{y}_\cdot - m]/\mathrm{SE}(\bar{y}_\cdot)$. The problem is in specifying what we mean by 'large'.

We reject the null hypothesis (disbelieve $\mu = m$) if the *test statistic*

$$\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)}$$

is greater than some positive cutoff value or less than some negative cutoff value. Very large and very small (large negative) values of the test statistic are those that are most inconsistent with $\mu = m$. The problem is in specifying the cutoff values. For example, we do not want to reject $\mu = 10$ if the data are consistent with $\mu = 10$. One of our basic assumptions is that we know the distribution of $[\bar{y}_\cdot - \mu]/\mathrm{SE}(\bar{y}_\cdot)$. Thus if $H_0 : \mu = 10$ is true, we know the distribution of the test statistic $[\bar{y}_\cdot - 10]/\mathrm{SE}(\bar{y}_\cdot)$, so we know what kind of data are consistent with $\mu = 10$. For instance, when $\mu = 10$, 95% of the possible values of $[\bar{y}_\cdot - 10]/\mathrm{SE}(\bar{y}_\cdot)$ are between $-K(.975)$ and $K(.975)$. Any values of $[\bar{y}_\cdot - 10]/\mathrm{SE}(\bar{y}_\cdot)$ that fall between these numbers are reasonably consistent with $\mu = 10$ and values outside the interval are defined as being inconsistent with $\mu = 10$. Thus values of $[\bar{y}_\cdot - 10]/\mathrm{SE}(\bar{y}_\cdot)$ greater than $K(.975)$ or less than $-K(.975)$ cause us to reject the null hypothesis. Note that we arbitrarily specified the central 95% of the distribution as being consistent with $\mu = 10$. That leaves a 5% chance of getting outside the central interval, so a 5% chance that

we will reject $\mu = 10$ even when it is true. In other words, even when $\mu = 10$, 5% of the time $[\bar{y}_\cdot - 10]/\mathrm{SE}(\bar{y}_\cdot)$ will be outside the limits. We could reduce this chance of error by specifying the central 99% of the distribution as consistent with $\mu = 10$. This reduces the chance of error to 1%, but then if $\mu \neq 10$, we are less likely to reject $\mu = 10$. Thus there are two types of possible errors that we need to play off against each other. *Type I error is rejecting $H_0$ when it is true. Type II error is not rejecting $H_0$ when it is not true. The probability of type I error is known as the $\alpha$ level of the test.*

EXAMPLE 2.2.3. For the dropout rate data, consider the null hypothesis $H_0 : \mu = 10$, i.e., that the mean dropout rate is 10%. The alternative hypothesis is $H_A : \mu \neq 10$. As discussed in the example on confidence intervals, these data are not normal, so we must hope that the sample size is large enough to justify use of the $N(0,1)$ distribution. If we choose a central 90% interval and thus a type I error rate of $\alpha = .10$, the upper cutoff value is $K\left(1-\frac{\alpha}{2}\right) = z\left(1-\frac{\alpha}{2}\right) = z(1-.05) = t(.95,\infty) = 1.645$.

The $\alpha = .10$ level test for $H_0 : \mu = 10$ versus $H_A : \mu \neq 10$ is to reject $H_0$ if

$$\frac{\bar{y}_\cdot - 10}{s/\sqrt{38}} > 1.645.$$

or if

$$\frac{\bar{y}_\cdot - 10}{s/\sqrt{38}} < -1.645.$$

The estimate of $\mu$ is $\bar{y}_\cdot = 13.11$ and the observed standard error is $s/\sqrt{n} = 10.32/\sqrt{38} = 1.67$, so *the observed value of the test statistic* is

$$\frac{13.11 - 10}{1.67} = 1.86.$$

Comparing this to the cutoff value of 1.645 we have $1.86 > 1.645$, so the null hypothesis is rejected. There is evidence at the $\alpha = .10$ level that the mean dropout rate is not 10%. In fact, since $\bar{y}_\cdot = 13.11 > 10$ there is the suggestion that the dropout rate is greater than 10%.

This conclusion depends on the choice of the $\alpha$ level. If we choose $\alpha = .05$, then the appropriate cutoff value is $z(.975) = 1.96$. Since the observed value of the test statistic is 1.86, which is neither greater than 1.96 nor less than $-1.96$, we do not reject the null hypothesis. When we do not reject $H_0$, we cannot say that the true mean dropout rate is 10%, but we can say that, at the $\alpha = .05$ level, the data are consistent with the (null) hypothesis that the true mean dropout rate is 10%. □

Generally, a test of hypothesis is based on controlling the probability of making an error when the null hypothesis is true. The $\alpha$ level of the test (the probability a type I error) is the probability of rejecting the null hypothesis (saying that it is false) when the null hypothesis is in fact true. The $\alpha$ level test for $H_0 : \mu = m$ versus $H_A : \mu \neq m$ is to reject $H_0$ if

$$\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} > K\left(1-\frac{\alpha}{2}\right)$$

or if

$$\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} < -K\left(1-\frac{\alpha}{2}\right).$$

This is equivalent to saying, reject $H_0$ if

$$\frac{|\bar{y}_\cdot - m|}{\mathrm{SE}(\bar{y}_\cdot)} > K\left(1-\frac{\alpha}{2}\right).$$

Note that if $H_0$ is true, the probability that we will reject $H_0$ is

$$\Pr\left[\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} > K\left(1-\frac{\alpha}{2}\right)\right] + \Pr\left[\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} < -K\left(1-\frac{\alpha}{2}\right)\right] = \alpha/2 + \alpha/2 = \alpha.$$

Also note that we are rejecting $H_0$ for those values of $[\bar{y}_\cdot - m]/\mathrm{SE}(\bar{y}_\cdot)$ that are most inconsistent with $H_0$, these being the values of the test statistic with large absolute values.

A null hypothesis should never be accepted; it is either rejected or not rejected. A better way to think of a test is that one concludes that the data are either consistent or inconsistent with the null hypothesis. The statement that the data are inconsistent with $H_0$ is a strong statement. It disproves $H_0$ in some specified degree. The statement that the data are consistent with $H_0$ is not a strong statement; it does not prove $H_0$. For example, the dropout data happen to be consistent with $H_0 : \mu = 12$; the test statistic

$$\frac{\bar{y}_\cdot - 12}{\mathrm{SE}(\bar{y}_\cdot)} = \frac{13.11 - 12}{1.67} = .66$$

is very small. However, the data are equally consistent with $\mu = 12.00001$. These data cannot possibly indicate that $\mu = 12$ rather than $\mu = 12.00001$. However, when the null hypothesis is $H_0 : \mu = 12$, the value $\mu = 12.00001$ is part of the alternative hypothesis $H_A : \mu \neq 12$, so clearly data that are consistent with $H_0$ are also consistent with some elements of the alternative. In fact, we established earlier that based on an $\alpha = .05$ test, these data are even consistent with $\mu = 10$. *Data that are consistent with $H_0$ do not imply that the alternative is false*.

With these data there is very little hope of distinguishing between $\mu = 12$ and $\mu = 12.00001$. The probability of getting data that lead to rejecting $H_0 : \mu = 12$ when $\mu = 12.00001$ is only just slightly more than the probability of getting data that lead to rejecting $H_0$ when $\mu = 12$. The probability of getting data that lead to rejecting $H_0 : \mu = 12$ when $\mu = 12.00001$ is called the *power* of the test when $\mu = 12.00001$. *The power is the probability of appropriately rejecting $H_0$ and depends on the particular value of $\mu$ ($\neq 12$).* The fact that the power is very small for detecting $\mu = 12.00001$ is not much of a problem because no one would really care about the difference between a dropout rate of 12 and a dropout rate of 12.00001. However, a small power for a difference that one cares about is a major concern. The power is directly related to the standard error and can be increased by reducing the standard error. One natural way to reduce the standard error $s/\sqrt{n}$ is by increasing the sample size $n$.

One of the difficulties in a general discussion of hypothesis testing is that the actual null hypothesis is always context specific. You cannot give general rules for what to use as a null hypothesis because the null hypothesis needs to be some interesting claim about the population mean $\mu$. When you sample different populations, the population mean differs, and interesting claims about the population mean depend on the exact nature of the population. The best practice for setting up null hypotheses is simply to look at lots of problems and ask yourself what claims about the population mean are of interest to you. As we examine more sophisticated data structures, some interesting hypotheses will arise from the structures themselves. For example, if we have two samples of similar measurements we might be interested in testing the null hypothesis that they have the same population means. Note that there are lots of ways in which the means could be different, but only one way in which they can be the same. Of course if the specific context suggests that one mean should be, say, 25 units greater than the other, we can use that as the null hypothesis. Similarly, if we have a sample of objects and two different measurements on each object, we might be interested in whether or not the measurements are related. In that case, an interesting null hypothesis is that the measurements are *not* related. Again, there is only one way in which measurements can be unrelated, but there are many ways for measurements to display a relationship.

We will see in the next chapter that there is a duality between testing and confidence intervals. Tests are used to examine whether a difference can be shown to exist between the hypothesized mean and the mean of the population being sampled. Confidence intervals are used to quantify what is known about the population mean. In particular, confidence intervals can be used to quantify how much difference exists between some hypothesized mean and the sampled population's mean. Of course, one must consider not only how much of a difference exists but also whether such a difference is meaningful in the context of the problem.

*One-sided tests*

Unless math classes were intentionally being used to weed out students (something I do not believe was true) high dropout rates are typically considered unfortunate. Math instructors might claim that dropout rates are 10% *or less* and students may want to test that claim. In such a case the claim is only contradicted by dropout rates greater than 10%

We can do one-sided tests in a similar manner to the two-sided testing discussed previously. The $\alpha$ level test for $H_0 : \mu \leq m$ versus $H_A : \mu > m$ is to reject $H_0$ if

$$\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} > K(1-\alpha).$$

Again, the value $m$ must be known; either someone tells it to you or you determine it from the subject being investigated. The alternative hypothesis is that $\mu$ is greater than something and the null hypothesis is rejected when the test statistic is greater than some cutoff value. *We reject the null hypothesis for those values of the test statistic that are most inconsistent with the null hypothesis and most consistent with the alternative hypothesis*. If the alternative is true, $\bar{y}_\cdot$ should be near $\mu$, which is greater than $m$, so large positive values of $\bar{y}_\cdot - m$ or, equivalently, large positive values of $[\bar{y}_\cdot - m]/\mathrm{SE}(\bar{y}_\cdot)$ are consistent with the alternative and inconsistent with the null hypothesis. Note that if $\mu = m$ is true, the probability of rejecting the test is

$$\Pr\left[\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} > K(1-\alpha)\right] = \alpha.$$

Moreover, it is easily seen that if $\mu < m$,

$$\Pr\left[\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} > K(1-\alpha)\right] < \alpha.$$

Thus when $H_0$ is true, i.e., when $\mu \leq m$, the probability of rejecting the null hypothesis is *at most* $\alpha$. As with the two-sided tests, we have controlled the probability of making an error when the null hypothesis is true.

EXAMPLE 2.2.4. The null hypothesis is that the dropout rate is 10% or less, i.e., $H_0 : \mu \leq 10$. The alternative is that the dropout rate is greater than 10%, i.e., $H_A : \mu > 10$. The $\alpha = .05$ level test rejects $H_0$ if

$$\frac{\bar{y}_\cdot - 10}{\mathrm{SE}(\bar{y}_\cdot)} > z(1-.05) = 1.645.$$

As seen earlier, the observed value of the test statistic is $1.86 > 1.645$, so the null hypothesis is rejected. Based on a one-sided $\alpha = .05$ test, we have evidence to reject the (null) hypothesis that the true dropout rate is 10% or less. In other words, we have evidence that the dropout rate is greater than 10%.

Students who are math averse might be interested in the claim that the dropout rate is *at least* 10%, i.e., $\mu \geq 10$. Setting this up as the null hypothesis is much less informative than the approach just demonstrated. In this case, the value of $\bar{y}_\cdot = 13.11$ is obviously consistent with $\mu$ being at least 10%. The question is whether $\bar{y}_\cdot$ is also inconsistent with $\mu \leq 10$. For $H_0 : \mu \geq 10$ a test will not be rejected. If you do not reject a test, $\alpha$ means very little. However, when you reject a test, $\alpha$ measures your chance of making an error. Setting up the test as we did allowed us to reject $H_0 : \mu \leq 10$ at $\alpha = .05$, which quantifies our chance for error. Accepting $H_0 : \mu \geq 10$ tells us nothing about the chance for error, so it is less informative. □

As we argued earlier, with a two-sided test you can *never* be sure that your $H_0$ claim is true. With a one-sided test, this is not the case. If the data are extreme enough, one hypothesis or the other is clearly indicated. In the dropout rate data example, with a standard error of 1.67, it is pretty

clear that $\bar{y}_\cdot = 4$ indicates $\mu \leq 10$ and $\bar{y}_\cdot = 16$ indicates $\mu \geq 10$, *assuming that all other assumptions are valid*. The problem occurs with $\bar{y}_\cdot$ values close to 10, say $\bar{y}_\cdot = 9$ or $\bar{y}_\cdot = 11$. If $\bar{y}_\cdot = 9$, we cannot be sure that $\mu \leq 10$ because $\mu$ could be 10 or a little larger and we would still have a reasonable chance of observing $\bar{y}_\cdot = 9$. Similarly, $\bar{y}_\cdot = 11$ is reasonably consistent with $\mu$ values of 10 or a little smaller. The only really hard problem is whether we are sure $\mu \neq 10$. If $\mu$ is different from 10, it is obvious whether $\mu < 10$ or $\mu > 10$. And if you are bothering to run this test at all, $\mu = 10$ must have some special significance and it should be of interest to establish which way $\mu$ might differ from 10. This is one of several reasons I have for preferring two-sided tests.

The $\alpha$ level test for $H_0 : \mu \geq m$ versus $H_A : \mu < m$ is to reject $H_0$ if

$$\frac{\bar{y}_\cdot - m}{\mathrm{SE}(\bar{y}_\cdot)} < -K(1-\alpha).$$

The alternative hypothesis is that $\mu$ is less than something and the null hypothesis is rejected when the test statistic is less than some cutoff value. Note that the form of the alternative determines the form of the rejection region. *In all cases we reject $H_0$ for the data that are most inconsistent with $H_0$.*

The one-sided null hypotheses involve inequalities, but $\mu = m$ is always part of the null hypothesis. The tests are set up assuming that $\mu = m$ and this needs to be part of the null hypothesis. In all cases, the test is set up so that if $\mu = m$, then the probability of making a mistake is $\alpha$.

*P values*

Rather than having formal rules for when to reject the null hypothesis, one can report the evidence against the null hypothesis. This is done by reporting the *significance level* of the test, also known as the *P value*. The $P$ value is computed under the assumption that $\mu = m$ and *is the probability of seeing data that are as extreme or more extreme than those that were actually observed*. In other words, it is the $\alpha$ level at which the test would just barely be rejected.

EXAMPLE 2.2.5. For $H_0 : \mu = 10$ versus $H_A : \mu \neq 10$ the observed value of the test statistic is 1.86. Clearly, data that give values of the test statistic that are greater than 1.86 are more extreme than the actual data. Also, by symmetry, data that give a test statistic of $-1.86$ are just as extreme as data that yield a 1.86. Finally, data that give values smaller than $-1.86$ are more extreme than data yielding a 1.86. As before, we use the standard normal distribution $z$. From a standard normal table or an appropriate computer program,

$$\begin{aligned} P &= \Pr[z \geq 1.86] + \Pr[z \leq -1.86] \\ &= .0314 + .0314 \\ &= .0628. \end{aligned}$$

Thus the approximate $P$ value is .06. The $P$ value is approximate because the use of the standard normal distribution is an approximation based on large samples. Note that

$$P = \Pr[z \geq 1.86] + \Pr[z \leq -1.86] = \Pr[|z| \geq |1.86|].$$

In the $t$ tables of Appendix B.1, the standard normal distribution corresponds to $t(\infty)$. Comparing $|1.86|$ to the tables, we see that

$$t(.95, \infty) = 1.645 < |1.86| < 1.96 = t(.975, \infty),$$

so for a two-sided test the $P$ value satisfies

$$2(1-.95) = .10 > P > .05 = 2(1-.975).$$

In other words, $t(.95,\infty)$ is the cutoff value for an $\alpha = .10$ test and $t(.975,\infty)$ is the cutoff value for an $\alpha = .05$ test; $|1.86|$ falls between these values, so the $P$ value is between .10 and .05. When only a $t$ table is available, $P$ values are most simply specified in terms of bounds such as these. □

The $P$ value is a measure of the evidence against the null hypothesis in which the smaller the $P$ value the more evidence against $H_0$. The $P$ value can be used to perform various $\alpha$ level tests. In the example, the $P$ value is .06. This is less than .10, so an $\alpha = .10$ level test of $H_0 : \mu = 10$ versus $H_A : \mu \neq 10$ will reject $H_0$. On the other hand, .06 is greater than .05, so an $\alpha = .05$ test does not reject $H_0 : \mu = 10$. Note that these are exactly the conclusions we reached in the earlier example on testing $H_0 : \mu = 10$ versus $H_A : \mu \neq 10$.

The $P$ value for a one-sided test, say, $H_0 : \mu \geq m$ versus $H_A : \mu < m$, is one half of the $P$ value from the test of $H_0 : \mu = m$ versus $H_A : \mu \neq m$ provided that $\bar{y}_\cdot < m$. If $\bar{y}_\cdot \geq m$, the $P$ value is at least .5.

## 2.3 Prediction intervals

In many situations, rather than trying to learn about $\mu$, it is more important to obtain information about future observations from the same process. With independent observations, the natural point prediction for a future observation is just the estimate of $\mu$, but a prediction interval with, say, 99% confidence of containing a future observation differs from a 99% confidence interval for $\mu$. Our ideas about where future observations will lie involves two sources of variability. First, there is the variability that a new observation $y$ displays about its mean value $\mu$. Second, we need to deal with the fact that we do not know $\mu$, so there is variability associated with $\bar{y}_\cdot$, our estimate of $\mu$. In the dropout rate example, $\bar{y}_\cdot = 13.11$ and $s^2 = 106.5$. If we could assume that the observations are normally distributed (which is a poor assumption), we could create a 99% prediction interval, i.e., an interval that contains a future observation with 99% confidence. The interval for the new observation is centered about $\bar{y}_\cdot$, our best point predictor, and is similar to a confidence interval but uses a standard error that is appropriate for prediction. The actual interval has endpoints

$$\bar{y}_\cdot \pm t(.995, n-1)\sqrt{s^2 + \frac{s^2}{n}}.$$

In our example, $n = 38$ and $t(.995,37) = 2.71$, so this becomes

$$13.11 \pm 2.71\sqrt{106.5 + \frac{106.5}{38}}$$

or

$$13.11 \pm 28.33$$

for an interval of $(-15.22, 41.44)$. In practice, dropout percentages cannot be less than 0, so a more practical interval is $(0, 41.44)$. To the limits of our assumptions, we can be 99% confident that the dropout rate for a new, similar math class will be between 0 and 41.5%. It is impossible to validate assumptions about future observations (as long as they remain in the future), thus the exact confidence levels of prediction intervals are always suspect.

The key difference between the 99% prediction interval and a 99% confidence interval is the standard error. In a confidence interval, the standard error is $\sqrt{s^2/n}$. In a prediction interval, we mentioned the need to account for two sources of variability and the corresponding standard error is $\sqrt{s^2 + s^2/n}$. The first term in the square root estimates the variance of the new observation, while the second term in the square root estimates the variance of $\bar{y}_\cdot$, the point predictor.

As mentioned earlier and as will be shown in the next section, the assumption of normality is pretty poor for the 38 observations on dropout rates. Even without the assumption of normality we can get an approximate evaluation of the interval. The interval uses the value $t(.995,37) = 2.71$, and

we will see below that even without the assumption of normality, the approximate confidence level of this prediction interval is at least

$$100\left(1-\frac{1}{(2.71)^2}\right)\% = 86\%.$$

*Theory*

In this chapter we assume that the observations $y_i$ are independent from a population with mean $\mu$ and variance $\sigma^2$. We have assumed that all our previous observations on the process have been independent, so it is reasonable to assume that the future observation $y$ is independent of the previous observations with the same mean and variance. The prediction interval is actually based on the difference $y-\bar{y}.$, i.e., we examine how far a new observation may reasonably be from our point predictor. Note that

$$\mathrm{E}(y-\bar{y}.) = \mu - \mu = 0.$$

To proceed we need a standard error for $y-\bar{y}.$ and a distribution that is symmetric about 0. The standard error of $y-\bar{y}.$ is just the standard deviation of $y-\bar{y}.$ when available or, more often, an estimate of the standard deviation. First we need to find the variance. As $\bar{y}.$ is computed from the previous observations, it is independent of $y$ and, using Proposition 1.2.11,

$$\mathrm{Var}(y-\bar{y}.) = \mathrm{Var}(y) + \mathrm{Var}(\bar{y}.) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left[1+\frac{1}{n}\right].$$

The standard deviation is the square root of the variance. Typically, $\sigma^2$ is unknown, so we estimate it with $s^2$ and our standard error becomes

$$\mathrm{SE}(y-\bar{y}.) = \sqrt{s^2+\frac{s^2}{n}} = \sqrt{s^2\left[1+\frac{1}{n}\right]} = s\sqrt{1+\frac{1}{n}}.$$

To get an appropriate distribution, we assume that all the observations are normally distributed. In this case,

$$\frac{y-\bar{y}.}{\mathrm{SE}(y-\bar{y}.)} \sim t(n-1).$$

The validity of the $t(n-1)$ distribution is established in Exercise 2.7.10. When the observations are not normally distributed, if we have a large sample we can use the law of large numbers and Chebyshev's inequality to approximate the worst case scenario.

Using the distribution based on normal observations, a 99% prediction interval is obtained from the following probability equalities:

$$\begin{aligned} .99 &= \Pr\left[-t(.995,n-1) < \frac{y-\bar{y}.}{\mathrm{SE}(y-\bar{y}.)} < t(.995,n-1)\right] \\ &= \Pr[\bar{y}. - t(.995,n-1)\mathrm{SE}(y-\bar{y}.) < y < \bar{y}. + t(.995,n-1)\mathrm{SE}(y-\bar{y}.)]. \end{aligned}$$

The key point is that the two sets of inequalities within the square brackets are algebraically equivalent. Based on the last equality, the 99% prediction interval consists of all $y$ values between $\bar{y}. - t(.995,n-1)\mathrm{SE}(y-\bar{y}.)$ and $\bar{y}. + t(.995,n-1)\mathrm{SE}(y-\bar{y}.)$. In other words, the 99% prediction interval has endpoints

$$\bar{y}. \pm t(.995,n-1)\mathrm{SE}(y-\bar{y}.).$$

This looks similar to a 99% confidence interval for $\mu$ but the standard error is very different. In the prediction interval, the endpoints are actually

$$\bar{y}. \pm t(.995,n-1)\, s\sqrt{\left[1+\frac{1}{n}\right]},$$

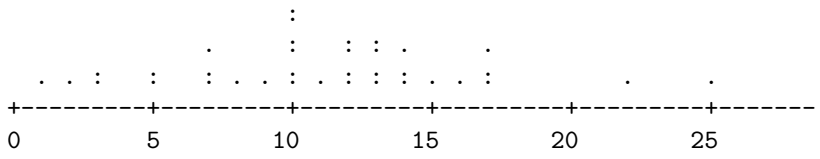



Figure 2.6: *Dot plot for drop rate percentage data: outliers deleted.*

while in a confidence interval the endpoints are

$$\bar{y}_\cdot \pm t(.995, n-1)\, s\sqrt{\frac{1}{n}}.$$

The standard error for the prediction interval is typically much larger than the standard error for the confidence interval. Moreover, unlike the confidence interval, the prediction interval cannot be made arbitrarily small by taking larger and larger sample sizes $n$. Of course to compute an arbitrary $(1-\alpha)100\%$ prediction interval, simply replace the value $t(.995, n-1)$ with $t(1-\alpha/2, n-1)$.

As mentioned above, even when the data are not normally distributed, we can obtain an approximate worst case result for large samples. The approximation comes from using the law of large numbers to justify treating $s$ as if it were the actual population standard deviation $\sigma$. With this approximation, Chebyshev's inequality states that

$$\begin{aligned} & 1 - \frac{1}{t(.995, n-1)^2} \\ & \leq \Pr\left[-t(.995, n-1) < \frac{y - \bar{y}_\cdot}{\mathrm{SE}(y - \bar{y}_\cdot)} < t(.995, n-1)\right] \\ & = \Pr[\bar{y}_\cdot - t(.995, n-1)\mathrm{SE}(y - \bar{y}_\cdot) < y < \bar{y}_\cdot + t(.995, n-1)\mathrm{SE}(y - \bar{y}_\cdot)], \end{aligned}$$

cf. Subsection 1.2.2. As mentioned above, the 99% prediction interval based on 38 normal observations has a confidence level of at least

$$\left(1 - \frac{1}{(2.71)^2}\right) 100\% = 86\%.$$

This assumes that the past observations and the future observation form a random sample from the same population and assumes that 38 observations is large enough to justify using the law of large numbers. Similarly, if we can apply the improved version of Chebyshev's inequality from Section 1.3, we get a lower bound of $1 - [1/2.25(2.71)^2] = 93.9\%$ on the confidence coefficient.

Throughout, we have assumed that the process of generating the data yields independent observations from some population. In quality control circles this is referred to as having a process that is under *statistical control*.

## 2.4 Checking normality

From Figure 2.1, we identified two *outliers* in the dropout rate data, the 40% and the 59% dropout rates. If we delete these two points from the data, the remaining data may have a more nearly normal distribution. The dot plot with the two cases deleted is given in Figure 2.6. This is much more nearly normally distributed, i.e., looks much more like a bell shaped curve, than the complete data.

Dot plots and other versions of histograms are not effective in evaluating normality. Very large amounts of data are needed before one can evaluate normality from a histogram. A more useful technique for evaluating the normality of small and moderate size samples is the construction of a *normal probability plot*, also known as a *normal plot* or a *rankit plot*. The idea is to order the data from smallest to largest and then to compare the ordered values to what one would expect the

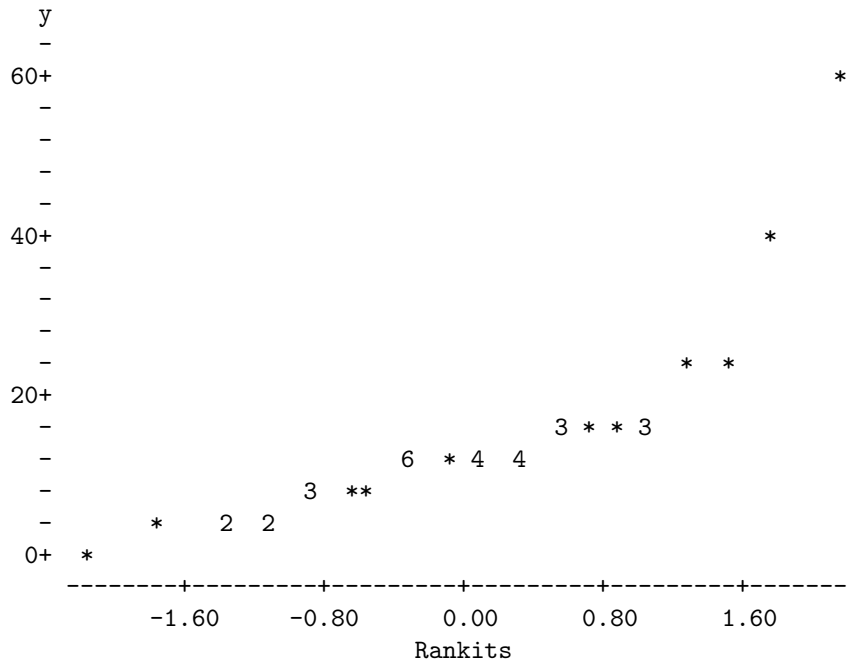



Figure 2.7: *Normal plot for drop rate percentage data: full data.*

ordered values to be if they were truly a random sample from a normal distribution. These pairs of values should be roughly equal, so if we plot the pairs we would expect to see a line with a slope of about 1 that goes through the origin.

The problem with this procedure is that finding the expected ordered values requires us to know the mean $\mu$ and standard deviation $\sigma$ of the appropriate population. These are generally not available. To avoid this problem, the expectations of the ordered values are computed assuming $\mu = 0$ and $\sigma = 1$. The expected ordered values from this standard normal distribution are called *normal scores* or *rankits*. Computing the expected values this way, we no longer anticipate a line with slope 1 and intercept 0. We now anticipate a line with slope $\sigma$ and intercept $\mu$. While it is possible to obtain estimates of the mean and standard deviation from a normal plot, our primary interest is in whether the plot looks like a line. A linear plot is consistent with normal data; a nonlinear plot is inconsistent with normal data. Christensen (1987, section XIII.2) gives a more detailed motivation for normal plots.

The normal scores are difficult to compute, so we generally get a computer program to do the work. In fact, just creating a plot is considerable work without a computer.

EXAMPLE 2.4.1. Consider the dropout rate data. Figure 2.7 contains the normal plot for the complete data. The two outliers cause the plot to be severely nonlinear. Figure 2.8 contains the normal plot for the dropout rate data with the two outliers deleted. It is certainly not horribly nonlinear. There is a little shoulder at the bottom end and some wiggling in the middle.

We can eliminate the shoulder in this plot by transforming the original data. Figure 2.9 contains a normal plot for the square roots of the data with the outliers deleted. While the plot no longer has a shoulder on the lower end, it seems to be a bit less well behaved in the middle.

We might now repeat our tests and confidence intervals for the 36 observations left when the outliers are deleted. We can do this for either the original data or the square roots of the original data. In either case, it now seems reasonable to treat the data as normal, so we can use a $t(36-1)$

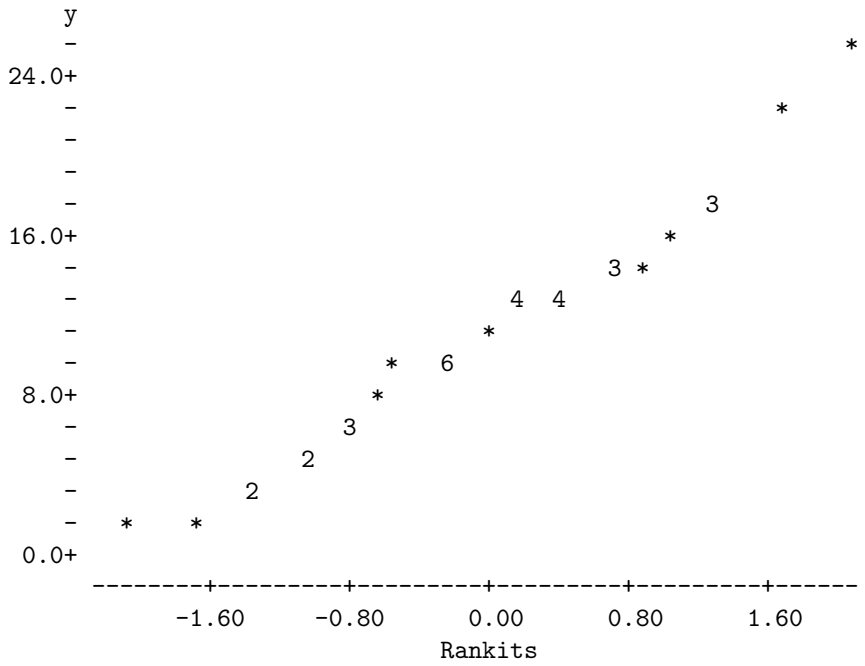



Figure 2.8: *Normal plot for drop rate percentage data: outliers deleted.*

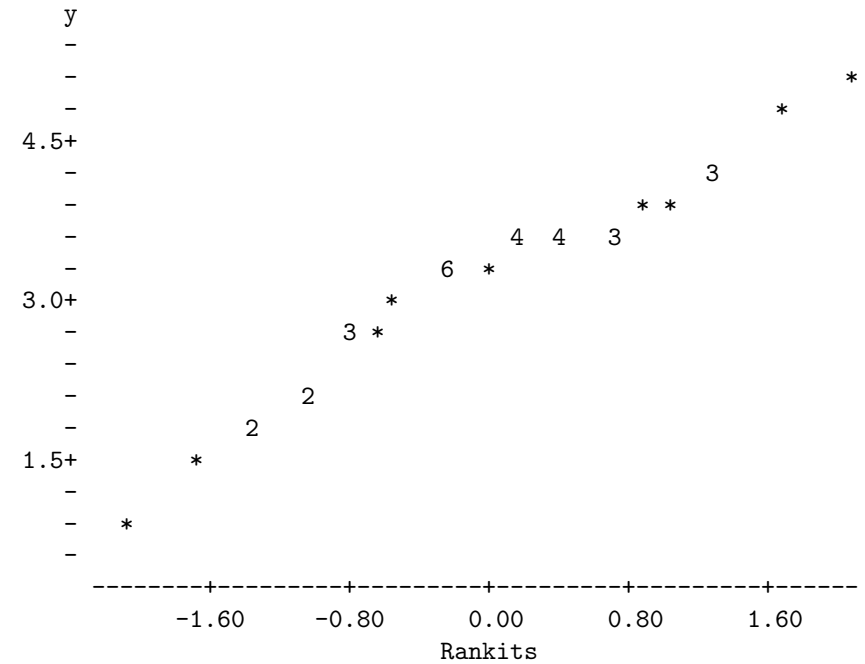



Figure 2.9: *Normal plot for square roots of drop rate percentage data: outliers deleted.*

distribution instead of hoping that the sample is large enough to justify use of the standard normal distribution. We will consider these tests and confidence intervals in the next chapter.

It is important to remember that *if outliers are deleted, the conclusions reached are not valid*

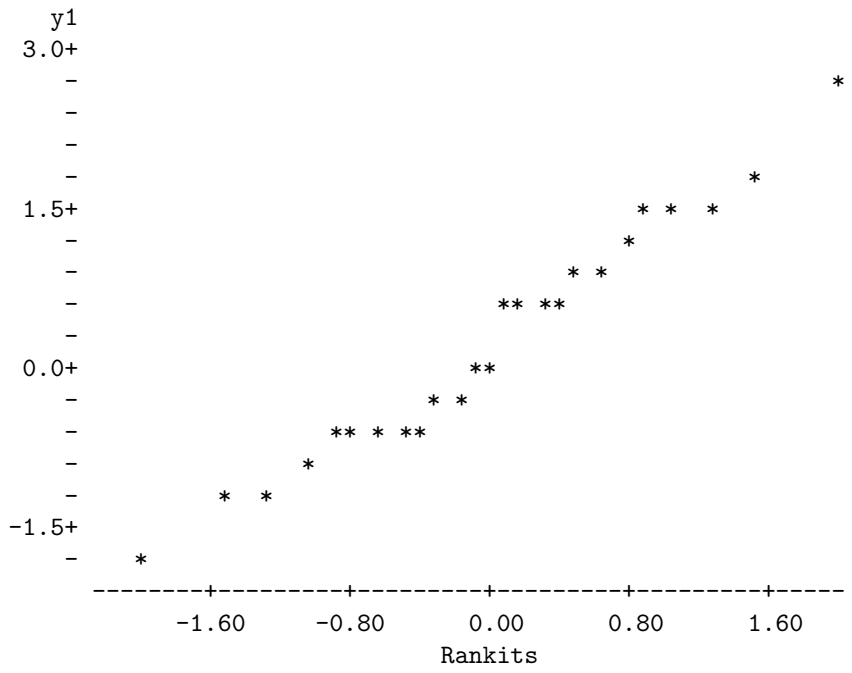



Figure 2.10: *Normal plot.*

*for data containing outliers*. For example, a confidence interval will be for the mean dropout rate excluding the occasional classes with extremely large dropout rates. If we are confident that any deleted outliers are not really part of the population of interest, this causes no problem. Thus, if we were sure that the large dropout rates were the result of clerical errors and did not provide any information about true dropout rates, our conclusions about the population should be based on the data excluding the outliers. More often though, we do not know that outliers are simple mistakes. *Often, outliers are true observations and often they are the most interesting and useful observations in the data*. If the outliers are true observations, systematically deleting them changes both the sample and the population of interest. In this case, the confidence interval is for the mean of a population implicitly defined by the process of deleting outliers. Admittedly, the idea of the mean dropout rate excluding the occasional outliers is not very clearly defined, but remember that the real population of interest is not too clearly defined either. We do not really want to learn about the clearly defined population of 1984–85 dropout rates, we really want to treat the dropout rate data as a sample from a population that allows us to draw useful inferences about current and future dropout rates. If we really cared about the fixed population, we could specify exactly what kinds of observations we would exclude and what we meant by the population mean of the observations that would be included. Given the nature of the true population of interest, I think that such technicalities are more trouble than they are worth at this point. □

Normal plots are subject to random variation because the data used in them are subject to random variation. Typically, normal plots are not perfectly straight. Figures 2.10 through 2.15 present six normal plots for which the data are in fact normally distributed. By comparison to these, Figures 2.8 and 2.9, the normal plots for the dropout rate data and the square root of the dropout rates both with outliers deleted, look reasonably normal. Of course, if the dropout rate data are truly normal, the square root of these data cannot be truly normal and vice versa. However, both are reasonably close to normal distributions.

Figures 2.10 through 2.15 contain normal plots based on 25 observations each. Normal plots

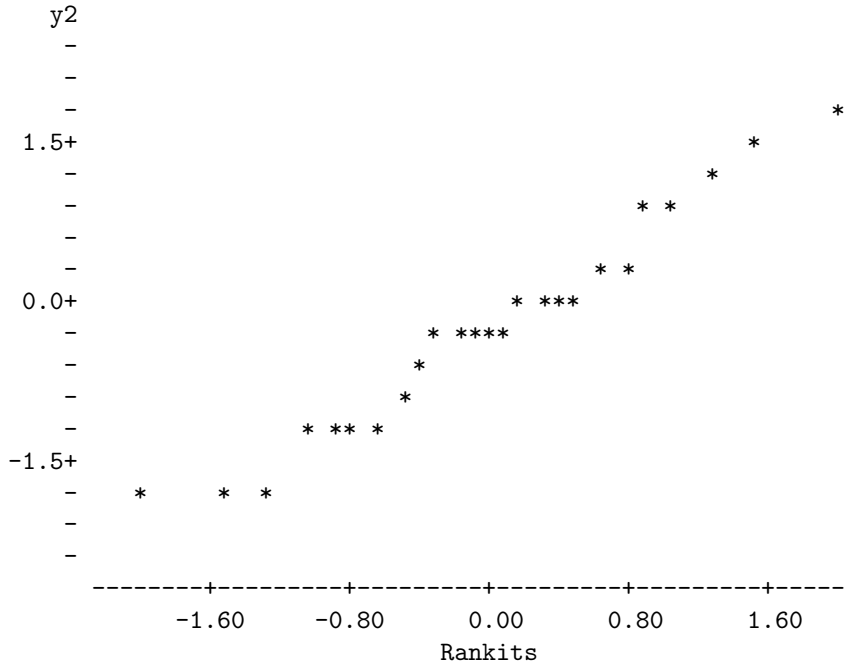



Figure 2.11: *Normal plot.*

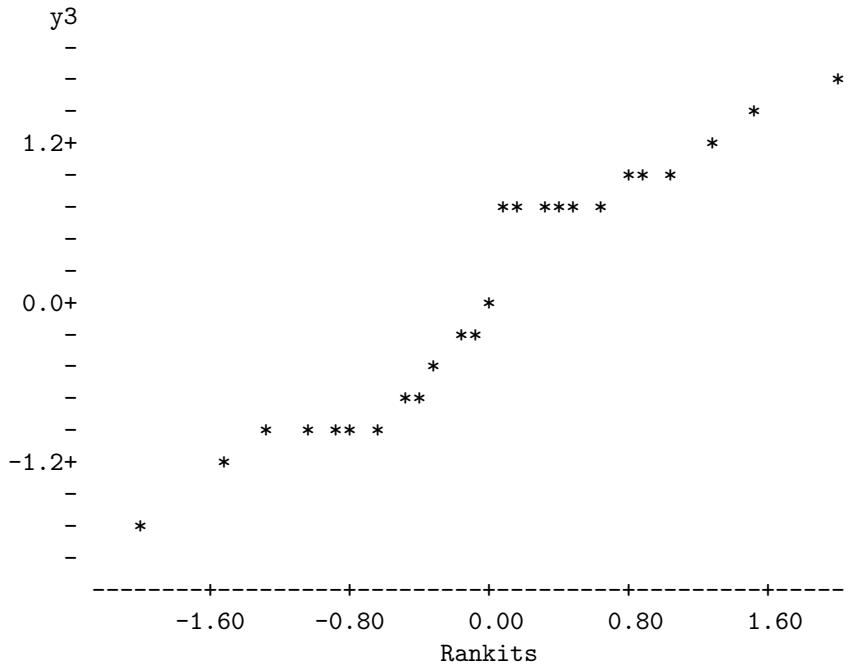



Figure 2.12: *Normal plot.*

based on larger normal samples tend to appear straighter than these. Normal plots based on smaller normal samples can look much more crooked.

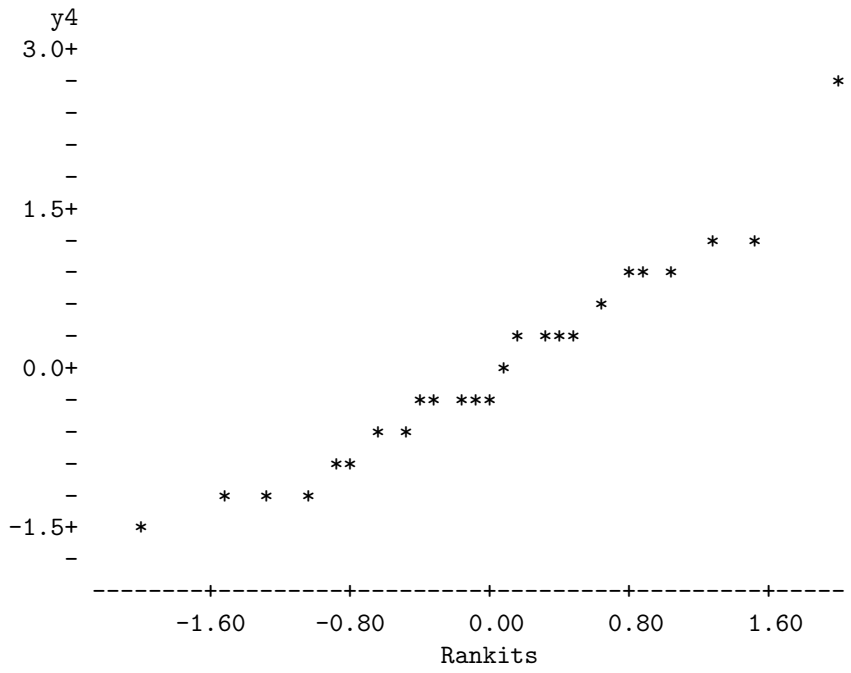



Figure 2.13: *Normal plot.*

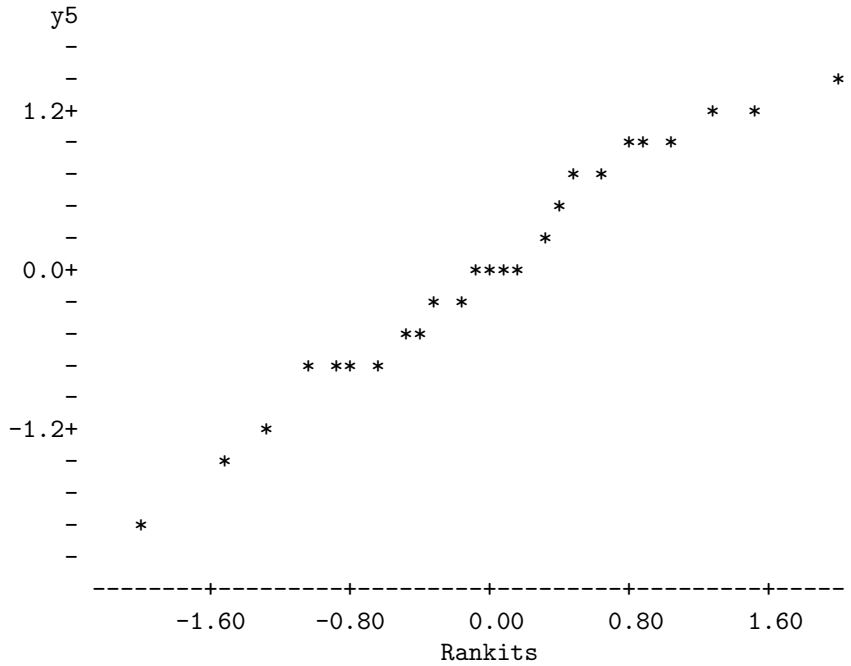



Figure 2.14: *Normal plot.*

*Testing normality*

In an attempt to quantify the straightness of a normal plot, Shapiro and Francia (1972) proposed the summary statistic $W'$, which is the squared sample correlation between the pairs of points in

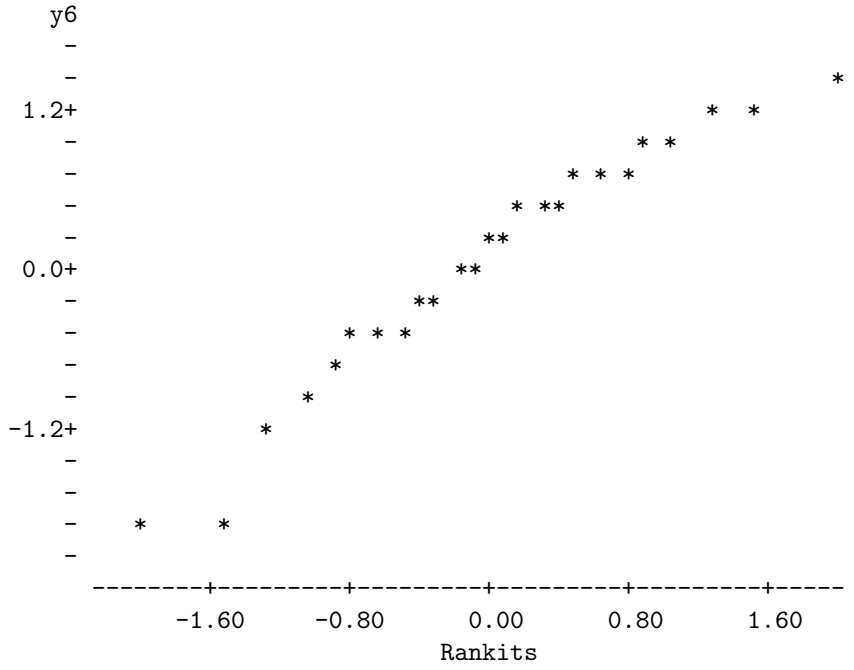



Figure 2.15: *Normal plot.*

the plots. The population correlation coefficient was introduced in Subsection 1.2.3. The sample correlation coefficient is introduced in Chapter 7. At this point, it is sufficient to know that sample correlation coefficients near 0 indicate very little linear relationship between two variables and sample correlation coefficients near 1 or $-1$ indicate a very strong linear relationship. Since you need a computer to get the normal scores (rankits) anyway, just rely on the computer to give you the squared sample correlation coefficient.

A sample correlation coefficient near 1 indicates a strong tendency of one variable to increase (linearly) as the other variable increases and sample correlation coefficients near $-1$ indicate a strong tendency for one variable to decrease (linearly) as the other variable increases. In normal plots we are looking for a strong tendency for one variable, the ordered data, to increase as the other variable, the rankits, increases, so normal data should display a sample correlation coefficient near 1 and thus the square of the sample correlation, $W'$, should be near 1. If $W'$ is too small, it indicates that the data are inconsistent with the assumption of normality. If $W'$ is smaller than, say, 95% of the values one would see from normally distributed data, it is substantial evidence that the data are not normally distributed. If $W'$ is smaller than, say, 99% of the values one would see from normally distributed data, it is strong evidence that the data are not normally distributed. Appendix B.3 presents tables of the values $W'(.05,n)$ and $W'(.01,n)$. These are the points above which fall, respectively, 95% and 99% of the $W'$ values one would see from normally distributed data. Of course the $W'$ percentiles are computed using not only the assumption of normality, but also the assumptions that the observations are independent with the same mean and variance. Note also that the values of these percentiles depend on the sample size $n$. The tabled values are consistent with our earlier observation that the plots are more crooked for smaller numbers of observations and straighter for larger numbers of observations in that the tabled values get larger with $n$. For comparison, we give the observed $W'$ values for the data used in Figures 2.10 through 2.15.

| Shapiro–Francia statistics | |
|---|---|
| Figure | $W'$ |
| 2.10 | 0.966 |
| 2.11 | 0.974 |
| 2.12 | 0.937 |
| 2.13 | 0.956 |
| 2.14 | 0.958 |
| 2.15 | 0.978 |

These should be compared to $W'(.05,25) \doteq .918$ and $W'(.01,25) \doteq .88$ from Appendix B.3. None of these six values is below the 5% point.

EXAMPLE 2.4.2. For the dropout rate data we have three normal plots. The complete, untransformed data yield a $W'$ value of .697. This value is inconsistent with the assumption that the dropout rate data has a normal distribution. Deleting the two outliers, $W'$ is .978 for the untransformed data and .960 for the square roots of the data. The tabled percentiles are $W'(.05,36) = .940$ and $W'(.01,36) = .91$, so the untransformed data and the square root data look alright. In addition, $W'$ was computed for the square roots of the complete data. Its value, .887, is still significantly low, but is a vast improvement over the untransformed complete data. The outliers are not nearly as strange when the square roots of the data are considered. Sometimes it is possible to find a transformation that eliminates outliers. □

*Minitab commands*

A computer program is necessary for finding the normal scores and convenient for plotting the data and computing $W'$. The following Minitab commands provide a normal plot and the $W'$ statistic for a variable in c1.

```
MTB > name c1 'y'
MTB > nscores c1 c2
MTB > plot c1 c2
MTB > corr c1 c2
MTB > note    The correlation is printed out, e.g., .987.
MTB > note    This correlation is used in the next command.
MTB > let k1=.987**2
MTB > note    k1 is W'
MTB > print k1
```

## 2.5 Transformations

In analyzing a collection of numbers, we assume that the observations are a random sample from some population. Often, the population from which the observations come is not as well defined as we might like. For example, if our observations are the yields of corn on 30 one acre plots of ground grown in the summer of 1990, what is the larger population from which this is a sample? Typically, we do not have a large number of one acre plots from which we randomly select 30. Even if we had a large collection of plots, these plots are subject to different weather conditions, have different fertilities, etc. Most importantly, we are rarely interested in corn grown in 1990 for its own sake. If we are studying corn grown in 1990, we are probably interested in predicting how that same type of corn would behave if we planted it at some time in the future. No population that currently exists could be completely appropriate for drawing conclusions about plant growths in a future year. Thus *the assumption that the observations are a random sample from some population is often only a useful approximation*.

When making approximations, it is often necessary to adjust things to make the approximations more accurate. In statistics, *two approximations we frequently make are that all the data have the same variance and that the data are normally distributed. Making numerical transformations of the data is a primary tool for improving the accuracy of these approximations*. When sampling from a fixed population, we are typically interested in transformations that improve the normality assumption because having different variances is not a problem associated with sampling from a fixed population. With a fixed population, the variance of an object is the variance of randomly choosing an object from the population. This is a constant regardless of which object we end up choosing. But data are rarely as simple as random samples from a fixed population. Once we have an object from the population, we have to obtain an observation (measurement or count) from the object. These observations on a given object are also subject to random error and the error may well depend on the specific object being observed.

We now examine the fact that observations often have different variances, depending on the object being observed. First consider taking length measurements using a 30 centimeter ruler that has millimeters marked on it. For measuring objects that are less than 30 centimeters long, like this book, we can make very accurate measurements. We should be able to measure things within half a millimeter. Now consider trying to measure the height of a dog house that is approximately 3.5 feet tall. Using the 30 cm ruler, we measure up from the base, mark 30 cm, measure from the mark up another 30 cm, make another mark, measure from the new mark up another 30 cm, mark again, and finally we measure from the last mark to the top of the house. With all the marking and moving of the ruler, we have much more opportunity for error than we have in measuring the length of the book. Obviously, if we try to measure the height of a house containing two floors, we will have much more error. If we try to measure the height of the Sears tower in Chicago using a 30 cm ruler, we will not only have a lot of error, but large psychiatric expenses as well. The moral of this tale is that, when making measurements, larger objects tend to have more variability. If the objects are about the same size, this causes little or no problem. One can probably measure female heights with approximately the same accuracy for all women in a sample. One probably cannot measure the weights of a large sample of marine animals with constant variability, especially if the sample includes both shrimp and blue whales. *When the observations are the measured* amounts *of something, often the standard deviation of an observation is proportional to its mean. When the standard deviation is proportional to the mean, analyzing the logarithms of the observations is more appropriate than analyzing the original data*.

Now consider the problem of counting up the net financial worth of a sample of people. For simplicity, let's think of just three people, me, my 10 year old son (at least he was 10 when I started writing this), and my rich uncle, Scrooge. In fact, let's just think of having a stack of one dollar bills in front of each person. My pile is of a decent size, my son's is small, and my uncle's is huge. When I count my pile, it is large enough that I could miscount somewhere and make a significant, but not major, error. When I count my son's pile, it is small enough that I should get it about right. When I count my uncle's pile, it is large enough that I will, almost inevitably, make several significant errors. As with measuring amounts of things, the larger the observation, the larger the potential error. However, the process of making these errors is very different than that described for measuring amounts. In such cases, the variance of the observations is often proportional to the mean of the observations. The standard corrective measure for counts is different from the standard corrective measure for amounts. *When the observations are* counts *of something, often the variance of the count is proportional to its mean. In this case, analyzing the square roots of the observations is more appropriate than analyzing the original data*.

Suppose we are looking at yearly sales for a sample of corporations. The sample may include both the corner gas (petrol) station and Exxon. It is difficult to argue that one can really *count* sales for a huge company such as Exxon. In fact, it may be difficult to count even yearly sales for a gas station. Although in theory one should be able to count sales, it may be better to think of yearly sales as measured amounts. It is not clear how to transform such data. Another example is age. We usually think of counting the years a person has been alive, but one could also argue that we are measuring the amount of time a person has been alive. *In practice, we often try both logarithmic and square root transformations and use the transformation that seems to work best*, even when the type of observation (count or amount) seems clear.

Finally, consider the proportion of times people drink a particular brand of soda pop, say, Dr. Pepper. The idea is simply that we ask a group of people what proportion of the time they drink Dr. Pepper. People who always drink Dr. Pepper are aware of that fact and should give a quite accurate proportion. Similarly, people who never drink Dr. Pepper should be able to give an accurate proportion. Moreover, people who drink Dr. Pepper about 90% of the time or about 10% of the time, can probably give a fairly accurate proportion. The people who will have a lot of variability in their replies are those who drink Dr. Pepper about half the time. They will have little idea whether they drink it 50% of the time, or 60%, or 40%, or just what. With observations that are counts or amounts, larger observations have larger variances. With observations that are proportions, observations near

0 and 1 have small variability and observations near .5 have large variability. Proportion data call for a completely different type of transformation. *The standard transformation for proportion data is the inverse sine (arcsine) of the square root of the proportion. When the observations are proportions, often the variance of the proportion is a constant times* $\mu(1-\mu)/N$*, where* $\mu$ *is the mean and N is the number of trials. In this case, analyzing the inverse sine (arcsine) of the square root of the proportion is more appropriate than analyzing the original data.*

In practice, the square root transformation is sometimes used with proportion data. After all, many proportions are obtained as a count divided by the total number of trials. For example, the best data we could get in the Dr. Pepper drinking example would be the count of the number of Dr. Peppers consumed divided by the total number of sodas devoured.

There is a subtle but important point that was glossed over in the previous paragraphs. If we take multiple measurements on a house, the variance depends on the true height, but the true height is the same for all observations. Such a dependence of the variance on the mean causes no problems. The problem arises when we measure a random sample of buildings each with a variance depending on its true height.

EXAMPLE 2.5.1. For the dropout rate data, we earlier considered the complete, untransformed data and after deleting two outliers, we looked at the untransformed data and the square roots of the data. In Examples 2.4.1 and 2.4.2 we saw that the untransformed data with the outliers deleted and the square roots of the data with the outliers deleted had approximate normal distributions. Based on the $W'$ statistic, the untransformed data seemed to be more nearly normal. The data are proportions of people who drop from a class, so our discussion in this section suggests transforming by the inverse sine of the square roots of the proportions. Recall that proportions are values between 0 and 1, while the dropout rates were reported as values between 0 and 100, so the reported rates need to be divided by 100. For the complete data, this transformation yields a $W'$ value of .85, which is much better than the untransformed value of .70, but worse than the value .89 obtained with the square root transformation. With the two outliers deleted, the inverse sine of the square roots of the proportions yields the respectable value $W' = .96$, but the square root transformation is simpler and gives almost the same value, while the untransformed data give a much better value of .98. Examination of the six normal plots (only three of which have been presented here) reinforce the conclusions given above.

With the outliers deleted, it seems reasonable to analyze the untransformed data and, to a lesser extent, the data after either transformation. *Other things being equal*, we prefer using the simplest transformation that seems to work. Simple transformations are easier to explain, justify, and interpret. The square root transformation is simpler, and thus better, than the inverse sine of the square roots of the proportions. Of course, not making a transformation seems to work best and not transforming is always the simplest transformation. Actually some people would point out, and it is undeniably true, that the act of deleting outliers is really a transformation of the data. However, we will not refer to it as such. □

*Minitab commands*

Minitab commands for the three transformations discussed here and for the cubed root power transformation are given below. The cubed root is just to illustrate a general power transformation.

```
MTB > name c1 'y'
MTB > let c2 = loge(c1)
MTB > let c3 = sqrt(c1)
MTB > let c4 = asin(sqrt(c1))
MTB > let c5 = c1**(1/3)
```

*Theory*

The standard transformations given above are referred to as *variance stabilizing transformations*. The idea is that each observation is a look at something with a different mean and variance, where the variance depends on the mean. For example, when we measure the height of a house, the house has some 'true' height and we simply take a measurement of it. The variability of the measurement depends on the true height of the house. Variance stabilizing transformations are designed to eliminate the dependence of the variance on the mean. Although variance stabilizing transformations are used quite generally for counts, amounts, and proportions, they are derived for certain assumptions about the relationship between the mean and the variance. These relationships are tied to theoretical distributions that are appropriate for some counts, amounts, and proportions. Rao (1973, section 6g) gives a nice discussion of the mathematical theory behind variance stabilizing transformations.

Proportions are related to the binomial distribution for the numbers of successes. We have a fixed number of trials; the proportion is the number of successes divided by the number of trials. The mean of a $\text{Bin}(N,p)$ distribution is $Np$ and the variance is $Np(1-p)$. This relationship between the mean and variance of a binomial leads to the inverse sine of the square root transformation.

Counts are related to the Poisson distribution. The Poisson distribution is an approximation used for binomials with a very large number of trials, each having a very small probability of success. Poisson data has the property that the variance equals the mean of the observation. This relationship leads to the square root as the variance stabilizing transformation.

For amounts, the log transformation comes from having the standard deviation proportional to the mean. The standard deviation divided by the mean is called the *coefficient of variation*, so the log transformation is appropriate for observations that have a constant coefficient of variation. (The square root transformation comes from having the variance, rather than the standard deviation, proportional to the mean.) A family of continuous distributions called the gamma distributions has constant coefficient of variation.

The variance stabilizing transformations are given below. In each case we assume $\text{E}(y_i) = \mu_i$ and $\text{Var}(y_i) = \sigma_i^2$. The symbol $\propto$ means 'proportional to.'

Variance stabilizing transformations

| Data | Distribution | Mean, variance relationship | Transformation |
|---|---|---|---|
| Count | Poisson | $\mu_i \propto \sigma_i^2$ | $\sqrt{y_i}$ |
| Amount | Gamma | $\mu_i \propto \sigma_i$ | $\log(y_i)$ |
| Proportion | Binomial$/N$ | $\frac{\mu_i(1-\mu_i)}{N} \propto \sigma_i^2$ | $\sin^{-1}\left(\sqrt{y_i}\right)$ |

I cannot honestly recommend using variance stabilizing transformations to analyze either binomial or Poisson data. In the past 20 years, a large body of statistical techniques has been developed specifically for analyzing binomial and Poisson data, see, for example, Christensen (1990b). I would recommend using these alternative methods. Many people would make a similar recommendation for gamma distributed data citing the applicability of generalized linear model theory, cf. McCullagh and Nelder (1989) or Christensen (1990b, chapter X). When applied to binomial, Poisson, or gamma distributed data, variance stabilizing transformations provide a way to force the methods developed for normally distributed data into giving a reasonable analysis for data that are not normally distributed. If you have a clear idea about the true distribution of the data, you should use methods developed specifically for that distribution. The problem is that we often have little idea of the appropriate distribution for a set of data. For example, if we simply ask people the proportion of times they drink Dr. Pepper, we have proportion data that is not binomial. In such cases, we seek a transformation that will make a normal theory analysis approximately correct. We often pick transformations by trial and error. *The variance stabilizing transformations provide little more than a place to start when considering transformations.*

At the beginning of this section, we mentioned two key approximations that we frequently make.

These are that all the data have the same variance and that the data are normally distributed. While the rationale given above for picking transformations was based on stabilizing variances, in practice we typically choose a transformation for a single sample to attain approximate normality. To evaluate whether a transformation really stabilizes the variance, we need more information than is contained in a single sample. Control chart methods can be used to evaluate variance stabilization for a single sample, cf. Shewhart (1931). Those methods require formation of rational subgroups and that requires additional information. We could also plot the sample against appropriately chosen variables to check variance stabilization, but finding appropriate variables can be quite difficult and would depend on properties of the particular sampling process. Variance stabilizing transformations are probably best suited to problems that compare samples from several populations, where the variance in each population depends on the mean of the population.

On the other hand, we already have examined methods for evaluating the normality of a single sample. Thus, since we cannot (actually, do not) evaluate variance stabilization in a single sample, if we think that the variance of observations should increase with their mean, we might try both the log and square root transformations and pick the one for which the transformed data best approximate normality.

## 2.6 Inference about $\sigma^2$

If the data are normally distributed, we can also perform confidence intervals and tests for the population variance $\sigma^2$. While these are not typically of primary importance, they can be useful. They also tend to be sensitive to the assumption of normality. The procedures do not follow the same pattern used for most inferences that involve 1) a parameter of interest, 2) an estimate of the parameter, 3) the standard error of the estimate, and 4) a known distribution symmetric about zero; however, there are similarities. Procedures for variances typically require a parameter, an estimate, and a known distribution.

The procedures discussed in this section actually apply to all the problems in this book that involve a single variance parameter $\sigma^2$. One need only substitute the relevant estimate of $\sigma^2$ and use its degrees of freedom. Applications to the data and models considered in Chapter 12 are not quite as straightforward because there the models involve more than one variance.

In the one-sample problem, the parameter is $\sigma^2$, the estimate is $s^2$, and the distribution, as discussed in equation (2.1.5), is

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

The notation $\chi^2(1-\alpha, n-1)$ is used to denote the point that cuts off the bottom $1-\alpha$ (top $\alpha$) of the $\chi^2$ distribution with $n-1$ degrees of freedom. Note that $(n-1)s^2/\sigma^2$ is nonnegative, so the curve in Figure 2.16 illustrating the $\chi^2$ distribution is also nonnegative. Figure 2.16 shows a central interval with probability $1-\alpha$ for a $\chi^2$ distribution.

A $(1-\alpha)100\%$ confidence interval for $\sigma^2$ is based on the following equality:

$$\begin{aligned} 1-\alpha &= \Pr\left[\chi^2\Big(\frac{\alpha}{2}, n-1\Big) < \frac{(n-1)s^2}{\sigma^2} < \chi^2\Big(1-\frac{\alpha}{2}, n-1\Big)\right] \\ &= \Pr\left[\frac{(n-1)s^2}{\chi^2\big(1-\frac{\alpha}{2}, n-1\big)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2\big(\frac{\alpha}{2}, n-1\big)}\right]. \end{aligned} \tag{2.6.1}$$

The first equality corresponds to Figure 2.16 and is just the definition of the percentage points $\chi^2\big(\frac{\alpha}{2}, n-1\big)$ and $\chi^2\big(1-\frac{\alpha}{2}, n-1\big)$. These are defined to be the points that cut out the middle $1-\alpha$ of the chi-squared distribution and are tabled in Appendix B.2. The second equality in (2.6.1) is based on algebraic manipulation of the terms in the square brackets. The actual derivation is given later in this section. The second equality gives an interval that contains $\sigma^2$. There is a probability of

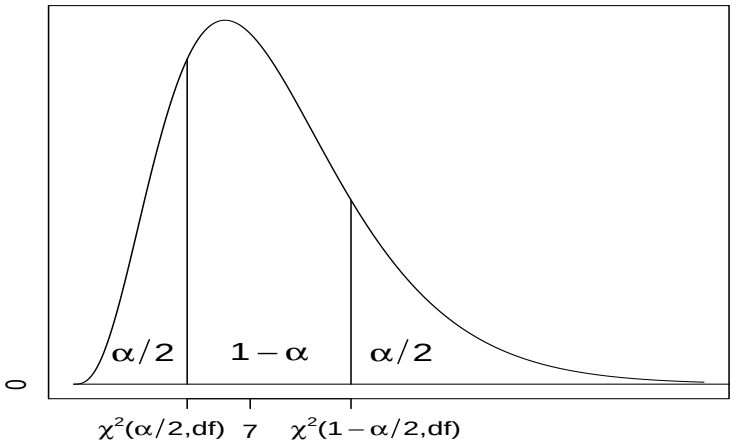



Figure 2.16: *Central $\chi^2$ interval with probability $1-\alpha$.*

$1-\alpha$ that $\sigma^2$ is *going to be* in the interval

$$\left(\frac{(n-1)s^2}{\chi^2\left(1-\frac{\alpha}{2},n-1\right)},\frac{(n-1)s^2}{\chi^2\left(\frac{\alpha}{2},n-1\right)}\right). \tag{2.6.2}$$

*The derivation of the confidence interval for $\sigma^2$ requires the data to be normally distributed. This assumption is more vital for inferences about $\sigma^2$ than it is for inferences about $\mu$.* For inferences about $\mu$, the central limit theorem indicates that the sample means are approximately normal even when the data are not normal. There is no similar result indicating that the sample variance is approximately $\chi^2$ even when the data are not normal.

EXAMPLE 2.6.1. Consider again the dropout rate data. We have seen that the complete data are not normal, but that after deleting the two outliers, the remaining data are reasonably normal. We find a 95% confidence interval for $\sigma^2$ from the deleted data. The deleted data contain 36 observations and $s^2$ for the deleted data is 27.45. The percentage points for the $\chi^2(36-1)$ distribution are $\chi^2(.025,35)=20.57$ and $\chi^2(.975,35)=53.20$. Applying (2.6.2), the 95% confidence interval is

$$\left(\frac{35(27.45)}{53.20},\frac{35(27.45)}{20.57},\right)$$

or equivalently $(18.1,46.7)$. We are 95% confident that the true variance is between 18.1 and 46.7, but remember that this is the true variance *after the deletion of outliers*. Again, when we delete outliers we are a little fuzzy about the exact definition of our parameter, but we are also being fuzzy about the exact population of interest. The exception to this is when we believe that the only outliers that exist are observations that are not really part of the population. □

It is the endpoints of the interval (2.6.2) that are random. To use the interval, we replace the random variable $s^2$ with the observed value of $s^2$ and replace the term 'probability $(1-\alpha)$' with '$(1-\alpha)100\%$ confidence.' *Once the observed value of $s^2$ is substituted into the interval, nothing about the interval is random any longer*, the fixed unknown value of $\sigma^2$ is either in the interval or it

is not; there is no probability associated with it. The probability statement about random variables is mystically transformed into a 'confidence' statement. This is not unreasonable, but the rationale is, to say the least, murky.

The $\alpha$ level test of $H_0 : \sigma^2 = \sigma_0^2$ versus $H_A : \sigma^2 \neq \sigma_0^2$ is again based on the first equality in equation (2.6.1). To actually perform a test, $\sigma_0^2$ *must be a known value*. As usual, we assume that the null hypothesis is true, i.e., $\sigma^2 = \sigma_0^2$, so under this assumption

$$1-\alpha = \Pr\left[\chi^2\left(\frac{\alpha}{2}, n-1\right) < \frac{(n-1)s^2}{\sigma_0^2} < \chi^2\left(1-\frac{\alpha}{2}, n-1\right)\right].$$

If we observe data yielding an $s^2$ such that $(n-1)s^2/\sigma_0^2$ is between the values $\chi^2\left(\frac{\alpha}{2}, n-1\right)$ and $\chi^2\left(1-\frac{\alpha}{2}, n-1\right)$, the data are consistent with the assumption that $\sigma^2 = \sigma_0^2$ at level $\alpha$. Conversely, we reject $H_0 : \sigma^2 = \sigma_0^2$ with a two-sided $\alpha$ level test if

$$\frac{(n-1)s^2}{\sigma_0^2} > \chi^2\left(1-\frac{\alpha}{2}, n-1\right)$$

or if

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi^2\left(\frac{\alpha}{2}, n-1\right).$$

A clear definition of 'confidence' can be given in terms of testing the hypothesis $H_0 : \sigma^2 = \sigma_0^2$ versus the alternative $H_A : \sigma^2 \neq \sigma_0^2$. The same algebraic manipulations that lead to equation (2.6.1) can be used to show that the $(1-\alpha)100\%$ confidence interval contains precisely those values of $\sigma_0^2$ that are consistent with the data when testing $H_0 : \sigma^2 = \sigma_0^2$ at level $\alpha$. This idea is discussed in more detail in Section 3.4.

EXAMPLE 2.6.2. For the dropout rate data consider testing $H_0 : \sigma^2 = 50$ versus $H_A : \sigma^2 \neq 50$ with $\alpha = .01$. Again, we use the data with the two outliers deleted, so our concept of the population variance $\sigma^2$ must account for our deletion of weird cases. The test statistic is

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{35(27.45)}{50} = 19.215.$$

The *critical region*, the region for which we reject $H_0$, contains all values greater than $\chi^2(.995, 35) = 60.275$ and all values less than $\chi^2(.005, 35) = 17.19$. The test statistic is certainly not greater than 60.275 and it is also not less than 17.19, so we have no basis for rejecting the null hypothesis at the $\alpha = .01$ level. At the .01 level, the data are consistent with the claim that $\sigma^2 = 50$.

The 95% confidence interval $(18.1, 46.7)$ from Example 2.6.1 contains all values of $\sigma^2$ that are consistent with the data as determined by a two-sided $\alpha = .05$ level test. The interval does not contain 50, so we do have evidence against $H_0 : \sigma^2 = 50$ at the $\alpha = .05$ level. □

While methods for drawing inferences about variances *do not* fit our standard pattern based on 1) a parameter of interest, 2) an estimate of the parameter, 3) the standard error of the estimate, and 4) a known distribution symmetric about zero, it should be noted that the basic logic behind these confidence intervals and tests is the same. Confidence intervals are based on a random interval that contains the parameter of interest with some specified probability. The unusable random interval is changed into a usable nonrandom interval by substituting the observed value of the random variable into the endpoints of the interval. The probability is then miraculously, if intuitively, turned into 'confidence.' Tests of hypotheses are based on evaluating whether the data are consistent with the null hypothesis. Consistency is defined in terms of a known distribution that applies when the null hypothesis is true. If the data are inconsistent with the null hypothesis, the null hypothesis is rejected as being inconsistent with the observed data.

Table 2.1: *Weights of rats*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 59 | 54 | 56 | 59 | 57 | 52 | 52 | 61 | 59 |
| 53 | 59 | 51 | 51 | 56 | 58 | 46 | 53 | 57 |
| 60 | 52 | 49 | 56 | 46 | 51 | 63 | 49 | 57 |

Below is a series of equalities that justify equation (2.6.1).

$$\begin{aligned}
1-\alpha &= \Pr\left[\chi^2\Big(\frac{\alpha}{2},n-1\Big) < \frac{(n-1)s^2}{\sigma^2} < \chi^2\Big(1-\frac{\alpha}{2},n-1\Big)\right] \\
&= \Pr\left[\frac{1}{\chi^2\big(\frac{\alpha}{2},n-1\big)} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{\chi^2\big(1-\frac{\alpha}{2},n-1\big)}\right] \\
&= \Pr\left[\frac{1}{\chi^2\big(1-\frac{\alpha}{2},n-1\big)} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi^2\big(\frac{\alpha}{2},n-1\big)}\right] \\
&= \Pr\left[\frac{(n-1)s^2}{\chi^2\big(1-\frac{\alpha}{2},n-1\big)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2\big(\frac{\alpha}{2},n-1\big)}\right].
\end{aligned}$$

## 2.7 Exercises

EXERCISE 2.7.1. Mulrow et al. (1988) presented data on the melting temperature of biphenyl as measured on a differential scanning calorimeter. The data are given below; they are the observed melting temperatures in Kelvin less 340.

$$3.02, 2.36, 3.35, 3.13, 3.33, 3.67, 3.54, 3.11, 3.31, 3.41, 3.84, 3.27, 3.28, 3.30$$

Compute the sample mean, variance, and standard deviation. Give a 99% confidence interval for the population mean melting temperature of biphenyl as measured by this machine. (Note that we don't know whether the calorimeter is accurately calibrated.)

EXERCISE 2.7.2. Box (1950) gave data on the weights of rats that were about to be used in an experiment. The data are repeated in Table 2.1. Assuming that these are a random sample from a broader population of rats, give a 95% confidence interval for the population mean weight. Test the null hypothesis that the population mean weight is 60 using a .01 level test.

EXERCISE 2.7.3. Fuchs and Kenett (1987) presented data on citrus juice for fruits grown during a specific season at a specific location. The sample size was 80 but many variables were measured on each sample. Sample statistics for some of these variables are given below

| Variable | BX | AC | SUG | K | FORM | PECT |
|---|---|---|---|---|---|---|
| Mean | 10.4 | 1.3 | 7.7 | 1180.0 | 22.2 | 451.0 |
| Variance | 0.38 | 0.036 | 0.260 | 43590.364 | 6.529 | 16553.996 |

The variables are BX – total soluble solids produced at 20°C, AC – acidity as citric acid unhydrons, SUG – total sugars after inversion, K – potassium, FORM – formol number, PECT – total pectin. Give a 99% confidence interval for the population mean of each variable. Give a 99% prediction interval for each variable. Test whether the mean of BX equals 10. Test whether the mean of SUG is less than or equal to 7.5. Use $\alpha = .01$ for each test.

EXERCISE 2.7.4. Jolicoeur and Mosimann (1960) gave data on female painted turtle shell

Table 2.2: *Female painted turtle shell lengths*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 98 | 138 | 123 | 155 | 105 | 147 | 133 | 159 |
| 103 | 138 | 133 | 155 | 109 | 149 | 134 | 162 |
| 103 | 141 | 133 | 158 | 123 | 153 | 136 | 177 |

Table 2.3: *Percentage of fathers with white collar jobs*

| | | | | |
|---|---|---|---|---|
| 28.87 | 20.10 | 69.05 | 65.40 | 29.59 |
| 44.82 | 77.37 | 24.67 | 65.01 | 9.99 |
| 12.20 | 22.55 | 14.30 | 31.79 | 11.60 |
| 68.47 | 42.64 | 16.70 | 86.27 | 76.73 |

lengths. The data are presented in Table 2.2. Give a 95% confidence interval for the population mean length. Give a 99% prediction interval for the shell length of a new female.

EXERCISE 2.7.5. Mosteller and Tukey (1977) extracted data from the *Coleman Report*. Among the variables considered was the percentage of sixth-graders who's fathers were employed in white collar jobs. Data for 20 New England schools are given in Table 2.3. Are the data reasonably normal? Do any of the standard transformations improve the normality? After finding an appropriate transformation (if necessary), test the null hypothesis that the percentage of white collar fathers is 50%. Use a .05 level test. Give a 99% confidence interval for the percentage of fathers with white collar jobs. If a transformation was needed, relate your conclusions back to the original measurement scale.

EXERCISE 2.7.6. Give a 95% confidence interval for the population variance associated with the data of Exercise 2.7.5. Remember that inferences about variances require the assumption of normality. Could the variance reasonably be 10?

EXERCISE 2.7.7. Give a 95% confidence interval for the population variance associated with the data of Exercise 2.7.4. Remember that the inferences about variances require the assumption of normality.

EXERCISE 2.7.8. Give 99% confidence intervals for the population variances of all the variables in Exercise 2.7.3. Assume that the original data were normally distributed. Using $\alpha = .01$, test whether the potassium variance could reasonably be 45000. Could the formol number variance be 8?

EXERCISE 2.7.9. Shewhart (1931, p. 62) reproduces Millikan's data on the charge of an election. These are repeated in Table 2.4. Check for outliers and nonnormality. Adjust the data appropriately if there are any problems. Give a 98% confidence interval for the population mean value. Give a 98% prediction interval for a new measurement. (Millikan argued that some adjustments were needed before these data could be used in an optimal fashion but we will ignore his suggestions.)

EXERCISE 2.7.10. Show that if $y, y_1, \ldots, y_n$ are independent $N(\mu, \sigma^2)$ random variables, $(y - \bar{y}_\cdot)/\sqrt{\sigma^2 + \sigma^2/n} \sim N(0,1)$. Recalling that $y$, $\bar{y}_\cdot$, and $s^2$ are independent and that $(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$, use Definition 2.1.3 to show that $(y - \bar{y}_\cdot)/\sqrt{s^2 + s^2/n} \sim t(n-1)$.

Table 2.4: *Observations on the charge of an electron*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4.781 | 4.764 | 4.777 | 4.809 | 4.761 | 4.769 | 4.795 | 4.776 |
| 4.765 | 4.790 | 4.792 | 4.806 | 4.769 | 4.771 | 4.785 | 4.779 |
| 4.758 | 4.779 | 4.792 | 4.789 | 4.805 | 4.788 | 4.764 | 4.785 |
| 4.779 | 4.772 | 4.768 | 4.772 | 4.810 | 4.790 | 4.775 | 4.789 |
| 4.801 | 4.791 | 4.799 | 4.777 | 4.772 | 4.764 | 4.785 | 4.788 |
| 4.779 | 4.749 | 4.791 | 4.774 | 4.783 | 4.783 | 4.797 | 4.781 |
| 4.782 | 4.778 | 4.808 | 4.740 | 4.790 | 4.767 | 4.791 | 4.771 |
| 4.775 | 4.747 | | | | | | |

Chapter 3

# A general theory for testing and confidence intervals

The most commonly used statistical tests and confidence intervals derive from a single theory. (Tests and confidence intervals about variances are an exception.) The basic ideas of this theory were illustrated in Chapter 2. The point of the current chapter is to present the theory in its general form and to reemphasize fundamental techniques. The general theory will then be used throughout the book. Because the theory is stated in quite general terms, some prior familiarity with the ideas, e.g., reading Sections 2.2 and 2.3, is highly recommended.

To use the general theory you need to know four things:

1. the parameter of interest, $Par$,
2. the estimate of the parameter, $Est$,
3. the standard error of the estimate, $\mathrm{SE}(Est)$, and
4. the appropriate reference distribution.

Specifically, what you need to know about the distribution is that

$$\frac{Est - Par}{\mathrm{SE}(Est)}$$

has a known (tabled) distribution that is symmetric about zero. The estimate $Est$ is taken to be a random variable. The standard error, $\mathrm{SE}(Est)$, is the standard deviation of the estimate if that is known, but more commonly it is an estimate of the standard deviation. If the $\mathrm{SE}(Est)$ is estimated, the known distribution is usually the $t$ distribution with some known number of degrees of freedom. If the $\mathrm{SE}(Est)$ is known, then the distribution is usually the standard normal distribution, i.e., mean 0, variance 1. In some problems, e.g., problems involving the binomial distribution, the central limit theorem is used to get an approximate distribution and inferences proceed as if that distribution is correct. When appealing to the central limit theorem, the known distribution is the standard normal.

Identifying a parameter of interest and an estimate of that parameter is relatively easy. The more complicated part of the procedure is obtaining the standard error. To do this, one typically derives the variance, estimates it (if necessary), and takes the square root. Obviously, rules for deriving variances play an important role in the process.

We need notation for the percentage points of the known reference distribution. In particular, we need a name for the point that cuts off the top $\alpha$ of the distribution. The point that cuts off the top $\alpha$ of the distribution also cuts off the bottom $1-\alpha$ of the distribution. These ideas are illustrated in Figure 3.1. The notation $K(1-\alpha)$ is used for the point that cuts off the top $\alpha$.

The illustration in Figure 3.1 is written formally as

$$\Pr\left[\frac{Est - Par}{\mathrm{SE}(Est)} > K(1-\alpha)\right] = \alpha.$$

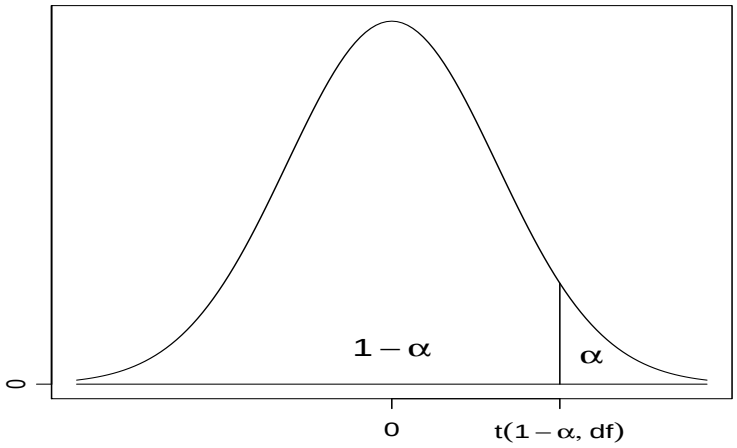



Figure 3.1: *Percentiles of $t(df)$ distributions.*

By the symmetry about zero we also have

$$\Pr\left[\frac{Est - Par}{\mathrm{SE}(Est)} < -K(1-\alpha)\right] = \alpha.$$

The value $K(1-\alpha)$ is called a percentile or percentage point; it is most often found from either a standard normal table or a $t$ table. For $t$ percentage points with $df$ degrees of freedom, we use the notation

$$t(1-\alpha, df) = K(1-\alpha)$$

and for standard normal percentage points we use

$$z(1-\alpha) = K(1-\alpha).$$

As the degrees of freedom get arbitrarily large, the $t$ distribution approximates the standard normal distribution. Thus we write

$$z(1-\alpha) = t(1-\alpha, \infty).$$

One can get a feeling for the quality of this approximation simply by examining the $t$ tables in Appendix B.1 and noting how quickly the $t$ percentiles approach the values given for infinite degrees of freedom.

### 3.1 Theory for confidence intervals

Confidence intervals are interval estimates of the parameter of interest. We have a specified 'confidence' that the parameter is in the interval. Confidence intervals are more valuable than simply reporting the estimate *Est* because confidence intervals provide an idea of the amount of error associated with the estimate.

A $(1-\alpha)100\%$ confidence interval for *Par* is based on the following probability equalities

$$\begin{aligned} 1-\alpha &= \Pr\left[-K\left(1-\frac{\alpha}{2}\right) < \frac{Est - Par}{\mathrm{SE}(Est)} < K\left(1-\frac{\alpha}{2}\right)\right] \\ &= \Pr\left[Est - K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est) < Par < Est + K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est)\right] \end{aligned} \tag{3.1.1}$$

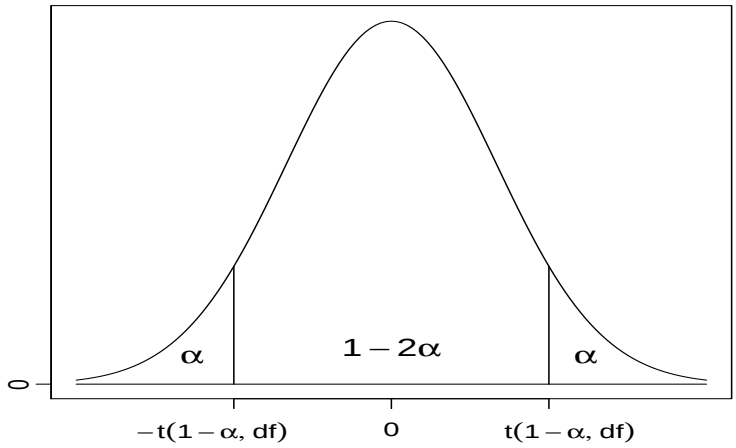



Figure 3.2: *Symmetry about 0 in the distribution of* $[Est - Par]/\mathrm{SE}(Est)$.

The first equality in (3.1.1) is simply a statement of the picture illustrated in Figure 3.2. It follows from the definition of $K\left(1-\frac{\alpha}{2}\right)$ and the symmetry of the distribution. The second equality follows from the fact that the statements within the two sets of square brackets are algebraically equivalent. *A proof of the equivalence is given in the appendix at the end of the chapter.*

The probability statement

$$1-\alpha = \Pr\left[Est - K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est) < Par < Est + K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est)\right].$$

is the basis for the confidence interval for *Par*. The $(1-\alpha)100\%$ confidence interval for *Par* is simply the interval within the square brackets, i.e., the points between $Est - K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est)$ and $Est + K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est)$. However, the confidence interval is obtained by substituting observed values for *Est* and $\mathrm{SE}(Est)$. We are $(1-\alpha)100\%$ 'confident' that *Par* is in this interval. The endpoints of the interval can be written succinctly as

$$Est \pm K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est).$$

I think everyone would agree with the statement 'The probability is $1-\alpha$ that you are *going to get* a confidence interval that covers what you are trying to estimate, *Par*.' I did not indicate that the probability that your actual interval covers *Par* is $1-\alpha$. The particular interval that you get uses the observed values of *Est* and $\mathrm{SE}(Est)$, so it is a fixed interval and either covers *Par* or does not. There is no probability associated with *Par* being in the interval. For this reason the result of a confidence interval is described as, 'We are $(1-\alpha)100\%$ *confident* that the true value of *Par* is in the interval.' I have no idea what this is supposed to mean, even though I find it intuitively appealing. I do, however, know of two acceptable interpretations for confidence intervals. As we will see in Section 3.4, a confidence interval contains all those parameter values that are consistent with the data. Consistency is measured by performing a statistical test with a specified error level $\alpha$. The $\alpha$ in the test plays the same role as the $\alpha$ in a confidence interval. Since I think I understand the philosophical basis of hypothesis tests, I am comfortable with this interpretation.

The confidence intervals obtained from the theory presented in this chapter can frequently be obtained by another approach using 'Bayesian' arguments. In the Bayesian justification, the *correct*

interpretation of a 95% confidence interval is that *the probability is 95% that the parameter is in the interval*. This is precisely the interpretation that most statistics students wish to adopt and that many statisticians strive so hard and so unsuccessfully to make their students reject. We will return to the issue of interpreting confidence intervals later in this section.

EXAMPLE 3.1.1. Years ago, 10 people were independently abducted by S.P.E.C.T.R.E after a Van Holland concert and forced to submit to psychological testing. Among the tests was a measure of audio acuity. From many past abductions in other circumstances, S.P.E.C.T.R.E knows that such observations form a normal population with variance 6. In this case, they found that $\bar{y}_\cdot$ was 17. They seek a 95% confidence interval for $\mu$, the mean of the population.

1) $Par = \mu$,
2) $Est = \bar{y}_\cdot$,
3) $\mathrm{SE}(Est) = \sqrt{6/10}$, in this case $\mathrm{SE}(Est)$ is known and not estimated.
4) $[Est - Par]/\mathrm{SE}(Est) = [\bar{y}_\cdot - \mu]/\sqrt{6/10}$ has a standard normal distribution.

To find the appropriate tabled values, observe that $(1-\alpha)100 = 95$, so $1-\alpha = .95$ and $\alpha = .05$. It follows that $K\big(1-\frac{\alpha}{2}\big) = K(.975) = z(.975) = 1.96$.

The limits of the 95% confidence interval are

$$\bar{y}_\cdot \pm 1.96\sqrt{6/10}$$

or, since $\bar{y}_\cdot = 17$,

$$17 \pm 1.96\sqrt{6/10}.$$

S.P.E.C.T.R.E. was 95% confident that the mean hearing score for people at this concert (or at least for the population they were considering for abduction) was between 15.5 and 18.5. □

EXAMPLE 3.1.2. In Chapter 2 we considered data on dropout rates for math classes. We found that the 38 observations on dropout rates were not normally distributed; they contained two outliers. Our parameter for these data is $\mu$, the population mean dropout rate for math classes, the estimate is the sample mean $\bar{y}_\cdot$, and the standard error is $\sqrt{s^2/38}$ where $s^2$ is the sample variance. Based on the central limit theorem and the law of large numbers, we used the *approximate* reference distribution

$$\frac{\bar{y}_\cdot - \mu}{\sqrt{s^2/38}} \sim N(0,1).$$

From the 38 observations, we computed $\bar{y}_\cdot = 13.11$ and $s^2 = 106.421$ and found a 95% confidence interval for the dropout rate of $(9.8, 16.4)$. The endpoints of the confidence interval are computed as

$$13.11 \pm 1.96(\sqrt{106.421/38}).$$

If we drop the two outliers, the remaining data seem to be normally distributed. Recomputing the sample mean and sample variance with the outliers deleted we get $\bar{y}_d = 11.083$ and $s_d^2 = 27.45$. Here the subscripts $d$ are used as a reminder that the outliers have been deleted. Without the outliers, we can use the reference distribution

$$\frac{\bar{y}_d - \mu_d}{\sqrt{s_d^2/36}} \sim t(35).$$

The $t$ distribution relies on the assumption of normality (which we have validated) rather than relying on the unvalidated large sample approximations from the central limit theorem and law of large numbers. The $t$ distribution should give more accurate results. For a 95% confidence interval based on the data without the outliers, we need to find the appropriate tabled values. Observe once

again that $(1-\alpha)100 = 95$, so $1-\alpha = .95$ and $\alpha = .05$. It follows that $K\left(1-\frac{\alpha}{2}\right) = K(.975) = t(.975,35) = 2.030$ and the confidence interval has endpoints

$$11.083 \pm 2.030(\sqrt{27.45/36}).$$

The actual interval is $(9.3, 12.9)$. *Excluding the extremely high values that occasionally occur*, we are 95% confident that the population mean dropout rate is between 9.3 and 12.9 percent. Remember, this is a confidence interval for the mean of math classes; it *does not* indicate that you can be 95% confident that your next math class will have a dropout rate between 9.3 and 12.9 percent. Such an inference requires a prediction interval. The interval $(9.3, 12.9)$ is much narrower than the one given in the previous paragraph, largely because our estimate of the variance is much smaller when the outliers have been deleted. Note also that with the outliers deleted, we are drawing inferences about a different parameter than when they are present. With the outliers deleted, our conclusions are only valid for the bulk of the observations. While occasional weird observations can be eliminated from our analysis, we cannot stop them from occurring.

In constructing the confidence interval we used the tabled value of 2.030 from the $t$ distribution. This is larger than the 1.96 we obtained earlier from the standard normal distribution. Using the larger $t$ value makes our confidence intervals wider. Other things being equal, we prefer narrower confidence intervals because they make more precise statements about the location of the mean. However, even though the value 1.96 is smaller than 2.030 and thus gives narrower intervals, we prefer to use the $t$ distribution. The $t$ distribution incorporates the fact that we do not know $\sigma^2$ and must estimate it. Thus an analysis using the $N(0,1)$ distribution is much cruder in that it treats the estimate of $\sigma^2$ as if it were really $\sigma^2$. *Whenever we can establish that the data are reasonably normal, we will use the t distribution because it should give more accurate results*.

In the previous chapter we discussed the use of transformations. In particular, we looked at the square roots of the dropout rate data. *We now consider the effect on confidence intervals of transforming the data*. With the two outliers deleted and taking square roots of the observations, we found earlier that the data are reasonably normal. The sample mean and variance of the transformed, deleted data are $\bar{y}_{rd} = 3.218$ and $s^2_{rd} = .749574$. Here the subscript $r$ reminds us that square *roots* have been taken and the subscript $d$ reminds us that outliers have been deleted. Using the reference distribution

$$\frac{\bar{y}_{rd} - \mu_{rd}}{\sqrt{s^2_{rd}/36}} \sim t(35),$$

we obtain a 95% confidence interval with endpoints

$$3.218 \pm 2.030\left(\sqrt{\frac{.749574}{36}}\right).$$

The confidence interval reduces to $(2.925, 3.511)$. This is a 95% confidence interval for the population mean of the square roots of the dropout rate percentages with 'outliers' removed from the population.

The confidence interval $(2.925, 3.511)$ does not really address the issue that we set out to investigate. We wanted some idea of the value of the population mean dropout rate. We have obtained a 95% confidence interval for the population mean of the square roots of the dropout rate percentages (with outliers removed from the population). There is no simple, direct relationship between the population mean dropout rate and the population mean of the square roots of the dropout rate percentages, but a simple device can be used to draw conclusions about typical values for mean dropout rates when the analysis is performed on the square roots of the dropout rates. Since $(2.925, 3.511)$ provides a 95% confidence interval from the *square roots* of the dropout rate percentages, we simply *square* all the values in the interval to draw conclusions about the dropout rate percentages. Squaring the endpoints of the interval gives the new interval $(2.925^2, 3.511^2) = (8.6, 12.3)$. We are now

95% confident that the *central value* of the population of dropout rates is between 8.6 and 12.3. The central value referred to here is really the square of the population mean of the square roots of the dropout rate percentages. We are using this central value as a surrogate for the population mean of the (outlier deleted) dropout rate percentages; generally this central value will not equal the mean of the (deleted) dropout rates. For the most part we ignore the difference between the surrogate and the parameter that we set out to investigate. Interestingly, we will see in Section 3.5 that prediction intervals do not share these difficulties associated with transforming the data.

Note that the retransformed interval $(8.6, 12.3)$ obtained from the transformed, deleted data is similar to the interval $(9.3, 12.9)$ obtained earlier from the untransformed data with the outliers deleted. When, as in this case, two distinct analyses both seem reasonably valid, I would be very hesitant about drawing practical conclusions that could not be justified from *both* analyses. □

*Interpreting confidence intervals*

The interpretation of confidence intervals is actually a quite profound issue that statisticians have been arguing about for decades. This subsection presents the author's point of view in the context of some relatively simple problems. Although the problems are simple, the issues being discussed are not.

The disquieting thing about confidence intervals is the logic (or lack thereof) behind the leap from the probability of $1-\alpha$ that a future interval will contain the parameter into a '$(1-\alpha)100\%$ confidence' that the parameter is in a particular observed interval. The problem is in defining the meaning of confidence.

The standard interpretation of $(1-\alpha)100\%$ confidence intervals is that if you repeatedly performed many similar independent confidence intervals, about $(1-\alpha)100\%$ would contain the true parameter. The repeated sampling interpretation is exactly the same idea as saying that since a future coin toss has probability .5 of turning up heads, if you actually make many independent tosses of a coin, about 50% will be heads. This interpretation is not really saying anything new nor does it solve any problems because it still only relates to things that may be observed in the future. The fundamental problem of *inverting probabilities* for future observables into confidence about parameters remains. Moreover, the repeated sampling interpretation rarely applies to interesting problems. If you are obtaining a confidence interval for the height of corn plants grown outdoors, there is no way to perform independent replications of the experiment because there is no way to reproduce the exact growing conditions. In such cases, not only will the data behave differently but even the parameter of interest is likely to have a different meaning and value.

An alternative interpretation of confidence intervals based on statistical tests of hypotheses is presented in Section 3.4. I feel comfortable with the logic behind testing, so I like this interpretation. However, this interpretation makes no appeal whatsoever to the intuitive idea that 95% confidence means something similar to 95% probability.

I personally do not think it is possible to define confidence as anything other than probability. Two simple examples illustrate my point. I am going to flip a coin; we agree that the probability is .5 that it will land heads up. I flip the coin, look at it, but refuse to show it to you. Undoubtedly, you would feel comfortable saying that you are 50% confident that the coin is heads. I cannot imagine what that would mean except that you believe the probability is .5 that the coin is heads. Note that the 50% confidence is a statement about your beliefs and not a statement about the coin. The outcome of the coin toss is fixed (and known by someone other than you). This example has neither a fixed parameter nor any observable data but we can modify the example to make it more like a confidence interval problem. I place a coin either heads up or tails up and hide it from you, this is the parameter. You are going to flip a coin but I exercise my well known psychic powers. When I do this, the probability is .75 that the coin face I chose will be the face on your coin. When you toss your coin it lands either heads or tails and you observe this datum. The observed outcome of your toss is no longer random and it either matches mine or does not. Intuitively, you may reasonably feel that the probability is still .75 that the coins match, regardless of how I set my coin. But now

the probability is no longer about what the outcome of your flip will be because you have seen your datum. The probability must now be about how you believe I set my coin. Such a probability can only exist in your head. (Of course, I have other ideas and probabilities because, having seen both coins, I know whether they match.) While the intuition behind this probability is appealing, the logic escapes me. Glossing over the problem by saying that you have confidence, but not probability, of .75 for matching my coin does nothing to clear up the real issue.

R. A. Fisher made an attempt to build a theory of inverting probabilities from future observables into probabilities for parameters using this sort of intuition that we all find appealing. While it was a noble effort, I do not know of anyone who thinks Fisher succeeded or anyone who thinks that such a theory can succeed. (Of course, I have a limited sphere of acquaintance.) For more information, see the discussion of fiducial probability in Fisher (1956).

Another method of inverting probabilities about future data into probabilities about parameters is the theory of Bayesian statistics. Let me briefly mention how a Bayesian could arrive at a probability of .75 for the second coin tossing example. The computations are illustrated in Exercise 3.7.1. I place my coin any way I want. To arrive at a probability, you need to decide on *your beliefs* about how I placed my coin. If you believe that I am equally likely to place it heads up or tails up, those are your prior beliefs. Your prior beliefs are then modified by any data. In this example, if your initial beliefs are that I was equally likely to place the coin heads up or tails up, using a result known as Bayes theorem the probability that your coin agrees with mine becomes .75, regardless of what face of the coin I chose to place upwards and regardless of what you actually saw on your flip.

Notice that there is a lot more structure here than the mere intuition referred to earlier. In the intuitive discussion, your personal probability of a 75 : 25 chance of matching exists regardless of how I set my coin. In this discussion, you need to specify your beliefs about how I set my coin and the final 75 : 25 chance is a result of your having chosen an initial 50 : 50 chance for how I set my coin. For example, if you thought I was four times more likely to select heads, the probability of matching would be $12/13$ if your coin turned up heads but only $3/7$ if it turned up tails. Note that these beliefs do not depend on how I *actually* set my coin because you cannot know that. These beliefs do depend on your knowledge of how the data relate to how I set my coin, i.e., what data are likely when I choose heads and what are likely when I choose tails.

Bayesian methods are often criticized for requiring you to specify your initial beliefs in terms of a probability distribution on the possible parameter values. The result of a Bayesian data analysis is then an updated version of *your beliefs*. Berger (1985), among many others, responds to such criticisms. Many of us think that Bayesian methods provide the only logically consistent (though I would *not* say the only useful) method for doing statistics.

As I see it, a person has three choices: one can ignore the problem of what confidence means, one can use the hypothesis testing interpretation of confidence intervals to be given later, or one can rely on Bayesian methods. As it turns out, the confidence intervals and prediction intervals used in this book can be obtained by reasonable Bayesian methods. In the Bayesian interpretation of these intervals, confidence simply means probability, as the data modify a particular set of prior beliefs that are chosen to have a minimum of influence on the results of the data analysis.

## 3.2 Theory for hypothesis tests

Hypothesis tests are used to check whether *Par* has some specified value. For some fixed known number $m$, we may want to test the *null hypothesis*

$$H_0 : Par = m$$

versus the *alternative hypothesis*

$$H_A : Par \neq m.$$

The number $m$ must be known; it is some number that is of interest for the specific data being analyzed. It is impossible to give general rules for picking $m$ because the choice must depend on

the context of the data. As mentioned in the previous chapter, *the structure of the data (but not the actual values of the data) sometimes suggests interesting hypotheses* such as testing whether two populations have the same mean or testing whether there is a relationship between two variables, but ultimately the researcher must determine what hypotheses are of interest and these hypotheses determine $m$. In any case, $m$ is never just an unspecified symbol; it must have meaning within the context of the problem. The test of $H_0 : Par = m$ versus $H_A : Par \neq m$ is based on the assumption that $H_0$ is true and consists of checking to see whether the data are inconsistent with that assumption.

To identify data that are inconsistent with the assumption that $Par = m$, we examine what happens when $Par \neq m$. Note that $Est$ is always an estimate of $Par$; this has nothing to do with any hypothesis. With $Est$ estimating $Par$, it follows that if $Par > m$ then $Est$ tends to be larger than $m$. Equivalently, $Est - m$, and thus $[Est - m]/\mathrm{SE}(Est)$, tend to be large positive numbers when $Par > m$ (larger than they would be if $H_0 : Par = m$ is true). On the other hand if $Par < m$, then $Est - m$ and $[Est - m]/\mathrm{SE}(Est)$ tend to be large negative numbers. Data that are inconsistent with the null hypothesis $Par = m$ are large positive and large negative values of the test statistic $[Est - m]/\mathrm{SE}(Est)$. The problem is in specifying what we mean by 'large.' In practice we conclude that the data contradict the null hypothesis $Par = m$ if we observe a value of $[Est - m]/\mathrm{SE}(Est)$ that is further from 0 than some cutoff values. The problem is to make an intelligent choice for the cutoff values. The solution is based on the fact that if $H_0$ is true, the *test statistic*

$$\frac{Est - m}{\mathrm{SE}(Est)}$$

has the known reference distribution that is symmetric about 0.

When we substitute the observed values of $Est$ and $\mathrm{SE}(Est)$ into the test statistic we get one observation on the random test statistic. When $H_0$ is true, this observation comes from the reference distribution. The question is whether it is reasonable to believe that this one observation came from the reference distribution. If so, the data are consistent with $H_0$. If the observation could not reasonably have come from the reference distribution, the data contradict $H_0$. Contradicting $H_0$ is a strong inference; it implies that $H_0$ is false. On the other hand, inferring that the data are consistent with $H_0$ does not suggest that $H_0$ is true. Such data can also be consistent with some aspects of the alternative.

Before we can state the test formally, i.e., give intelligent cutoff values to determine the test, we need to consider the concept of error. Even if $H_0$ is true, it is usually possible (not probable but possible) to get any value at all for $[Est - m]/\mathrm{SE}(Est)$. For that reason, no matter what we conclude about the null hypothesis, there is a possibility of error. A test of hypothesis is based on controlling the probability of making an error when the null hypothesis is true. We define the $\alpha$ level of the test as the probability of rejecting the null hypothesis (saying that it is false) when the null hypothesis is in fact true. *The $\alpha$ level is also called the probability of a type I error, with a type I error being the rejection of a true null hypothesis*.

The $\alpha$ level determines the cutoff values for testing. The $\alpha$ level test for $H_0 : Par = m$ versus $H_A : Par \neq m$ is to reject $H_0$ if

$$\frac{Est - m}{\mathrm{SE}(Est)} > K\left(1 - \frac{\alpha}{2}\right)$$

or if

$$\frac{Est - m}{\mathrm{SE}(Est)} < -K\left(1 - \frac{\alpha}{2}\right).$$

This is equivalent to saying, reject $H_0$ if

$$\frac{|Est - m|}{\mathrm{SE}(Est)} > K\left(1 - \frac{\alpha}{2}\right).$$

To see that using $K\left(1 - \frac{\alpha}{2}\right)$ and $-K\left(1 - \frac{\alpha}{2}\right)$ as cutoff values gives an $\alpha$ level test, observe that if $H_0$

is true, the probability that we will reject $H_0$ is

$$\Pr\left[\frac{Est-m}{\mathrm{SE}(Est)} > K\Big(1-\frac{\alpha}{2}\Big)\right] + \Pr\left[\frac{Est-m}{\mathrm{SE}(Est)} < -K\Big(1-\frac{\alpha}{2}\Big)\right] = \alpha/2+\alpha/2 = \alpha,$$

see Figure 3.2. Also note that we are rejecting $H_0$ for those values of $[Est-m]/\mathrm{SE}(Est)$ that are most inconsistent with $H_0$, these being the values far from zero.

Actually, this test could be developed without any reference to the alternative hypothesis whatsoever. (In fact, I much prefer such a development since I believe that if you are willing to specify an alternative you should probably do a Bayesian analysis.) The only place where we used the alternative hypothesis was in determining which values of the test statistic were inconsistent with $H_0$. A different approach simply uses Figure 3.2 to decide which values of the test statistic are inconsistent. We can define the values that are most inconsistent as those that are the least likely to occur. The values that are least likely to occur are those where the density (i.e., the curve) is lowest. In Figure 3.2, the lowest values of the density are those corresponding to values of the test statistic that are far from 0. The density is symmetric, so our test should be symmetric. Thus an $\alpha$ level test has exactly the form given above. Of course this analysis relies on Figure 3.2 being an accurate portrayal of the distribution under $H_0$, but for all of our applications it is.

EXAMPLE 3.2.1. In Example 3.1.1 we considered past data on audio acuity in a post-rock environment. Those data were collected on fans of the group Van Holland in their Lee David Rothschild days. The nefarious organization responsible for this study found it necessary to update their findings after Rothschild was replaced by Slammy Hagar-Slacks. This time they abducted for themselves 16 independent observations and they were positive that the data would continue to follow a normal distribution. (Such arrogance is probably responsible for the failure of S.P.E.C.T.R.E.'s plans of world domination. In any case, their resident statistician was in no position to question this assumption.) The observed values of $\bar{y}_\cdot$ and $s^2$ were 22 and .25 respectively for the audio acuity scores. Now the purpose of all this is that S.P.E.C.T.R.E. had a long standing plot that required the use of a loud rock band. They had been planning to use the group Audially Disadvantaged Leopard but Van Holland's fans offered certain properties they preferred, provided that those fans audio acuity scores were satisfactory. From extremely long experience with abducting Audially Disadvantaged Leopard fans, S.P.E.C.T.R.E. knows that they have a population mean of 20 on the audio acuity test. S.P.E.C.T.R.E. wishes to know whether Van Holland fans differ from this value. Naturally, they tested $H_0 : \mu = 20$ versus $H_A : \mu \neq 20$ and they chose an $\alpha$ level of .01.

1) $Par = \mu$
2) $Est = \bar{y}_\cdot$
3) $\mathrm{SE}(Est) = s/\sqrt{16}$. In this case the $\mathrm{SE}(Est)$ is estimated.
4) $[Est-Par]/\mathrm{SE}(Est) = [\bar{y}_\cdot - \mu]/[s/\sqrt{16}]$ has a $t(15)$ distribution. This follows because the data are normally distributed and the standard error is estimated using $s$.

The $\alpha = .01$ test is to reject $H_0$ if

$$\frac{|\bar{y}_\cdot - 20|}{s/\sqrt{16}} > 2.947 = t(.995, 15).$$

Note that the sample size is $n = 16$ and $K(1-\alpha/2) = K(1-.005) = t(.995,15)$. Since $\bar{y}_\cdot = 22$ and $s^2 = .25$ we reject $H_0$ if

$$\frac{|22-20|}{\sqrt{.25/16}} > 2.947.$$

Since $|22-20|/\sqrt{.25/16} = 16$ is greater than 2.947, we reject the null hypothesis at the $\alpha = .01$ level. There is clear (indeed, overwhelming) evidence that the Van Holland fans have higher scores.

(Unfortunately, my masters will not let me inform you whether high scores mean better hearing or worse.) □

EXAMPLE 3.2.2. The National Association for the Abuse of Student Yahoos (also known as NAASTY) has established guidelines indicating that university dropout rates for math classes should be 15%. Based on an $\alpha = .05$ test, we wish to know if the University of New Mexico (UNM) meets these guidelines when treating the 1984–85 academic year data as a random sample. As is typical in such cases, NAASTY has specified that the central value of the distribution of dropout rates should be 15% but it has not stated a specific definition of the central value. We interpret the central value to be the population mean of the dropout rates and test the null hypothesis $H_0 : \mu = 15\%$ against the two-sided alternative $H_A : \mu \neq 15\%$.

The complete data consist of 38 observations from which we compute $\bar{y}_\cdot = 13.11$ and $s^2 = 106.421$. The data are nonnormal, so we have little choice but to hope that 38 observations constitute a sufficiently large sample to justify the use of

$$\frac{\bar{y}_\cdot - \mu}{\sqrt{s^2/38}} \sim N(0,1)$$

as an approximate reference distribution. With an $\alpha$ level of .05 and the standard normal distribution, the two-sided test rejects $H_0$ if

$$\frac{\bar{y}_\cdot - 15}{\sqrt{s^2/38}} > 1.96 = z(.975) = z\Big(1 - \frac{\alpha}{2}\Big)$$

or if

$$\frac{\bar{y}_\cdot - 15}{\sqrt{s^2/38}} < -1.96.$$

Substituting the observed values for $\bar{y}_\cdot$ and $s^2$ gives the observed value of the test statistic

$$\frac{13.11 - 15}{\sqrt{106.421/38}} = -1.13.$$

The value of $-1.13$ is neither greater than 1.96 nor less than $-1.96$, so the null hypothesis cannot be rejected at the .05 level. The 1984–85 data provide no evidence that UNM violates the NAASTY guidelines.

If we delete the two outliers, the analysis changes somewhat. Without the outliers, the data are approximately normal and we can use the reference distribution

$$\frac{\bar{y}_d - \mu_d}{\sqrt{s_d^2/36}} \sim t(35).$$

For this reference distribution the two-sided $\alpha = .05$ test rejects $H_0 : \mu_d = 15$ if

$$\frac{\bar{y}_d - 15}{\sqrt{s_d^2/36}} > 2.030 = t(.975, 35)$$

or if

$$\frac{\bar{y}_d - 15}{\sqrt{s_d^2/36}} < -2.030 = -t(.975, 35).$$

With $\bar{y}_d = 11.083$ and $s_d^2 = 27.45$ from the data without the outliers, the observed value of the test statistic is

$$\frac{11.083 - 15}{\sqrt{27.45/36}} = -4.49.$$

The absolute value of $-4.49$ is greater than 2.030, i.e., $-4.49 < -2.030$, so we reject the null hypothesis of $H_0 : \mu_d = 15\%$ at the .05 level. When we exclude the two extremely high observations, we have evidence that the typical dropout rate was different from 15%. In particular, *since the test statistic is negative*, we have evidence that the population mean dropout rate with outliers deleted was actually *less than* 15%. Obviously, most of the UNM math faculty during 1984–85 were not sufficiently nasty.

Finally, we consider the role of transformations in testing. We again consider the square roots of the dropout rates with the two outliers deleted. As discussed earlier, NAASTY has specified that the central value of the distribution of dropout rates should be 15% but has not stated a specific definition of the central value. We are reasonably free to interpret their guideline and we now interpret it as though the population mean of the square roots of the dropout rates should be $\sqrt{15}$. This interpretation leads us to the null hypothesis $H_0 : \mu_{rd} = \sqrt{15}$ and the alternative $H_A : \mu_{rd} \neq \sqrt{15}$. As discussed earlier, a reasonably appropriate reference distribution is

$$\frac{\bar{y}_{rd} - \mu_{rd}}{\sqrt{s_{rd}^2/36}} \sim t(35),$$

so the test rejects $H_0$ if

$$\frac{|\bar{y}_{rd} - \sqrt{15}|}{\sqrt{s_{rd}^2/36}} > 2.030 = t(.975, 35).$$

The sample mean and variance of the transformed, deleted data are $\bar{y}_{rd} = 3.218$ and $s_{rd}^2 = .749574$, so the observed value of the test statistic is

$$\frac{3.218 - 3.873}{\sqrt{.749574/36}} = -4.54.$$

The test statistic is similar to that in the previous paragraph. The null hypothesis is again rejected and all conclusions drawn from the rejection are essentially the same. As stated earlier, I believe that when two analyses both appear to be valid, either the practical conclusions agree or neither analysis should be trusted. □

*One-sided tests*

We can do one-sided tests in a similar manner. The $\alpha$ level test for $H_0 : Par \leq m$ versus $H_A : Par > m$ is to reject $H_0$ if

$$\frac{Est - m}{\mathrm{SE}(Est)} > K(1-\alpha).$$

The alternative hypothesis is that *Par* is greater than something and the null hypothesis is rejected when the test statistic is greater than some cutoff value. We reject the null hypothesis for the values of the test statistic that are most inconsistent with the null hypothesis and thus most consistent with the alternative hypothesis. If the alternative is true, $Est$ should be near *Par*, which is greater than $m$, so large positive values of $Est - m$ or, equivalently, large positive values of $[Est - m]/\mathrm{SE}(Est)$ are consistent with the alternative and inconsistent with the null hypothesis.

The $\alpha$ level test for $H_0 : Par \geq m$ versus $H_A : Par < m$ is to reject $H_0$ if

$$\frac{Est - m}{\mathrm{SE}(Est)} < -K(1-\alpha).$$

The alternative hypothesis is that *Par* is less than something and the null hypothesis is rejected when the test statistic is less than some cutoff value. *The form of the alternative determines the form of the rejection region*. In both cases we reject $H_0$ for the data that are most inconsistent with $H_0$

*The null hypotheses involve inequalities but $Par = m$ is always part of the null hypotheses. The tests are set up assuming that $Par = m$ and this needs to be part of any null hypothesis. In both cases, if $Par = m$ then the probability of making a mistake is $\alpha$ and, more generally, if $H_0$ is true, the probability of making a mistake is no greater than $\alpha$.*

EXAMPLE 3.2.3. Again consider the Slammy Hagar-Slacks era Van Holland audio data. Recall that there are 16 independent observations taken from a normal population with observed statistics of $\bar{y}_\cdot = 22$ and $s^2 = .25$. This time I have been required to perform a one-sided test to see whether I can prove that the Van Holland mean audio acuity scores are lower than the Audially Disadvantaged Leopard mean. I now test $H_0 : \mu \geq 20$ versus $H_A : \mu < 20$ with $\alpha = .01$. Here I am claiming that the scores are not lower and check to see whether the data contradict this. If they do, then my claim must be false and I have proven that the scores must be lower. If I initially claimed that the scores were lower, I would not be able to prove it; I could only establish that the data were consistent with my claim. As before,

1) $Par = \mu$

2) $Est = \bar{y}_\cdot$

3) $\mathrm{SE}(Est) = s/\sqrt{16}$. In this case the $\mathrm{SE}(Est)$ is estimated.

4) $[Est - Par]\big/\mathrm{SE}(Est) = [\bar{y}_\cdot - \mu]\big/[s/\sqrt{16}]$ has a $t(15)$ distribution.

The $\alpha = .01$ test is to reject $H_0 : \mu \geq 20$ if

$$\frac{\bar{y}_\cdot - 20}{s/\sqrt{16}} < -2.602 = -t(.99, 15).$$

Note that with a sample size of $n = 16$ we get $K(1-\alpha) = K(1-.01) = t(.99, 15)$. With $\bar{y}_\cdot = 22$ and $s^2 = .25$, we reject if

$$\frac{22 - 20}{\sqrt{.25/16}} < -2.602$$

Since $(22 - 20)\big/\sqrt{.25/16} = 16$ is greater than $-2.602$ we do not reject the null hypothesis at the $\alpha = .01$ level. There is no evidence that the Van Holland mean is lower than the Audially Disadvantaged Leopard mean. Observe that with the alternative $\mu < 20$, i.e., $\mu$ *less than* something, $H_0$ is only rejected when the test statistic is *less than* some cutoff value.

If you stop and think about it, we really did not have to go to all this trouble to discover the conclusion of this test. The null hypothesis is that $\mu \geq 20$. The observed $\bar{y}_\cdot$ value of 22 is obviously consistent with the hypothesis that the mean is greater than or equal to 20. Only $\bar{y}_\cdot$ values that are less than 20 could possibly contradict the null hypothesis. The only point at issue is how far $\bar{y}_\cdot$ must be below 20 before we can claim that $\bar{y}_\cdot$ contradicts the null hypothesis. As discussed in Example 2.2.4, given a choice it would be more informative to reverse the inequalities in $H_0$ and $H_A$ for this problem. □

EXAMPLE 3.2.4. A colleague of mine claims that, excluding classes with outrageous dropout rates, the math dropout rate at UNM was never more than 9% in any year during the 1980s. We now test this claim using the only data we have, that from the 1984–85 school year. My colleague excluded classes with outrageous dropout rates, so we use only the data with the outliers deleted. We again use $\alpha = .05$.

Based on the untransformed data, the null hypothesis is simply my colleague's claim, i.e., $H_0 : \mu \leq 9$. The alternative is $H_A : \mu > 9$. With $\alpha = .05$, the test is rejected if

$$\frac{\bar{y}_d - 9}{\sqrt{s_d^2/36}} > 1.690 = t(.95, 35).$$

With $\bar{y}_d = 11.083$ and $s_d^2 = 27.45$, the observed test statistic is

$$\frac{11.083 - 9}{\sqrt{27.45/36}} = 2.39,$$

so the test is easily, but not overwhelmingly, rejected.

Using the square roots of the data, the null hypothesis becomes $H_0 : \mu_{rd} \leq \sqrt{9}$. The alternative is $H_A : \mu_{rd} > \sqrt{9}$. With $\alpha = .05$, the test is rejected if

$$\frac{\bar{y}_{rd} - \sqrt{9}}{\sqrt{s_{rd}^2/36}} > 1.690 = t(.95, 35).$$

The sample mean and variance of the transformed, deleted data are $\bar{y}_{rd} = 3.218$ and $s_{rd}^2 = .749574$, so the observed value of the test statistic is

$$\frac{3.218 - 3}{\sqrt{.749574/36}} = 1.51.$$

The observed value is not greater than 1.690, so the test cannot be rejected at the .05 level.

In this case the two tests disagree. The untransformed data rejects the .05 level test easily. The transformed data does not quite achieve significance at the .05 level. To me, the data seem inconclusive. There is certainly some reason to suspect that the true dropout rate during 1984–85 was greater than 9%; one test rejected the null hypothesis and the other came somewhat close to being rejected. However, both analyses seem reasonable, so I cannot place great confidence in the rejection obtained using the untransformed data when the result is not fully corroborated by the transformed data. □

*P values*

Rather than having formal rules for when to reject the null hypothesis, one can report the evidence against the null hypothesis. This is done by reporting the $P$ value. The $P$ value is computed under the assumption that $Par = m$. It is the probability of seeing data that are as extreme or more extreme than those that were actually observed. Formally, we write $t_{obs}$ for the observed value of the test statistic, computed from the *observed* values of $Est$ and $\mathrm{SE}(Est)$. Thus $t_{obs}$ is our summary of the data that were actually observed. Recalling our earlier discussion of which values of $Est$ would be most inconsistent with $Par = m$, the probability of seeing something as or more extreme than we actually saw is

$$P = \Pr\left[\left|\frac{Est - m}{\mathrm{SE}(Est)}\right| \geq |t_{obs}|\right]$$

where $Est$ (and usually $\mathrm{SE}(Est)$) are viewed as random and it is assumed that $Par = m$. Under these conditions $(Est - m)/\mathrm{SE}(Est)$ has the known reference distribution and $t_{obs}$ is a known number, so we can actually compute $P$. The basic idea is that for, say, $t_{obs}$ positive, any value of $(Est - m)/\mathrm{SE}(Est)$ greater than $t_{obs}$ is more extreme than $t_{obs}$. Any data that yield $(Est - m)/\mathrm{SE}(Est) = -t_{obs}$ are just as extreme as $t_{obs}$ and values of $(Est - m)/\mathrm{SE}(Est)$ less than $-t_{obs}$ are more extreme than observing $t_{obs}$.

EXAMPLE 3.2.5. Again consider the Slammy Hagar-Slacks era Van Holland data. We have 16 observations taken from a normal population and we wish to test $H_0 : \mu = 20$ versus $H_A : \mu \neq 20$. As before, 1) $Par = \mu$, 2) $Est = \bar{y}_\cdot$, 3) $\mathrm{SE}(Est) = s/\sqrt{16}$, and 4) $[Est - Par]/\mathrm{SE}(Est) = [\bar{y}_\cdot - \mu]/[s/\sqrt{16}]$ has a $t(15)$ distribution. This time we take $\bar{y}_\cdot = 19.78$ and $s^2 = .25$, so the observed test statistic is

$$t_{obs} = \frac{19.78 - 20}{\sqrt{.25/16}} = -1.76.$$

From a $t$ table, $t(.95,15) = 1.75$, so

$$P = \Pr[|t(15)| \geq |-1.76|] \doteq \Pr[|t(15)| \geq 1.75] = .10.$$

Alternatively, $t(.95,15) \doteq |1.76|$, so $P \doteq 2(1-.95)$. □

Equivalently, the $P$ value is the smallest $\alpha$ level for which the test would be rejected. With this definition, if we perform an $\alpha$ level test where $\alpha$ *is less than the P value, we can conclude immediately that the null hypothesis is not rejected*. If we perform an $\alpha$ level test where $\alpha$ *is greater than the P value, we know immediately that the null hypothesis is rejected*. Thus computing a $P$ value eliminates the need to go through the formal testing procedures described above. Knowing the $P$ value immediately gives the test results for any choice of $\alpha$. The $P$ value is a measure of how consistent the data are with $H_0$. Large values (near 1) indicate great consistency. Small values (near 0) indicate data that are inconsistent with $H_0$.

EXAMPLE 3.2.6. In Example 3.2.2 we considered two-sided tests for the drop rate data. Using the complete untransformed data, the null hypothesis $H_0 : \mu = 15$, and the alternative $H_A : \mu \neq 15$, we observed the test statistic

$$t_{obs} = \frac{13.11 - 15}{\sqrt{106.421/38}} = -1.13.$$

Using a standard normal table or a computer program, we can compute

$$P = \Pr[|z| \geq |-1.13|] = .26.$$

An $\alpha = .26$ test would be just barely rejected by these data. Any test with an $\alpha$ level smaller than .26 is more stringent (the cutoff values are farther from 0 than 1.13) and would not be rejected. Thus the standard $\alpha = .05$ and $\alpha = .01$ tests would not be rejected. Similarly, any test with an $\alpha$ level greater than .26 is less stringent and would be rejected. Of course, it is extremely rare that one would use a test with an $\alpha$ level greater than .26.

Using the untransformed data with outliers deleted, the null hypothesis $H_0 : \mu_d = 15$, and the alternative $H_A : \mu_d \neq 15$, we observed the test statistic

$$\frac{11.083 - 15}{\sqrt{27.45/36}} = -4.49.$$

We compute

$$P = \Pr[|t(35)| \geq |-4.49|] = .000.$$

This $P$ value is not really zero; it is a number that is so small that when we round it off to three decimal places the number is zero. In any case, the test is rejected for any reasonable choice of $\alpha$. In other words, the test is rejected for any choice of $\alpha$ that is greater than .000. (Actually for any $\alpha$ greater than .0005 because of the round off problem.)

Using the square roots of the data with outliers deleted, the null hypothesis $H_0 : \mu_{rd} = \sqrt{15}$, and the alternative $H_A : \mu_{rd} \neq \sqrt{15}$, the observed value of the test statistic is

$$\frac{3.218 - 3.873}{\sqrt{.749574/36}} = -4.54.$$

We compute

$$P = \Pr[|t(35)| \geq |-4.54|] = .000.$$

Once again, the test result is highly significant. □

EXAMPLE 3.2.7. In Example 3.2.4 we considered one-sided tests for the drop rate data. Using

the deleted untransformed data, the null hypothesis $H_0 : \mu_d \leq 9$, and the alternative $H_A : \mu_d > 9$, we observed the test statistic

$$\frac{11.083 - 9}{\sqrt{27.45/36}} = 2.39.$$

Using Minitab, we compute

$$P = \Pr[t(35) \geq 2.39] = .011.$$

The probability is only for large positive values because negative values of the test statistic are consistent with $H_0$. The $P$ value of .011 is just greater than .01, so we would not be able to reject an $\alpha = .01$ test. We can of course reject any test with $\alpha$ greater than .011. The $P$ value for the one-sided test is exactly half of what the $P$ value would be for testing $H_0 : \mu_d = 9$ versus $H_A : \mu_d \neq 9$.

Using the square roots of the data, the null hypothesis became $H_0 : \mu_{rd} \leq \sqrt{9}$ with the alternative $H_A : \mu_{rd} > \sqrt{9}$. The observed value of the test statistic was

$$\frac{3.218 - 3}{\sqrt{.749574/36}} = 1.51.$$

We compute

$$P = \Pr[t(35) \geq 1.51] = .07.$$

The $P$ value here is small, .07, but not small enough to reject an $\alpha = .05$ test. There is some indication that the null hypothesis is not true but the indication is not very strong. To be precise, if we repeated this test procedure many times when the null hypothesis is true, 7% of the time we would expect to get results that are at least this suggestive of the incorrect conclusion that the null hypothesis is false. □

*Minitab commands*

To find a $P$ value using Minitab when the reference distribution is a $t$, start with the number $-|t_{obs}|$, where $t_{obs}$ is the observed value of the test statistic. In other words, find the observed test statistic and make it a negative number. Then simply use this number with the 'cdf' command, specifying the $t$ distribution and the degrees of freedom in the subcommand. The procedure for $t_{obs} = 1.51$ is illustrated below. The probability given by the cdf command is the appropriate $P$ value for one-sided tests but *must be doubled if the test is two-sided*.

```
MTB > cdf -1.51;
SUBC> t 35.
```

*Conclusion*

To keep this discussion as simple as possible, the examples have been restricted to one-sample normal theory. However, the results of this section and Section 3.1 apply to more complicated problems such as two-sample problems, testing contrasts in analysis of variance, and testing coefficients in regression. All of these applications will be considered in later chapters.

### 3.3 Validity of tests and confidence intervals

In testing an hypothesis, we make an assumption, namely the null hypothesis, and check to see whether the data are consistent with the assumption or inconsistent with it. If the data are consistent with the null hypothesis, that is all that we can say. If the data are inconsistent with the null hypothesis, it suggests that our assumption was wrong. (This is very similar to the mathematical idea of a proof by contradiction.)

One of the problems with testing hypotheses is that we are really making a series of assumptions. The null hypothesis is one of these, but there are many others. Typically we assume that observations

are independent. In most tests that we will consider, we assume that the data have normal distributions. As we consider more complicated data structures, we will need to make more assumptions. The proper conclusion from a test of hypothesis is that either the data are consistent with our assumptions or the data are inconsistent with our assumptions. If the data are inconsistent with the assumptions, it suggests that at least one of them is invalid. In particular, if the data are inconsistent with the assumptions, it does not necessarily imply that the particular assumption embodied in the null hypothesis is the one that is invalid. Before we can reasonably conclude that the null hypothesis is untrue, we need to ensure that the other assumptions are reasonable. Thus it is crucial to check our assumptions as fully as we can. Plotting the data plays a vital role in checking assumptions. Plots are used throughout the book, but special emphasis on plotting is given in Chapter 7.

Typically, it is quite easy to define parameters *Par* and estimates *Est*. The role of the assumptions is crucial in obtaining a valid SE(*Est*) and an appropriate reference distribution. If our assumptions are reasonably valid, our SE(*Est*) and reference distribution will be reasonably valid and the procedures outlined here for performing statistical inferences will be reasonably valid. This applies not only to testing but to confidence intervals as well. Of course the assumptions that need to be checked depend on the precise nature of the analysis being performed.

### 3.4 The relationship between confidence intervals and tests

The two most commonly used tools in statistical inference are tests and confidence intervals. Tests determine whether a difference can be established between an hypothesized parameter value and the true parameter for the data. Typically, one must consider not only whether a difference exists, but how much difference exists, and whether such a difference is important within the context of the problem. Confidence intervals are used to quantify what is known about the true parameter and thus can be used to quantify how much of a difference may exist. In particular, confidence intervals give all the possible parameter values that seem to be consistent with the data. Tests and confidence intervals are very closely related inferential tools and in this section we explore their relationship.

As discussed earlier, the term 'confidence' as used in confidence intervals is rather nebulously defined. Confidence intervals are based on the unusable probability statement

$$1-\alpha = \Pr\left[Est - K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est) < Par < Est + K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est)\right],$$

which is a statement about the unknown (unobserved) random variables *Est* and SE(*Est*). It is a highly intuitive idea that this probability statement generates a usable interval for *Par*,

$$Est - K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est) < Par < Est + K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est),$$

in which the *observed* values of *Est* and SE(*Est*) are used to define a known interval. However, the logic behind this intuitive idea is not clear and so we are left with an unclear definition of 'confidence.'

A clear definition of confidence can be made in terms of testing hypotheses. *The* $(1-\alpha)100\%$ *confidence interval for Par*,

$$Est - K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est) < Par < Est + K\left(1-\frac{\alpha}{2}\right)\mathrm{SE}(Est),$$

*consists of all the values m that* would not *be rejected by an* $\alpha$ *level test of* $H_0 : Par = m$ *versus* $H_A : Par \neq m$. To see this recall that the $\alpha$ level test is rejected when

$$\frac{Est - m}{\mathrm{SE}(Est)} > K\left(1-\frac{\alpha}{2}\right)$$

or

$$\frac{Est - m}{\mathrm{SE}(Est)} < -K\left(1-\frac{\alpha}{2}\right).$$

Conversely, the $\alpha$ level test is not rejected when

$$-K\Big(1-\frac{\alpha}{2}\Big) \le \frac{Est-m}{\mathrm{SE}(Est)} \le K\Big(1-\frac{\alpha}{2}\Big).$$

Exactly the same algebraic manipulations that lead to equation (3.1.1) also lead to the conclusion that the test is not rejected when

$$Est - K\Big(1-\frac{\alpha}{2}\Big)\mathrm{SE}(Est) < m < Est + K\Big(1-\frac{\alpha}{2}\Big)\mathrm{SE}(Est).$$

Thus the confidence interval consists of all values of $m$ for which the $\alpha$ level test of $H_0 : Par = m$ versus $H_A : Par \neq m$ is not rejected. In other words, *a* $(1-\alpha)100\%$ *confidence interval consists of all parameter values that are consistent with the data as judged by an* $\alpha$ *level test*.

*We have now established that there is little point in performing the fixed* $\alpha$*, fixed m testing procedures discussed in Section 3.2. P values give the results of testing* $H_0 : Par = m$ *versus* $H_A : Par \neq m$ *for a fixed m but* every *choice of* $\alpha$*. Confidence intervals give the results of testing* $H_0 : Par = m$ *versus* $H_A : Par \neq m$ *for a fixed* $\alpha$ *but* every *choice of m*.

EXAMPLE 3.4.1. In Example 3.2.1 we considered audio acuity data for Van Holland fans and tested whether their mean score differed from fans of Audially Disadvantaged Leopard. In this example we test whether their mean score differs from that of Tangled Female Sibling fans. Recall that the observed values of $n$, $\bar{y}.$, and $s^2$ for Van Holland fans were 16, 22, and .25, respectively and that the data were normal. Tangled Female Sibling fans have a population mean score of 22.325, so we test $H_0 : \mu = 22.325$ versus $H_A : \mu \neq 22.325$. The test statistic is $(22-22.325)/\sqrt{.25/16} = -2.6$. If we do an $\alpha = .05$ test, $|-2.6| > 2.13 = t(.975, 15)$, so we reject $H_0$, but if we do an $\alpha = .01$ test, $|-2.6| < 2.95 = t(.995, 15)$, so we do not reject $H_0$. In fact, $|-2.6| \doteq t(.99, 15)$, so the $P$ value is essentially .02. The $P$ value is larger than .01, so the .01 test does not reject $H_0$; the $P$ value is less than .05, so the test rejects $H_0$ at the .05 level.

If we consider confidence intervals, the 99% interval has endpoints $22 \pm 2.95\sqrt{.25/16}$ for an interval of $(21.631, 22.369)$ and the 95% interval has endpoints $22 \pm 2.13\sqrt{.25/16}$ for an interval of $(21.734, 22.266)$. Notice that the hypothesized value of 22.325 is inside the 99% interval, so it is not rejected by a .01 level test, but 22.325 is outside the 95% interval, so a .05 two-sided test rejects $H_0 : \mu = 22.325$. The 98% interval has endpoints $22 \pm 2.60\sqrt{.25/16}$ for an interval of $(21.675, 22.325)$ and the hypothesized value is on the edge of the interval.

□

## 3.5 Theory of prediction intervals

Some slight modifications of the general theory allow us to construct prediction intervals. Many of us would argue that the fundamental purpose of science is making accurate predictions of things that could be observed in the future. As with estimation, predicting the occurrence of a particular value (point prediction) is less valuable than interval prediction because a point prediction gives no idea of the variability associated with the prediction.

In constructing prediction intervals for a new observation $y$, we make a number of assumptions. The observations, including the new one, are assumed to be independent and normally distributed. Moreover, we take as our parameter $Par = \mathrm{E}(y)$. $\mathrm{E}(y)$ would be a reasonable point prediction for $y$ but we do not know the value of $\mathrm{E}(y)$. $Est$ depends only on the observations other than $y$ and it estimates $\mathrm{E}(y)$, so $Est$ makes a reasonable point prediction of $y$. We also assume that $\mathrm{Var}(y) = \sigma^2$, that $\sigma^2$ has an estimate $\hat{\sigma}^2$, that $\mathrm{SE}(Est) = \hat{\sigma}A$ for some known constant $A$, and that $(Est - Par)/\mathrm{SE}(Est)$ has a $t$ distribution with, say, $r$ degrees of freedom. (Technically, we need $Est$ to have a normal distribution, $r(\hat{\sigma}^2/\sigma^2)$ to have a $\chi^2(r)$ distribution, and independence of $Est$ and $\hat{\sigma}^2$.) In some applications, these methods are used with the approximation $r \doteq \infty$, i.e., we act as if we know the variance and the appropriate distribution is taken to be a standard normal.

A prediction interval for $y$ is based on the distribution of $y - Est$ because we need to evaluate how far $y$ can reasonably be from our point prediction of $y$. The value of the future observation $y$ is independent of the past observations and thus of $Est$. It follows that the variance of $y - Est$ is

$$\text{Var}(y - Est) = \text{Var}(y) + \text{Var}(Est) = \sigma^2 + \text{Var}(Est)$$

and that the standard error of $y - Est$ is

$$\text{SE}(y - Est) = \sqrt{\hat{\sigma}^2 + [\text{SE}(Est)]^2}. \tag{3.5.1}$$

One can then show that

$$\frac{y - Est}{\text{SE}(y - Est)} \sim t(r).$$

A $(1-\alpha)100\%$ prediction interval is based on the probability equality,

$$1 - \alpha = \Pr\left[-t\left(1 - \frac{\alpha}{2}, r\right) < \frac{y - Est}{\text{SE}(y - Est)} < t\left(1 - \frac{\alpha}{2}, r\right)\right].$$

Rearranging the terms within the square brackets leads to the equality

$$1 - \alpha = \Pr\left[Est - t\left(1 - \frac{\alpha}{2}, r\right)\text{SE}(y - Est) < y < Est + t\left(1 - \frac{\alpha}{2}, r\right)\text{SE}(y - Est)\right].$$

The prediction interval consists of all $y$ values that fall between the two observable limits in the probability statement. The endpoints of the interval are generally written

$$Est \pm t\left(1 - \frac{\alpha}{2}, r\right)\text{SE}(y - Est).$$

Of course, it is impossible to validate assumptions about observations to be taken in the future, so the confidence levels of prediction intervals are always suspect.

From the form of $\text{SE}(y - Est)$ given in (3.5.1), we see that

$$\text{SE}(y - Est) = \sqrt{\hat{\sigma}^2 + [\text{SE}(Est)]^2} \geq \text{SE}(Est).$$

Typically, the prediction standard error is much larger than the standard error of the estimate, so prediction intervals are much wider than confidence intervals. In particular, increasing the number of observations typically decreases the standard error of the estimate but has a *relatively* minor effect on the standard error of prediction. Increasing the sample size is not intended to make $\hat{\sigma}^2$ smaller, it only makes $\hat{\sigma}^2$ a more accurate estimate of $\sigma^2$.

EXAMPLE 3.5.1. As in Example 3.1.2, we eliminate the two outliers from the dropout rate data. The 36 remaining observations are approximately normal. A 95% confidence interval for the mean had endpoints

$$11.083 \pm 2.030\sqrt{27.45/36}.$$

A 95% prediction interval has endpoints

$$11.083 \pm 2.030\sqrt{27.45 + \frac{27.45}{36}}$$

or

$$11.083 \pm 10.782.$$

The prediction interval is $(.301, 21.865)$, which is much wider than the confidence interval of $(9.3, 12.9)$. We are 95% confident that the dropout rate for a new math class would be between

.3% and 21.9%. We are 95% confident that the population mean dropout rate for math classes is between 9% and 13%. Of course the prediction interval assumes that the new class is from a population similar to the 1984–85 math classes with huge dropout rates deleted. Such assumptions are almost impossible to validate. Moreover, there is some chance that the new observation will be one with a huge dropout rate and this interval says nothing about such observations.

In Example 3.1.2 we also considered the square roots of the dropout rate data with the two outliers eliminated. To predict the square root of a new observation, we use the 95% interval

$$3.218 \pm 2.030 \left( \sqrt{.749574 + \frac{.749574}{36}} \right),$$

which reduces to $(1.436, 5.000)$. This is a prediction interval for the square root of a new observation, so we are 95% confident that the actual value of the new observation will fall between $(1.436^2, 5.000^2)$, i.e., $(2.1, 25)$. Retransforming a prediction interval back into the original scale causes no problems of interpretation whatsoever. This prediction interval and the one in the previous paragraph are comparable. Both include values from near 0 up to the low to mid twenties. □

We have criticized commonly used definitions of the word 'confidence' but to this point the motivation for a prediction interval is exactly analogous to the motivation for confidence intervals. The endpoints of a prediction interval are obtained by taking a probability statement about random variables, substituting observed values for the random variables, and replacing 'probability' by 'confidence'. For some reason, explicitly stating that a 95% prediction interval gives 95% *confidence* that a future observation will fall within the interval seems to be a somewhat rare occurrence. Once again, a solution to the problem of defining confidence can be obtained by testing. If we wanted to test whether a new observation $y$ was consistent with the old observations we could set up an $\alpha$ level test that would reject if $(y - Est)/\mathrm{SE}(y - Est)$ was too far from zero, i.e., if its absolute value was greater than $K(1-\alpha/2)$. Analogous to the relationship between tests of parameters and confidence intervals, this test of a new observation will not be rejected precisely when $y$ is within the prediction interval. Thus the $(1-\alpha)100\%$ prediction interval consists of all values of $y$ that are consistent with the other data as determined by an $\alpha$ level test. Moreover, the testing approach gives some insight into why prediction intervals are based on the distribution of $y - Est$, i.e., because we are comparing the new observation $y$ to the old data as summarized by $Est$.

*Lower bounds on prediction confidence*

If the normal and $\chi^2$ distributional assumptions stated at the beginning of the section break down, the prediction interval based on the $t$ distribution is invalid. Relying primarily on the independence assumptions and there being sufficient data to use $\hat{\sigma}^2$ as an approximation to $\sigma^2$, we can find an approximate lower bound for the confidence that a new observation is in the prediction interval. Chebyshev's inequality from Subsection 1.2.2 gives

$$1 - t\left(1 - \frac{\alpha}{2}, r\right)^{-2} \leq \Pr\left[ -t\left(1 - \frac{\alpha}{2}, r\right) < \frac{y - Est}{\mathrm{SE}(y - Est)} < t\left(1 - \frac{\alpha}{2}, r\right) \right]$$

or equivalently

$$1 - t\left(1 - \frac{\alpha}{2}, r\right)^{-2} \leq \Pr\Big[ Est - t\left(1 - \frac{\alpha}{2}, r\right) \mathrm{SE}(y - Est) < y < Est + t\left(1 - \frac{\alpha}{2}, r\right) \mathrm{SE}(y - Est) \Big].$$

This indicates that the confidence coefficient for the prediction interval given by

$$Est \pm t\left(1 - \frac{\alpha}{2}, r\right) \mathrm{SE}(y - Est)$$

is (approximately) at least

$$\left[1 - t\left(1 - \frac{\alpha}{2}, r\right)^{-2}\right] 100\%.$$

If we can use the improved version of Chebyshev's inequality from Section 1.3, we can raise the confidence coefficient to

$$\left[1 - (2.25)^{-1} t\left(1 - \frac{\alpha}{2}, r\right)^{-2}\right] 100\%.$$

EXAMPLE 3.5.2. Assuming that a sample of 36 observations is enough to ensure that $s^2$ is essentially equal to $\sigma^2$, the nominal 95% prediction interval given in Example 3.5.1 for dropout rates has a confidence level, regardless of the distribution of the data, that is at least

$$\left(1 - \frac{1}{2.030^2}\right) = 76\% \text{ or even } \left(1 - \frac{1}{2.25(2.030)^2}\right) = 89\%.$$

### 3.6 Sample size determination and power

Suppose we wish to estimate the mean height of the men officially enrolled in statistics classes at the University of New Mexico on Thursday, February 4, 1993 at 3 pm. How many observations should we take? The answer to that question depends on how accurate our estimate needs to be and on our having some idea of the variability in the population.

To get a rough indication of the variability we argue as follows. Generally, men have a mean height of about 69 inches and I would guess that about 95% of them are between 63 inches and 75 inches. The probability that a $N(\mu, \sigma^2)$ random variable is between $\mu \pm 2\sigma$ is approximately .95, which suggests that $\sigma = [(\mu + 2\sigma) - (\mu - 2\sigma)]/4$ may be about $(75 - 63)/4 = 3$ for a typical population of men.

Before proceeding with sample size determination, observe that sample sizes have a real effect on the usefulness of confidence intervals. Suppose $\bar{y}_\cdot = 72$ and $n = 9$, so the 95% confidence interval for mean height has endpoints of roughly $72 \pm 2(3/\sqrt{9})$, or $72 \pm 2$, with an interval of $(70, 74)$. Here we use 3 as a rough indication of $\sigma$ in the standard error and 2 as a rough indication of the tabled value for a 95% interval. If having an estimate that is off by 1 inch is a big deal, the confidence interval is totally inadequate. There is little point in collecting the data, because regardless of the value of $\bar{y}_\cdot$, we do not have enough accuracy to draw interesting conclusions. For example, if I claimed that the true mean height for this population was 71 inches and I cared whether my claim was off by an inch, the data are not only consistent with my claim but also with the claims that the true mean height is 70 inches and 72 inches and even 74 inches. The data are inadequate for my purposes. Now suppose $\bar{y}_\cdot = 72$ and $n = 3600$, the confidence interval has endpoints $72 \pm 2(3/\sqrt{3600})$ or $72 \pm .1$ with an interval of $(71.9, 72.1)$. We can tell that the population mean may be 72 inches but we are quite confident that it is not 72.11 inches. Would anyone really care about the difference between a mean height of 72 inches and a mean height of 72.11 inches? Three thousand six hundred observations gives us more information that we really need. We would like to find a middle ground.

Now suppose we wish to learn the mean height to within 1 inch with 95% confidence. From a sample of size $n$, a 95% confidence interval for the mean has endpoints that are roughly $\bar{y}_\cdot \pm 2(3/\sqrt{n})$. With 95% confidence, the mean height could be as high as $\bar{y}_\cdot + 2(3/\sqrt{n})$ or as low as $\bar{y}_\cdot - 2(3/\sqrt{n})$. We want the difference between these numbers to be no more than 1 inch. The difference between the two numbers is $12/\sqrt{n}$, so for the required difference of 1 inch set $1 = 12/\sqrt{n}$, so that $\sqrt{n} = 12/1$ or $n = 144$.

The semantics of these problems can be a bit tricky. We asked for an interval that would tell us the mean height to within 1 inch with 95% confidence. If instead we specified that we wanted our estimate to be off by no more than 1 inch, the estimate is in the middle of the interval, so the distance from the middle to the endpoint needs to be 1 inch. In other words, $1 = 2(3/\sqrt{n})$, so $\sqrt{n} = 6/1$ or

$n = 36$. Note that learning the parameter to within 1 inch is the same as having an estimate that is off by no more than $1/2$ inch.

The concepts illustrated above work quite generally. Typically an observation $y$ has $\text{Var}(y) = \sigma^2$ and $Est$ has $\text{SE}(Est) = \sigma A$. The constant $A$ in $\text{SE}(Est)$ is a known function of the sample size (or sample sizes in situations involving more than one sample). In inference problems we replace $\sigma$ in the standard error with an estimate of $\sigma$ obtained from the data. In determining sample sizes, the data have not yet been observed, so $\sigma$ has to be approximated from previous data or knowledge. The length of a $(1-\alpha)100\%$ confidence interval is

$$[Est + K(1-\alpha/2)\text{SE}(Est)] - [Est - K(1-\alpha/2)\text{SE}(Est)] \\ = 2K(1-\alpha/2)\text{SE}(Est) = 2K(1-\alpha/2)\sigma A.$$

The tabled value $K(1-\alpha/2)$ can be approximated by $t(1-\alpha/2,\infty)$. If we specify that the confidence interval is to be $w$ units wide, set

$$w = 2t(1-\alpha/2,\infty)\sigma A \tag{3.6.1}$$

and solve for the (approximate) appropriate sample size. In equation (3.6.1), $w$, $t(1-\alpha/2,\infty)$, and $\sigma$ are all known and $A$ is a known function of the sample size.

Unfortunately it is not possible to take equation (3.6.1) any further and show directly how it determines the sample size. The discussion given here is general and thus the ultimate solution depends on the type of data being examined. In the only case we have examined as yet, there is one-sample, $Par = \mu$, $Est = \bar{y}.$, and $\text{SE}(Est) = \sigma/\sqrt{n}$. Thus, $A = 1/\sqrt{n}$ and equation (3.6.1) becomes

$$w = 2t(1-\alpha/2,\infty)\sigma/\sqrt{n}.$$

Rearranging this gives

$$\sqrt{n} = 2t(1-\alpha/2,\infty)\sigma/w$$

and

$$n = \left(2t(1-\alpha/2,\infty)\sigma/w\right)^2.$$

*But this formula only applies to one sample problems*. For other problems considered in this book, e.g., comparing two independent samples, comparing more than two independent samples, and simple linear regression, equation (3.6.1) continues to apply but the constant $A$ becomes more complicated. In cases where there is more than one sample involved, the various sample sizes are typically assumed to all be the same, and in general their relative sizes need to be specified, e.g., we could specify that the first sample will have 10 more observations than the second or that the first sample will have twice as many observations as the second.

Another approach to determining approximate sample sizes is based on the power of an $\alpha$ level test. Tests are set up assuming that, say, $H_0 : Par = m_0$ is true. Power is computed assuming that $Par \neq m_0$. Suppose that $Par = m_A \neq m_0$, then *the power when* $Par = m_A$ *is the probability that the* $(1-\alpha)100\%$ *confidence interval will not contain* $m_0$. Another way of saying that the confidence interval does not contain $m_0$ is saying that an $\alpha$ level two-sided test of $H_0 : Par = m_0$ rejects $H_0$. In determining sample sizes, you need to pick $m_A$ as some value you care about. You need to care about it in the sense that if $Par = m_A$ rather than $Par = m_0$, you would like to have a reasonably good chance of rejecting $H_0 : Par = m_0$.

Cox (1958, p. 176) points out that it often works well to choose the sample size so that

$$|m_A - m_0| \doteq 3\text{SE}(Est). \tag{3.6.2}$$

Cox shows that this procedure gives reasonable powers for common choices of $\alpha$. Here $m_A$ and $m_0$ are known and $\text{SE}(Est) = \sigma A$, where $\sigma$ is known and $A$ is a known function of sample size. Also note that this suggestion does not depend on the $\alpha$ level of the test. As with equation (3.6.1),

equation (3.6.2) can be solved to give $n$ in particular cases, but a general solution for $n$ is not possible because it depends on the exact nature of the value $A$.

Consider again the problem of determining the mean height. If my null hypothesis is $H_0 : \mu = 72$ and I want a reasonable chance of rejecting $H_0$ when $\mu = 73$, Cox's rule suggests that I should have $1 = |73 - 72| \doteq 3\,(3/\sqrt{n})$ so that $\sqrt{n} \doteq 9$ or $n \doteq 81$.

It is important to remember that these are only rough guides for sample sizes. They involve several approximations, the most important of which is approximating $\sigma$. If there is more than one parameter of interest in a study, sample size computations can be performed for each and a compromise sample size can be selected.

For the past ten years I've been amazed at my own lack of interest in teaching students about statistical power. Cox (1958, p. 161) finally explained it for me. He points out that power is very important in planning investigations but it is not very important in analyzing them. I might even go so far as to say that once the data have been collected, a power analysis can at best tell you whether you have been wasting your time. In other words, a power analysis will only tell you how likely you were to find differences given the design of your experiment and the choice of test.

*Appendix: derivation of confidence intervals*

We wish to establish the validity of equation (3.1.1), i.e.,

$$\begin{aligned} 1-\alpha &= \Pr\left[-K\Big(1-\frac{\alpha}{2}\Big) < \frac{Est-Par}{\mathrm{SE}(Est)} < K\Big(1-\frac{\alpha}{2}\Big)\right] \\ &= \Pr\left[Est - K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) < Par < Est + K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est)\right] \end{aligned}$$

and in particular we wish to show that the expressions in the square brackets are equivalent. We do this by establishing a series of equivalences. The justifications for the equivalences are given at the end.

$$-K\Big(1-\frac{\alpha}{2}\Big) < \frac{Est-Par}{\mathrm{SE}(Est)} < K\Big(1-\frac{\alpha}{2}\Big) \tag{1}$$

if and only if

$$-K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) < Est - Par < K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) \tag{2}$$

if and only if

$$K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) > -Est + Par > -K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) \tag{3}$$

if and only if

$$Est + K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) > Par > Est - K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) \tag{4}$$

if and only if

$$Est - K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est) < Par < Est + K\Big(1-\frac{\alpha}{2}\Big)\,\mathrm{SE}(Est). \tag{5}$$

JUSTIFICATION OF STEPS.

For (1) iff (2): if $c > 0$, then $a < b$ if and only if $ac < bc$.
For (2) iff (3): $a < b$ if and only if $-a > -b$.
For (3) iff (4): $a < b$ if and only if $a + c < b + c$.
For (4) iff (5): $a > b$ if and only if $b < a$.

## 3.7 Exercises

EXERCISE 3.7.1. This exercise illustrates the Bayesian computations discussed in the subsection of 3.1 on interpreting confidence intervals. I place a coin either heads up or tails up and hide it from you. Because of my psychic powers, when you subsequently flip a coin the probability is .75 that your coin face will be the same as mine. The four things of interest here are the outcomes that I have tails ($IT$), I have heads ($IH$), you have tails ($YT$), and you have heads ($YH$).

The computations involve ideas of conditional probability. For example, the probability that you get tails given that my coin was placed tails up is defined to be $\Pr(YT|IT) \equiv \Pr(YT \text{ and } IT)/\Pr(IT)$

Bayes' theorem relates different conditional probabilities. It states that

$$\Pr(IT|YT) = \frac{\Pr(YT|IT)\Pr(IT)}{\Pr(YT|IT)\Pr(IT) + \Pr(YT|IH)\Pr(IH)}.$$

Similarly,

$$\Pr(IH|YH) = \frac{\Pr(YH|IH)\Pr(IH)}{\Pr(YH|IH)\Pr(IH) + \Pr(YH|IT)\Pr(IT)}.$$

Clearly this problem is set up so that $\Pr(YT|IT) = \Pr(YH|IH) = .75$. Show that if your prior probability is $\Pr(IT) = \Pr(IH) = .5$, then $\Pr(IT|YT) = \Pr(IH|YH) = .75$ as claimed in the earlier discussion.

The earlier discussion also mentioned prior probabilities that were four times greater for me placing my coin heads up than tails up. In this case, $\Pr(IT) = 1/5$ and $\Pr(IH) = 4/5$. Find $\Pr(IT|YT)$ and $\Pr(IH|YH)$ and check whether these agree with the values given in Section 3.1.

Obviously, you should show all of your work.

EXERCISE 3.7.2. Identify the parameter, estimate, standard error of the estimate, and reference distribution for Exercise 2.7.1.

EXERCISE 3.7.3. Identify the parameter, estimate, standard error of the estimate, and reference distribution for Exercise 2.7.2.

EXERCISE 3.7.4. Identify the parameter, estimate, standard error of the estimate, and reference distribution for Exercise 2.7.4.

EXERCISE 3.7.5. Consider that I am collecting (normally distributed) data with a variance of 4 and I want to test a null hypothesis of $H_0 : \mu = 10$. What sample size should I take according to Cox's rule if I want a reasonable chance of rejecting $H_0$ when $\mu = 13$? What if I want a reasonable chance of rejecting $H_0$ when $\mu = 12$? What sample size should I take if I want a 95% confidence interval that is no more than 2 units long? What if I want a 99% confidence interval that is no more than 2 units long?

EXERCISE 3.7.6. The turtle shell data of Jolicoeur and Mosimann (1960) given in Exercise 2.7.4 has a standard deviation of about 21.25. If we were to collect a new sample, how large should the sample size be in order to have a 95% confidence interval with a length of (about) four units? According to Cox's rule, what sample size should I take if I want a reasonable chance of rejecting $H_0 : \mu = 130$ when $\mu = 140$?

EXERCISE 3.7.7. With reference to Exercise 2.7.3, give the approximate number of observations necessary to estimate the mean of BX to within .01 units with 99% confidence. How large a sample is needed to get a reasonable test of $H_0 : \mu = 10$ when $\mu = 11$ using Cox's rule?

EXERCISE 3.7.8. With reference to Exercise 2.7.3, give the approximate number of observations necessary to get a 99% confidence for the mean of K that has a length of 60. How large a sample

is needed to get a reasonable test of $H_0 : \mu = 1200$ when $\mu = 1190$ using Cox's rule? What is the number when $\mu = 1150$?

EXERCISE 3.7.9. With reference to Exercise 2.7.3, give the approximate number of observations necessary to estimate the mean of FORM to within .5 units with 95% confidence. How large a sample is needed to get a reasonable test of $H_0 : \mu = 20$ when $\mu = 20.2$ using Cox's rule?

EXERCISE 3.7.10. With reference to Exercise 2.7.2, give the approximate number of observations necessary to estimate the mean rat weight to within 1 unit with 95% confidence. How large a sample is needed to get a reasonable test of $H_0 : \mu = 55$ when $\mu = 54$ using Cox's rule?

Chapter 4

# Two sample problems

In this chapter we consider several situations where it is of interest to compare two samples. First we consider two samples of correlated data. These are data that consist of pairs of observations measuring comparable quantities. Next we consider two independent samples from populations with the same variance. We then examine two independent samples from populations with different variances. Finally we consider the problem of testing whether the variances of two populations are equal.

## 4.1 Two correlated samples: paired comparisons

Paired comparisons involve pairs of observations on similar variables. Often these are two observations taken on the same object under different circumstances or two observations taken on related objects. No new statistical methods are needed for analyzing such data.

EXAMPLE 4.1.1. Shewhart (1931, p. 324) presents data on the hardness of an item produced by welding two parts together. Table 4.1 gives the hardness measurements for each of the two parts. The hardness of part 1 is denoted $y_1$ and the hardness of part 2 is denoted $y_2$. For $i = 1, 2$, the data for part $i$ are denoted $y_{ij}$, $j = 1, \ldots, 27$. The data are actually a subset of the data presented by Shewhart.

We are interested in the difference between $\mu_1$, the population mean for part one, and $\mu_2$, the population mean for part two. In other words, the parameter of interest is $Par = \mu_1 - \mu_2$. Note that if there is no difference between the population means, $\mu_1 - \mu_2 = 0$. The natural estimate of this parameter is the difference between the sample means, i.e., $Est = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$. Here we use the $\cdot$ subscript to indicate averaging over the second subscript in $\bar{y}_{i\cdot} = (y_{i1} + \cdots + y_{i27})/27$.

To perform statistical inferences, we need the standard error of the estimate, i.e., $\mathrm{SE}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})$.

Table 4.1: *Shewhart's hardness data*

| Case | $y_1$ | $y_2$ | $d = y_1 - y_2$ | Case | $y_1$ | $y_2$ | $d = y_1 - y_2$ |
|---|---|---|---|---|---|---|---|
| 1 | 50.9 | 44.3 | 6.6 | 15 | 46.6 | 31.5 | 15.1 |
| 2 | 44.8 | 25.7 | 19.1 | 16 | 50.4 | 38.1 | 12.3 |
| 3 | 51.6 | 39.5 | 12.1 | 17 | 45.9 | 35.2 | 10.7 |
| 4 | 43.8 | 19.3 | 24.5 | 18 | 47.3 | 33.4 | 13.9 |
| 5 | 49.0 | 43.2 | 5.8 | 19 | 46.6 | 30.7 | 15.9 |
| 6 | 45.4 | 26.9 | 18.5 | 20 | 47.3 | 36.8 | 10.5 |
| 7 | 44.9 | 34.5 | 10.4 | 21 | 48.7 | 36.8 | 11.9 |
| 8 | 49.0 | 37.4 | 11.6 | 22 | 44.9 | 36.7 | 8.2 |
| 9 | 53.4 | 38.1 | 15.3 | 23 | 46.8 | 37.1 | 9.7 |
| 10 | 48.5 | 33.0 | 15.5 | 24 | 49.6 | 37.8 | 11.8 |
| 11 | 46.0 | 32.6 | 13.4 | 25 | 51.4 | 33.5 | 17.9 |
| 12 | 49.0 | 35.4 | 13.6 | 26 | 45.8 | 37.5 | 8.3 |
| 13 | 43.4 | 36.2 | 7.2 | 27 | 48.5 | 38.3 | 10.2 |
| 14 | 44.4 | 32.5 | 11.9 | | | | |

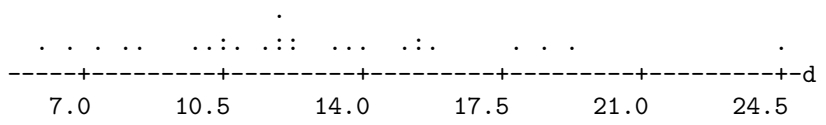



Figure 4.1: *Dot plot of differences.*

As indicated earlier, finding an appropriate standard error is often the most difficult aspect of statistical inference. In problems such as this, where the data are paired, finding the standard error is complicated by the fact that the two observations in each pair are not independent. In data such as these, *different pairs are often independent but observations within a pair are not*.

In paired comparisons, we use a trick to reduce the problem to one sample. It is a simple algebraic fact that the difference of the sample means, $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ is the same as the sample mean of the differences $d_j = y_{1j} - y_{2j}$, i.e., $\bar{d} = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$. Thus $\bar{d}$ is an estimate of the parameter of interest $\mu_1 - \mu_2$. The differences are given in Table 4.1 along with the data. Summary statistics are listed below for each variable and the differences. Note that for the hardness data, $\bar{d} = 12.663 = 47.552 - 34.889 = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$. In particular, if the positive value for $\bar{d}$ means anything (other than random variation), it indicates that part one is harder than part two.

Sample statistics

| Variable | $N_i$ | Mean | Variance | Std. dev. |
|---|---|---|---|---|
| $y_1$ | 27 | 47.552 | 6.79028 | 2.606 |
| $y_2$ | 27 | 34.889 | 26.51641 | 5.149 |
| $d = y_1 - y_2$ | 27 | 12.663 | 17.77165 | 4.216 |

Given that $\bar{d}$ is an estimate of $\mu_1 - \mu_2$, we can base the entire analysis on the differences. The differences constitute a single sample of data, so the standard error of $\bar{d}$ is simply the usual one-sample standard error,

$$\mathrm{SE}(\bar{d}) = s_d/\sqrt{27},$$

where $s_d$ is the sample standard deviation as computed from the 27 differences. The differences are plotted in Figure 4.1. Note that there is one potential outlier. We leave it as an exercise to reanalyze the data with the possible outlier removed.

We now have *Par*, *Est*, and SE(*Est*); it remains to find the appropriate distribution. Figure 4.2 gives a normal plot for the differences. While there is an upward curve at the top due to the possible outlier, the curve is otherwise reasonably straight. The Wilk–Francia statistic of $W' = 0.955$ is above the fifth percentile of the null distribution. With normal data we use the reference distribution

$$\frac{\bar{d} - (\mu_1 - \mu_2)}{s_d/\sqrt{27}} \sim t(27-1)$$

and we are now in a position to perform statistical inferences.

Our observed values of the mean and standard error are $\bar{d} = 12.663$ and $\mathrm{SE}(\bar{d}) = 4.216/\sqrt{27} = 0.811$. From a $t(26)$ distribution, we find $t(.995, 26) = 2.78$. A 99% confidence interval for the difference in hardness has endpoints

$$12.663 \pm 2.78(.811),$$

which gives an interval of, roughly, $(10.41, 14.92)$. We are 99% confident that the population mean hardness for part 1 is between 10.41 and 14.92 units harder than that for part 2.

We can also get a 99% prediction interval for the difference in hardness to be observed on a new welded piece. The prediction interval has endpoints of

$$12.663 \pm 2.78\sqrt{4.216^2 + .811^2}$$

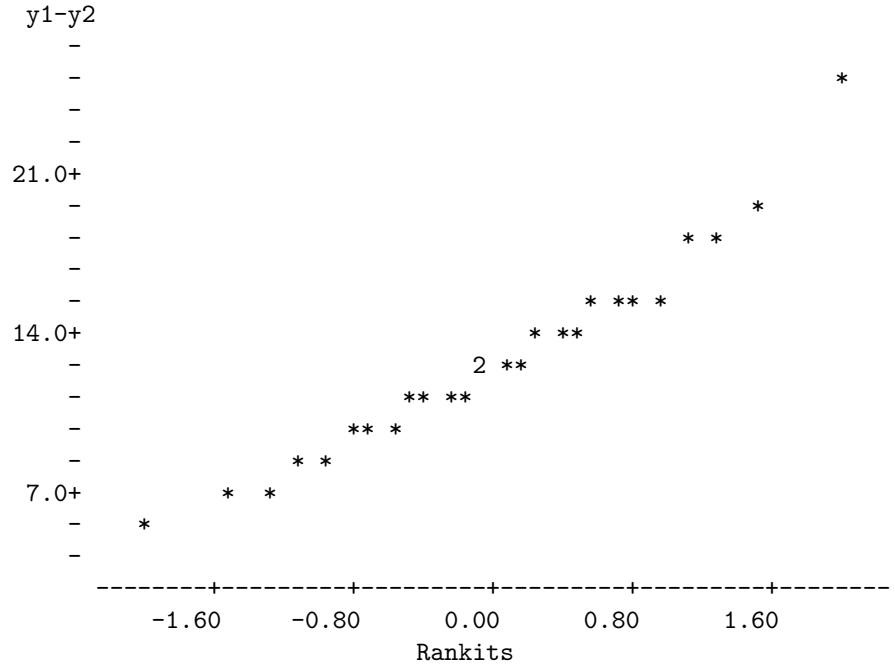



Figure 4.2: *Normal plot of differences,* $W' = .955$.

for an interval of $(.73, 24.60)$.

To test the hypothesis that the two parts have the same hardness, we set up the hypotheses $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$, or equivalently, $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$. The test statistic is

$$\frac{12.663 - 0}{.811} = 15.61.$$

This is far from zero, so the data are inconsistent with the null hypothesis. In other words, there is strong evidence that the hardness of part 1 is different than the hardness of part 2. Since the test statistic is positive, we conclude that $\mu_1 - \mu_2 > 0$ and that part 1 is harder than part 2. Note that this is consistent with our 99% confidence interval $(10.41, 14.92)$, which contains only positive values for $\mu_1 - \mu_2$.

Inferences and predictions for an individual population are made ignoring the other population, i.e., they are made using methods for one sample. For example, using the sample statistics for $y_1$ gives a 99% confidence interval for $\mu_1$, the population mean hardness for part 1, with endpoints

$$47.552 \pm 2.78\sqrt{\frac{6.79028}{27}}$$

and a 99% prediction interval for the hardness of a new piece of part 1 has endpoints

$$47.552 \pm 2.78\sqrt{6.79028 + \frac{6.79028}{27}}$$

and interval $(40.175, 59.929)$. Of course, the use of the $t(26)$ distribution requires that we validate the assumption that the observations on part 1 are a random sample from a normal distribution.

When finding a prediction interval for $y_1$, we can typically improve the interval if we know the corresponding value of $y_2$. As we saw earlier, the 99% prediction interval for a new difference $d = y_1 - y_2$ has $.73 < y_1 - y_2 < 24.60$. If we happen to know that, say, $y_2 = 35$, the interval becomes

$.73 < y_1 - 35 < 24.60$ or $35.73 < y_1 < 59.60$. As it turns out, with these data the new 99% prediction interval for $y_1$ is not an improvement over the interval in the previous paragraph. The new interval is noticeably wider. However, these data are somewhat atypical. Typically in paired data, the two measurements are highly correlated, so that the sample variance of the differences is substantially less than the sample variance of the individual measurements. In such situations, the new interval will be substantially narrower. In these data, the sample variance for the differences is 17.77165 and is actually much larger than the sample variance of 6.79028 for $y_1$. □

The trick of looking at differences between pairs is necessary because the two observations in a pair are not independent. While different pairs of welded parts are assumed to behave independently, it seems unreasonable to *assume* that two hardness measurements on a single item that has been welded together would behave independently. This lack of independence makes it difficult to find a standard error for comparing the sample means unless we look at the differences. In the remainder of this chapter, we consider two-sample problems in which all of the observations are assumed to be independent. The observations in each sample are independent of each other and independent of all the observations in the other sample. Paired comparison problems almost fit those assumptions but they break down at one key point. In a paired comparison, we assume that every observation is independent of the other observations in the same sample and that each observation is independent of all the observations in the other sample *except* for the observation in the other sample that it is paired with. When analyzing two samples, if we can find any reason to identify individuals as being part of a pair, that fact is sufficient to make us treat the data as a paired comparison.

The method of paired comparisons is also the name of a totally different statistical procedure. Suppose one wishes to compare five brands of chocolate chip cookies: $A$, $B$, $C$, $D$, $E$. It would be difficult to taste all five and order them appropriately. As an alternative, one can taste test pairs of cookies, e.g., $(A,B)$, $(A,C)$, $(A,D)$, $(A,E)$, $(B,C)$, $(B,D)$, etc. and identify the better of the two. The benefit of this procedure is that it is much easier to rate two cookies than to rate five. See David (1988) for a survey and discussion of procedures developed to analyze such data.

## 4.2 Two independent samples with equal variances

The most commonly used two-sample technique consists of comparing independent samples from two populations with the same variance. The sample sizes for the two groups are possibly different, say, $N_1$ and $N_2$, and we write the common variance as $\sigma^2$.

EXAMPLE 4.2.1. The data in Table 4.2 are final point totals for an introductory statistics class. The data are divided by the sex of the student. We investigate whether the data display sex differences. The data are plotted in Figure 4.3. Figures 4.4 and 4.5 contain normal plots for the two sets of data. Figure 4.4 is quite straight but Figure 4.5 looks curved. Our analysis is not particularly sensitive to nonnormality and the $W'$ statistic for Figure 4.5 is .937, which is well above the fifth percentile, so we proceed under the assumption that both samples are normal. We also assume that all of the observations are independent. This assumption may be questionable because some students probably studied together, nonetheless, independence seems like a reasonable working assumption. □

The methods in this section rely on the assumption that the two populations are normally distributed and have the same variance. In particular, we assume two independent samples

| Sample | Data | | Distribution |
|---|---|---|---|
| 1 | $y_{11}, y_{12}, \ldots, y_{1N_1}$ | iid | $N(\mu_1, \sigma^2)$ |
| 2 | $y_{21}, y_{22}, \ldots, y_{2N_2}$ | iid | $N(\mu_2, \sigma^2)$ |

and compute summary statistics from the samples. The summary statistics are just the sample mean and the sample variance for each individual sample.

Table 4.2: *Final point totals for an introductory statistics class*

| Females | | | | | Males | | |
|---|---|---|---|---|---|---|---|
| 140 | 125 | 90 | 105 | 145 | 165 | 175 | 135 |
| 135 | 155 | 170 | 140 | 85 | 175 | 160 | 165 |
| 150 | 115 | 125 | 95 | | 170 | 115 | 150 |
| 135 | 145 | 110 | 135 | | 150 | 85 | 130 |
| 110 | 120 | 140 | 145 | | 90 | 95 | 125 |

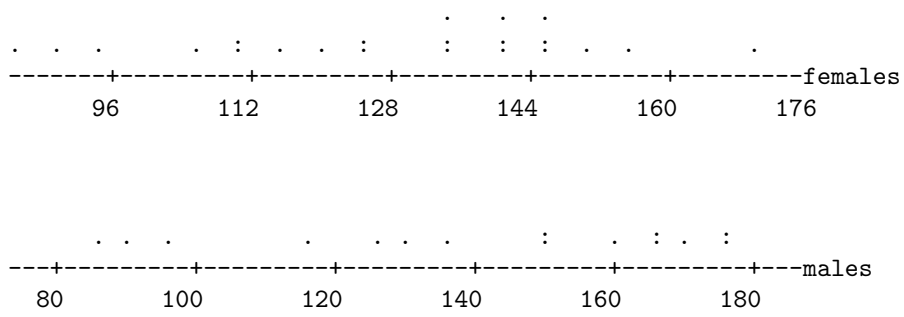



Figure 4.3: *Dot plots for final point totals.*

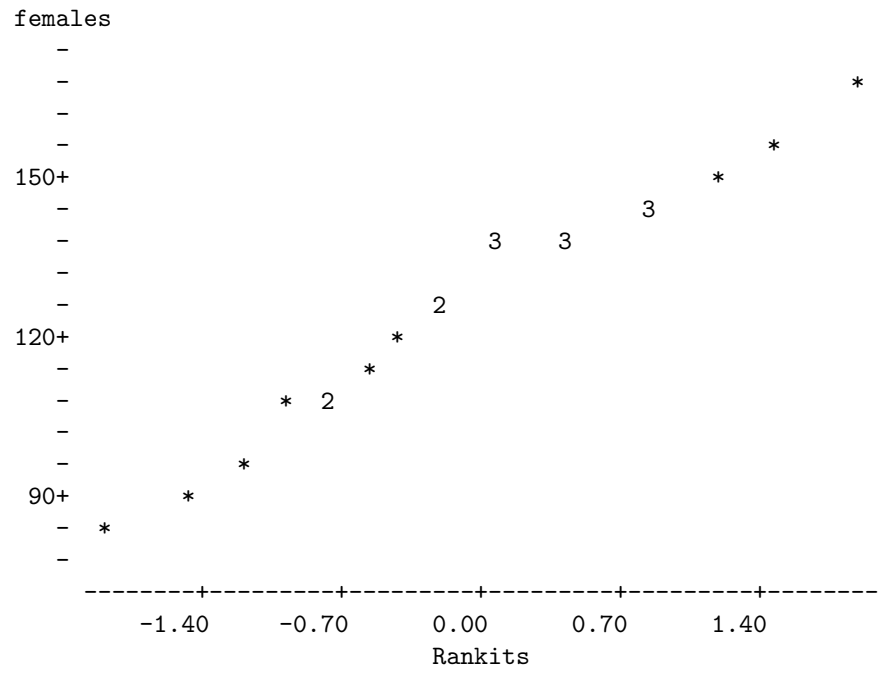



Figure 4.4: *Normal plot for females,* $W' = .974$.

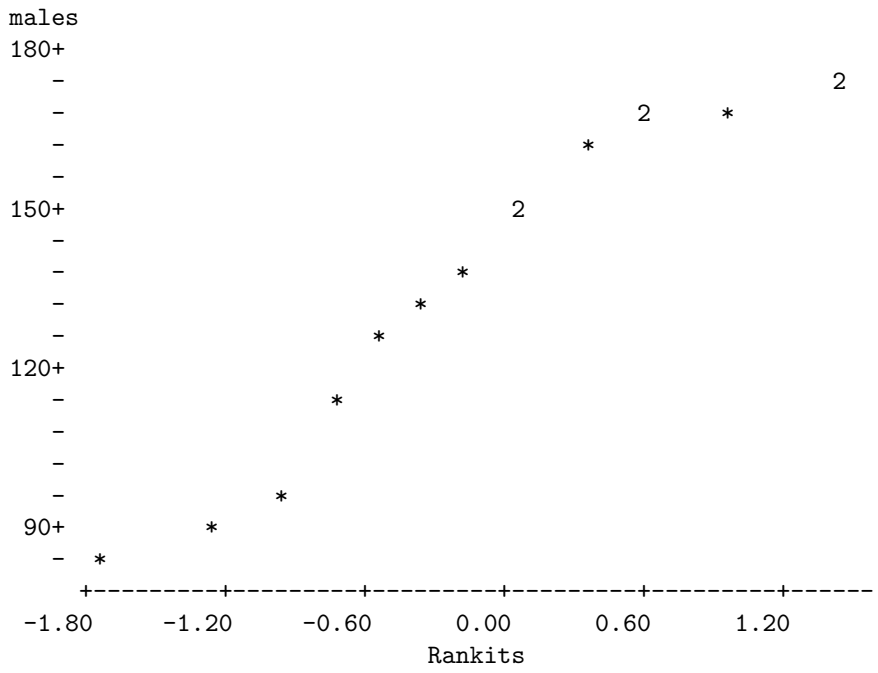



Figure 4.5: *Normal plot for males, $W' = .937$.*

Sample statistics

| Sample | Size | Mean | Variance |
|---|---|---|---|
| 1 | $N_1$ | $\bar{y}_{1\cdot}$ | $s_1^2$ |
| 2 | $N_2$ | $\bar{y}_{2\cdot}$ | $s_2^2$ |

Except for checking the validity of our assumptions, these summary statistics are more than sufficient for the entire analysis. Algebraically, the sample mean for population $i$, $i = 1,2$, is

$$\bar{y}_{i\cdot} \equiv \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij} = \frac{1}{N_i}\left[y_{i1} + y_{i2} + \cdots + y_{iN_i}\right]$$

where the $\cdot$ in $\bar{y}_{i\cdot}$ indicates that the mean is obtained by averaging over $j$, the second subscript in the $y_{ij}$s. The sample means, $\bar{y}_{1\cdot}$ and $\bar{y}_{2\cdot}$, are estimates of $\mu_1$ and $\mu_2$.

The sample variance for population $i$, $i = 1,2$, is

$$\begin{aligned} s_i^2 &= \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_{i\cdot})^2 \\ &= \frac{1}{N_i - 1}\left[(y_{i1} - \bar{y}_{i\cdot})^2 + (y_{i2} - \bar{y}_{i\cdot})^2 + \cdots + (y_{iN_i} - \bar{y}_{i\cdot})^2\right]. \end{aligned}$$

The $s_i^2$s both estimate $\sigma^2$. Combining the $s_i^2$s can yield a better estimate of $\sigma^2$ than either individual estimate. We form a pooled estimate of the variance, say $s_p^2$, by averaging $s_1^2$ and $s_2^2$. With unequal sample sizes an efficient pooled estimate of $\sigma^2$ must be a weighted average of the $s_i^2$s. Obviously, if we have $N_1 = 100\,000$ observations in the first sample and only $N_2 = 10$ observations in the second sample, the variance estimate $s_1^2$ is much better than $s_2^2$ and we want to give it more weight. The weights are the degrees of freedom associated with the estimates. The pooled estimate of the variance is

$$s_p^2 \equiv \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 - 1) + (N_2 - 1)}$$

$$= \frac{1}{N_1+N_2-2}\left[\sum_{j=1}^{N_1}(\bar{y}_{1j}-\bar{y}_{1\cdot})^2+\sum_{j=1}^{N_2}(\bar{y}_{2j}-\bar{y}_{2\cdot})^2\right]$$
$$= \frac{1}{N_1+N_2-2}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(\bar{y}_{ij}-\bar{y}_{i\cdot})^2.$$

The degrees of freedom for $s_p^2$ are $N_1+N_2-2=(N_1-1)+(N_2-1)$, i.e., the sum of the degrees of freedom for the individual estimates $s_i^2$.

EXAMPLE 4.2.2. For the data on final point totals, the sample statistics are given below.

Sample Statistics

| Sample | $N_i$ | $\bar{y}_{i\cdot}$ | $s_i^2$ | $s_i$ |
|---|---|---|---|---|
| females | 22 | 127.954545 | 487.2835498 | 22.07 |
| males | 15 | 139.000000 | 979.2857143 | 31.29 |

From these values, we obtain the pooled estimate of the variance,

$$s_p^2 = \frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2} = \frac{(21)487.28+(14)979.29}{35} = 684.08. \qquad \square$$

We are now in a position to draw statistical inferences about the $\mu_i$s. The main problem in obtaining tests and confidence intervals is in finding appropriate standard errors. The crucial fact is that the samples are independent so that the $\bar{y}_{i\cdot}$s are independent.

For inferences about the difference between the two means, say, $\mu_1-\mu_2$, use the general procedure of Chapter 3 with

$$Par = \mu_1-\mu_2$$

and

$$Est = \bar{y}_{1\cdot}-\bar{y}_{2\cdot}.$$

Note that $\bar{y}_{1\cdot}-\bar{y}_{2\cdot}$ is unbiased for estimating $\mu_1-\mu_2$ because

$$\mathrm{E}(\bar{y}_{1\cdot}-\bar{y}_{2\cdot}) = \mathrm{E}(\bar{y}_{1\cdot})-\mathrm{E}(\bar{y}_{2\cdot}) = \mu_1-\mu_2.$$

The two means are independent, so the variance of $\bar{y}_{1\cdot}-\bar{y}_{2\cdot}$ is the variance of $\bar{y}_{1\cdot}$ plus the variance of $\bar{y}_{2\cdot}$, i.e.,

$$\mathrm{Var}(\bar{y}_{1\cdot}-\bar{y}_{2\cdot}) = \mathrm{Var}(\bar{y}_{1\cdot})+\mathrm{Var}(\bar{y}_{2\cdot}) = \frac{\sigma^2}{N_1}+\frac{\sigma^2}{N_2} = \sigma^2\left[\frac{1}{N_1}+\frac{1}{N_2}\right].$$

The standard error of $\bar{y}_{1\cdot}-\bar{y}_{2\cdot}$ is the estimated standard deviation of $\bar{y}_{1\cdot}-\bar{y}_{2\cdot}$,

$$\mathrm{SE}(\bar{y}_{1\cdot}-\bar{y}_{2\cdot}) = \sqrt{s_p^2\left[\frac{1}{N_1}+\frac{1}{N_2}\right]}.$$

Under our assumption that the original data are normal, the reference distribution is

$$\frac{(\bar{y}_{1\cdot}-\bar{y}_{2\cdot})-(\mu_1-\mu_2)}{\sqrt{s_p^2\left[\frac{1}{N_1}+\frac{1}{N_2}\right]}} \sim t(N_1+N_2-2).$$

The degrees of freedom for the $t$ distribution are the degrees of freedom for $s_p^2$. For nonnormal data with large sample sizes, the reference distribution is $N(0,1)$.

Having identified the parameter, estimate, standard error, and distribution, inferences follow the usual pattern. A 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) \pm t(.975, N_1 + N_2 - 2)\sqrt{s_p^2\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}.$$

A test of hypothesis that the means are equal, say

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_A : \mu_1 \neq \mu_2$$

can be converted into the equivalent hypothesis involving $Par = \mu_1 - \mu_2$, namely

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_A : \mu_1 - \mu_2 \neq 0.$$

The test is handled in the usual way. An $\alpha = .01$ test rejects $H_0$ if

$$\frac{|(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) - 0|}{\sqrt{s_p^2\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}} > t(.995, N_1 + N_2 - 2).$$

In our discussion of comparing differences, we have defined the parameter as $\mu_1 - \mu_2$. We could just as well have defined the parameter as $\mu_2 - \mu_1$. This would have given an entirely equivalent analysis.

Inferences about a single mean, say, $\mu_2$, use the general procedures with $Par = \mu_2$ and $Est = \bar{y}_{2\cdot}$. The variance of $\bar{y}_{2\cdot}$ is $\sigma^2/N_2$, so $\mathrm{SE}(\bar{y}_{2\cdot}) = \sqrt{s_p^2/N_2}$. Note the use of $s_p^2$ rather than $s_2^2$. The reference distribution is $[\bar{y}_{2\cdot} - \mu_2]/\mathrm{SE}(\bar{y}_{2\cdot}) \sim t(N_1 + N_2 - 2)$. A 95% confidence interval for $\mu_2$ is

$$\bar{y}_{2\cdot} \pm t(.975, N_1 + N_2 - 2)\sqrt{s_p^2/N_2}.$$

A 95% prediction interval for a new observation on variable $y_2$ is

$$\bar{y}_{2\cdot} \pm t(.975, N_1 + N_2 - 2)\sqrt{s_p^2 + \frac{s_p^2}{N_2}}.$$

An $\alpha = .01$ test of the hypothesis, say

$$H_0 : \mu_2 = 5 \quad \text{versus} \quad H_A : \mu_2 \neq 5,$$

rejects $H_0$ if

$$\frac{|\bar{y}_{2\cdot} - 5|}{\sqrt{s_p^2/N_2}} > t(.995, N_1 + N_2 - 2).$$

EXAMPLE 4.2.3. For comparing females and males on final point totals, the parameter of interest is

$$Par = \mu_1 - \mu_2$$

where $\mu_1$ indicates the population mean final point total for females and $\mu_2$ indicates the population mean final point total for males. The estimate of the parameter is

$$Est = \bar{y}_{1\cdot} - \bar{y}_{2\cdot} = 127.95 - 139.00 = -11.05\,.$$

The pooled estimate of the variance is $s_p^2 = 684.08$, so the standard error is

$$\mathrm{SE}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \sqrt{s_p^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{684.08\left(\frac{1}{22} + \frac{1}{15}\right)} = 8.7578\,.$$

The data have reasonably normal distributions and the variances are not too different (more on this later), so the reference distribution is taken as

$$\frac{(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{22} + \frac{1}{15}\right)}} \sim t(35)$$

where $35 = N_1 + N_2 - 2$. The tabled value for finding 95% confidence intervals and $\alpha = .05$ two-sided tests is

$$t(.975, 35) = 2.030\,.$$

A 95% confidence interval for $\mu_1 - \mu_2$ has endpoints

$$-11.05 \pm (2.030)8.7578$$

which yields an interval $(-28.8, 6.7)$. We are 95% confident that the population mean scores are between, roughly, 29 points *less* for females and 7 points *more* for females.

An $\alpha = .05$ two-sided test of $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$ is not rejected because 0, the hypothesized value of $\mu_1 - \mu_2$, is contained in the 95% confidence interval for $\mu_1 - \mu_2$. The $P$ value for the test is based on the observed value of the test statistic

$$t_{obs} = \frac{(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) - 0}{\sqrt{s_p^2\left(\frac{1}{22} + \frac{1}{15}\right)}} = \frac{-11.05 - 0}{8.7578} = -1.26\,.$$

The probability of obtaining an observation from a $t(35)$ distribution that is as extreme or more extreme than $|-1.26|$ is 0.216. There is very little evidence that the population mean final point total for females is different (smaller) than the population mean final point total for males. The $P$ value is greater than .2, so, as we established earlier, neither an $\alpha = .05$ nor an $\alpha = .01$ test is rejected. If we were silly enough to do an $\alpha = .25$ test, we would then reject the null hypothesis.

If one claimed that, for whatever reason, females tend to do worse than males in statistics classes, a two-sided test would probably be inappropriate. To test $H_0 : \mu_1 - \mu_2 \leq 0$ versus $H_A : \mu_1 - \mu_2 > 0$, the test statistic is the same but the interpretation is very different. The negative value of the test statistic is consistent with the null hypothesis. The $P$ value is the very large value $1 - .216/2 = .892$. Claiming that females do better would give the opposite one-sided test with a $P$ value of $.216/2 = .108$.

A 95% confidence interval for $\mu_1$, the mean of the females, has endpoints

$$127.95 \pm (2.030)\sqrt{684.08/22}$$

which gives the interval $(116.6, 139.3)$. We are 95% confident that the mean of the final point totals for females is between 117 and 139. A 95% prediction interval for a new observation on a female has endpoints

$$127.95 \pm (2.030)\sqrt{684.08 + \frac{684.08}{22}}$$

which gives the interval $(73.7, 182.2)$. We are 95% confident that a new observation on a female will be between 74 and 182. This assumes that the new observation is randomly sampled from the same population as the previous data.

A test of the assumption of equal variances is left for the final section but we will see in the next section that the results for these data do not depend substantially on the equality of the variances. □

Table 4.3: *Turtle shell heights*

| Female | | | | Male | | | |
|---|---|---|---|---|---|---|---|
| 51 | 38 | 63 | 46 | 39 | 42 | 37 | 43 |
| 51 | 38 | 60 | 51 | 39 | 45 | 35 | 41 |
| 53 | 42 | 62 | 51 | 38 | 45 | 35 | 41 |
| 57 | 42 | 63 | 51 | 40 | 45 | 39 | 41 |
| 55 | 44 | 61 | 48 | 40 | 46 | 38 | 40 |
| 56 | 50 | 67 | 49 | 40 | 47 | 37 | 44 |

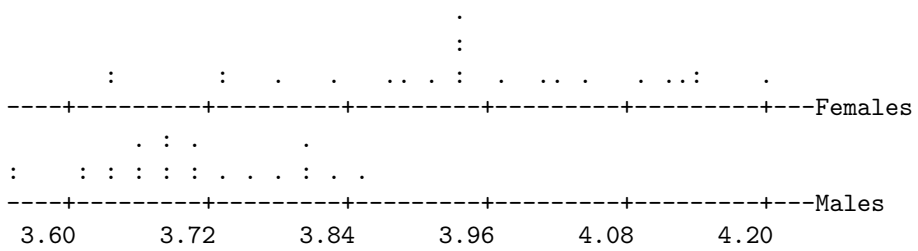



Figure 4.6: *Plot of turtle shell log heights.*

### 4.3 Two independent samples with unequal variances

We now consider two independent samples with unequal variances $\sigma_1^2$ and $\sigma_2^2$. In this section we examine inferences about the means of the two populations. While inferences about means are important, some care is required when drawing practical conclusions about populations with unequal variances. For example, if you want to produce gasoline with an octane of at least 87, you may have a choice between two processes. One process $y_1$ gives octanes distributed as $N(89,4)$ and the other $y_2$ gives $N(90,4)$. The two processes have the same variance, so the process with the higher mean gives more gas with an octane of at least 87. On the other hand, if $y_1$ gives $N(89,4)$ and $y_2$ gives $N(90,16)$, the $y_1$ process with mean 89 has a higher probability of achieving an octane of 87 than the $y_2$ process with mean 90, see Exercise 4.5.10. This is a direct result of the $y_2$ process having more variability. Having given this warning, we proceed with our discussion on drawing statistical inferences for the means.

EXAMPLE 4.3.1. Jolicoeur and Mosimann (1960) present data on the sizes of turtle shells (carapaces). Table 4.3 presents data on the shell heights for 24 females and 24 males. These data are not paired; it is simply a caprice that 24 carapaces were measured for each sex. Our interest centers on estimating the population means for female and male heights, estimating the difference between the heights, and testing whether the difference is zero.

Following Christensen (1990a) and others, we take natural logarithms of the data, i.e.,

$$y_1 = log(\text{female height}) \qquad y_2 = log(\text{male height}).$$

(*All logarithms in this book are natural logarithms.*) The log data are plotted in Figure 4.6. The female heights give the impression of being both larger and more spread out. Figures 4.7 and 4.8 contain normal plots for the females and males respectively. Neither is exceptionally straight but they do not seem too bad. Summary statistics are given below; they are consistent with the visual impressions given by Figure 4.6. The summary statistics will be used in later examples as the basis for our statistical inferences.

| Group | Size | Mean | Variance | Standard deviation |
|---|---|---|---|---|
| Females | 24 | 3.9403 | 0.02493979 | 0.1579 |
| Males | 24 | 3.7032 | 0.00677276 | 0.0823 |

□

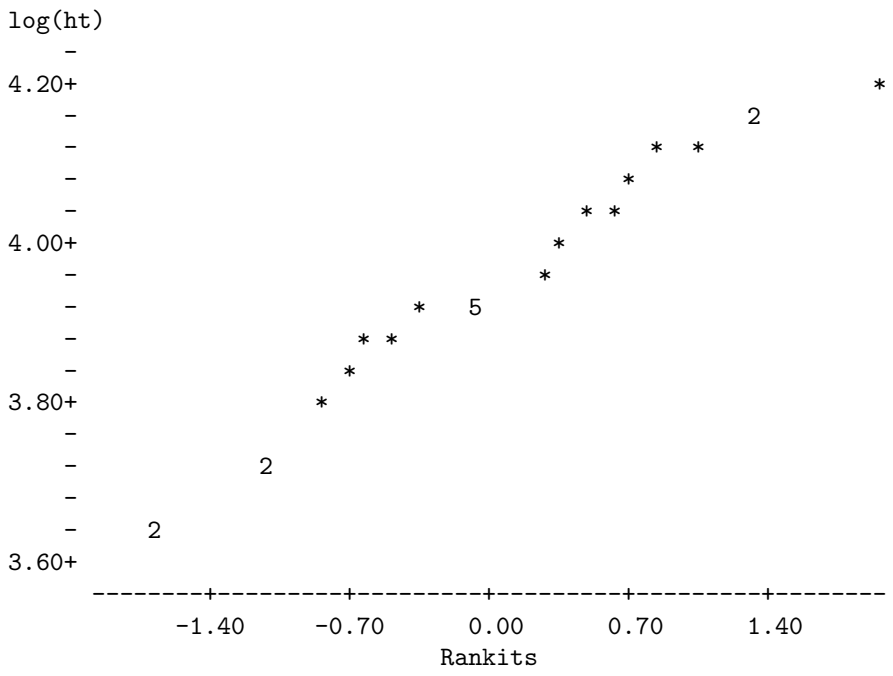



Figure 4.7: *Normal plot for female turtle shell log heights.*

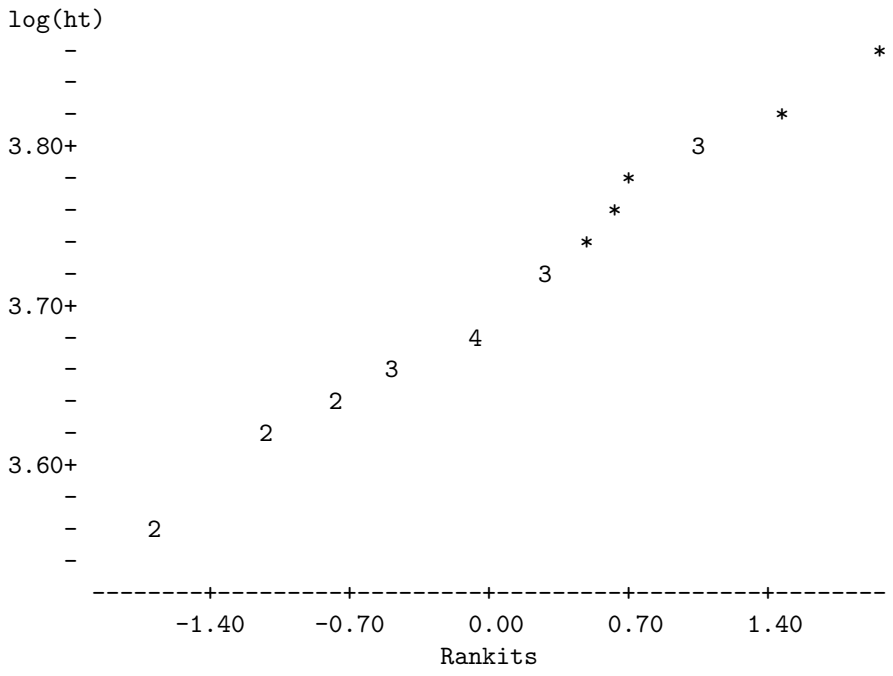



Figure 4.8: *Normal plot for male turtle shell log heights.*

In general we assume two independent samples

| Sample | Data | | Distribution |
|---|---|---|---|
| 1 | $y_{11}, y_{12}, \ldots, y_{1N_1}$ | iid | $N(\mu_1, \sigma_1^2)$ |
| 2 | $y_{21}, y_{22}, \ldots, y_{2N_2}$ | iid | $N(\mu_2, \sigma_2^2)$ |

and compute summary statistics from the samples.

| Sample | Size | Mean | Variance |
|---|---|---|---|
| 1 | $N_1$ | $\bar{y}_{1\cdot}$ | $s_1^2$ |
| 2 | $N_2$ | $\bar{y}_{2\cdot}$ | $s_2^2$ |

Again, the sample means, $\bar{y}_{1\cdot}$ and $\bar{y}_{2\cdot}$, are estimates of $\mu_1$ and $\mu_2$, but now $s_1^2$ and $s_2^2$ estimate $\sigma_1^2$ and $\sigma_2^2$. We have two different variances, so it is inappropriate to pool the variance estimates. Once again, the crucial fact in obtaining a standard error is that the samples are independent.

For inferences about the difference between the two means, say, $\mu_1 - \mu_2$, again use the general procedure with

$$Par = \mu_1 - \mu_2$$

and

$$Est = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}.$$

Just as before, $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ is unbiased for estimating $\mu_1 - \mu_2$. The two sample means are independent so

$$\mathrm{Var}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \mathrm{Var}(\bar{y}_{1\cdot}) + \mathrm{Var}(\bar{y}_{2\cdot}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}.$$

The standard error of $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ is

$$\mathrm{SE}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}.$$

Even when the original data are normal, the appropriate reference distribution is not a $t$ distribution. As a matter of fact, the appropriate reference distribution is not known. However, a good approximate distribution is

$$\frac{(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \sim t(\nu)$$

where

$$\nu \equiv \frac{\left(s_1^2/N_1 + s_2^2/N_2\right)^2}{\left(s_1^2/N_1\right)^2/(N_1 - 1) + \left(s_2^2/N_2\right)^2/(N_2 - 1)} \tag{4.3.1}$$

is an approximate number of degrees of freedom. This approximate distribution was proposed by Satterthwaite (1946) and is discussed by Snedecor and Cochran (1980).

For nonnormal data with large sample sizes, the reference distribution can be taken as $N(0,1)$. Having identified the parameter, estimate, standard error and reference distribution, inferences follow the usual pattern.

EXAMPLE 4.3.2. Consider the turtle data. Recall that

| Group | Size | Mean | Variance | Standard deviation |
|---|---|---|---|---|
| Females | 24 | 3.9403 | 0.02493979 | 0.1579 |
| Males | 24 | 3.7032 | 0.00677276 | 0.0823 |

We begin by considering a test of $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$ or equivalently $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$. As before, $Par = \mu_1 - \mu_2$ and $Est = 3.9403 - 3.7032 = .2371$. The standard error is now

$$\mathrm{SE}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \sqrt{\frac{0.02493979}{24} + \frac{0.00677276}{24}} = .03635.$$

Using $s_1^2/N_1 = 0.02493979/24 = .001039158$ and $s_2^2/N_2 = 0.00677276/24 = .000282198$ in equation (4.3.1), the approximate degrees of freedom are

$$\nu = \frac{(.001039158 + .000282198)^2}{(.001039158)^2/23 + (.000282198)^2/23} = 34.6.$$

An $\alpha = .01$ test is rejected if the observed value of the test statistic is farther from zero than the cutoff value $t(.995, 34.6) \doteq t(.995, 35) = 2.72$. The observed value of the test statistic is

$$t_{obs} = \frac{.2371 - 0}{.03635} = 6.523$$

which is greater than the cutoff value, so the test is rejected. There is evidence at the .01 level that the mean shell height for females is different from the mean shell height for males. Obviously, since $\bar{y}_{1\cdot} - \bar{y}_{2\cdot} = .2371$ is positive, there is evidence that the females have shells of greater height. Actually, the conclusion is that the means of the $log(\text{heights})$ are different, but if these are different we conclude that the mean heights are different.

The 95% confidence interval for the difference between mean log shell heights for females and males, i.e., $\mu_1 - \mu_2$, uses $t(.975, 34.6) \doteq t(.975, 35) = 2.03$. The endpoints are

$$.2371 \pm 2.03\,(.03635),$$

and the interval is $(.163, .311)$. We took logs of the data, so if we transform back to the original scale the interval is $(e^{.163}, e^{.311})$ or $(1.18, 1.36)$. We are 95% confident that the population center for females is, roughly, between one and a sixth and one and a third *times* the shell heights for males. Note that a difference between .163 and .311 on the log scale transforms into a *multiplicative effect* between 1.18 and 1.36 on the original scale. This idea is discussed in more detail in Example 5.1.1.

It is inappropriate to pool the variance estimates, so inferences about $\mu_1$ and $\mu_2$ are performed just as for one sample. The 95% confidence interval for the mean shell height for females, $\mu_1$, uses the estimate $\bar{y}_{1\cdot}$, the standard error $s_1/\sqrt{24}$, and the tabled value $t(.975, 24-1) = 2.069$. It has endpoints

$$3.9403 \pm 2.069\left(0.1579/\sqrt{24}\right)$$

which gives the interval $(3.87, 4.01)$. Transforming to the original scale gives the interval $(47.9, 55.1)$. We are 95% confident that the 'average' height for females' shells is between, roughly, 48 and 55 millimeters. Males also have 24 observations, so the interval for $\mu_2$ also uses $t(.975, 24-1)$, has endpoints

$$3.7032 \pm 2.069\left(0.0823/\sqrt{24}\right),$$

and an interval $(3.67, 3.74)$. Transforming the interval back to the original scale gives $(39.3, 42.1)$. We are 95% confident that the 'average' height for males's shells is between, roughly, 39 and 42 millimeters. The 95% prediction interval for the transformed shell height of a future male has endpoints

$$3.7032 \pm 2.069\left(0.0823\sqrt{1 + \frac{1}{24}}\right),$$

which gives the interval $(3.529, 3.877)$. Transforming the prediction interval back to the original

scale gives $(34.1, 48.3)$. Transforming a prediction interval back to the original scale creates no problems of interpretation. □

EXAMPLE 4.3.3. Reconsider the final point totals data of Section 4.2. Without the assumption of equal variances, the standard error is

$$\mathrm{SE}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \sqrt{\frac{487.28}{22} + \frac{979.29}{15}} = 9.3507\,.$$

From equation (4.3.1), the degrees of freedom for the approximate $t$ distribution are 23. A 95% confidence interval for the difference is $(-30.4, 8.3)$ and the observed value of the statistic for testing equal means is $t_{obs} = -1.18$. This gives a $P$ value for a two-sided test of 0.22. These values are all quite close to those obtained using the equal variance assumption.

□

It is an algebraic fact that if $N_1 = N_2$, the observed value of the test statistic for $H_0 : \mu_1 = \mu_2$ based on unequal variances is the same as that based on equal variances. In the turtle example, the sample sizes are both 24 and the test statistic of 6.523 is the same as the equal variances test statistic. The algebraic equivalence occurs because with equal sample sizes, the standard errors from the two procedures are the same. With equal sample sizes, the only practical difference between the two procedures for examining $Par = \mu_1 - \mu_2$ is in the choice of degrees of freedom for the $t$ distribution. In the turtle example above, the unequal variances procedure had approximately 35 degrees of freedom, while the equal variance procedure has 46 degrees of freedom. The degrees of freedom are sufficiently close that the substantive results of the turtle analysis are essentially the same, regardless of method. The other fact that should be recalled is that the reference distribution associated with $\mu_1 - \mu_2$ for the equal variance method is exactly correct for data that satisfy the assumptions. Even for data that satisfy the unequal variance method assumptions, the reference distribution is just an approximation.

### 4.4 Testing equality of the variances

Throughout this section we assume that the original data are normally distributed and that the two samples are independent. Our goal is to test the hypothesis that the variances are equal, i.e.,

$$H_0 : \sigma_2^2 = \sigma_1^2 \quad \text{versus} \quad H_A : \sigma_2^2 \neq \sigma_1^2.$$

The hypotheses can be converted into equivalent hypotheses,

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = 1 \quad \text{versus} \quad H_A : \frac{\sigma_2^2}{\sigma_1^2} \neq 1.$$

An obvious test statistic is

$$\frac{s_2^2}{s_1^2}.$$

We will reject the hypothesis of equal variances if the test statistic is too much greater than 1 or too much less than 1. As always, the problem is in identifying a precise meaning for 'too much'. To do this, we need to know the distribution of the test statistic when the variances are equal. The distribution is known as an $F$ distribution, i.e., if $H_0$ is true

$$\frac{s_2^2}{s_1^2} \sim F(N_2 - 1, N_1 - 1).$$

The distribution depends on the degrees of freedom for the two estimates. The first parameter in $F(N_2 - 1, N_1 - 1)$ is $N_2 - 1$, the degrees of freedom for the variance estimate in the numerator of

$s_2^2/s_1^2$, and the second parameter is $N_1 - 1$, the degrees of freedom for the variance estimate in the denominator. The test statistic $s_2^2/s_1^2$ is nonnegative, so our reference distribution $F(N_2 - 1, N_1 - 1)$ is nonnegative. Tables are given in Appendix B.

In some sense, the $F$ distribution is 'centered' around one and we reject $H_0$ if $s_2^2/s_1^2$ is too large or too small to have reasonably come from an $F(N_2 - 1, N_1 - 1)$ distribution. An $\alpha = .01$ level test is rejected, i.e., we conclude that $\sigma_2^2 \neq \sigma_1^2$, if

$$\frac{s_2^2}{s_1^2} > F(.995, N_2 - 1, N_1 - 1)$$

or if

$$\frac{s_2^2}{s_1^2} < F(.005, N_2 - 1, N_1 - 1)$$

where $F(.995, N_2 - 1, N_1 - 1)$ cuts off the top .005 of the distribution and $F(.005, N_2 - 1, N_1 - 1)$ cuts off the bottom .005 of the distribution. It is rare that one finds the bottom percentiles of an $F$ distribution tabled but they can be obtained from the top percentiles. In particular,

$$F(.005, N_2 - 1, N_1 - 1) = \frac{1}{F(.995, N_1 - 1, N_2 - 1)}.$$

Note that the degrees of freedom have been reversed in the right-hand side of the equality.

The procedure for this test does not fit within the general procedures outlined in Chapter 3. It has been indicated all along that results for variances do not fit the general pattern. Although we have a parameter, $\sigma_2^2/\sigma_1^2$, and an estimate of the parameter, $s_2^2/s_1^2$, we do not have a standard error or a reference distribution that is symmetric about zero. In fact, the $F$ distribution is not symmetric though we rely on it being 'centered' about 1.

EXAMPLE 4.4.1. We again consider the log turtle height data. The sample variance of log female heights is $s_1^2 = 0.02493979$ and the sample variance of log male heights is $s_2^2 = 0.00677276$. An $\alpha = .01$ level test is rejected, i.e., we conclude that $\sigma_2^2 \neq \sigma_1^2$, if

$$.2716 = \frac{0.00677276}{0.02493979} = \frac{s_2^2}{s_1^2} > F(.995, 23, 23) = 3.04$$

or if

$$.2716 < F(.005, 23, 23) = \frac{1}{F(.995, 23, 23)} = \frac{1}{3.04} = .33.$$

The second of these inequalities is true, so the null hypothesis of equal variances is rejected at the .01 level. We have evidence that $\sigma_2^2 \neq \sigma_1^2$ and, since the statistic is less than one, evidence that $\sigma_2^2 < \sigma_1^2$. □

EXAMPLE 4.4.2. Consider again the final point total data. The sample variance for females is $s_1^2 = 487.28$ and the sample variance for males is $s_2^2 = 979.29$. The test statistic is

$$\frac{s_1^2}{s_2^2} = \frac{487.28}{979.29} = 0.498\,.$$

Naturally, it does not matter which variance estimate we put in the numerator as long as we keep the degrees of freedom straight. The observed test statistic is not less than $1/F(.95, 14, 21) = 1/2.197 = .455$ nor greater than $F(.95, 21, 14) = 2.377$, so the test cannot be rejected at the $\alpha = .10$ level. □

In practice, *tests for the equality of variances are rarely performed*. Typically, the main emphasis is on drawing conclusions about the $\mu_i$s; the motivation for testing equality of variances is

frequently to justify the use of the pooled estimate of the variance. The test assumes that the null hypothesis of equal variances is true and data that are inconsistent with the assumptions indicate that the assumptions are false. We generally take this to indicate that the assumption about the null hypothesis is false, but, in fact, unusual data may be obtained if any of the assumptions are invalid. The equal variances test assumes that the data are independent and normal and that the variances are equal. Minor deviations from normality may cause the test to be rejected. While procedures for comparing $\mu_i$s based on the pooled estimate of the variance are sensitive to unequal variances, they are not particularly sensitive to nonnormality. The test for equality of variances is so sensitive to nonnormality that when rejecting this test one has little idea if the problem is really unequal variances or if it is nonnormality. Thus one has little idea whether there is a problem with the pooled estimate procedures or not. Since the test is not very informative, it is rarely performed. However, *studying this test prepares one for examining the important analysis of variance F test that is treated in the next chapter*.

*Minitab commands*

Minitab can be used to get the $F$ percentiles reported in Example 4.4.1.

```
MTB > invcdf .995
SUBC> f 23 23.
MTB > invcdf .005
SUBC> f 23 23.
```

*Theory*

The $F$ distribution used here is related to the fact that for normal data

$$\frac{(N_i-1)s_i^2}{\sigma_i^2} \sim \chi^2(N_i-1).$$

**Definition 4.4.3.** An $F$ distribution is the ratio of two independent chi-squared random variables divided by their degrees of freedom. The numerator and denominator degrees of freedom for the $F$ distribution are the degrees of freedom for the respective chi-squares.

In this problem, the two chi-squared random variables divided by their degrees of freedom are

$$\frac{(N_i-1)s_i^2/\sigma_i^2}{N_i-1} = \frac{s_i^2}{\sigma_i^2}$$

$i = 1,2$. They are independent because they are taken from independent samples and their ratio is

$$\frac{s_2^2}{\sigma_2^2} \Big/ \frac{s_1^2}{\sigma_1^2} = \frac{s_2^2}{s_1^2}\frac{\sigma_1^2}{\sigma_2^2}.$$

When the null hypothesis is true, i.e., $\sigma_2^2/\sigma_1^2 = 1$, by definition, we get

$$\frac{s_2^2}{s_1^2} \sim F(N_2-1, N_1-1),$$

so the test statistic has an $F$ distribution under the null hypothesis.

Note that we could equally well have reversed the roles of the two groups and set the test up as

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{versus} \quad H_A : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Table 4.4: *Weights of rats on thiouracil*

| Rat | Start | Finish | Rat | Start | Finish |
|---|---|---|---|---|---|
| 1 | 61 | 129 | 6 | 51 | 119 |
| 2 | 59 | 122 | 7 | 56 | 108 |
| 3 | 53 | 133 | 8 | 58 | 138 |
| 4 | 59 | 122 | 9 | 46 | 107 |
| 5 | 51 | 140 | 10 | 53 | 122 |

Table 4.5: *Weight gain comparison*

| Control | | Thyroxin | |
|---|---|---|---|
| 115 | 107 | 132 | 88 |
| 117 | 90 | 84 | 119 |
| 133 | 91 | 133 | |
| 115 | 91 | 118 | |
| 95 | 112 | 87 | |

with the test statistic

$$\frac{s_1^2}{s_2^2}.$$

An $\alpha$ level test is rejected if

$$\frac{s_1^2}{s_2^2} > F\left(1-\frac{\alpha}{2}, N_1-1, N_2-1\right)$$

or if

$$\frac{s_1^2}{s_2^2} < F\left(\frac{\alpha}{2}, N_1-1, N_2-1\right).$$

Using the fact that for any $\alpha$ between zero and one and any degrees of freedom $r$ and $s$,

$$F(\alpha, r, s) = \frac{1}{F(1-\alpha, s, r)}, \tag{4.4.1}$$

it is easily seen that this test is equivalent to the one we constructed. Relation (4.4.1) is a result of the fact that with equal variances both $s_2^2/s_1^2$ and $s_1^2/s_2^2$ have $F$ distributions. Clearly, the smallest, say, 5% of values from $s_2^2/s_1^2$ are also the largest 5% of the values of $s_1^2/s_2^2$.

## 4.5 Exercises

EXERCISE 4.5.1. Box (1950) gave data on the weights of rats that were given the drug Thiouracil. The rats were measured at the start of the experiment and at the end of the experiment. The data are given in Table 4.4. Give a 99% confidence interval for the difference in weights between the finish and the start. Test the null hypothesis that the population mean weight gain was less than or equal to 50 with $\alpha = .02$.

EXERCISE 4.5.2. Box (1950) also considered data on rats given Thyroxin and a control group of rats. The weight *gains* are given in Table 4.5. Give a 95% confidence interval for the difference in weight gains between the Thyroxin group and the control group. Give an $\alpha = .05$ test of whether the control group has weight gains no greater than the Thyroxin group.

EXERCISE 4.5.3. Conover (1971, p. 226) considered data on the physical fitness of male seniors in a particular high school. The seniors were divided into two groups based on whether they lived

Table 4.6: *Physical fitness of male high school seniors*

| | | | | | | |
|---|---|---|---|---|---|---|
| Town | 12.7 | 16.9 | 7.6 | 2.4 | 6.2 | 9.9 |
| Boys | 14.2 | 7.9 | 11.3 | 6.4 | 6.1 | 10.6 |
| | 12.6 | 16.0 | 8.3 | 9.1 | 15.3 | 14.8 |
| | 2.1 | 10.6 | 6.7 | 6.7 | 10.6 | 5.0 |
| | 17.7 | 5.6 | 3.6 | 18.6 | 1.8 | 2.6 |
| | 11.8 | 5.6 | 1.0 | 3.2 | 5.9 | 4.0 |
| Farm | 14.8 | 7.3 | 5.6 | 6.3 | 9.0 | 4.2 |
| Boys | 10.6 | 12.5 | 12.9 | 16.1 | 11.4 | 2.7 |

Table 4.7: *Turtle lengths*

| Females | | | | Males | | | |
|---|---|---|---|---|---|---|---|
| 98 | 138 | 123 | 155 | 121 | 104 | 116 | 93 |
| 103 | 138 | 133 | 155 | 125 | 106 | 117 | 94 |
| 103 | 141 | 133 | 158 | 127 | 107 | 117 | 96 |
| 105 | 147 | 133 | 159 | 128 | 112 | 119 | 101 |
| 109 | 149 | 134 | 162 | 131 | 113 | 120 | 102 |
| 123 | 153 | 136 | 177 | 135 | 114 | 120 | 103 |

on a farm or in town. The results in Table 4.6 are from a physical fitness test administered to the students. High scores indicate that an individual is physically fit. Give a 95% confidence interval for the difference in mean fitness scores between the town and farm students. Test the hypothesis of no difference at the $\alpha = .10$ level. Give a 99% confidence interval for the mean fitness of town boys. Give a 99% prediction interval for a future fitness score for a farm boy.

EXERCISE 4.5.4. Use the data of Exercise 4.5.3 to test whether the fitness scores for farm boys are more or less variable than fitness scores for town boys.

EXERCISE 4.5.5. Jolicoeur and Mosimann (1960) gave data on turtle shell *lengths*. The data for females and males are given in Table 4.7. Explore the need for a transformation. Test whether there is a difference in lengths using $\alpha = .01$. Give a 95% confidence interval for the difference in lengths.

EXERCISE 4.5.6. Koopmans (1987) gave the data in Table 4.8 on verbal ability test scores for 8 year-olds and 10 year-olds. Test whether the two groups have the same mean with $\alpha = .01$ and give a 95% confidence interval for the difference in means. Give a 95% prediction interval for a new 10 year old. Check your assumptions.

EXERCISE 4.5.7. Burt (1966) and Weisberg (1985) presented data on IQ scores for identical twins that were raised apart, one by foster parents and one by the genetic parents. Variable $y_1$ is the IQ score for a twin raised by foster parents, while $y_2$ is the corresponding IQ score for the twin raised by the genetic parents. The data are given in Table 4.9.

We are interested in the difference between $\mu_1$, the population mean for twins raised by foster

Table 4.8: *Verbal ability test scores*

| 8 yr. olds | | | 10 yr. olds | | |
|---|---|---|---|---|---|
| 324 | 344 | 448 | 428 | 399 | 414 |
| 366 | 390 | 372 | 366 | 412 | 396 |
| 322 | 434 | 364 | 386 | 436 | |
| 398 | 350 | | 404 | 452 | |

Table 4.9: *Burt's IQ data*

| Case | $y_1$ | $y_2$ | Case | $y_1$ | $y_2$ | Case | $y_1$ | $y_2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 82 | 82 | 10 | 93 | 82 | 19 | 97 | 87 |
| 2 | 80 | 90 | 11 | 95 | 97 | 20 | 87 | 93 |
| 3 | 88 | 91 | 12 | 88 | 100 | 21 | 94 | 94 |
| 4 | 108 | 115 | 13 | 111 | 107 | 22 | 96 | 95 |
| 5 | 116 | 115 | 14 | 63 | 68 | 23 | 112 | 97 |
| 6 | 117 | 129 | 15 | 77 | 73 | 24 | 113 | 97 |
| 7 | 132 | 131 | 16 | 86 | 81 | 25 | 106 | 103 |
| 8 | 71 | 78 | 17 | 83 | 85 | 26 | 107 | 106 |
| 9 | 75 | 79 | 18 | 93 | 87 | 27 | 98 | 111 |

Table 4.10: *Atomic weights in 1931 and 1936*

| Compound | 1931 | 1936 | Compound | 1931 | 1936 |
|---|---|---|---|---|---|
| Arsenic | 74.93 | 74.91 | Lanthanum | 138.90 | 138.92 |
| Caesium | 132.81 | 132.91 | Osmium | 190.8 | 191.5 |
| Columbium | 93.3 | 92.91 | Potassium | 39.10 | 39.096 |
| Iodine | 126.932 | 126.92 | Radium | 225.97 | 226.05 |
| Krypton | 82.9 | 83.7 | Ytterbium | 173.5 | 173.04 |

parents, and $\mu_2$, the population mean for twins raised by genetic parents. Analyze the data. Check your assumptions.

EXERCISE 4.5.8. Table 4.10 presents data given by Shewhart (1939, p. 118) on various atomic weights as reported in 1931 and again in 1936. Analyze the data. Check your assumptions.

EXERCISE 4.5.9. Reanalyze the data of Example 4.1.1 after deleting the one possible outlier. Does the analysis change much? If so, how?

EXERCISE 4.5.10. Let $y_1 \sim N(89,4)$ and $y_2 \sim N(90,16)$. Show that $\Pr[y_1 \geq 87] > \Pr[y_2 \geq 87]$, so that the population with the lower mean has a higher probability of exceeding 87. Recall that $(y_1 - 89)/\sqrt{4} \sim N(0,1)$ with a similar result for $y_2$ so that both probabilities can be rewritten in terms of a $N(0,1)$.

EXERCISE 4.5.11. Mandel (1972) reported stress test data on elongation for a certain type of rubber. Four pieces of rubber sent to one laboratory yielded a sample mean and variance of 56.50 and 5.66, respectively. Four different pieces of rubber sent to another laboratory yielded a sample mean and variance of 52.50 and 6.33, respectively. Are the data two independent samples or a paired comparison? Is the assumption of equal variances reasonable? Give a 99% confidence interval for the difference in population means and give an approximate $P$ value for testing that there is no difference between population means.

EXERCISE 4.5.12. Bethea et al. (1985) reported data on the peel-strengths of adhesives. Some of the data are presented in Table 4.11. Give an approximate $P$ value for testing no difference between adhesives, a 95% confidence interval for the difference between mean peel-strengths, and a 95% prediction interval for a new observation on Adhesive A.

EXERCISE 4.5.13. Garner (1956) presented data on the tensile strength of fabrics. Here we consider a subset of the data. The complete data and a more extensive discussion of the experimental procedure are given in Exercise 11.5.2. The experiment involved testing fabric strengths on different machines. Eight homogeneous strips of cloth were divided into samples and each machine was used

Table 4.11: *Peel-strengths*

| Adhesive | Observations | | | | | |
|---|---|---|---|---|---|---|
| A | 60 | 63 | 57 | 53 | 56 | 57 |
| B | 52 | 53 | 44 | 48 | 48 | 53 |

Table 4.12: *Tensile strength*

| Strip | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $m_1$ | 18 | 9 | 7 | 6 | 10 | 7 | 13 | 1 |
| $m_2$ | 7 | 11 | 11 | 4 | 8 | 12 | 5 | 11 |

on a sample from each strip. The data are given in Table 4.12. Are the data two independent samples or a paired comparison? Give a 98% confidence interval for the difference in population means. Give an approximate $P$ value for testing that there is no difference between population means. What is the result of an $\alpha = .05$ test?

EXERCISE 4.5.14. Snedecor and Cochran (1967) presented data on the number of acres planted in corn for two sizes of farms. Size was measured in acres. Some of the data are given in Table 4.13. Are the data two independent samples or a paired comparison? Is the assumption of equal variances reasonable? Test for differences between the farms of different sizes. Clearly state your $\alpha$ level. Give a 98% confidence interval for the mean difference between different farms.

EXERCISE 4.5.15. Snedecor and Haber (1946) presented data on cutting dates of asparagus. On two plots of land, asparagus was grown every year from 1929 to 1938. On the first plot the asparagus was cut on June 1, while on the second plot the asparagus was cut on June 15. Note that growing conditions will vary considerably from year to year. Also note that the data presented have cutting dates confounded with the plots of land. If one plot of land is intrinsically better for growing asparagus than the other, there will be no way of separating that effect from the effect of cutting dates. Are the data two independent samples or a paired comparison? Give a 95% confidence interval for the difference in population means and give an approximate $P$ value for testing that there is no difference between population means. Give a 95% prediction interval for the difference in a new year. The data are given in Table 4.14.

EXERCISE 4.5.16. Snedecor (1945b) presented data on a pesticide spray. The treatments were the number of units of active ingredient contained in the spray. Several different sources for breeding mediums were used and each spray was applied on each distinct breeding medium. The data consisted of numbers of dead adults flies found in cages that were set over the breeding medium

Table 4.13: *Acreage in corn for different farm acreages*

| Size | Corn acreage | | | | |
|---|---|---|---|---|---|
| 240 | 65 | 80 | 65 | 85 | 30 |
| 400 | 75 | 35 | 140 | 90 | 110 |

Table 4.14: *Cutting dates*

| Year | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 |
|---|---|---|---|---|---|---|---|---|---|---|
| June 1 | 201 | 230 | 324 | 512 | 399 | 891 | 449 | 595 | 632 | 527 |
| June 15 | 301 | 296 | 543 | 778 | 644 | 1147 | 585 | 807 | 804 | 749 |

Table 4.15: *Dead adult flies*

| Medium | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 0 units | 423 | 326 | 246 | 141 | 208 | 303 | 256 |
| 8 units | 414 | 127 | 206 | 78 | 172 | 45 | 103 |

containers. Some of the data are presented in Table 4.15. Give a 95% confidence interval for the difference in population means. Give an approximate $P$ value for testing that there is no difference between population means and an $\alpha = .05$ test. Give a 95% prediction interval for a new observation with 8 units. Give a 95% prediction interval for a new observation with 8 units when the corresponding 0 unit value is 300.

EXERCISE 4.5.17. Using the data of Example 4.2.1 give a 95% prediction interval for the difference in total points between a new female and a new male. This was not discussed earlier so it requires a deeper understanding of Section 3.5.

Chapter 5

# One-way analysis of variance

Analysis of variance (ANOVA) involves comparing random samples from several populations. Often the samples arise from observing experimental units with different treatments applied to them and we refer to the populations as treatment groups. The sample sizes for the treatment groups are possibly different, say, $N_i$ and we assume that the samples are all independent. Moreover, we assume that each population has the same variance and is normally distributed.

## 5.1 Introduction and examples

EXAMPLE 5.1.1. Table 5.1 gives data from Koopmans (1987, p. 409) on the ages at which suicides were committed in Albuquerque during 1978. Ages are listed by ethnic group. The data are plotted in Figure 5.1. The assumption is that the observations in each group are a random sample from some population. While it is not clear what these populations would be, we proceed to examine the data. Note that there are fewer Native Americans in the study than either Hispanics or non-Hispanic Caucasians; moreover the ages for Native Americans seem to be both lower and less variable than for the other groups. The ages for Hispanics seem to be a bit lower than for non-Hispanic Caucasians.

Summary statistics are given below for the three groups.

Sample statistics: suicide ages

| Group | $N_i$ | $\bar{y}_{i\cdot}$ | $s_i^2$ | $s_i$ |
|---|---|---|---|---|
| Caucasians | 44 | 41.66 | 282.9 | 16.82 |
| Hispanics | 34 | 35.06 | 268.3 | 16.38 |
| Native Am. | 15 | 25.07 | 74.4 | 8.51 |

The sample standard deviation for the Native Americans is about half the size of the others. To

Table 5.1: *Suicide ages*

| Non-Hispanic Caucasians | | | | Hispanics | | | Native Americans | |
|---|---|---|---|---|---|---|---|---|
| 21 | 31 | 28 | 52 | 50 | 27 | 45 | 26 | 23 |
| 55 | 31 | 24 | 27 | 31 | 22 | 57 | 17 | 25 |
| 42 | 32 | 53 | 76 | 29 | 20 | 22 | 24 | 23 |
| 25 | 43 | 66 | 44 | 21 | 51 | 48 | 22 | 22 |
| 48 | 57 | 90 | 35 | 27 | 60 | 48 | 16 | |
| 22 | 42 | 27 | 32 | 34 | 15 | 14 | 21 | |
| 42 | 34 | 48 | 26 | 76 | 19 | 52 | 36 | |
| 53 | 39 | 47 | 51 | 35 | 24 | 29 | 18 | |
| 21 | 24 | 49 | 19 | 55 | 24 | 21 | 48 | |
| 21 | 79 | 53 | 27 | 24 | 18 | 28 | 20 | |
| 31 | 46 | 62 | 58 | 68 | 43 | 17 | 35 | |
| | | | | 38 | | | | |

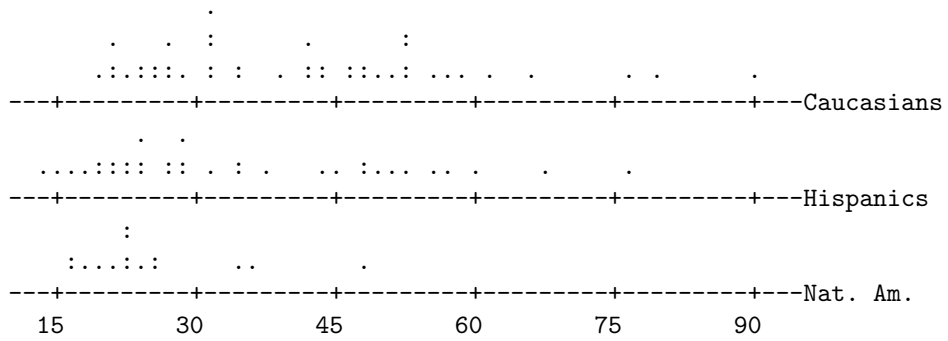



Figure 5.1: *Dot plots of suicide age data.*

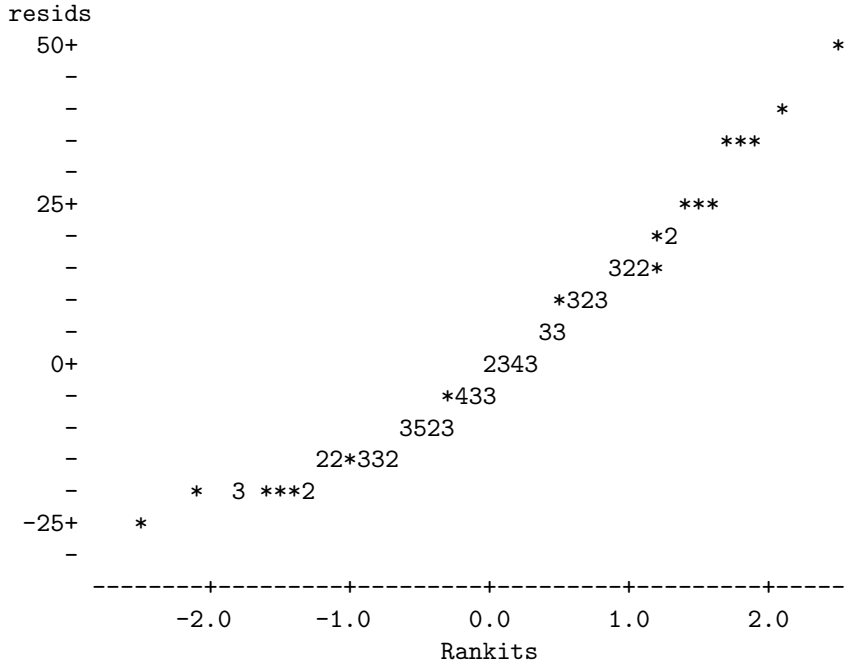



Figure 5.2: *Normal plot of suicide residuals,* $W' = .945$.

evaluate the combined normality of the data, we subtracted the appropriate group mean from each observation, i.e., we computed *residuals*

$$\hat{\varepsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot},$$

where $y_{ij}$ is the $j$th observation in the $i$th group and $\bar{y}_{i\cdot}$ is the sample mean from the $i$th group. We then did a normal plot of the residuals. One normal plot for all of the $y_{ij}$s would not be appropriate because they have different means, $\mu_i$. The residuals adjust for the different means. Of course with the reasonably large samples available here for each group, it would be permissible to do three separate normal plots, but in other situations with small samples for each group, individual normal plots would not contain enough observations to be of any value. The normal plot for the residuals is given in Figure 5.2. The plot is based on $n = 44 + 34 + 15 = 93$ observations. This is quite a large number, so if the data are normal the plot should be quite straight. In fact, the plot seems reasonably curved.

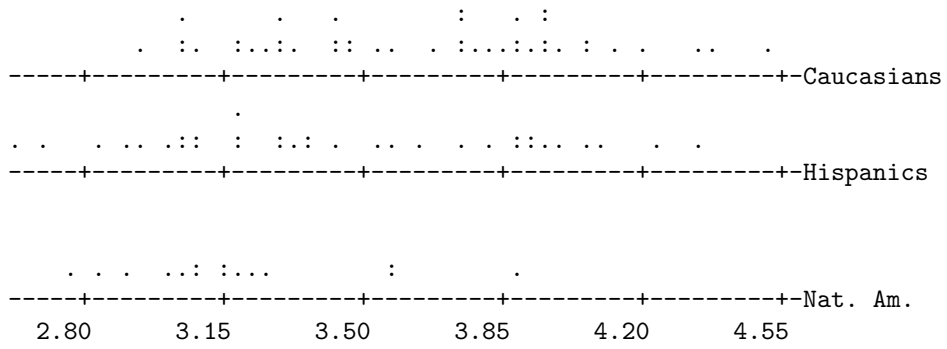



Figure 5.3: *Dotplots of log suicide age data.*

In order to improve the quality of the assumptions of equal variances and normality, we consider transformations of the data. In particular, consider transforming to $\log(y_{ij})$. Figure 5.3 contains the plot of the transformed data. The variability in the groups seems more nearly the same. This is confirmed by the sample statistics given below.

Sample statistics: log of suicide ages

| Group | $N_i$ | $\bar{y}_{i\cdot}$ | $s_i^2$ | $s_i$ |
|---|---|---|---|---|
| Caucasians | 44 | 3.6521 | 0.1590 | 0.3987 |
| Hispanics | 34 | 3.4538 | 0.2127 | 0.4612 |
| Native Am. | 15 | 3.1770 | 0.0879 | 0.2965 |

The largest sample standard deviation is only about 1.5 times the smallest. The normal plot of residuals for the transformed data is given in Figure 5.4; it seems considerably straighter than the normal plot for the untransformed data.

All in all, the logs of the original data seem to satisfy the assumptions reasonably well and considerably better than the untransformed data. The square roots of the data were also examined as a possible transformation. While the square roots seem to be an improvement over the original scale, they do not seem to satisfy the assumptions nearly as well as the log transformed data.

A basic assumption in analysis of variance is that the variance is the same for all populations. As we did for two independent samples with the same variance, we can compute a pooled estimate of the variance. Again, this is a weighted average of the variance estimates from the individual groups with weights that are the individual degrees of freedom. In analysis of variance, the pooled estimate of the variance is called the *mean squared error* (*MSE*). For the logs of the suicide age data, the mean squared error is

$$MSE = \frac{(44-1)(.1590)+(34-1)(.2127)+(15-1)(.0879)}{(44-1)+(34-1)+(15-1)} = .168.$$

The degrees of freedom for this estimate are the sum of the degrees of freedom for the individual estimates; the degrees of freedom for error ($dfE$) are

$$dfE = (44-1)+(34-1)+(15-1) = 44+34+15-3 = 90.$$

The data have an approximate normal distribution, so we can use $t(90)$ as the reference distribution for statistical inference.

We can now perform statistical inferences for a variety of parameters using our standard procedure involving a *Par*, an *Est*, a $\mathrm{SE}(Est)$, and a known distribution symmetric about 0 for $[Est-Par]/\mathrm{SE}(Est)$. In this example, perhaps the most useful things to look at are simply whether

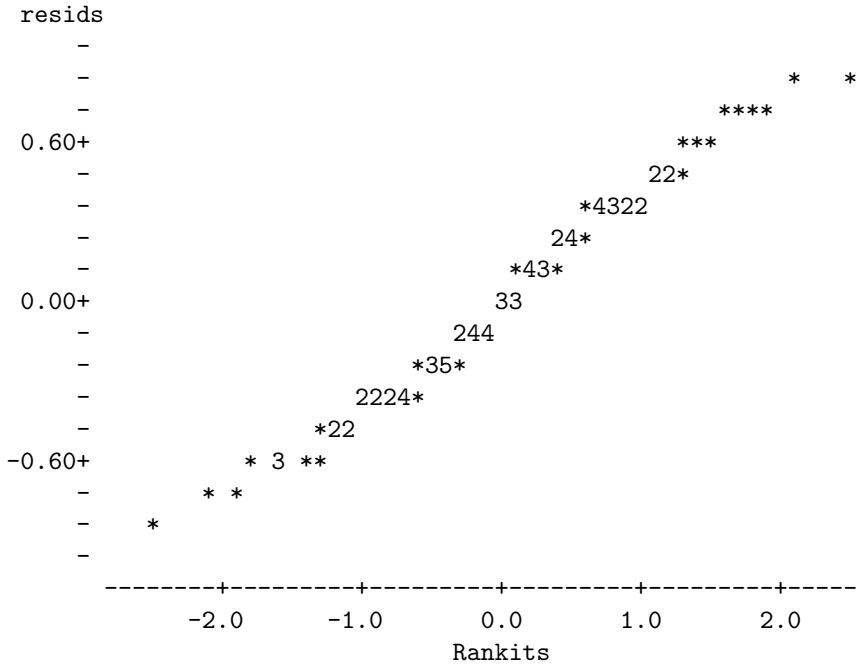



Figure 5.4: *Normal plot of suicide residuals, log data,* $W' = .986$.

there is evidence of any age differences in the three groups. Let $\mu_C$, $\mu_H$, and $\mu_N$ denote the population means for the log ages of the non-Hispanic Caucasian, Hispanic, and Native American groups respectively. Parameters of interest, with their estimates and the variances of the estimates, are given below.

| *Par* | *Est* | Var(*Est*) |
|---|---|---|
| $\mu_C - \mu_H$ | $3.6521 - 3.4538$ | $\sigma^2\left(\frac{1}{44} + \frac{1}{34}\right)$ |
| $\mu_C - \mu_N$ | $3.6521 - 3.1770$ | $\sigma^2\left(\frac{1}{44} + \frac{1}{15}\right)$ |
| $\mu_H - \mu_N$ | $3.4538 - 3.1770$ | $\sigma^2\left(\frac{1}{34} + \frac{1}{15}\right)$ |

The estimates and variances are obtained exactly as in Section 4.2. The standard errors of the estimates are obtained by substituting *MSE* for $\sigma^2$ in the variance formula and taking the square root. Below are given the estimates, standard errors, the $t_{obs}$ values for testing $H_0 : Par = 0$, the two-sided test $P$ values, and the 99% confidence intervals for *Par*. The confidence intervals require the value $t(.995, 90) = 2.631$. This $t$ table value appears repeatedly in our discussion.

| *Par* | *Est* | SE(*Est*) | $t_{obs}$ | $P$ | 99% CI |
|---|---|---|---|---|---|
| $\mu_C - \mu_H$ | .1983 | .0936 | 2.12 | .037 | $(-.04796, .44456)$ |
| $\mu_C - \mu_N$ | .4751 | .1225 | 3.88 | .000 | $(.15280, .79740)$ |
| $\mu_H - \mu_N$ | .2768 | .1270 | 2.18 | .032 | $(-.05734, .61094)$ |

Note that while the estimated difference between Hispanics and Native Americans is half again as large as the difference between non-Hispanic Caucasians and Hispanics, the $t_{obs}$ values, and thus the significance levels of the differences, are almost identical. This occurs because the standard errors are substantially different. The standard error for the estimate of $\mu_C - \mu_H$ involves only the reasonably large samples for non-Hispanic Caucasians and Hispanics; the standard error for the estimate of $\mu_H - \mu_N$ involves the comparatively small sample of Native Americans, which is why

this standard error is larger. On the other hand, the standards errors for the estimates of $\mu_C - \mu_N$ and $\mu_H - \mu_N$ are very similar. The difference in the standard error between having a sample of 34 or 44 is minor by comparison to the effect on the standard error of having a sample size of only 15.

The hypothesis $H_0 : \mu_C - \mu_H = 0$, or equivalently $H_0 : \mu_C = \mu_H$, is the only one rejected at the .01 level. Summarizing the results of the tests at the .01 level, we have no strong evidence of a difference between the ages at which non-Hispanic Caucasians and Hispanics commit suicide, we have no strong evidence of a difference between the ages at which Hispanics and Native Americans commit suicide, but we do have strong evidence that there is a difference in the ages at which non-Hispanic Caucasians and Native Americans commit suicide.

Note that establishing a difference between non-Hispanic Caucasians and Native Americans does little to explain why that difference exists. The reason that Native Americans committed suicide at younger ages could be some complicated function of socio-economic factors or it could be simply that there were many more young Native Americans than old ones in Albuquerque at the time. The test only indicates that the two groups were different, it says nothing about why the groups were different.

The confidence interval for the difference between non-Hispanic Caucasians and Native Americans was constructed on the log scale. Transforming the interval gives $(e^{.1528}, e^{.7974})$ or $(1.2, 2.2)$. We are 99% confident that the average age of suicides is between 1.2 and 2.2 times higher for non-Hispanic Caucasians than for Native Americans. Note that examining differences in log ages transforms to the original scale as a multiplicative factor between groups. The parameters $\mu_C$ and $\mu_N$ are means for the logs of the suicide ages. When we transform the interval $(.1528, .7974)$ for $\mu_C - \mu_N$ into the interval $(e^{.1528}, e^{.7974})$, we obtain a confidence interval for $e^{\mu_C - \mu_N}$ or equivalently for $e^{\mu_C}/e^{\mu_N}$. We can think of $e^{\mu_C}$ and $e^{\mu_N}$ as 'average' values for the age distributions of the non-Hispanic Caucasians and Native Americans although they are not the expected values of the distributions. Obviously, $e^{\mu_C} = (e^{\mu_C}/e^{\mu_N})\, e^{\mu_N}$, so $e^{\mu_C}/e^{\mu_N}$ is the number of times greater the average suicide age is for non-Hispanic Caucasians. That is the basis for the interpretation of the interval $(e^{.1528}, e^{.7974})$.

With these data, the tests for differences in means do not depend crucially on the log transformation but interpretations of the confidence intervals do. For the untransformed data, the mean squared error is $MSE_u = 245$ and the observed value of the test statistic for comparing non-Hispanic Caucasians and Native Americans is

$$t_u = 3.54 = \frac{41.66 - 25.07}{\sqrt{245\left(\frac{1}{44} + \frac{1}{15}\right)}},$$

which is not far from the transformed value 3.88. However, the untransformed 99% confidence interval is $(4.3, 28.9)$, indicating a 4 to 29 year higher age for the mean non-Hispanic Caucasian suicide, rather than the transformed interval $(1.2, 2.2)$, indicating that typical non-Hispanic Caucasian suicide ages are 1.2 to 2.2 times greater than those for Native Americans.

The data do not strongly suggest that the means for Hispanics and Native Americans are different, so we *might* wish to compare the mean of the non-Hispanic Caucasians with the average of these groups. Typically, *averaging means will only be of interest if we feel comfortable treating the means as the same*. The parameter of interest is $Par = \mu_C - (\mu_H + \mu_N)/2$ or

$$Par = \mu_C - \frac{1}{2}\mu_H - \frac{1}{2}\mu_N$$

with

$$Est = \bar{y}_C - \frac{1}{2}\bar{y}_H - \frac{1}{2}\bar{y}_N = 3.6521 - \frac{1}{2}3.4538 - \frac{1}{2}3.1770 = .3367.$$

It is not appropriate to use our standard methods to test this *contrast* between the means because the contrast was suggested by the data. Nonetheless, we will illustrate the standard methods. From the

independence of the data in the three groups and Proposition 1.2.11, the variance of the estimate is

$$\begin{aligned}
&\mathrm{Var}\left(\bar{y}_C - \frac{1}{2}\bar{y}_H - \frac{1}{2}\bar{y}_N\right) \\
&= \mathrm{Var}(\bar{y}_C) + \left(\frac{-1}{2}\right)^2 \mathrm{Var}(\bar{y}_H) + \left(\frac{-1}{2}\right)^2 \mathrm{Var}(\bar{y}_N) \\
&= \frac{\sigma^2}{44} + \left(\frac{-1}{2}\right)^2 \frac{\sigma^2}{34} + \left(\frac{-1}{2}\right)^2 \frac{\sigma^2}{15} \\
&= \sigma^2 \left[\frac{1}{44} + \left(\frac{-1}{2}\right)^2 \frac{1}{34} + \left(\frac{-1}{2}\right)^2 \frac{1}{15}\right].
\end{aligned}$$

Substituting the *MSE* for $\sigma^2$ and taking the square root, the standard error is

$$.0886 = \sqrt{.168\left[\frac{1}{44} + \left(\frac{-1}{2}\right)^2 \frac{1}{34} + \left(\frac{-1}{2}\right)^2 \frac{1}{15}\right]}.$$

Note that the standard error happens to be smaller than any of those we have considered when comparing pairs of means. To test the null hypothesis that the mean for non-Hispanic Caucasians equals the average of the other groups, i.e., $H_0 : \mu_C - \frac{1}{2}\mu_H - \frac{1}{2}\mu_N = 0$, the test statistic is $[.3367 - 0]/.0886 = 3.80$, so the null hypothesis is easily rejected. This is an appropriate test statistic for evaluating $H_0$, but when letting the data suggest the contrast, the $t(90)$ distribution is no longer appropriate for quantifying the level of significance. Similarly, we could construct the 99% confidence interval

$$.3367 \pm 2.631(.0886)$$

but again, the confidence coefficient 99% is not really appropriate for a contrast suggested by the data.

While the parameter $\mu_C - \frac{1}{2}\mu_H - \frac{1}{2}\mu_N$ was suggested by the data, the theory of inference in Chapter 3 assumes that the parameter of interest does not depend on the data. In particular, the reference distributions we have used are invalid when the parameters depend on the data. Moreover, performing numerous inferential procedures complicates the analysis. Our standard tests are set up to check on one particular hypothesis. In the course of analyzing these data we have performed several tests. Thus we have had multiple opportunities to commit errors. In fact, the reason we have been discussing .01 level tests rather than .05 level tests is to help limit the number of errors made when all of the null hypotheses are true. In Chapter 6, we discuss methods of dealing with the problems that arise from making *multiple comparisons* among the means.

To this point, we have considered contrasts (comparisons) among the means. In constructing confidence intervals, prediction intervals, or tests for an individual mean, we continue to use the *MSE* and the $t(dfE)$ distribution. For example, the endpoints of a 99% confidence interval for $\mu_H$, the mean of the log suicide age for this Hispanic population, are

$$3.4538 \pm 2.631\sqrt{\frac{.168}{34}}$$

for an interval of $(3.269, 3.639)$. Transforming the interval back to the original scale gives $(26.3, 38.1)$, i.e., we are 99% confident that the average age of suicides for this Hispanic population is between 26.3 years old and 38.1 years old. The word 'average' is used because this is not a confidence interval for the expected value of the suicide ages, it is a confidence interval for the exponential transformation of the expected value of the log suicide age. A 99% prediction interval for the age of a future suicide from this Hispanic population has endpoints

$$3.4538 \pm 2.631\sqrt{.168 + \frac{.168}{34}}$$

for an interval of $(2.360, 4.548)$. Transforming the interval back to the original scale gives $(10.6, 94.4)$, i.e., we are 99% confident that a future suicide from this Hispanic population would be between 10.6 years old and 94.4 years old. This interval happens to include all of the observed suicide ages for Hispanics in Table 5.1; that seems reasonable, if not terribly informative. □

### *5.1.1 Theory*

In analysis of variance, we assume that we have independent observations on, say, $a$ different normal populations with the same variance. In particular, we assume the following data structure.

| Sample | Data | | Distribution |
|---|---|---|---|
| 1 | $y_{11}, y_{12}, \ldots, y_{1N_1}$ | iid | $N(\mu_1, \sigma^2)$ |
| 2 | $y_{21}, y_{22}, \ldots, y_{2N_2}$ | iid | $N(\mu_2, \sigma^2)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a$ | $y_{a1}, y_{a2}, \ldots, y_{aN_a}$ | iid | $N(\mu_a, \sigma^2)$ |

Here each sample is independent of the other samples. These assumptions can be written more succinctly as the one-way analysis of variance model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij}\text{s independent } N(0, \sigma^2) \tag{5.1.1}$$

$i = 1, \ldots, a$, $j = 1, \ldots, N_i$. The $\varepsilon_{ij}$s are unobservable random errors. We are writing each observation as its mean plus some random error. Alternatively, model (5.1.1) is often written as

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij}\text{s independent } N(0, \sigma^2) \tag{5.1.2}$$

where $\mu_i = \mu + \alpha_i$. The parameter $\mu$ is viewed as a grand mean, while $\alpha_i$ is an effect for the $i$th treatment group. The $\mu$ and $\alpha_i$ parameters are not well defined. In model (5.1.2) they only occur as the sum $\mu + \alpha_i$, so for any choice of $\mu$ and $\alpha_i$ the choices, say, $\mu + 5$ and $\alpha_i - 5$ are equally valid. The 5 can be replaced by any number we choose. The parameters $\mu$ and $\alpha_i$ are not completely specified by the model. There would seem to be little point in messing around with model (5.1.2) except that it has useful relationships with other models that will be considered later.

To analyze the data, we compute summary statistics from each sample. These are the sample means and sample variances. For the $i$th group of observations, the sample mean is

$$\bar{y}_{i\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$$

and the sample variance is

$$s_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{i\cdot})^2 .$$

*With independent normal errors having the same variance,* all *of the summary statistics are independent of one another*. Except for checking the validity of our assumptions, these summary statistics are more than sufficient for the entire analysis. Typically, we present the summary statistics in tabular form.

Sample statistics

| Group | Size | Mean | Variance |
|---|---|---|---|
| 1 | $N_1$ | $\bar{y}_{1\cdot}$ | $s_1^2$ |
| 2 | $N_2$ | $\bar{y}_{2\cdot}$ | $s_2^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a$ | $N_a$ | $\bar{y}_{a\cdot}$ | $s_a^2$ |

The sample means, the $\bar{y}_{i\cdot}$s, are estimates of the corresponding $\mu_i$s and the $s_i^2$s all estimate the common population variance $\sigma^2$. With unequal sample sizes an efficient pooled estimate of $\sigma^2$ must be a weighted average of the $s_i^2$s. The weights are the degrees of freedom associated with the various estimates. The pooled estimate of $\sigma^2$ is called the *mean squared error* (*MSE*),

$$\begin{aligned} MSE \equiv s_p^2 &\equiv \frac{(N_1-1)s_1^2+(N_2-1)s_2^2+\cdots+(N_a-1)s_a^2}{\sum_{i=1}^{a}(N_i-1)} \\ &= \frac{1}{(n-a)}\sum_{i=1}^{a}\sum_{j=1}^{N_i}(y_{ij}-\bar{y}_{i\cdot})^2 \end{aligned}$$

where $n=\sum_{i=1}^{a}N_i$ is the total sample size. The degrees of freedom for the *MSE* are the *degrees of freedom for error*,

$$dfE \equiv n-a=\sum_{i=1}^{a}(N_i-1).$$

This is the sum of the degrees of freedom for the individual variance estimates. Note that the *MSE* depends only on the sample variances, so, with independent normal errors having the same variance, *MSE is independent of the* $\bar{y}_{i\cdot}s$.

A simple average of the sample variances $s_i^2$ is not reasonable. If we had $N_1=1\,000\,000$ observations in the first sample and only $N_2=5$ observations in the second sample, obviously the variance estimate from the first sample is much better than that from the second and we want to give it more weight.

We need to check the validity of our assumptions. The errors in models (1) and (2) are assumed to be independent normals with mean 0 and variance $\sigma^2$, so we would like to use them to evaluate the distributional assumptions, e.g., equal variances and normality. Unfortunately, the errors are unobservable, we only see the $y_{ij}$s and we do not know the $\mu_i$s, so we cannot compute the $\varepsilon_{ij}$s. However, since $\varepsilon_{ij}=y_{ij}-\mu_i$ and we can estimate $\mu_i$, we can estimate the errors with the *residuals*,

$$\hat{\varepsilon}_{ij}=y_{ij}-\bar{y}_{i\cdot}.$$

The residuals $y_{ij}-\bar{y}_{i\cdot}$ can be plotted against *predicted values* $\bar{y}_{i\cdot}$ to check whether the variance depends in some way on the means $\mu_i$. They can also be plotted against rankits (normal scores) to check the normality assumption.

Using residuals to evaluate assumptions is a fundamental part of modern statistical data analysis. However, complications can arise. In later chapters we will discuss reasons for using standardized residuals rather than these raw residuals. Standardized residuals will be discussed in connection with regression analysis. In balanced analysis of variance, i.e., situations with equal numbers of observations on each group, the complications disappear. Thus, the unstandardized residuals are adequate for evaluating the assumptions in a balanced analysis of variance. In other analysis of variance situations, the problems are *relatively* minor.

If we are satisfied with the assumptions, we proceed to examine the parameters of interest. The basic parameters of interest in analysis of variance are the $\mu_i$s, which have natural estimates, the $\bar{y}_{i\cdot}$s. We also have an estimate of $\sigma^2$, so we are in a position to draw a variety of statistical inferences. The main problem in obtaining tests and confidence intervals is in finding appropriate standard errors. To do this we need to observe that each of the $a$ samples are independent. The $\bar{y}_{i\cdot}$s are computed from different samples, so they are independent of each other. Moreover, $\bar{y}_{i\cdot}$ is the sample mean of $N_i$ observations, so

$$\bar{y}_{i\cdot}\sim N\left(\mu_i,\frac{\sigma^2}{N_i}\right).$$

For inferences about a single mean, say, $\mu_2$, use the general procedures with $Par=\mu_2$ and $Est=\bar{y}_{2\cdot}$. The variance of $\bar{y}_{2\cdot}$ is $\sigma^2/N_2$, so $\mathrm{SE}(\bar{y}_{2\cdot})=\sqrt{MSE/N_2}$. The reference distribution is $[\bar{y}_{2\cdot}-\mu_2]/\mathrm{SE}(\bar{y}_{2\cdot})\sim t(dfE)$. Note that the degrees of freedom for the $t$ distribution are precisely the

degrees of freedom for the $MSE$. The general procedures also provide prediction intervals using the $MSE$ and $t(dfE)$ distribution.

For inferences about the difference between two means, say, $\mu_2 - \mu_1$, use the general procedures with $Par = \mu_2 - \mu_1$ and $Est = \bar{y}_{2\cdot} - \bar{y}_{1\cdot}$. The two means are independent, so the variance of $\bar{y}_{2\cdot} - \bar{y}_{1\cdot}$ is the variance of $\bar{y}_{2\cdot}$ plus the variance of $\bar{y}_{1\cdot}$, i.e., $\sigma^2/N_2 + \sigma^2/N_1$. The standard error of $\bar{y}_{2\cdot} - \bar{y}_{1\cdot}$ is

$$\mathrm{SE}(\bar{y}_{2\cdot} - \bar{y}_{1\cdot}) = \sqrt{\frac{MSE}{N_2} + \frac{MSE}{N_1}} = \sqrt{MSE\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}.$$

The reference distribution is

$$\frac{(\bar{y}_{2\cdot} - \bar{y}_{1\cdot}) - (\mu_2 - \mu_1)}{\sqrt{MSE\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}} \sim t(dfE).$$

We might wish to compare one mean, $\mu_1$, with the average of two other means, $(\mu_2 + \mu_3)/2$. In this case, the parameter can be taken as $Par = \mu_1 - (\mu_2 + \mu_3)/2 = \mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3$. The estimate is $Est = \bar{y}_{1\cdot} - \frac{1}{2}\bar{y}_{2\cdot} - \frac{1}{2}\bar{y}_{3\cdot}$. By the independence of the sample means, the variance of the estimate is

$$\begin{aligned}
\mathrm{Var}\left(\bar{y}_{1\cdot} - \frac{1}{2}\bar{y}_{2\cdot} - \frac{1}{2}\bar{y}_{3\cdot}\right) &= \mathrm{Var}(\bar{y}_{1\cdot}) + \mathrm{Var}\left(\frac{-1}{2}\bar{y}_{2\cdot}\right) + \mathrm{Var}\left(\frac{-1}{2}\bar{y}_{3\cdot}\right) \\
&= \frac{\sigma^2}{N_1} + \left(\frac{-1}{2}\right)^2 \frac{\sigma^2}{N_2} + \left(\frac{-1}{2}\right)^2 \frac{\sigma^2}{N_3} \\
&= \sigma^2\left[\frac{1}{N_1} + \frac{1}{4}\frac{1}{N_2} + \frac{1}{4}\frac{1}{N_3}\right].
\end{aligned}$$

The standard error is

$$\mathrm{SE}\left(\bar{y}_{1\cdot} - \frac{1}{2}\bar{y}_{2\cdot} - \frac{1}{2}\bar{y}_{3\cdot}\right) = \sqrt{MSE\left[\frac{1}{N_1} + \frac{1}{4N_2} + \frac{1}{4N_3}\right]}.$$

The reference distribution is

$$\frac{\left(\bar{y}_{1\cdot} - \frac{1}{2}\bar{y}_{2\cdot} - \frac{1}{2}\bar{y}_{3\cdot}\right) - \left(\mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3\right)}{\sqrt{MSE\left[\frac{1}{N_1} + \frac{1}{4N_2} + \frac{1}{4N_3}\right]}} \sim t(dfE).$$

Typically, in analysis of variance we are concerned with parameters that are *contrasts* (comparisons) among the $\mu_i$s. For *known* coefficients $\lambda_1, \ldots, \lambda_a$ with $\sum_{i=1}^a \lambda_i = 0$, a contrast is defined by $\sum_{i=1}^a \lambda_i \mu_i$. For example, $\mu_2 - \mu_1$ has $\lambda_1 = -1$, $\lambda_2 = 1$, and all other $\lambda_i$s equal to 0. The contrast $\mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3$ has $\lambda_1 = 1$, $\lambda_2 = -1/2$, $\lambda_3 = -1/2$, and all other $\lambda_i$s equal to 0. The natural estimate of $\sum_{i=1}^a \lambda_i \mu_i$ substitutes the sample means for the population means, i.e., the natural estimate is $\sum_{i=1}^a \lambda_i \bar{y}_{i\cdot}$. In fact, Proposition 1.2.11 gives

$$\mathrm{E}\left(\sum_{i=1}^a \lambda_i \bar{y}_{i\cdot}\right) = \sum_{i=1}^a \lambda_i \mathrm{E}(\bar{y}_{i\cdot}) = \sum_{i=1}^a \lambda_i \mu_i,$$

so by definition this is an *unbiased* estimate of the contrast. Using the independence of the sample means and Proposition 1.2.11,

$$\mathrm{Var}\left(\sum_{i=1}^a \lambda_i \bar{y}_{i\cdot}\right) = \sum_{i=1}^a \lambda_i^2 \mathrm{Var}(\bar{y}_{i\cdot})$$

$$\begin{aligned} &= \sum_{i=1}^{a} \lambda_i^2 \frac{\sigma^2}{N_i} \\ &= \sigma^2 \sum_{i=1}^{a} \frac{\lambda_i^2}{N_i}. \end{aligned}$$

The standard error is

$$\mathrm{SE}\left(\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot}\right) = \sqrt{MSE \sum_{i=1}^{a} \frac{\lambda_i^2}{N_i}}$$

and the reference distribution is

$$\frac{(\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot}) - (\sum_{i=1}^{a} \lambda_i \mu_i)}{\sqrt{MSE \sum_{i=1}^{a} \lambda_i^2 / N_i}} \sim t(dfE),$$

see Exercise 5.7.14. If the independence and equal variance assumptions hold, then the central limit theorem and law of large numbers can be used to justify a $N(0,1)$ reference distribution even when the data are not normal. Moreover, in one-way ANOVA all of these results hold even when $\sum_i \lambda_i \neq 0$, so they hold for linear combinations of the $\mu_i$s that are not contrasts. Nonetheless, our primary interest is in contrasts.

Having identified a parameter, an estimate, a standard error, and an appropriate reference distribution, inferences follow the usual pattern. A 95% confidence interval for $\sum_{i=1}^{a} \lambda_i \mu_i$ has endpoints

$$\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} \pm t(.975, dfE) \sqrt{MSE \sum_{i=1}^{a} \lambda_i^2 / N_i}.$$

An $\alpha = .05$ test of $H_0 : \sum_{i=1}^{a} \lambda_i \mu_i = 0$ versus $H_A : \sum_{i=1}^{a} \lambda_i \mu_i \neq 0$ rejects $H_0$ if

$$\frac{|\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} - 0|}{\sqrt{MSE \sum_{i=1}^{a} \lambda_i^2 / N_i}} > t(.975, dfE) \tag{5.1.3}$$

An equivalent procedure to the test in (5.1.3) is often useful. If we square both sides of (5.1.3), the test rejects if

$$\left( \frac{|\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} - 0|}{\sqrt{MSE \sum_{i=1}^{a} \lambda_i^2 / N_i}} \right)^2 > \left(t(.975, dfE)\right)^2.$$

The square of the test statistic leads to another statistic that will be useful later, the sum of squares for the contrast. Rewrite the test statistic as

$$\begin{aligned} \left( \frac{|\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} - 0|}{\sqrt{MSE \sum_{i=1}^{a} \lambda_i^2 / N_i}} \right)^2 &= \frac{(\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} - 0)^2}{MSE \sum_{i=1}^{a} \lambda_i^2 / N_i} \\ &= \frac{(\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot})^2 \big/ \sum_{i=1}^{a} \lambda_i^2 / N_i}{MSE} \end{aligned}$$

and define the *sum of squares for the contrast* as

$$SS\left(\sum_{i=1}^{a} \lambda_i \mu_i\right) \equiv \frac{(\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot})^2}{\sum_{i=1}^{a} \lambda_i^2 / N_i}. \tag{5.1.4}$$

The $\alpha = .05$ $t$ test of $H_0 : \sum_{i=1}^{a} \lambda_i \mu_i = 0$ versus $H_A : \sum_{i=1}^{a} \lambda_i \mu_i \neq 0$ is equivalent to rejecting $H_0$ if

$$\frac{SS(\sum_{i=1}^{a} \lambda_i \mu_i)}{MSE} > \left[t(.975, dfE)\right]^2.$$

It is a mathematical fact that for any $\alpha$ between 0 and 1 and any $dfE$,

$$\left[t\left(1-\frac{\alpha}{2},dfE\right)\right]^2 = F(1-\alpha,1,dfE).$$

Thus the test based on the sum of squares for the contrast is an $F$ test with 1 degree of freedom in the numerator. *Any contrast has 1 degree of freedom associated with it.*

A notational matter needs to be mentioned. Contrasts, by definition, have $\sum_{i=1}^{a}\lambda_i = 0$. If we use model (5.1.2) rather than model (5.1.1) we get

$$\sum_{i=1}^{a}\lambda_i\mu_i = \sum_{i=1}^{a}\lambda_i(\mu+\alpha_i) = \mu\sum_{i=1}^{a}\lambda_i + \sum_{i=1}^{a}\lambda_i\alpha_i = \sum_{i=1}^{a}\lambda_i\alpha_i.$$

Thus contrasts in model (5.1.2) involve only the treatment effects. This is of some importance later when dealing with more complicated models.

In our first example we transformed the suicide age data so that they better satisfy the assumptions of equal variances and normal distributions. In fact, analysis of variance tests and confidence intervals are frequently useful even when these assumptions are violated. Scheffé (1959, p. 345) concludes that (a) nonnormality is not a serious problem for inferences about means but it is a serious problem for inferences about variances, (b) unequal variances are not a serious problem for inferences about means from samples of the same size but are a serious problem for inferences about means from samples of unequal sizes, and (c) lack of independence can be a serious problem. Of course any such rules depend on just how bad the nonnormality is, how unequal the variances are, and how bad the lack of independence is. My own interpretation of these rules is that if you check the assumptions and they do not look too bad, you can probably proceed with a fair amount of assurance.

### *5.1.2 Balanced ANOVA: introductory example*

We now consider an example of a balanced one-way ANOVA. A balanced one-way ANOVA has equal numbers of observations in each group, say, $N = N_1 = \cdots = N_a$.

EXAMPLE 5.1.2. Ott (1949) presented data on an electrical characteristic associated with ceramic components for a phonograph. Ott and Schilling (1990) and Ryan (1989) have also considered these data. Ceramic pieces were cut from strips, each of which could provide 25 pieces. It was decided to take 7 pieces from each strip, manufacture the 7 ceramic phonograph components, and measure the electrical characteristic on each. The data from 4 strips are given below. (These are actually the third through sixth of the strips reported by Ott.)

| Strip | Observations | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 17.3 | 15.8 | 16.8 | 17.2 | 16.2 | 16.9 | 14.9 |
| 2 | 16.9 | 15.8 | 16.9 | 16.8 | 16.6 | 16.0 | 16.6 |
| 3 | 15.5 | 16.6 | 15.9 | 16.5 | 16.1 | 16.2 | 15.7 |
| 4 | 13.5 | 14.5 | 16.0 | 15.9 | 13.7 | 15.2 | 15.9 |

In the current analysis, we act as if the four strips are of intrinsic interest and investigate whether there are differences among them. In Subsection 13.4.2 we will consider an analysis in which we assume that the strips are themselves a random sample from some wider population. The data are displayed in Figure 5.5 and summary statistics follow.

Sample statistics: electrical characteristics

| Strip | $N$ | $\bar{y}_{i\cdot}$ | $s_i^2$ | $s_i$ |
|---|---|---|---|---|
| 1 | 7 | 16.4429 | 0.749524 | 0.866 |
| 2 | 7 | 16.5143 | 0.194762 | 0.441 |
| 3 | 7 | 16.0714 | 0.162381 | 0.403 |
| 4 | 7 | 14.9571 | 1.139524 | 1.067 |

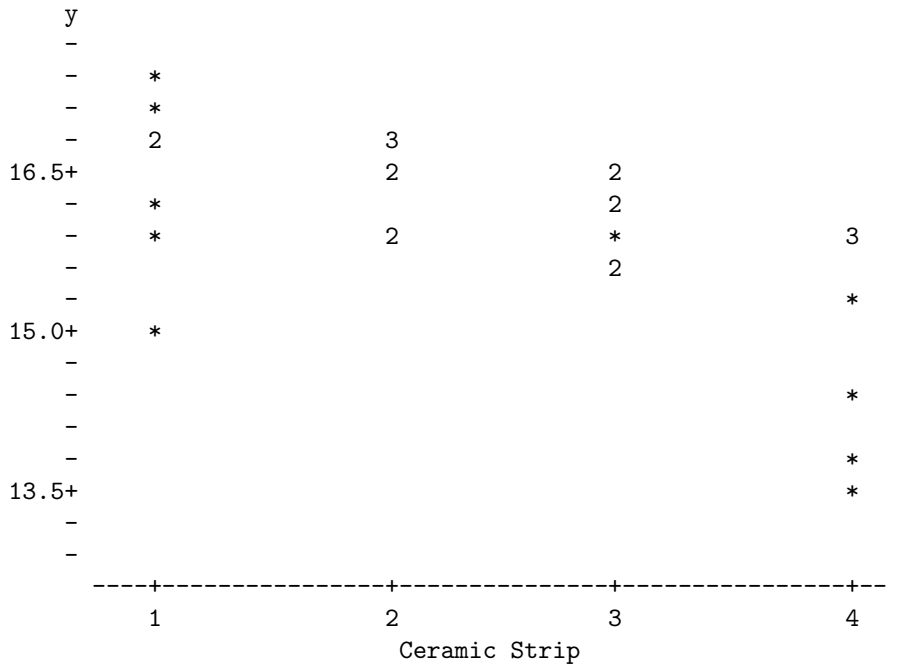



Figure 5.5: *Plot of electrical characteristics data.*

The electrical characteristic appears to be lowest for strip 4 and highest for strips 1 and 2, but we need to use formal inferential procedures to establish whether the differences could be reasonably ascribed to random variation. The sample standard deviations, and thus the sample variances, are comparable. The ratio of the largest to the smallest standard deviation is just over 2.5, which is not small but which is also not large enough to cause major concern. As in Section 4.4, we could do $F$ tests to determine whether any pairs of variances differ. The largest of these $F$ tests is not significant at the .02 level and, after considering that there are six pairs to test, we conclude that there is no cause for major concern. Figure 5.5 is poorly suited to evaluate the variances visually because in Figure 5.5 the plot involves any differences in means as well as differences in variance. A better plot from which to evaluate the variances is given as Figure 5.6. Figure 5.6 is a plot of the residuals $\hat{\varepsilon}_{ij} \equiv y_{ij} - \bar{y}_{i\cdot}$ against the appropriate group. The residuals have been adjusted for their different means, so residuals, and thus residual plots, are centered at 0. Figure 5.6 is not wonderful in that we see differences in variability for the four groups, but it is also not outlandishly inconsistent with the assumption of equal variances. (Note that if one group had many more observations than another, the spread for that group would be greater even if the population variances were the same.) Figure 5.7 contains a normal plot of the residuals. The plot looks fairly reasonable, although it tails off at the top. The $W'$ statistic of .956 gives a $P$ value for the hypothesis of normality that is larger than .05 and in any case, analysis of variance procedures are not particularly sensitive to nonnormality.

With equal sample sizes in each group, the $MSE$ reduces to the simple average of the sample variances.

$$\begin{aligned} MSE &= \frac{(7-1).74952+(7-1).19476+(7-1).16238+(7-1)1.13952}{7+7+7+7-4} \\ &= \frac{.74952+.19476+.16238+1.13952}{4} \\ &= .56155 \end{aligned}$$

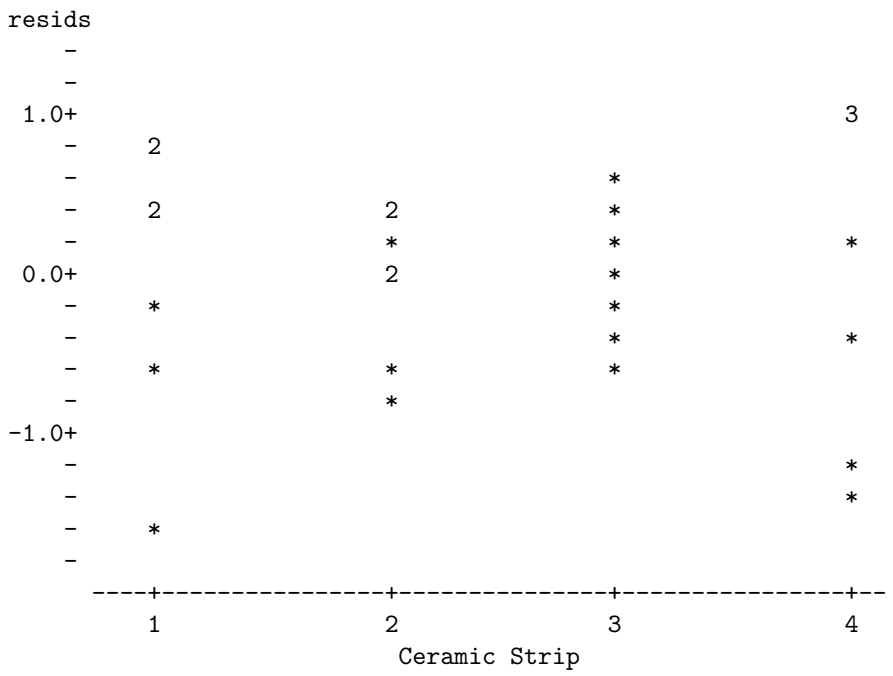



Figure 5.6: *Residual plot.*

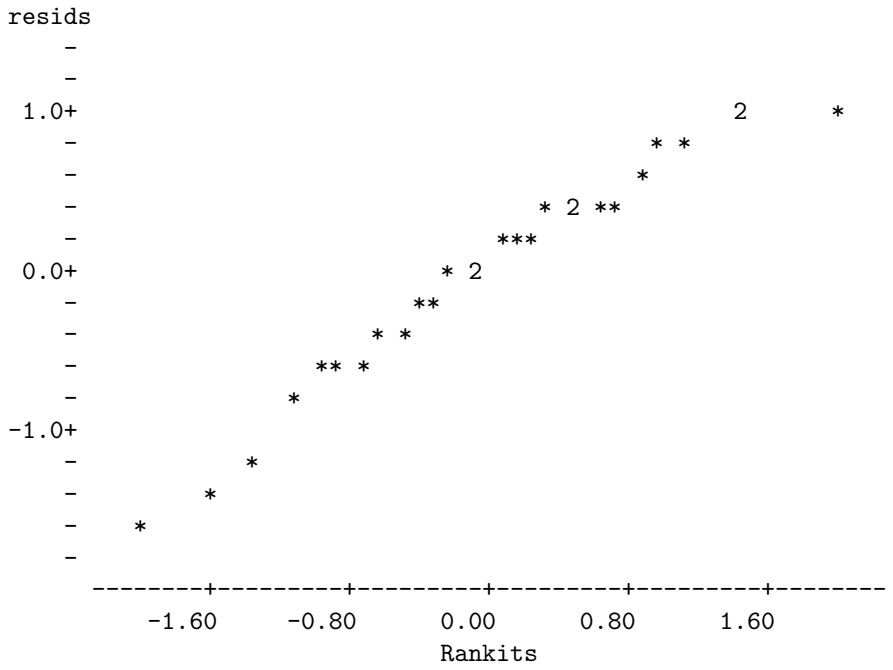



Figure 5.7: *Normal plot of residuals,* $W' = 0.956$.

and has error degrees of freedom $dfE = 7+7+7+7-4 = 24$. Again, we compare all pairs of means. The value $t(.995,24) = 2.797$ is required for constructing 99% confidence intervals. These intervals and two-sided tests of $H_0 : Par = 0$ are given below.

| *Par* | *Est* | SE(*Est*) | $t_{obs}$ | *P* | 99% CI |
|---|---|---|---|---|---|
| $\mu_1 - \mu_4$ | 1.4858 | 0.4006 | 3.709 | .001 | $(0.37, 2.61)$ |
| $\mu_2 - \mu_4$ | 1.5572 | 0.4006 | 3.887 | .001 | $(0.44, 2.68)$ |
| $\mu_3 - \mu_4$ | 1.1143 | 0.4006 | 2.782 | .010 | $(-0.01, 2.23)$ |
| $\mu_1 - \mu_2$ | −0.0714 | 0.4006 | −0.178 | .860 | $(-1.19, 1.05)$ |
| $\mu_1 - \mu_3$ | 0.3715 | 0.4006 | 0.927 | .363 | $(-0.75, 1.49)$ |
| $\mu_2 - \mu_3$ | 0.4429 | 0.4006 | 1.106 | .280 | $(-0.68, 1.56)$ |

Note that with equal numbers of observations on each group, the standard errors are the same for each comparison of two means. Based on $\alpha = .01$ tests, the electrical characteristic for strip 4 differs significantly from those for strips 1 and 2, the decision for strip 3 is essentially a toss-up, and no other differences are significant. Even for strips 1 and 2, the 99% confidence intervals indicate that the data are consistent with differences from strip 4 as small as .37 and .44 respectively. Such differences may or may not be of practical importance. Clearly, the main source of differences among these data is that strip 4 tends to give smaller values than the other strips. In fact, the *P* values for comparisons among the other three strips are all quite large.

Using formula (5.1.4), the sum of squares for $\mu_1 - \mu_4$ is

$$SS(\mu_1 - \mu_4) = \frac{(1.4858)^2}{(1)^2/7 + (-1)^2/7} = 7.7266.$$

The table below gives the sums of squares and *F* tests for equality between all pairs of means.

| *Par* | *SS* | $F_{obs}$ | *P* |
|---|---|---|---|
| $\mu_1 - \mu_4$ | 7.727 | 13.76 | .001 |
| $\mu_2 - \mu_4$ | 8.487 | 15.11 | .001 |
| $\mu_3 - \mu_4$ | 4.346 | 7.74 | .010 |
| $\mu_1 - \mu_2$ | 0.018 | 0.03 | .860 |
| $\mu_1 - \mu_3$ | 0.483 | 0.86 | .363 |
| $\mu_2 - \mu_3$ | 0.687 | 1.22 | .280 |

Note that the *F* statistics are just the sums of squares divided by the *MSE*. They equal the squares of the *t* statistics given earlier and the *P* values are identical. □

### *5.1.3 Analytic and enumerative studies*

In one-sample, two-sample, and one-way ANOVA problems, we assume that we have random samples from various populations. In the more sophisticated models treated later, we continue to assume that at least the errors are a random sample from a $N(0, \sigma^2)$ population. The statistical inferences we draw are valid for the populations that were sampled. Often it is not clear what the sampled populations are. What are the populations from which the Albuquerque suicide ages were sampled? Presumably, our data were all of the suicides reported in 1978 for these ethnic groups. The electrical characteristic data has four ceramic strips divided into 25 pieces, of which seven pieces are taken. Are the seven pieces a random sample from the 25? They could be. Is the collection of 25 pieces the population that we really care about? Doubtful! What we really care about is whether the differences in ceramic strips are large enough to cause problems in the production of phonographs. (Not that anyone makes phonographs anymore.)

When we analyze data, we assume that the measurements are subject to errors and that the errors are consistent with our models. However, the populations from which these samples are taken

may be nothing more than mental constructs. In such cases, it requires extrastatistical reasoning to justify applying the statistical conclusions to whatever issues we really wish to address. Moreover, the desire to predict the future underlies virtually all studies and, unfortunately, one can never be sure that data collected now will apply to the conditions of the future. So what can you do? Only your best. You can try to make your data as relevant as possible to your anticipation of future conditions. You can try to collect data for which the assumptions will be reasonably true. You can try to validate your assumptions. Studies in which it is not clear that the data are random samples from the population of immediate interest are often called *analytic studies*.

About the only time one can be really sure that statistical conclusions apply directly to the population of interest is when one has control of the population of interest. If we have a list of all the elements in the population, we can choose a random sample from the population. Of course, choosing a random sample is still very different from obtaining a random sample of observations. Without control or total cooperation, we may not be able to take measurements on the sample. (Even when you can find people that you want for a sample, many will not submit to a measurement process.) Studies in which one can arrange to have the assumptions met are often called *enumerative studies*. See Hahn and Meeker (1993) and Deming (1986) for additional discussion of these issues.

## 5.2 Balanced one-way analysis of variance: theory

We now examine in detail the important special case of one-way analysis of variance in which the numbers of observations for each sample are the same, cf. Subsection 5.1.2. In this case, the analysis of variance is referred to as *balanced*. Balanced one-way ANOVA is important because it is both understandable and extendable. The logic behind analysis of variance is much clearer when dealing with balanced samples and the standard methods for multifactor analysis of variance are extensions of the techniques developed for balanced one-way ANOVA. The standard methods for multifactor ANOVA also assume equal numbers of observations on all treatments.

For balanced analysis of variance, let $N \equiv N_1 = \cdots = N_a$ be the number of observations in each sample. In particular, we assume the data structure

| Sample | Data | | Distribution |
|---|---|---|---|
| 1 | $y_{11}, y_{12}, \ldots, y_{1N}$ | iid | $N(\mu_1, \sigma^2)$ |
| 2 | $y_{21}, y_{22}, \ldots, y_{2N}$ | iid | $N(\mu_2, \sigma^2)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a$ | $y_{a1}, y_{a2}, \ldots, y_{aN}$ | iid | $N(\mu_a, \sigma^2)$ |

with all samples independent. The data structure can be rewritten as the balanced one-way ANOVA model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij}\text{s independent } N(0, \sigma^2)$$

$i = 1, \ldots, a$, $j = 1, \ldots, N$. Again, we have assumed the same variance $\sigma^2$ for each sample.

In this section, we focus on testing the (null) hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a.$$

This is a test of whether there are *any* differences among the groups. If we use model (5.1.2), the null hypothesis can be written as $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a$. To perform the test, first compute summary statistics from the samples.

Sample statistics

| Group | Size | Mean | Variance |
|---|---|---|---|
| 1 | $N$ | $\bar{y}_{1\cdot}$ | $s_1^2$ |
| 2 | $N$ | $\bar{y}_{2\cdot}$ | $s_2^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a$ | $N$ | $\bar{y}_{a\cdot}$ | $s_a^2$ |

As before, the sample means, the $\bar{y}_{i\cdot}$s, are estimates of the $\mu_i$s and the $s_i^2$s all estimate $\sigma^2$.

The test of $H_0$ is based on estimating $\sigma^2$. We construct two estimates of the variance. The first estimate is always valid, assuming of course that our initial assumptions were correct. The second estimate is valid *only* when $\mu_1 = \mu_2 = \cdots = \mu_a$. If the $\mu_i$s are all equal, we have two estimates of $\sigma^2$, so they should be about the same. If the $\mu_i$s are not all equal, the second estimate tends to be bigger than $\sigma^2$, so it should be larger than the first estimate. We conclude that the data are consistent with $\mu_1 = \mu_2 = \cdots = \mu_a$ when the two estimates seem to be about the same and conclude that the $\mu_i$s are not all equal when the second estimate is substantially larger than the first. As usual, when the estimates are about the same we conclude that the data are consistent with the $\mu_i$s all being equal; *we do not conclude that the $\mu_i$s are really all equal*. If the $\mu_i$s are not quite equal but are very nearly so, we cannot expect to be able to detect the differences. On the other hand, two widely different variance estimates give substantial proof that the $\mu_i$s are not all the same.

The easy part of the process is creating the first estimate of the variance, the one that is always valid. From each sample, regardless of the value of $\mu_i$, we have an estimate of $\sigma^2$, namely $s_i^2$. Obviously, the average of the $s_i^2$s must also be an estimate of $\sigma^2$. The average is the pooled estimate of the variance, i.e., the mean squared error is

$$\begin{aligned} MSE &\equiv \frac{s_1^2 + s_2^2 + \cdots + s_a^2}{a} \\ &= \frac{1}{a(N-1)} \sum_{i=1}^{a} \sum_{j=1}^{N} (y_{ij} - \bar{y}_{i\cdot})^2 . \end{aligned}$$

As discussed earlier, a simple average such as this is not always appropriate. The simple average is only reasonable because we have the same number of observations in each sample.

Recall that each $s_i^2$ has $N-1$ degrees of freedom. Each $s_i^2$ is based on $N$ observations but is functionally based on $N-1$ observations because of the need to estimate $\mu_i$ before estimating the variance. By pooling together the variance estimates, we also get to pool the degrees of freedom. We have combined $a$ independent estimates of $\sigma^2$, each with $N-1$ degrees of freedom, so the pooled estimate has $a(N-1)$ degrees of freedom. In other words, the $MSE$ is functionally based on $a(N-1)$ observations. The degrees of freedom associated with the $MSE$ are the degrees of freedom for error ($dfE$), so we have

$$dfE = a(N-1).$$

The data, the $y_{ij}$s, are random, so the $MSE$, which is computed from them, must also be random. If we collected another set of similar data we would not expect to get exactly the same value for the $MSE$. If we are to evaluate whether this estimate of $\sigma^2$ is similar to another estimate, we need to have some idea of the variability in the $MSE$. Under the assumptions we have made, the distribution of the $MSE$ depends only on $dfE$ and $\sigma^2$. The distribution is related to the $\chi^2$ family of distributions. In particular,

$$\frac{dfE \times MSE}{\sigma^2} \sim \chi^2(dfE)$$

where, on the right hand side, $dfE$ indicates the particular member of the $\chi^2$ family that is appropriate. A commonly used terminology in analysis of variance is the sum of squares for error ($SSE$). This is defined to be

$$SSE \equiv dfE \times MSE = \sum_{i=1}^{a} \sum_{j=1}^{N} (y_{ij} - \bar{y}_{i\cdot})^2 . \tag{5.2.1}$$

Note that $SSE/\sigma^2 \sim \chi^2(dfE)$. Note also that the *SSE is the sum of the squared residuals*, the residuals being

$$\hat{\varepsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot} .$$

The second estimate of $\sigma^2$ is to be valid only when $\mu_1 = \mu_2 = \cdots = \mu_a$. We have already used the sample variances $s_i^2$ in constructing the $MSE$, so we use the rest of our summary statistics, the

$\bar{y}_{i\cdot}$s, in constructing the second estimate of $\sigma^2$. In fact, the $\bar{y}_{i\cdot}$s are estimates of the $\mu_i$s, so it is only reasonable to use the $\bar{y}_{i\cdot}$s when trying to draw conclusions about the $\mu_i$s. Consider the distributions of the $\bar{y}_{i\cdot}$s. Each is the sample mean of $N$ observations, so each has the distribution of a sample mean. In particular,

$$\begin{aligned}
\bar{y}_{1\cdot} &\sim N\left(\mu_1, \frac{\sigma^2}{N}\right) \\
\bar{y}_{2\cdot} &\sim N\left(\mu_2, \frac{\sigma^2}{N}\right) \\
&\vdots \\
\bar{y}_{a\cdot} &\sim N\left(\mu_a, \frac{\sigma^2}{N}\right)
\end{aligned}$$

The $a$ different samples are independent of each other, so $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \ldots, \bar{y}_{a\cdot}$ are all independent. They all have the same variance, $\sigma^2/N$, and they all have normal distributions. In fact, the only thing keeping them from having independent and identical distributions is that they have different means $\mu_i$. If we assume that $\mu_1 = \mu_2 = \cdots = \mu_a$, they have independent and identical distributions and thus form a random sample from a population. Balanced analysis of variance is based on the fact that if the $\mu_i$s are the same, the $\bar{y}_{i\cdot}$s can be treated as a random sample. If the $\bar{y}_{i\cdot}$s are a random sample, we can compute their sample variance to get an estimate of the variance of the $\bar{y}_{i\cdot}$s. The variance of the $\bar{y}_{i\cdot}$s is $\sigma^2/N$ and the sample variance of the $\bar{y}_{i\cdot}$s is

$$s^2_{\bar{y}} = \frac{1}{a-1}\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

where

$$\bar{y}_{\cdot\cdot} \equiv \frac{1}{a}\sum_{i=1}^{a}\bar{y}_{i\cdot}$$

is the sample mean of the $\bar{y}_{i\cdot}$s. We have $s^2_{\bar{y}}$ as an estimate of $\sigma^2/N$ but we set out to find an estimate of $\sigma^2$. The obvious choice is

$$MSTrts \equiv N s^2_{\bar{y}} = \frac{N}{a-1}\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

where $MSTrts$ abbreviates the commonly used term *mean squared treatments*. The estimate $s^2_{\bar{y}}$ is based on a sample of size $a$, so it, and thus $MSTrts$, has $a-1$ degrees of freedom. These are referred to as the degrees of freedom for treatments ($dfTrts$). The sum of squares for treatments is defined as

$$SSTrts \equiv dfTrts \times MSTrts = N\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2. \tag{5.2.2}$$

Just as the $MSE$ is random, the $MSTrts$ is also random. The estimate $s^2_{\bar{y}}$ is the sample variance of a random sample of size $a$ from a normal population with variance $\sigma^2/N$, so

$$\frac{(a-1)s^2_{\bar{y}}}{\sigma^2/N} = \frac{(a-1)MSTrts}{\sigma^2} \sim \chi^2(a-1).$$

The discussion above is based on the assumption that $\mu_1 = \mu_2 = \cdots = \mu_a$. If this is not true, the $\bar{y}_{i\cdot}$s do not form a random sample and $s^2_{\bar{y}}$ does not estimate $\sigma^2/N$. Actually, it estimates $\sigma^2/N$ plus the 'variance' of the $\mu_i$s. Algebraically, $s^2_{\bar{y}}$ estimates

$$\mathrm{E}(s^2_{\bar{y}}) = \frac{\sigma^2}{N} + \frac{1}{a-1}\sum_{i=1}^{a}(\mu_i - \bar{\mu}_{\cdot})^2$$

where $\bar{\mu}_\cdot \equiv \sum_{i=1}^{a} \mu_i/a$ is the mean of the $\mu_i$s. Multiplying by $N$ gives

$$\mathrm{E}(MSTrts) = \mathrm{E}(Ns_{\bar{y}}^2) = \sigma^2 + \frac{N}{a-1}\sum_{i=1}^{a}(\mu_i - \bar{\mu}_\cdot)^2, \tag{5.2.3}$$

so $MSTrts$ is an estimate of $\sigma^2$ plus something that is always nonnegative. If $\mu_1 = \mu_2 = \cdots = \mu_a$, the $\mu_i$s are all equal to their average $\bar{\mu}_\cdot$, thus $(\mu_i - \bar{\mu}_\cdot)^2 = 0$ for all $i$, and

$$\frac{N}{a-1}\sum_{i=1}^{a}(\mu_i - \bar{\mu}_\cdot)^2 = 0.$$

As advertised earlier, if $\mu_1 = \cdots = \mu_a$, $MSTrts$ is an estimate of $\sigma^2$. If the $\mu_i$s are not all the same, $[N/(a-1)]\sum_{i=1}^{a}(\mu_i - \bar{\mu}_\cdot)^2$ is positive. The larger this term is, the easier it is to conclude that the treatment means are different. The term increases when $N$, the number of observations in each group, increases and when the variability of the $\mu_i$s increases, i.e., when $\sum_{i=1}^{a}(\mu_i - \bar{\mu}_\cdot)^2/(a-1)$ increases.

A decision regarding the validity of the claim $\mu_1 = \mu_2 = \cdots = \mu_a$ is based on comparing $MSTrts$ with $MSE$. If they are about the same, or equivalently if

$$F \equiv \frac{MSTrts}{MSE} \tag{5.2.4}$$

is about 1, the data are consistent with the idea that $MSTrts$ and $MSE$ both estimate the same (unknown) quantity $\sigma^2$ and thus are consistent with $\mu_1 = \mu_2 = \cdots = \mu_a$. The alternative is that the $\mu_i$s are not all equal, in which case $MSTrts$ is estimating something larger than $\sigma^2$, while $MSE$ continues to estimate $\sigma^2$. In this case, the ratio $F = MSTrts/MSE$ estimates something greater than 1. If $F$ is much greater than 1, it provides clear evidence that the statistics are not estimating the same thing and thus that the $\mu_i$s are not all equal.

The nature of this evidence is probabilistic and one cannot eliminate the possibility of error. Although they are very unlikely to occur, $F$ ratios much greater than 1 can arise even when the $\mu_i$s are all equal. Assuming that model (5.1.1) is appropriate, when the data yield a very large $F$ ratio, the correct conclusion is either that the assumption of equal treatment means is violated or that the means are equal and a very rare event has occurred. The rarer the event, the stronger the suggestion of unequal treatment means. While we cannot directly quantify the strength of the suggestion of unequal treatment means, we can quantify it indirectly by evaluating how rarely large $F$ ratios occur when the treatment means are equal. Under the assumption that $\mu_1 = \mu_2 = \cdots = \mu_a$, the $F$ ratio is random and has an $F(a-1, a(N-1))$ distribution. (This distribution is called an $F$ distribution in honor of the originator of analysis of variance, R. A. Fisher.)

The $F$ distribution determines those values of the $F$ ratio in (5.2.4) that commonly occur with equal treatment means. If the observed $F$ ratio is so large as to be an uncommon occurrence when $\mu_1 = \mu_2 = \cdots = \mu_a$, we conclude that the $\mu_i$s are not all equal. To measure the strength of this conclusion, compute the probability of obtaining an $F$ ratio as large or larger than that actually obtained from the data. This probability is called the *P value* or the *significance level* of the test. The smaller the $P$ value, the more inconsistent the observed $F$ ratio is with the assumption that the $\mu_i$s are all equal.

On occasion, it may be desired to have a fixed decision rule as to whether the data are inconsistent with the (null) hypothesis of equal means. One may decide that, with equal treatment means, common occurrences of the $F$ ratio include 95% or 99% or more generally $(1-\alpha)100\%$ of the possible $F$ values. Thus uncommon occurrences constitute 5% or 1% or $100\alpha\%$ of the observations. The hypothesis $\mu_1 = \mu_2 = \cdots = \mu_a$ is rejected at the $\alpha$ level if

$$\frac{MSTrts}{MSE} \geq F(1-\alpha, a-1, dfE).$$

Table 5.2: *Analysis of variance*

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Treatments | $a-1$ | $N\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$ | $SSTrts/(a-1)$ | $\frac{MSTrts}{MSE}$ |
| Error | $aN-a$ | $\sum_{i=1}^{a}\sum_{j=1}^{N}(y_{ij}-\bar{y}_{i\cdot})^2$ | $SSE/(n-a)$ | |
| Total$-C$ | $aN-1$ | $\sum_{i=1}^{a}\sum_{j=1}^{N}(y_{ij}-\bar{y}_{\cdot\cdot})^2$ | | |

Here $F(1-\alpha,a-1,dfE)$ is the number below which fall $(1-\alpha)100\%$ of the possible $F$ ratios when the $\mu_i$s are all equal. There is a possibility that our data would yield an $F$ ratio at least this large when the $\mu_i$s are all equal but it is pretty slim, $\alpha$. We consider it more reasonable that the assumption of equal $\mu_i$s is violated. The number $F(1-\alpha,a-1,dfE)$ can be obtained from tables of the $F$ distribution, see Appendix B.7. The number depends not only on the choice of $\alpha$ but also on the degrees of freedom for the estimate in the numerator of the ratio, $a-1$, and the degrees of freedom for the estimate in the denominator of the ratio, $dfE$. If we do not reject the hypothesis, the data are consistent with the hypothesis. Again, just because the data are consistent with the hypothesis does not mean that the hypothesis is true.

Fixed $\alpha$ level tests are easy to perform if the $P$ value is available. To perform, say, an $\alpha = .05$ test, just compare the $P$ value with .05. If the $P$ value is greater than .05, a .05 level test does not reject the hypothesis of equal treatment means $\mu_i$. If the $P$ value is less than .05, a .05 test rejects the hypothesis.

### *5.2.1 The analysis of variance table*

The computations for the analysis of variance $F$ test can be summarized in an analysis of variance table. The columns of the table are sources, degrees of freedom ($df$), sums of squares ($SS$), mean squares ($MS$), and $F$. There are rows for treatments, error, and total (corrected for the grand mean). The commonly used form for the analysis of variance table is given in Table 5.2. The sums of squares for error and treatments are just those given in equations (5.2.1) and (5.2.2). In each row, the mean square is the sum of squares divided by the degrees of freedom. The degrees of freedom and sums of squares for treatments and error can be added together to give the degrees of freedom and sum of squares total (corrected for the grand mean) respectively. Note that the sum of squares total divided by the degrees of freedom total is $s_y^2$, the sample variance of all $aN$ observations computed without reference to any treatment groups. The degrees of freedom in the total line are just the degrees of freedom associated with the sample variance based on all $aN$ observations. Traditionally, the total line does not include a mean square. The sample variance of all $aN$ observations, and thus the total line, involves adjusting each observation for the grand mean. This can be accomplished as indicated in Table 5.2 or, alternatively, by the use of a *correction factor*. The correction factor is $C \equiv aN\bar{y}_{\cdot\cdot}^2$, so that $SSTot - C = \sum_{i=1}^{a}\sum_{j=1}^{N} y_{ij}^2 - C$, which is the sum of the squares of all the observations minus the correction factor.

A less commonly used form for the analysis of variance table, but one I prefer, is presented in Table 5.3. In this form, the total degrees of freedom consist of one degree of freedom for every observation, the sum of squares total is the sum of all of the squared observations, and an extra row has been added for the grand mean. The degrees of freedom and sums of squares for the grand mean, treatments, and error can be added together to obtain the degrees of freedom and sums of square total. In spite of my preference for Table 5.3, I will bow to tradition and generally use Table 5.2 with the $-C$ notation deleted from the Total line.

EXAMPLE 5.2.1. We now examine the analysis of variance table for the electrical characteristic data of Example 5.1.2. The summary statistics for the four samples are repeated below.

Table 5.3: *Analysis of variance*

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Grand mean | 1 | $aN\bar{y}_{\cdot\cdot}^2 \equiv C$ | $aN\bar{y}_{\cdot\cdot}^2$ | |
| Treatments | $a-1$ | $N\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$ | $SSTrts/(a-1)$ | $\frac{MSTrts}{MSE}$ |
| Error | $aN-a$ | $\sum_{i=1}^{a}\sum_{j=1}^{N}(y_{ij}-\bar{y}_{i\cdot})^2$ | $SSE/(n-a)$ | |
| Total | $aN$ | $\sum_{i=1}^{a}\sum_{j=1}^{N}y_{ij}^2$ | | |

Table 5.4: *Analysis of variance table: electrical characteristic data*

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Treatments | 3 | 10.873 | 3.624 | 6.45 | 0.002 |
| Error | 24 | 13.477 | 0.562 | | |
| Total | 27 | 24.350 | | | |

Sample statistics: electrical characteristics

| Strip | $N$ | $\bar{y}_{i\cdot}$ | $s_i^2$ |
|---|---|---|---|
| 1 | 7 | 16.4429 | 0.749524 |
| 2 | 7 | 16.5143 | 0.194762 |
| 3 | 7 | 16.0714 | 0.162381 |
| 4 | 7 | 14.9571 | 1.139524 |

The $MSE$ for balanced data is the simple average of the $s_i^2$s,

$$MSE = \frac{.74952 + .19476 + .16238 + 1.13952}{4} = .56155.$$

The sample mean of the $\bar{y}_{i\cdot}$s is

$$\bar{y}_{\cdot\cdot} = \frac{16.4429 + 16.5143 + 16.0714 + 14.9571}{4} = 15.996425$$

and the sample variance of the $\bar{y}_{i\cdot}$s is

$$s_{\bar{y}}^2 = \frac{1}{4-1}\big[(16.4429 - 15.996425)^2 + (16.5143 - 15.996425)^2 + (16.0714 - 15.996425)^2 + (14.9571 - 15.996425)^2\big] = .517784.$$

The mean square treatments is the sample variance of the $\bar{y}_{i\cdot}$s times the number of observations in each $\bar{y}_{i\cdot}$,

$$MSTrts = Ns_{\bar{y}}^2 = 7(.517784) = 3.6245.$$

The analysis of variance table is given as Table 5.4. As discussed earlier in this section, all of the table entries are easily computed given the $MSE$ and the $MSTrts$.

The $F$ statistic for these data is substantial and the $P$ value is quite small. There is strong evidence that the treatments do not have the same mean. In other words, strips 1, 2, 3, and 4 do not have the same mean value for the electrical characteristic. The analysis of variance $F$ test tells us that the means are not all equal but it does not tell us which particular means are unequal. Examining individual contrasts is required to answer more specific questions about the means. □

*Distribution theory*

It has been stated that when there are no differences between the treatment means, the test statistic $F = MSTrts/MSE$ has an $F(a-1, dfE)$ distribution. We now briefly expand on that statement. By Definition 4.4.3, an $F$ distribution is constructed from two independent $\chi^2$ distributions. If $W_1 \sim \chi^2(r)$ and $W_2 \sim \chi^2(s)$ with $W_1$ and $W_2$ independent, then by definition

$$\frac{W_1/r}{W_2/s} \sim F(r,s).$$

In analysis of variance with the usual assumptions, the $\bar{y}_{i\cdot}$s and $s_i^2$s are all independent of each other. The $MSE$ is computed from the $s_i^2$s and the $MSTrts$ is computed from the $\bar{y}_{i\cdot}$s, so the $MSE$ is independent of the $MSTrts$. We mentioned earlier that when the means are all equal

$$\frac{(a-1)MSTrts}{\sigma^2} \sim \chi^2(a-1)$$

and regardless of the mean structure

$$\frac{dfE \times MSE}{\sigma^2} \sim \chi^2(dfE),$$

so it follows from the definition of the $F$ distribution that, when the means are all equal,

$$\frac{MSTrts}{MSE} = \frac{\left[(a-1)MSTrts/\sigma^2\right]/(a-1)}{\left[(dfE)MSE/\sigma^2\right]/dfE} \sim F(a-1, dfE).$$

When the treatment means are not all equal, the distribution of $MSTrts$ depends on the value of

$$\frac{N}{(a-1)\sigma^2}\sum_{i=1}^{a}(\mu_i - \bar{\mu}_\cdot)^2.$$

Note the similarity of this number to the expected value of $MSTrts$ given in (5.2.3).

## 5.3 Unbalanced analysis of variance

In unbalanced analysis of variance we allow different numbers $N_i$ of observations on the groups. The analysis is slightly more difficult but it follows the same pattern as in Section 5.2. In particular, we assume that

| Sample | Data | | Distribution |
|---|---|---|---|
| 1 | $y_{11}, y_{12}, \ldots, y_{1N_1}$ | iid | $N(\mu_1, \sigma^2)$ |
| 2 | $y_{21}, y_{22}, \ldots, y_{2N_2}$ | iid | $N(\mu_2, \sigma^2)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a$ | $y_{a1}, y_{a2}, \ldots, y_{aN_a}$ | iid | $N(\mu_a, \sigma^2)$ |

with independent samples and the same variance $\sigma^2$ for each sample. In other words, we assume

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij}\text{s independent } N(0, \sigma^2)$$

$i = 1, \ldots, a$ and $j = 1, \ldots, N_i$. The total number of observations is denoted $n = \sum_{i=1}^{a} N_i$. We wish to examine the (null) hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a.$$

Again we compute summary statistics from the samples.

Sample statistics

| Group | Size | Mean | Variance |
|---|---|---|---|
| 1 | $N_1$ | $\bar{y}_{1\cdot}$ | $s_1^2$ |
| 2 | $N_2$ | $\bar{y}_{2\cdot}$ | $s_2^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a$ | $N_a$ | $\bar{y}_{a\cdot}$ | $s_a^2$ |

As before, the sample means, the $\bar{y}_{i\cdot}$s, are estimates of the corresponding $\mu_i$s and the $s_i^2$s all estimate $\sigma^2$. As discussed earlier, with unequal sample sizes an efficient pooled estimate of $\sigma^2$ must be a weighted average of the $s_i^2$s. The weights are the degrees of freedom associated with various estimates.

$$\begin{aligned} MSE &\equiv \frac{(N_1-1)s_1^2+(N_2-1)s_2^2+\cdots+(N_a-1)s_a^2}{\sum_{i=1}^{a}(N_i-1)} \\ &= \frac{1}{(n-a)}\sum_{i=1}^{a}\sum_{j=1}^{N_i}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2. \end{aligned}$$

As before, $dfE = n-a$ and $SSE = (dfE)MSE$.

The second estimate of $\sigma^2$, the one based on the $\bar{y}_{i\cdot}$s, is not particularly intuitive. The $\bar{y}_{i\cdot}$s do not all have the same variance, so even when the $\mu_i$s are all equal, the $\bar{y}_{i\cdot}$s do not form a random sample. To get a variance estimate, the $\bar{y}_{i\cdot}$s must be weighted appropriately. It turns out that the appropriate estimate of $\sigma^2$ is

$$MSTrts = \frac{1}{a-1}\sum_{i=1}^{a} N_i\left(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2$$

where

$$\bar{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{a} N_i\bar{y}_{i\cdot} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{N_i} y_{ij}.$$

Thus $\bar{y}_{\cdot\cdot}$ is the sample mean of all $n$ observations, ignoring the treatment structure. As in the balanced case, the degrees of freedom are $a-1$ and $SSTrts = (a-1)MSTrts$. In general, $MSTrts$ is an estimate of

$$\mathrm{E}(MSTrts) = \sigma^2 + \frac{1}{a-1}\sum_{i=1}^{a} N_i\left(\mu_i-\bar{\mu}_{\cdot}\right)^2$$

where

$$\bar{\mu}_{\cdot} \equiv \frac{1}{n}\sum_{i=1}^{a} N_i\mu_i$$

is the weighted mean of the $\mu_i$s. Once again, if the $\mu_i$s are all equal, $\mu_i = \bar{\mu}_{\cdot}$ for every $i$ and $MSTrts$ is an estimate of $\sigma^2$. If the means are not all equal, $MSTrts$ is an estimate of something larger than $\sigma^2$. Values of $MSTrts/MSE$ that are much larger than 1 call in question the hypothesis of equal population means. Note that the computations for balanced data are just a special, simpler case of the computations for unbalanced data. In particular, the balanced case has $N_i = N$ and $n = aN$.

The computations are again summarized in an analysis of variance table. The commonly used form for the analysis of variance table is given below.

Analysis of variance

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Treatments | $a-1$ | $\sum_{i=1}^{a} N_i\left(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2$ | $SSTrts/(a-1)$ | $\frac{MSTrts}{MSE}$ |
| Error | $n-a$ | $\sum_{i=1}^{a}\sum_{j=1}^{N_i}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2$ | $SSE/(n-a)$ | |
| Total | $n-1$ | $\sum_{i=1}^{a}\sum_{j=1}^{N_i}\left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2$ | | |

Table 5.5: *Analysis of variance, logs of suicide age data*

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Groups | 2 | 2.655 | 1.328 | 7.92 | 0.001 |
| Error | 90 | 15.088 | 0.168 | | |
| Total | 92 | 17.743 | | | |

The degrees of freedom and sums of squares for treatments and error can be added together to give the degrees of freedom and sum of squares total (corrected for the grand mean). Again,

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{a} \sum_{j=1}^{N_i} \hat{\varepsilon}_{ij}^2,$$

establishing that the sum of squares error is the sum of the squared residuals. Moreover, $SSTot/dfTot = s_y^2$, the sample variance of all $n$ observations computed without reference to treatment groups. The degrees of freedom in the total line are the degrees of freedom associated with the sample variance based on all $n$ observations. The total line is corrected for the grand mean, so that $SSTot = \sum_{i=1}^{a} \sum_{j=1}^{N_i} y_{ij}^2 - C$, which is the sum of the squares of all the observations minus the correction factor, $C \equiv n\bar{y}_{\cdot\cdot}^2$.

EXAMPLE 5.3.1. We now consider construction of the analysis of variance table for the logs of the suicide data. The sample statistics are repeated below.

Sample statistics: log of suicide ages

| Group | $N_i$ | $\bar{y}_{i\cdot}$ | $s_i^2$ |
|---|---|---|---|
| Caucasians | 44 | 3.6521 | 0.1590 |
| Hispanics | 34 | 3.4538 | 0.2127 |
| Native Am. | 15 | 3.1770 | 0.0879 |

The mean squared error was computed earlier as .168. The sum of squares error is just the degrees of freedom error, 90, times the $MSE$. The sum of squares treatments is

$$SSTrts = 2.655 = 44(3.6521 - 3.5030)^2 + 34(3.4538 - 3.5030)^2 + 15(3.1770 - 3.5030)^2$$

where

$$3.5030 = \bar{y}_{\cdot\cdot} = \frac{44(3.6521) + 34(3.4538) + 15(3.1770)}{44 + 34 + 15}.$$

The ANOVA table is presented as Table 5.5.

The extremely small $P$ value for the analysis of variance $F$ test establishes a clear difference between the mean log suicide ages. Again, more detailed comparisons are needed to identify which particular groups are different. We established earlier that at the .01 level, only non-Hispanic Caucasians and Native Americans display a pairwise difference. □

## 5.4 Choosing contrasts

You may be wondering why statisticians make a big fuss about analysis of variance. The procedures discussed in Sections 5.2 and 5.3 are not really of much use. The analysis of variance test involves only one hypothesis, that of equal treatment means $\mu_i$. The more interesting issue of identifying which means are different is handled with a pooled estimate of the variance and the usual techniques involving a $Par$, an $Est$, a SE($Est$), and a known distribution symmetric about zero for $[Est - Par]/\mathrm{SE}(Est)$. Actually, 'analysis of variance' is used as a name for the entire package of techniques used to compare more than two samples. The analysis of variance $F$ test, from which the name

devolves, is only one small part of the package. There are two reasons for examining the $F$ test in detail. In more complicated situations than one-way ANOVA, the analysis of variance table becomes a very useful tool for identifying aspects of a complicated problem that deserve more attention. The other reason is that it introduces the $SSTrts$ as a measure of treatment differences.

The $SSTrts$ can be broken into components corresponding to the sums of squares for individual *orthogonal* contrasts. These components of $SSTrts$ can then be used to explain the differences in the means. Recall that a contrast is a parameter $\sum_{i=1}^{a}\lambda_i\mu_i$ where the $\lambda_i$s satisfy $\sum_{i=1}^{a}\lambda_i = 0$. The appropriate estimate and standard error were discussed earlier and the sum of squares for a contrast was given in (5.1.4) as

$$SS\left(\sum_{i=1}^{a}\lambda_i\mu_i\right) \equiv \frac{\left(\sum_{i=1}^{a}\lambda_i\bar{y}_{i\cdot}\right)^2}{\sum_{i=1}^{a}\lambda_i^2/N_i}.$$

In the balanced case with $N = N_i$ for all $i$,

$$SS\left(\sum_{i=1}^{a}\lambda_i\mu_i\right) = \frac{\left(\sum_{i=1}^{a}\lambda_i\bar{y}_{i\cdot}\right)^2}{\left(\sum_{i=1}^{a}\lambda_i^2\right)/N}.$$

The $F$ test for $H_0 : \sum_{i=1}^{a}\lambda_i\mu_i = 0$ versus $H_A : \sum_{i=1}^{a}\lambda_i\mu_i \neq 0$ rejects $H_0$ for large values of $SS\left(\sum_{i=1}^{a}\lambda_i\mu_i\right)/MSE$.

Two contrasts $\sum_{i=1}^{a}\lambda_{i1}\mu_i$ and $\sum_{i=1}^{a}\lambda_{i2}\mu_i$ are defined to be *orthogonal* if

$$\sum_{i=1}^{a}\frac{\lambda_{i1}\lambda_{i2}}{N_i} = 0.$$

In balanced problems, $N_i = N$ for all $i$, so the condition of orthogonality becomes $\sum_{i=1}^{a}\lambda_{i1}\lambda_{i2}/N = 0$ or equivalently

$$\sum_{i=1}^{a}\lambda_{i1}\lambda_{i2} = 0.$$

Contrasts are only of interest when they define interesting functions of the $\mu_i$s. Orthogonal contrasts are most useful in balanced problems because a set of orthogonal contrasts can retain interesting interpretations. In unbalanced cases, orthogonality depends on the unequal $N_i$s, so there is rarely more than one interpretable contrast in a set of orthogonal contrasts.

EXAMPLE 5.4.1. Consider again the electrical characteristic data. The sample statistics are

Sample statistics: electrical characteristics

| Strip | $N$ | $\bar{y}_{i\cdot}$ | $s_i^2$ |
|---|---|---|---|
| 1 | 7 | 16.4429 | 0.749524 |
| 2 | 7 | 16.5143 | 0.194762 |
| 3 | 7 | 16.0714 | 0.162381 |
| 4 | 7 | 14.9571 | 1.139524 |

with $MSE = .56155$. We examine four contrasts

$$C_1 \equiv (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 = \mu_1 - \mu_2,$$

$$C_2 \equiv (1/2)\mu_1 + (1/2)\mu_2 + (-1)\mu_3 + (0)\mu_4 = \frac{\mu_1+\mu_2}{2} - \mu_3,$$

$$C_3 \equiv (1/3)\mu_1 + (1/3)\mu_2 + (1/3)\mu_3 + (-1)\mu_4 = \frac{\mu_1+\mu_2+\mu_3}{3} - \mu_4,$$

and

$$C_4 \equiv (-1)\mu_1 + (-1)\mu_2 + (2)\mu_3 + (0)\mu_4.$$

Contrasts $C_1$ and $C_2$ are orthogonal because

$$(1)(1/2)+(-1)(1/2)+(0)(-1)+(0)(0)=0.$$

Similarly, $C_1$ and $C_3$ are orthogonal and $C_2$ and $C_3$ are orthogonal. We have previously examined the contrast $C_1$ and found the sum of squares to be

$$SS(C_1)=\frac{[(1)16.4429+(-1)16.5143+(0)16.0714+(0)14.9571]^2}{[1^2+(-1)^2+0^2+0^2]/7}=0.0178\,.$$

The sum of squares for $C_2$ is

$$SS(C_2)=\frac{[(1/2)16.4429+(1/2)16.5143+(-1)16.0714+(0)14.9571]^2}{[(1/2)^2+(1/2)^2+(-1)^2+0^2]/7}=0.7738.$$

The sum of squares for $C_3$ is

$$\begin{aligned} SS(C_3) &= \frac{[(1/3)16.4429+(1/3)16.5143+(1/3)16.0714+(-1)14.9571]^2}{[(1/3)^2+(1/3)^2+(1/3)^2+(-1)^2]/7} \\ &= 10.0818. \end{aligned}$$

The decomposition referred to earlier follows from the fact that

$$10.873=SSTrts=SS(C_1)+SS(C_2)+SS(C_3)=0.0178+0.7738+10.0818\,.$$

$SSTrts$ is a measure of the evidence for differences between means. Almost all of the $SSTrts$ is accounted for by $C_3$. Thus, almost all of the differences between the means can be accounted for by the difference between $\mu_4$ and the average of $\mu_1$, $\mu_2$, and $\mu_3$. Almost none of the sum of squares for treatments is due to the difference between $\mu_1$ and $\mu_2$. A small amount is due to the difference between $\mu_3$ and the average of $\mu_1$ and $\mu_2$. The data are consistent with the idea that the means for strips 1, 2, and 3 are the same.

The contrast $C_4$ was introduced to illustrate the fact that *multiplying a contrast by a constant has no real effect on the contrast*. Observe that

$$C_4=-2C_2\,.$$

In particular, $C_4=0$ if and only if $C_2=0$. Note that

$$SS(C_4)=\frac{[(-1)16.4429+(-1)16.5143+(2)16.0714+(0)14.9571]^2}{[(-1)^2+(-1)^2+2^2+0^2]/7}=0.7738,$$

so $SS(C_4)=SS(C_2)$ and the $F$ test for $C_2=0$ is identical to the $F$ test for $C_4=0$. It is also easily seen that a two-sided $t$ test for $H_0:C_4=0$ is identical to that for $H_0:C_2=0$. The factor of $-2$ must be accounted for in estimation and in tests of $C_2$ and $C_4$ other than testing that they are zero, but, after suitable adjustment, estimation and testing are equivalent. The virtue of using $C_4$ rather than $C_2$ is that the $\lambda_i$s in $C_4$ are all integers, so computations are simpler with $C_4$.

There are many ways to pick a set of orthogonal contrasts. We established that the data are consistent with the idea that ceramic strip 4 is different from the other strips and that there are no differences between the other strips. The data are even more consistent with another set of orthogonal contrasts. Consider the claim that the value for strip 4 is the average of the values for strips 1 and 2, i.e., $\mu_4=(\mu_1+\mu_2)/2$ or equivalently

$$C_5\equiv(1)\mu_1+(1)\mu_2+(0)\mu_3+(-2)\mu_4=0.$$

A contrast orthogonal to $C_5$ is $C_1$, considered earlier. A contrast orthogonal to both $C_5$ and $C_1$ is

$$C_6 \equiv (1)\mu_1 + (1)\mu_2 + (-3)\mu_3 + (1)\mu_4.$$

The sum of squares for $C_5$ is

$$SS(C_5) = \frac{[(1)16.4429 + (1)16.5143 + (0)16.0714 + (-2)14.9571]^2}{[1^2 + 1^2 + 0^2 + (-2)^2]/7} = 10.803.$$

The sum of squares for $C_1$ was given earlier, $SS(C_1) = .018$. The sum of squares for $C_6$ is

$$SS(C_6) = \frac{[(1)16.4429 + (1)16.5143 + (-3)16.0714 + (1)14.9571]^2}{[1^2 + 1^2 + (-3)^2 + 1^2]/7} = .052.$$

As before, with orthogonal contrasts

$$10.873 = SSTrts = SS(C_5) + SS(C_1) + SS(C_6) = 10.803 + .018 + .052.$$

For all practical purposes, these data are *totally* consistent with the claims $C_6 = 0$ and $C_1 = 0$ because $SS(C_6) \doteq 0 \doteq SS(C_1)$. Essentially, all the differences in means can be attributed to $C_5$ because $SS(C_5) \doteq SSTrts$. □

It is a mathematical fact that *there is always one contrast that accounts for all of SSTrts*, however, this contrast rarely has a simple interpretation because the coefficients of this contrast depend on the sample means. In a balanced one-way analysis of variance with, say, four treatments, the coefficients of the contrast that accounts for the entire $SSTrts$ are $\lambda_1 = \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot}$, $\lambda_2 = \bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot}$, $\lambda_3 = \bar{y}_{3\cdot} - \bar{y}_{\cdot\cdot}$, and $\lambda_4 = \bar{y}_{4\cdot} - \bar{y}_{\cdot\cdot}$. *Typically, a contrast with these coefficients will be difficult to interpret*. In Example 5.4.1, $C_5$ was constructed in this way, but to simplify the discussion we rounded the coefficients off. Rounding the coefficients helps to make the contrast more interpretable. In the case of $C_5$, the contrast became very simple. When rounding the coefficients, the contrast will not contain quite all of the sum of squares for treatments.

One reasonable approach to analysis of variance is to identify the contrast that accounts for all of the $SSTrts$ and to try to interpret it. I prefer to look at the data and try to identify a contrast or a few orthogonal contrasts that are interpretable and account for most of $SSTrts$. Either of these approaches involves looking at the data to identify contrasts of interest. In such a situation, using the standard $F(1, dfE)$ or $t(dfE)$ distributions for statistical inference is inappropriate. Appropriate statistical methods are discussed in the next chapter.

In some situations, the structure of the treatments suggests orthogonal contrasts that are both interesting and interpretable. When the structure of the treatments, rather than the data, suggests the contrasts, standard methods of inference apply.

The key fact about orthogonal contrasts is that if $C_1, \ldots, C_{a-1}$ is any set of contrasts with each orthogonal to every other one, then

$$SSTrts = SS(C_1) + \cdots + SS(C_{a-1}).$$

In our example, $a = 4$, so there were sets of $a - 1 = 3$ orthogonal contrasts that decompose the $SSTrts$. We gave two such sets of contrasts. There are an infinite number of other ways to choose sets of orthogonal contrasts.

With $a$ treatments, a set of orthogonal contrasts can contain no more than $a - 1$ elements. There can be at most $a - 1$ orthogonal contrasts but one can also choose sets of orthogonal contrasts with, say, $q < a - 1$ elements. In such a case,

$$SSTrts \geq SS(C_1) + \cdots + SS(C_q).$$

In particular, any one contrast $C$ can be viewed as a set with $q = 1$, so

$$SSTrts \geq SS(C). \tag{5.4.1}$$

Interesting contrasts are determined by the structure of the treatments. We now illustrate this fact with an example.

EXAMPLE 5.4.2. Five diets were investigated to determine their effects on the growth of animals. If the diets do not have any recognizable structure, about the only interesting set of contrasts is to compare all pairs of population means. The collection of contrasts is $\mu_i - \mu_{i'}$ for $i, i' = 1,2,3,4,5$ with $i \neq i'$. Note that $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ if and only if all 10 of these contrast are zero, i.e., if $\mu_i - \mu_{i'} = 0$ for all $i \neq i'$. These contrasts are not orthogonal. There can be at most $5 - 1 = 4$ members in a set of orthogonal contrasts; this collection of contrasts has 10 members. In fact, many of these 10 contrasts are redundant. For example, if $\mu_1 - \mu_2 = 0$ and $\mu_2 - \mu_3 = 0$, then, of course, $\mu_1 - \mu_3 = 0$. More generally, if you know the value of $\mu_i - \mu_j$ and the value of $\mu_j - \mu_k$, you also know the value of $\mu_i - \mu_k$.

Although these 10 contrasts may be redundant, statistical inferences about them may not be. For example, failing to reject $H_0 : \mu_1 - \mu_2 = 0$ and $H_0 : \mu_2 - \mu_3 = 0$ in no way implies the we will fail to reject $H_0 : \mu_1 - \mu_3 = 0$. Similarly, rejecting $H_0 : \mu_1 - \mu_2 = 0$ and $H_0 : \mu_2 - \mu_3 = 0$ does not imply that we will reject $H_0 : \mu_1 - \mu_3 = 0$.

Now suppose we are told that treatment 1 is the standard diet and that the other four treatments are new, experimental diets. In this case, the structure of the treatments suggests that we might examine only the contrasts $\mu_1 - \mu_i$ for $i = 2,3,4,5$. These contrasts are not redundant. Knowing two or three of them will never tell you the values of any others. For example, if $\mu_1 - \mu_2 = 0$, $\mu_1 - \mu_3 = 0$, and $\mu_1 - \mu_4 = 0$, we still do not know the value of $\mu_1 - \mu_5$. On the other hand, if all 4 of the contrasts equal 0, we must have $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, and if the treatment means are all equal, every contrast must be zero. These four contrasts are not orthogonal in any ANOVA.

*Contrasts that are not redundant are said to be* linearly independent*. With a treatments, one can have at most $a-1$ linearly independent contrasts. Nontrivial orthogonal contrasts are always linearly independent. (The trivial contrast has $\lambda_i = 0$ for all i.) If any set of $a-1$ linearly independent contrasts are all equal to 0, then $\mu_1 = \mu_2 = \cdots = \mu_a$.*

Additional structure on the treatments may suggest other contrasts. Suppose that the four new diets are, in order, two based on beef, one based on pork, and one based on soybeans. In this case contrasts with the following coefficients seem interesting.

| | Diet treatments | | | | |
|---|---|---|---|---|---|
| | Control | Beef | Beef | Pork | Beans |
| Contrast | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ |
| Ctrl vs others | 4 | −1 | −1 | −1 | −1 |
| Beef vs beef | 0 | 1 | −1 | 0 | 0 |
| Beef vs pork | 0 | 1 | 1 | −2 | 0 |
| Meat vs beans | 0 | 1 | 1 | 1 | −3 |

The first contrast, Ctrl vs others, compares the control (standard diet) to the average of the other four diets. This contrast would actually be

$$\mu_1 - \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4}$$

but multiplying the contrast by 4 gives the equivalent contrast

$$4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5$$

which is the one tabled. The tabled contrast is simpler to work with because its contrast coefficients

are all integers. The other three contrasts compare the two beef diets, the average of the beef diets with the pork diet, and the average of the meat diets with the soybean diet. In a balanced ANOVA, these four contrasts are all orthogonal to each other.

If the structure of the treatments was different, say, the first beef diet was instead a diet based on lima beans, the interesting orthogonal contrasts change.

| | Diet treatments | | | | |
|---|---|---|---|---|---|
| | Control | Lima | Beef | Pork | Soy |
| Contrast | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ |
| Ctrl vs others | 4 | −1 | −1 | −1 | −1 |
| Beef vs pork | 0 | 0 | 1 | −1 | 0 |
| Lima vs soy | 0 | 1 | 0 | 0 | −1 |
| Meat vs beans | 0 | −1 | 1 | 1 | −1 |

These contrasts compare the control to the average of the other four diets, the two meat diets, the two bean diets, and the average of the meat diets with the average of the bean diets. Again, the contrasts are all orthogonal in a balanced ANOVA. □

## 5.5 Comparing models

The hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$$

can be viewed as imposing a change in the analysis of variance model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \tag{5.5.1}$$

$i = 1,\ldots,a,\ j = 1,\ldots,N_i$. If for some value $\mu$, $\mu = \mu_1 = \mu_2 = \cdots = \mu_a$ the analysis of variance model can be rewritten as

$$y_{ij} = \mu + \varepsilon_{ij}, \tag{5.5.2}$$

which involves only a grand mean $\mu$. This is just the special case of the analysis of variance model in which the $\mu_i$s do not really depend on the value of $i$. In (5.1.2) we wrote the analysis of variance model as $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. Model (5.5.2) is the special case obtained by dropping the $\alpha_i$s. For simplicity, in this section models (5.5.1) and (5.5.2) will be referred to as models (1) and (2), respectively.

We wish to evaluate how well model (2) fits as compared to how well model (1) fits. A measure of how well any model fits is the sum of squared errors; a poor fitting model has much larger errors and thus a much larger $SSE$. In $y_{ij} = \mu_i + \varepsilon_{ij}$, the errors are $\varepsilon_{ij} = y_{ij} - \mu_i$ and the estimated errors (*residuals*) are $\hat{\varepsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot}$. The sum of squares error in model (1) is the usual analysis of variance sum of squares error,

$$SSE(1) = \sum_{i=1}^{a}\sum_{j=1}^{N_i} \hat{\varepsilon}_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{i\cdot})^2 .$$

Recall that $MSE(1) = SSE(1)/(n-a)$ is an estimate of $\sigma^2$ and denote the error degrees of freedom

$$dfE(1) = n - a.$$

Model (2) treats all $n$ observations as a random sample from one population with mean $\mu$. Under model (2), an estimate of $\sigma^2$ is $s_y^2$, the sample variance of all $n$ observations, so

$$MSE(2) \equiv s_y^2 = \frac{1}{n-1}\sum_{i=1}^{a}\sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2$$

with error degrees of freedom

$$dfE(2) = n - 1.$$

We define the sum of squares error from model (2) to be

$$\begin{aligned} SSE(2) &= dfE(2) \times MSE(2) \\ &= \sum_{i=1}^{a} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{..})^2 . \end{aligned}$$

Since model (2) is a special case of model (1), the error from model (2) must be as large as the error from model (1), i.e., $SSE(2) \geq SSE(1)$. However, if $SSE(2)$ is much greater than $SSE(1)$, it suggests that the special case, model (2), is an inadequate substitute for the *full* model (1). In particular, large values of $SSE(2) - SSE(1)$ suggest that the *reduced* model (2) is inadequate to explain the data that, by assumption, were adequately explained using the full model (1). It can be established that, if the reduced model is true, the statistic

$$MSTest \equiv \frac{SSE(2) - SSE(1)}{dfE(2) - dfE(1)}$$

is an estimate of $\sigma^2$, which is independent of the estimate from the full model, $MSE(1)$. If the reduced model is not true, $MSTest$ estimates $\sigma^2$ plus a positive number. A test of whether model (2) is an adequate substitute for model (1) is rejected if

$$F = \frac{[SSE(2) - SSE(1)] \big/ [dfE(2) - dfE(1)]}{MSE(1)} \tag{5.5.3}$$

is too much larger than 1. In particular, an $\alpha$ level test rejects the adequacy of model (2) when

$$\frac{[SSE(2) - SSE(1)] \big/ [dfE(2) - dfE(1)]}{MSE(1)} > F\left(1 - \alpha, dfE(2) - dfE(1), dfE(1)\right). \tag{5.5.4}$$

To see that the numerator in (5.5.3) is a reasonable estimate of $\sigma^2$ when model (2) holds, write

$$\begin{aligned} MSE(2) &= \frac{1}{dfE(2)} [SSE(2) - SSE(1) + SSE(1)] \\ &= \frac{dfE(2) - dfE(1)}{dfE(2)} \left( \frac{SSE(2) - SSE(1)}{dfE(2) - dfE(1)} \right) + \frac{dfE(1)}{dfE(2)} MSE(1). \end{aligned}$$

$MSE(2)$ is a weighted average of $MSTest$ and $MSE(1)$. $MSE(1)$ is certainly a reasonable estimate of $\sigma^2$ and, if the means are all equal, $MSE(2)$ is also a reasonable estimate of $\sigma^2$. Thus, if the means are all equal, $[SSE(2) - SSE(1)]/[dfE(2) - dfE(1)]$ must be a reasonable estimate of $\sigma^2$ because if it were not, a weighted average of it and $MSE(1)$ would not be a reasonable estimate of $\sigma^2$.

The $F$ statistic in (5.5.3) is exactly the analysis of variance table $F$ statistic. This follows because, relative to the analysis of variance table,

$$\begin{aligned} SSTot &= SSE(2) \\ dfTot &= dfE(2) \\ SSE &= SSE(1) \\ dfE &= dfE(1) \\ MSE &= SSE(1)/dfE(1) \\ SSTrts &= SSE(2) - SSE(1) \\ dfTrts &= dfE(2) - dfE(1) \\ MSTrts &= [SSE(2) - SSE(1)] \big/ [dfE(2) - dfE(1)] \end{aligned}$$

*This technique of testing the adequacy of a reduced (special case) model by comparing the error sum of squares for the full model and the reduced model is applicable very generally*. In more sophisticated unbalanced analysis of variance situations and in regression analysis, this is a primary method used to test hypotheses. In particular, *the test in (5.5.4) applies for any ANOVA or regression model (2) that is a special case of any ANOVA or regression model (1) as long as the errors are independent* $N(0,\sigma^2)$.

## 5.6 The power of the analysis of variance $F$ test

The power of a test is the probability of rejecting the null hypothesis when the null hypothesis is false. Thus, the power of the analysis of variance $F$ test is the probability of correctly concluding that the $\mu_i$s are not all the same when they are in fact not all the same. In this section we give some intuition for the power of the analysis of variance $F$ test. For simplicity we discuss only balanced analysis of variance.

As discussed in Section 5.2, whenever the analysis of variance model is correct, the $MSE$ is an unbiased estimate of

$$\mathrm{E}(MSE) = \sigma^2 \tag{5.6.1}$$

and $MSTrts$ is an unbiased estimate of

$$\mathrm{E}(MSTrts) = \sigma^2 + \frac{N}{a-1}\sum_{i=1}^{a}(\mu_i - \bar{\mu}_\cdot)^2. \tag{5.6.2}$$

Write

$$s_\mu^2 = \frac{1}{a-1}\sum_{i=1}^{a}(\mu_i - \bar{\mu}_\cdot)^2, \tag{5.6.3}$$

so $s_\mu^2$ is the 'sample' variance of the $\mu_i$s. The word sample is in quotation marks because we do not really have a sample of $\mu_i$s, in fact we *never* get to observe the $\mu_i$s. $s_\mu^2$ is a sample variance only in the sense that the computational formula (5.6.3) is identical to that for a sample variance. With the new notation, we can rewrite (5.6.2) as

$$\mathrm{E}(MSTrts) = \sigma^2 + Ns_\mu^2. \tag{5.6.4}$$

The analysis of variance $F$ statistic is defined as

$$F = \frac{MSTrts}{MSE}.$$

Since the $MSE$ and the $MSTrts$ estimate (5.6.1) and (5.6.4) respectively, by substitution we see that $F$ is an estimate of

$$\frac{\sigma^2 + Ns_\mu^2}{\sigma^2} = 1 + \frac{N}{\sigma^2}s_\mu^2. \tag{5.6.5}$$

($F$ is *not* an unbiased estimate of this quantity but $F$ is a reasonable estimate of it.)

The behavior of the $F$ test depends crucially on the quantity that $F$ estimates. First notice that if the $\mu_i$s are all equal, they have no variability and $s_\mu^2 = 0$. In fact *the $\mu_i$s are all equal if and only if* $s_\mu^2 = 0$. The statistic $F$ always estimates the value in (5.6.5) and when $s_\mu^2 = 0$ that value is 1. Thus an $F$ statistic that is too far above 1 suggests that $s_\mu^2 \neq 0$. Alternatively, when the $\mu_i$s are not all equal, they have positive variability and $s_\mu^2 > 0$. In this case, $F$ is estimating a value in (5.6.5) that is greater than 1, so values of $F$ substantially greater than 1 lead us to suspect that $s_\mu^2 > 0$ and hence that the $\mu_i$s are not all equal.

Remember that even when $s_\mu^2 = 0$, $F$ is only an estimate of 1; it has a natural variability about 1. To reject the idea that $s_\mu^2 = 0$, an observed $F$ value must be larger than would normally be experienced when $s_\mu^2 = 0$.

When $s_\mu^2 = 0$, the statistic $F$ has an $F(dfTrts, dfE)$ distribution. This distribution specifies the values of $F$ that would normally be experienced. Thus an $\alpha$ level test is rejected when the observed $F$ value is larger than all but $100\alpha\%$ of the observations that normally occur, i.e., larger than $F(1-\alpha, dfTrts, dfE)$. When $s_\mu^2 = 0$ there is only a probability of $\alpha$ that the observed $F$ value will exceed $F(1-\alpha, dfTrts, dfE)$.

Note that values of $F$ that are much smaller than 1 *do not* suggest that $s_\mu^2 > 0$. Within the present discussion, values of $F$ that are smaller than 1 are most consistent with $s_\mu^2 = 0$ and will not be considered further. It should be noted, however, that very small values of $F$ are suggestive. In terms of modeling, they are suggestive of something fairly complicated, cf. Christensen (1989, 1991). Very small test statistics have also been known to occur when someone has manufactured data in order to justify a null hypothesis. For example, some data reported by Mendel that supported his theories of genetic inheritance were too good to be true.

We reject the hypothesis of equal $\mu_i$s for $F$ values that are substantially greater than 1. It is natural to ask what causes $F$ to take on values that are substantially greater than 1. In other words, what causes the test to have high power for detecting differences in the $\mu_i$s? Obviously, $F$ will tend to be substantially greater than 1 when it is estimating something that is substantially greater than 1, i.e., when

$$1 + \frac{N}{\sigma^2} s_\mu^2$$

is substantially greater than 1. There are three items involved. To make $1 + Ns_\mu^2/\sigma^2$ much larger than 1 we need some combination of $N$ large, $\sigma^2$ small, and $s_\mu^2$ large. The first two items are somewhat controllable. To increase the power of the $F$ test we can increase $N$, the size of the various samples. The second item, $\sigma^2$, is a parameter, so we will never know it exactly, but improving one's experimental methods can make it smaller. For example, measuring the height of a house with a meter stick rather than a 30 centimeter ruler is likely to yield a much more accurate value for the height. In later chapters we discuss some general methods for designing experiments that enable us to reduce $\sigma^2$. The third item above, $s_\mu^2$, we are simply stuck with. There is little we can do with the $\mu_i$s to cause a test to be powerful. If the differences among the $\mu_i$s are small, the $\mu_i$s have little variability and $s_\mu^2$ is near zero. Other things being equal, it is unlikely that we will correctly reject the $F$ test when $s_\mu^2$ is near zero. More accurately, it is unlikely that we will correctly reject the $F$ test when $s_\mu^2$ is so small that $Ns_\mu^2/\sigma^2$ is near zero. Even when $s_\mu^2$ is small in absolute terms, if $N$ is large or $\sigma^2$ is much smaller than $s_\mu^2$, we have a good chance of correctly identifying that there are differences in the $\mu_i$s.

For specified values of $Ns_\mu^2/\sigma^2$ it is possible to compute the probability of rejecting the $F$ test. To specify $Ns_\mu^2/\sigma^2$ one needs to know $N$ and some approximation for $\sigma^2$; these are often available. The most difficult part of computing the power of an $F$ test is in specifying a reasonable value for $s_\mu^2$. In specifying a value for $s_\mu^2$ we need both to specify a pattern for the differences in the $\mu_i$s and to quantify the extent of the differences. For example, our interest may be in detecting differences in the $\mu_i$s when all of the $\mu_i$s are equal except one, which is, say, $d$ units larger than the others. We can compute the value for $s_\mu^2$ by specifying $d$. Similarly, with an even number of treatments our interest may be in detecting differences in the $\mu_i$s in which half the $\mu_i$s equal one value and the other half equal a different value, with the two values $d$ units apart. Again we can compute a value of $s_\mu^2$ for any difference $d$ but the value of $s_\mu^2$ depends on $d$ in a very different manner than in the first case.

## 5.7 Exercises

EXERCISE 5.7.1. In a study of stress at 600% elongation for a certain type of rubber, Mandel (1972) reported stress test data from five different laboratories. Summary statistics are given in Table 5.6. Compute the analysis of variance table and test for differences in means between all pairs

Table 5.6: *Rubber stress at five laboratories*

| Lab. | Sample size | Sample mean | Sample variance |
|---|---|---|---|
| 1 | 4 | 57.00 | 32.00 |
| 2 | 4 | 67.50 | 46.33 |
| 3 | 4 | 40.25 | 14.25 |
| 4 | 4 | 56.50 | 5.66 |
| 5 | 4 | 52.50 | 6.33 |

Table 5.7: *Acreage in corn for different sized farms*

| Farm acres | Sample size | Sample mean | Sample std. dev. |
|---|---|---|---|
| 80 | 5 | 2.9957 | 0.4333 |
| 160 | 5 | 3.6282 | 0.4056 |
| 240 | 5 | 4.1149 | 0.4169 |
| 320 | 5 | 4.0904 | 0.4688 |
| 400 | 5 | 4.4030 | 0.5277 |

of labs. Use $\alpha = .01$. Is there any reason to worry about the assumptions of the analysis of variance model?

EXERCISE 5.7.2. Snedecor and Cochran (1967, section 6.18) presented data obtained in 1942 from South Dakota on the relationship between the size of farms (in acres) and the number of acres planted in corn. Summary statistics are presented in Table 5.7. Note that the sample standard deviations rather than the sample variances are given. In addition, the pooled standard deviation is 0.4526.

(a) Give the one-way analysis of variance model with all of its assumptions. Can any problems with the assumptions be identified?

(b) Give the analysis of variance table for these data. Test whether there are any differences in corn acreages due to the different size farms. Use $\alpha = .01$.

(c) Test for differences between all pairs of farm sizes using $\alpha = .01$ tests.

(d) Find the sum of squares for the following contrast:

| Farm | 80 | 160 | 240 | 320 | 400 |
|---|---|---|---|---|---|
| Coeff. | $-2$ | $-1$ | 0 | 1 | 2 |

What percentage is this of the treatment sum of squares?

e) Give 95% confidence and prediction intervals for the number of acres in corn for each farm size.

EXERCISE 5.7.3. Table 5.8 gives data on heights and weights of people. Give the analysis of variance table and test for differences among the four groups. Give a 99% confidence interval for the mean weight of people in the 72 inch height group.

EXERCISE 5.7.4. Conover (1971, p. 326) presented data on the amount of iron found in the livers of white rats. Fifty rats were randomly divided into five groups of ten and each group was given a different diet. We analyze the logs of the original data. The total sample variance of the 50 observations is 0.521767 and the means for each diet are given below.

Table 5.8: *Weights (in pounds) for various heights (in inches)*

| Height | Sample size | Sample mean | Sample variance |
|---|---|---|---|
| 63 | 3 | $121.6\bar{6}$ | $158.33\bar{3}$ |
| 65 | 4 | 131.25 | $72.91\bar{3}$ |
| 66 | 2 | 142.50 | 112.500 |
| 72 | 3 | $171.6\bar{6}$ | $158.33\bar{3}$ |

Table 5.9: *Peel-strength of various adhesive systems*

| Adhesive system | Observations | | | | | |
|---|---|---|---|---|---|---|
| 1 | 60 | 63 | 57 | 53 | 56 | 57 |
| 2 | 57 | 52 | 55 | 59 | 56 | 54 |
| 3 | 19.8 | 19.5 | 19.7 | 21.6 | 21.1 | 19.3 |
| 4 | 52 | 53 | 44 | 48 | 48 | 53 |

| Diet | A | B | C | D | E |
|---|---|---|---|---|---|
| Mean | 1.6517 | 0.87413 | 0.89390 | 0.40557 | 0.025882 |

Compute the analysis of variance table and test whether there are differences due to diet.

If diets A and B emphasize beef and pork respectively, diet C emphasizes poultry, and diets D and E are based on dried beans and oats, the following contrasts may be of interest.

| | Diet | | | | |
|---|---|---|---|---|---|
| Contrast | A | B | C | D | E |
| Beef vs. pork | 1 | $-1$ | 0 | 0 | 0 |
| Mammals vs. poultry | 1 | 1 | $-2$ | 0 | 0 |
| Beans vs. oats | 0 | 0 | 0 | 1 | $-1$ |
| Animal vs. vegetable | 2 | 2 | 2 | $-3$ | $-3$ |

Show that the contrasts are orthogonal and compute sums of squares for each contrast. Interpret your results and draw conclusions about the data.

EXERCISE 5.7.5. In addition to the data discussed earlier, Mandel (1972) reported data from one laboratory on four different types of rubber. Four observations were taken on each type of rubber. The means are given below.

| Material | A | B | C | D |
|---|---|---|---|---|
| Mean | 26.4425 | 26.0225 | 23.5325 | 29.9600 |

The sample variance of the 16 observations is 14.730793. Compute the analysis of variance table, the overall $F$ test, and test for differences between each pair of rubber types. Use $\alpha = .05$.

EXERCISE 5.7.6. In Exercise 5.7.5 on the stress of four types of rubber, the observations on material B were 22.96, 22.93, 22.49, and 35.71. Redo the analysis, eliminating the outlier. The sample variance of the 15 remaining observations is 9.3052838.

EXERCISE 5.7.7. Bethea et al. (1985) reported data on an experiment to determine the effectiveness of four adhesive systems for bonding insulation to a chamber. The data are a measure of the peel-strength of the adhesives and are presented in Table 5.9. A disturbing aspect of these data is that the values for adhesive system 3 are reported with an extra digit.

Table 5.10: *Weight gains of rats*

| Thyroxin | Thiouracil | | Control | |
|---|---|---|---|---|
| 132 | 68 | 68 | 107 | 115 |
| 84 | 63 | 52 | 90 | 117 |
| 133 | 80 | 80 | 91 | 133 |
| 118 | 63 | 61 | 91 | 115 |
| 87 | 89 | 69 | 112 | 95 |
| 88 | | | | |
| 119 | | | | |

Table 5.11: *Tetrahydrocortisone values for patients with Cushing's syndrome*

| *a* | *b* | | *c* |
|---|---|---|---|
| 3.1 | 8.3 | 15.4 | 10.2 |
| 3.0 | 3.8 | 7.7 | 9.2 |
| 1.9 | 3.9 | 6.5 | 9.6 |
| 3.8 | 7.8 | 5.7 | 53.8 |
| 4.1 | 9.1 | 13.6 | 15.8 |
| 1.9 | | | |

(a) Compute the sample means and variances for each group. Give the one-way analysis of variance model with all of its assumptions. Are there problems with the assumptions? If so, does an analysis on the square roots or logs of the data reduce these problems?

(b) Give the analysis of variance table for these (possibly transformed) data. Test whether there are any differences in adhesive systems. Use $\alpha = .01$.

(c) Test for differences between all pairs of adhesive systems using $\alpha = .01$ tests.

(d) Find the sums of squares i) for comparing system 1 with system 4 and ii) for comparing system 2 with system 3.

(e) Perform a .01 level $F$ test for whether the mean peel-strength of systems 1 and 4 differs from the mean peel-strength of systems 2 and 3.

(f) What property is displayed by the sums of squares computed in (d) and (e)? Why do they have this property?

(g) Give a 99% confidence interval for the mean of every adhesive system.

(h) Give a 99% prediction interval for every adhesive system.

(i) Give a 95% confidence interval for the difference between systems 1 and 2.

EXERCISE 5.7.8. Table 5.10 contains weight gains of rats from Box (1950). The rats were given either Thyroxin or Thiouracil or were in a control group. Do a complete analysis of variance on the data. Give the model, check assumptions, make residual plots, give the ANOVA table, and examine appropriate contrasts.

EXERCISE 5.7.9. Aitchison and Dunsmore (1975) presented data on Cushing's syndrome. Cushing's syndrome is a condition in which the adrenal cortex overproduces cortisol. Patients are divided into one of three groups based on the cause of the syndrome: $a$ – adenoma, $b$ – bilateral hyperplasia, and $c$ – carcinoma. The data are amounts of tetrahydrocortisone in the urine of the patients. The data are given in Table 5.11. Give a complete analysis.

EXERCISE 5.7.10. Draper and Smith (1966, p. 41) considered data on the relationship between the age of truck tractors (in years) and the cost (in dollars) of maintaining them over a six month period. The data are given in Table 5.12.

Table 5.12: *Age and costs of maintenance for truck tractors*

| Age | Costs | | |
|---|---|---|---|
| 0.5 | 163 | 182 | |
| 1.0 | 978 | 466 | 549 |
| 4.0 | 495 | 723 | 681 |
| 4.5 | 619 | 1049 | 1033 |
| 5.0 | 890 | 1522 | 1194 |
| 5.5 | 987 | | |
| 6.0 | 764 | 1373 | |

Note that there is only one observation at 5.5 years of age. This group does not yield an estimate of the variance and can be ignored for the purpose of computing the mean squared error. In the weighted average of variance estimates, the variance of this group is undefined but the variance gets 0 weight, so there is no problem.

Give the analysis of variance table for these data. Does cost differ with age? Is there a significant difference between the cost at 0.5 years as opposed to 1.0 year? Use several contrasts to determine whether there are any differences between costs at 4, 4.5, 5, 5.5, and 6 years. How much of the sum of squares for treatments is due to the following contrast?

| Age | 0.5 | 1.0 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 |
|---|---|---|---|---|---|---|---|
| Coeff. | $-5$ | $-5$ | 2 | 2 | 2 | 2 | 2 |

What is the sum of squares for the contrast that compares the average of 0.5 and 1.0 with the averages of 4, 4.5, 5, 5.5, and 6?

EXERCISE 5.7.11. George Snedecor (1945a) asked for the appropriate variance estimate in the following problem. One of six treatments was applied to the 10 hens contained in each of 12 cages. Each treatment was randomly assigned to two cages. The data were the number of eggs laid by each hen.

(a) What should you tell Snedecor? Were the treatments applied to the hens or to the cages? How will the analysis differ depending on the answer to this question?

(b) The mean of the 12 sample variances computed from the 10 hens in each cage was 297.8. The average of the 6 sample variances computed from the two cage means for each treatment was 57.59. The sample variance of the 6 treatment means was 53.725. How should you construct an $F$ test? Remember that the numbers reported above are not necessarily mean squares.

EXERCISE 5.7.12. Lehmann (1975), citing Heyl (1930) and Brownlee (1960), considered data on determining the gravitational constant of three elements: gold, platinum, and glass. The data Lehmann gives are the third and fourth decimal places in five determinations of the gravitational constant. They are presented below. Analyze the data.

| Gold | Platinum | Glass |
|---|---|---|
| 83 | 61 | 78 |
| 81 | 61 | 71 |
| 76 | 67 | 75 |
| 79 | 67 | 72 |
| 76 | 64 | 74 |

EXERCISE 5.7.13. Shewhart (1939, p. 69) also presented the gravitational constant data of Heyl (1930) that was considered in the previous problem, but Shewhart reports six observations for gold instead of five. Shewhart's data are given below. Analyze these data and compare your results to those of the previous exercise.

| Gold | Platinum | Glass |
|---|---|---|
| 83 | 61 | 78 |
| 81 | 61 | 71 |
| 76 | 67 | 75 |
| 79 | 67 | 72 |
| 78 | 64 | 74 |
| 72 | | |

EXERCISE 5.7.14. Recall that if $Z \sim N(0,1)$ and $W \sim \chi^2(r)$ with $Z$ and $W$ independent, then by Definition 2.1.3 $Z/\sqrt{W/r}$ has a $t(r)$ distribution. Also recall that in a one-way ANOVA with independent normal errors, a contrast has

$$\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} \sim N\left(\sum_{i=1}^{a} \lambda_i \mu_i, \sigma^2 \sum_{i=1}^{a} \frac{\lambda_i^2}{N_i}\right),$$

$$\frac{SSE}{\sigma^2} \sim \chi^2(dfE),$$

and $MSE$ independent of all the $\bar{y}_{i\cdot}$s. Show that

$$\frac{\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} - \sum_{i=1}^{a} \lambda_i \mu_i}{\sqrt{MSE \sum_{i=1}^{a} \lambda_i^2 / N_i}} \sim t(dfE).$$

Chapter 6

# Multiple comparison methods

As illustrated in Section 5.1, the most useful information from a one-way ANOVA is obtained through examining contrasts. The trick is in picking interesting contrasts to consider. Interesting contrasts are determined by the structure of the treatments or are suggested by the data.

The structure of the treatments often suggests a fixed group of contrasts that are of interest. For example, if one of the treatments is a standard treatment or a control, it is of interest to compare all of the other treatments to the standard. With $a$ treatments, this leads to $a-1$ contrasts. (These will not be orthogonal.) In Chapter 11 we will consider factorial treatment structures. These include cases such as four fertilizer treatments, say,

$$n_0p_0 \quad n_0p_1 \quad n_1p_0 \quad n_1p_1$$

where $n_0p_0$ is no fertilizer, $n_0p_1$ consists of no nitrogen fertilizer but application of a phosphorous fertilizer, $n_1p_0$ consists of a nitrogen fertilizer but no phosphorous fertilizer, and $n_1p_1$ indicates both types of fertilizer. Again the treatment structure suggests a fixed group of contrasts to examine. One interesting contrast compares the two treatments having nitrogen fertilizer against the two without nitrogen fertilizer, another compares the two treatments having phosphorous fertilizer against the two without phosphorous fertilizer, and a third contrast compares the effect of nitrogen fertilizer when phosphorous is not applied with the effect of nitrogen fertilizer when phosphorous is applied. Again, we have $a$ treatments and $a-1$ contrasts. In a balanced ANOVA, these $a-1$ contrasts are orthogonal. Even when there is an apparent lack of structure in the treatments, the very lack of structure suggests a fixed group of contrasts. If there is no apparent structure, the obvious thing to do is compare all of the treatments with all of the other treatments. With three treatments, there are three distinct pairs of treatments to compare. With four treatments, there are six distinct pairs of treatments to compare. With five treatments, there are ten pairs. With seven treatments, there are 21 pairs. With 13 treatments, there are 78 pairs.

One problem is that, with a moderate number of treatment groups, there are many contrasts to look at. When we do tests or confidence intervals, there is a built in chance for error. The more statistical inferences we perform, the more likely we are to commit an error. The purpose of the multiple comparison methods examined in this chapter is to control the probability of making a specific type of error. When testing many contrasts, we have many null hypotheses. This chapter considers *multiple comparison methods that control (i.e., limit) the probability of making an error in any of the tests, when* all *of the null hypotheses are correct*. Limiting this probability is referred to as weak control of the *experimentwise error rate*. It is referred to as weak control because the control only applies under the very stringent assumption that all null hypotheses are correct. Some authors consider a different approach and define strong control of the experimentwise error rate as control of the probability of falsely rejecting any null hypothesis. Thus strong control limits the probability of false rejections even when some of the null hypotheses are false. Not everybody distinguishes between weak and strong control, so the definition of experimentwise error rate depends on whose work you are reading. One argument against weak control of the experimentwise error rate is that in designed experiments, you choose treatments that you expect to have different effects. In such cases,

Table 6.1: *Mandel's data on thirteen laboratories with summary statistics for the logs of the data*

| Lab | Observations | | | | $N$ | $\bar{y}_{i\cdot}$ | $s_i^2$ | $s_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 133 | 129 | 123 | 156 | 4 | 4.9031 | 0.01061315 | 0.1030 |
| 2 | 129 | 125 | 136 | 127 | 4 | 4.8612 | 0.00134015 | 0.0366 |
| 3 | 121 | 125 | 109 | 128 | 4 | 4.7919 | 0.00502248 | 0.0709 |
| 4 | 57 | 58 | 59 | 67 | 4 | 4.0964 | 0.00540738 | 0.0735 |
| 5 | 122 | 98 | 107 | 110 | 4 | 4.6906 | 0.00814531 | 0.0903 |
| 6 | 109 | 120 | 112 | 107 | 4 | 4.7175 | 0.00252643 | 0.0503 |
| 7 | 80 | 72 | 76 | 64 | 4 | 4.2871 | 0.00915446 | 0.0957 |
| 8 | 135 | 151 | 143 | 142 | 4 | 4.9603 | 0.00210031 | 0.0458 |
| 9 | 69 | 69 | 73 | 70 | 4 | 4.2518 | 0.00071054 | 0.0267 |
| 10 | 132 | 129 | 141 | 137 | 4 | 4.9028 | 0.00155179 | 0.0394 |
| 11 | 118 | 109 | 115 | 106 | 4 | 4.7176 | 0.00239586 | 0.0489 |
| 12 | 133 | 133 | 129 | 128 | 4 | 4.8731 | 0.00040518 | 0.0201 |
| 13 | 86 | 84 | 96 | 81 | 4 | 4.4610 | 0.00535505 | 0.0732 |

it makes little sense to concentrate on controlling the error under the assumption that all treatments have the same effect. On the other hand, strong control is more difficult to establish.

Our discussion of multiple comparisons focuses on testing whether contrasts are equal to 0. In all but one of the methods considered in this chapter, the experimentwise error rate is (weakly) controlled by first doing a test of the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_a$. If this test is not rejected, we do not claim that any individual contrast is different from 0. In particular, if $\mu_1 = \mu_2 = \cdots = \mu_a$, any contrast among the means must equal 0, so all of the null hypotheses are correct. Since the error rate for the test of $\mu_1 = \mu_2 = \cdots = \mu_a$ is controlled, the weak experimentwise error rate for the contrasts is also controlled.

Many multiple testing procedures can be adjusted to provide multiple confidence intervals that have a guaranteed simultaneous coverage. Several such methods will be presented in this chapter.

Besides the treatment structure suggesting contrasts, the other source of interesting contrasts is having the data suggest them. If the data suggest a contrast, then the 'parameter' in our standard theory for statistical inferences is a function of the data and not a parameter in the usual sense of the word. When the data suggest the parameter, the standard theory for inferences does not apply. To handle such situations we can often include the contrasts suggested by the data in a broader class of contrasts and develop a procedure that applies to *all* contrasts in the class. In such cases we can ignore the fact that the data suggested particular contrasts of interest because these are still contrasts in the class and the method applies for all contrasts in the class. Of the methods considered in the current chapter, only Scheffé's method (discussed in Section 6.4) is generally considered appropriate for this kind of data dredging.

Recently, a number of books have been published on multiple comparison methods, e.g., Hochberg and Tamhane (1987). A classic discussion is Miller (1981), who also focuses on weak control of the experimentwise error rate, cf. Miller's section 1.2.

We present multiple comparison methods in the context of the one-way ANOVA model (5.1.1) but the methods extend easily to many other situations. We will use a single numerical example to illustrate most of the methods discussed in this chapter. The data are introduced in Example 6.0.1.

EXAMPLE 6.0.1. Mandel (1972) presented data on the stress at 600% elongation for natural rubber with a 40 minute cure at 140 °C. Stress was measured four times by each of 13 laboratories. The units for the data are kilograms per centimeter squared ($kg/cm^2$). The data are presented in Table 6.1. While an analysis of these data on the original scale is not unreasonable, the assumptions of equal variances and normality seem to be more nearly satisfied on the logarithmic scale. The standard summary statistics for computing the analysis of variance on the natural logs of the data are also given in Table 6.1.

This is a balanced one-way ANOVA, so the simple average of the 13 $s_i^2$s gives the $MSE$. There

Table 6.2: *Analysis of variance table for logs of Mandel's data*

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Trts | 12 | 3.92678 | 0.32723 | 77.73 | 0.000 |
| Error | 39 | 0.16418 | 0.00421 | | |
| Total | 51 | 4.09097 | | | |

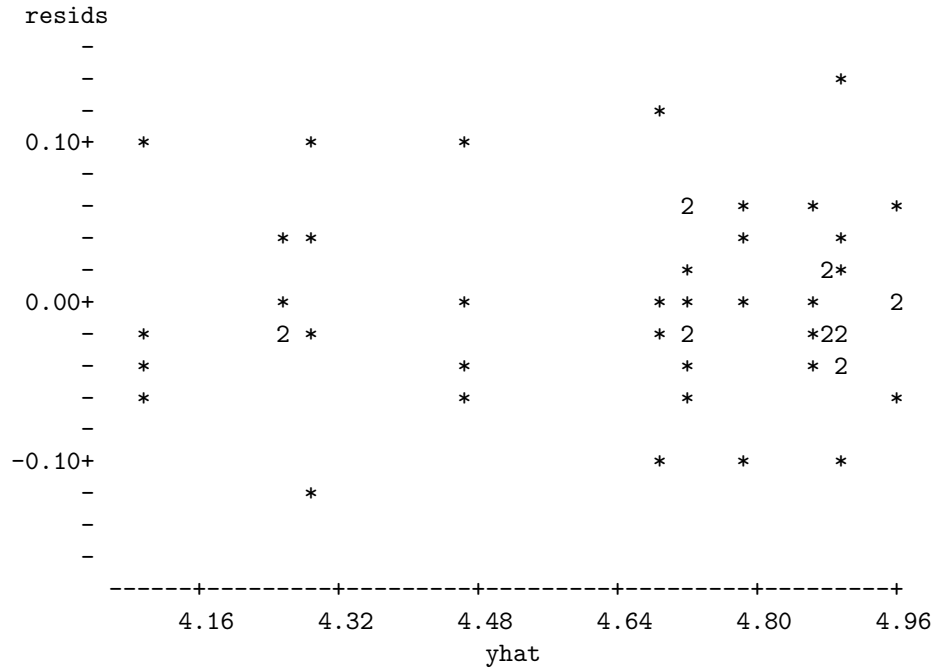



Figure 6.1: *Plot of residuals versus predicted values.*

are three degrees of freedom for the variance estimate from each laboratory, so with 13 laboratories there are a total of $13(3) = 39$ degrees of freedom for error. The mean squared error times the degrees of freedom for error gives the sum of squares for error. The sample variance of the 13 $\bar{y}_{i\cdot}$s is $s^2_{\bar{y}} = .081806429$. Multiplying this by the number of observations in each group, 4, gives the $MSTrts$. The $MSTrts$ times $(13-1)$ gives the $SSTrts$. The sum of squares total is the sample variance of the logs of all 52 observations times $(52-1)$. The degrees of freedom total are $52-1$. These calculations are summarized in the analysis of variance table given in Table 6.2.

Figures 6.1, 6.2, and 6.3 give residual plots. Figure 6.1 is a plot of the residuals versus the predicted values. The group mean $\bar{y}_{i\cdot}$ is the predicted value for an observation from group $i$. Figure 6.1 shows no particular trend in the variabilities. Figure 6.2 is a plot of the residuals versus indicators of the 13 laboratories. Again, there are no obvious problems. Figure 6.3 gives a normal plot of the residuals; the plot looks quite straight.

For pedagogical purposes, on some occasions we consider only the first seven of the 13 treatment groups. We are not selecting these laboratories based on the data and we will continue to use the $MSE$ and $dfE$ from the full data. □

## 6.1 Fisher's least significant difference method

The easiest way to adjust for multiple comparisons is to use R. A. Fisher's least significant difference method. To put it as simply as possible, with this method you first look at the analysis of variance

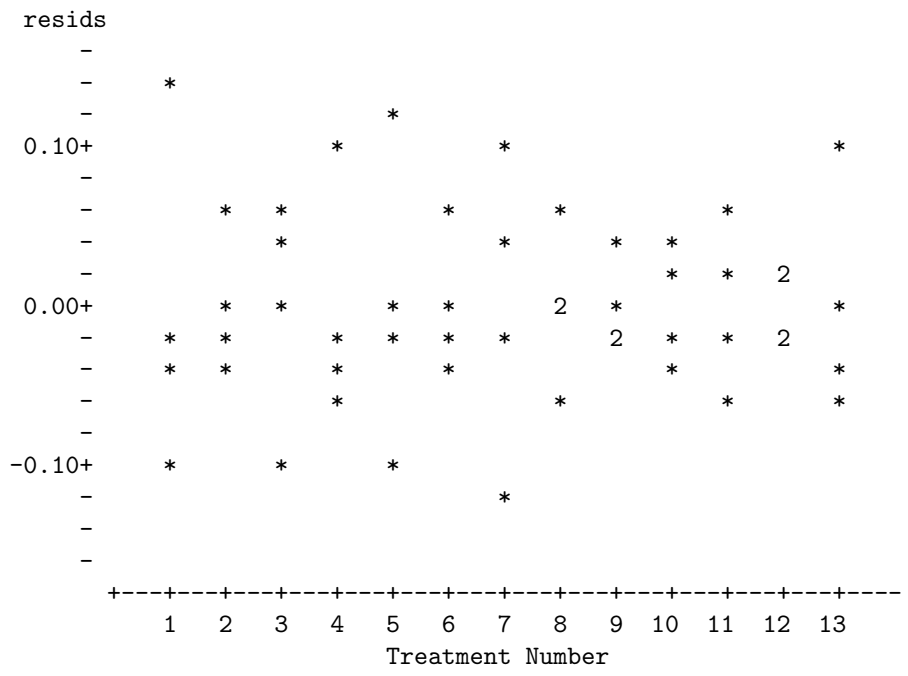



Figure 6.2: *Plot of residuals versus treatment number.*

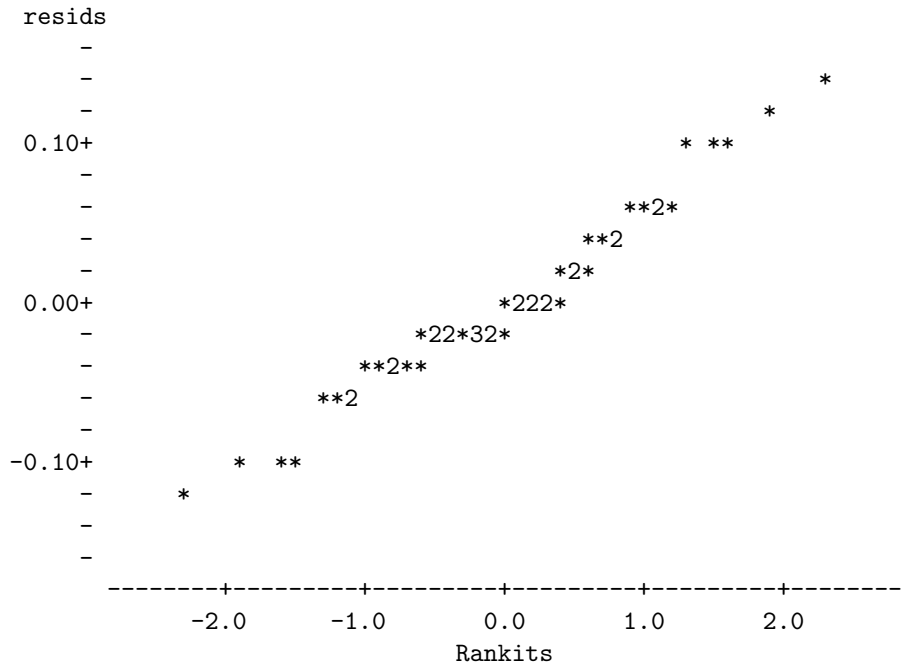



Figure 6.3: *Normal plot of residuals,* $W' = 0.976$.

$F$ test for whether there are differences between the groups. If this test provides no evidence of differences, you quit and go home. If the test is significant at, say, the $\alpha = .05$ level, you just ignore the multiple comparison problem and do all other tests in the usual way at the .05 level. *This method is generally considered inappropriate for use with contrasts suggested by the data*. While the theoretical basis for excluding contrasts suggested by the data is not clear (at least relative to weak control of the experimentwise error rate), experience indicates that the method rejects far too many individual null hypotheses if this exclusion is not applied. In addition, many people would not apply the method unless the number of comparisons to be made was quite small.

The term 'least significant difference' comes from comparing pairs of means in a balanced ANOVA. There is a number, the least significant difference (*LSD*), such that the difference between two means must be greater than the *LSD* for the corresponding treatments to be considered significantly different. Generally, we have a significant difference between $\mu_i$ and $\mu_j$ if

$$\frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|}{\sqrt{MSE\left[\frac{1}{N} + \frac{1}{N}\right]}} > t\left(1 - \frac{\alpha}{2}, dfE\right).$$

Multiplying both sides by the standard error leads to rejection if

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > t\left(1 - \frac{\alpha}{2}, dfE\right)\sqrt{MSE\left[\frac{1}{N} + \frac{1}{N}\right]}.$$

The number on the right is defined as the least significant difference,

$$LSD \equiv t\left(1 - \frac{\alpha}{2}, dfE\right)\sqrt{MSE\frac{2}{N}}.$$

Note that the *LSD* depends on the choice of $\alpha$ but does not depend on which means are being examined. If the absolute difference between two sample means is greater than the *LSD* the population means are declared significantly different. Recall, however, that these comparisons are never attempted unless the analysis of variance $F$ test is rejected at the $\alpha$ level. The reason that a single number exists for comparing all pairs of means is that in a balanced ANOVA the standard error is the same for any comparison between a pair of means.

EXAMPLE 6.1.1. For Mandel's laboratory data, the analysis of variance $F$ test is highly significant, so we can proceed to make individual comparisons among pairs of means. With $\alpha = .05$,

$$LSD = t(.975, 39)\sqrt{.00421\left[\frac{1}{4} + \frac{1}{4}\right]} = 2.023(.0459) = .093$$

Means that are greater than .093 apart are significantly different. Means that are less than .093 apart are not significantly different. We display the results visually. Order the sample means from smallest to largest and indicate groups of means that are not significantly different by underlining the group. Such a display is given below for comparing laboratories 1 through 7.

| Lab. | 4 | 7 | 5 | 6 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|
| Mean | 4.0964 | 4.2871 | 4.6906 | 4.7175 | 4.7919 | 4.8612 | 4.9031 |

Laboratories 4 and 7 are distinct from all other laboratories. All the other consecutive pairs of labs are insignificantly different. Thus labs 5 and 6 cannot be distinguished. Similarly, labs 6 and 3 cannot be distinguished, 3 and 2 cannot be distinguished, and labs 2 and 1 cannot be distinguished.

However, lab 5 is significantly different from labs 3, 2, and 1. Lab 6 is significantly different from labs 2 and 1. Also, lab 3 is different from lab 1.

To be completely correct, *when comparing just the first 7 laboratories the LSD method should be based on an F test for just those 7 laboratories* rather than the $F$ test from Table 6.2 which is based on all 13 laboratories. This can be done by computing a $MSTrts$ in the usual way from the sample means of just the first 7 labs. The resulting $F$ test has 6 degrees of freedom in the numerator and is highly significant ($F = 89.84$). Unfortunately, this point is often ignored in practice.

We can also use the *LSD* to compare all 13 laboratories. Again, we use a visual display, but with more means we list the ordered means vertically and use letters, rather than lines, to indicate groups that are not significantly different.

| Lab. | Mean | | | |
|---|---|---|---|---|
| 4 | 4.0964 | A | | |
| 9 | 4.2518 | B | | |
| 7 | 4.2871 | B | | |
| 13 | 4.4610 | C | | |
| 5 | 4.6906 | D | | |
| 6 | 4.7175 | D | E | |
| 11 | 4.7176 | D | E | |
| 3 | 4.7919 | F | E | |
| 2 | 4.8612 | F | G | |
| 12 | 4.8731 | F | G | H |
| 10 | 4.9028 | | G | H |
| 1 | 4.9031 | | G | H |
| 8 | 4.9603 | | | H |

For example, labs 12, 10, 1, and 8 all share the letter H, so there are no significant differences declared among those four labs. □

For testing a group of contrasts that are 1) not just comparisons between pairs of means or 2) not from a balanced ANOVA, first perform the analysis of variance $F$ test at the $\alpha$ level and if it is rejected, test $H_0 : \sum_i \lambda_i \mu_i = 0$ by rejecting if

$$\frac{SS(\sum_i \lambda_i \mu_i)}{MSE} > F(1-\alpha, 1, dfE).$$

Alternatively, one can use the equivalent $t$ tests for the contrasts.

EXAMPLE 6.1.2. Suppose that in Mandel's data the first two laboratories are in San Francisco, the second two are in Seattle, the fifth is in New York, and the sixth and seventh are in Boston. This structure to the treatments suggests some interesting orthogonal contrasts. We can compare the average of the labs on the West Coast with the average of the labs on the East Coast. On the West Coast we can compare the average of the San Francisco labs with the average of the Seattle labs, we can compare the San Francisco labs with each other and the Seattle labs with each other. On the East Coast we can compare the New York lab with the average of the Boston labs and the Boston labs with each other. The contrast coefficients along with estimates and sums of squares are given in Table 6.3. The contrasts involving averages have been multiplied by appropriate constants to get simple integer contrast coefficients.

Recalling that the overall $F$ test is highly significant for the first 7 labs, to perform the $\alpha = .05$ level *LSD* method on the contrasts of Table 6.3, just divide each sum of squares by $MSE = .00421$ to get an $F$ statistic and compare the $F$ statistics to $F(.95, 1, 39) = 4.09$. The $F$ statistics are given below.

| Contrast | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|
| $F$ | 15.68 | 182.28 | 0.83 | 229.76 | 22.45 | 88.03 |

Table 6.3: *Orthogonal contrasts for the first seven laboratories in Mandel's data*

| | Contrast coefficients | | | | | |
|---|---|---|---|---|---|---|
| Lab. | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
| 1 | 3 | 1 | 1 | 0 | 0 | 0 |
| 2 | 3 | 1 | −1 | 0 | 0 | 0 |
| 3 | 3 | −1 | 0 | 1 | 0 | 0 |
| 4 | 3 | −1 | 0 | −1 | 0 | 0 |
| 5 | −4 | 0 | 0 | 0 | 2 | 0 |
| 6 | −4 | 0 | 0 | 0 | −1 | 1 |
| 7 | −4 | 0 | 0 | 0 | −1 | −1 |
| *Est* | 1.177 | .8760 | .0418 | .6954 | .3765 | .4305 |
| *SS* | .0660 | .7674 | .0035 | .9673 | .0945 | .3706 |

All of the contrasts are significantly different from zero except $C_3$, the comparison between the two labs in San Francisco. □

Apparently some people have taken to calling this method the Fisher significant difference ($FSD$) method. One suspects that this is a reaction to another meaning commonly associated with the letters LSD. I, for one, would *never* suggest that only people who are hallucinating would believe all differences declared by LSD are real.

## 6.2 Bonferroni adjustments

The Bonferroni method is the one method we consider that *does not* stem from a test of $\mu_1 = \mu_2 = \cdots = \mu_a$. Rather, it controls the experimentwise error rate by employing a simple adjustment to the significance level of each individual test. If you have planned to do $s$ tests, you just perform each test at the $\alpha/s$ level rather than at the $\alpha$ level. This method is *absolutely not appropriate for contrasts that are suggested by the data*.

The justification for Bonferroni's method relies on a very simple result from probability: for two events, the probability that one or the other event occurs is no more than the sum of the probabilities for the individual events. Thus with two tests, say $A$ and $B$, the probability that we reject $A$ or reject $B$ is less than or equal to the probability of rejecting $A$ plus the probability of rejecting $B$. In particular, if we fix the probability of rejecting $A$ at $\alpha/2$ and the probability of rejecting $B$ at $\alpha/2$, then the probability of rejecting $A$ or $B$ is no more than $\alpha/2+\alpha/2=\alpha$. More generally, if we have $s$ tests and control the probability of type I error for each test at $\alpha/s$, then the probability of rejecting any of the tests when all $s$ null hypotheses are true is no more than $\alpha/s+\cdots+\alpha/s=\alpha$.

To compare pairs of means in a balanced ANOVA, as with the least significant difference method, there is a single number to which we can compare the differences in means. For a fixed $\alpha$, this number is called the *Bonferroni significant difference* and takes on the value

$$BSD \equiv t\left(1-\frac{\alpha}{2s}, dfE\right)\sqrt{MSE\left[\frac{1}{N}+\frac{1}{N}\right]}.$$

Recall for comparison that with the least significant difference method, the necessary tabled value is $t(1-\alpha/2, dfE)$, which is always smaller than the tabled value for the $BSD$. Thus the $BSD$ is always larger than the $LSD$ and the $BSD$ tends to display fewer differences among the means than the $LSD$.

When testing a group of contrasts that are not just comparisons between pairs of means in a balanced ANOVA, reject a particular contrast hypothesis $H_0 : \sum_i \lambda_i \mu_i = 0$ if

$$\frac{SS(\sum_i \lambda_i \mu_i)}{MSE} > F\left(1-\frac{\alpha}{s}, 1, dfE\right).$$

Equivalent adjustments can be made when performing $t$ rather than $F$ tests.

Bonferroni adjustments can also be used to obtain confidence intervals that have a simultaneous confidence of $(1-\alpha)100\%$ for covering all of the contrasts. The endpoints of these intervals are

$$\sum_{i=1}^{a}\lambda_i\bar{y}_{i\cdot} \pm t\left(1-\frac{\alpha}{2s},dfE\right)\,\mathrm{SE}\left(\sum_{i=1}^{a}\lambda_i\bar{y}_{i\cdot}\right).$$

Recall that for an unbalanced ANOVA,

$$\mathrm{SE}\left(\sum_{i=1}^{a}\lambda_i\bar{y}_{i\cdot}\right)=\sqrt{MSE\sum_{i=1}^{a}\frac{\lambda_i^2}{N_i}}.$$

Only the tabled value distinguishes this interval from a standard confidence interval for $\sum_{i=1}^{a}\lambda_i\mu_i$. In the special case of comparing pairs of means in a balanced ANOVA, the Bonferroni confidence interval for, say, $\mu_i-\mu_j$ reduces to

$$(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})\pm BSD.$$

For these intervals, we are $(1-\alpha)100\%$ confident that the collection of all such intervals simultaneously contain all of the corresponding differences between pairs of population means.

EXAMPLE 6.2.1. In comparing the first 7 laboratories, we have $\binom{7}{2}=21$ pairs of laboratories to contrast. The Bonferroni significant difference for $\alpha=.05$ is

$$BSD=t\left(1-\frac{.025}{21},39\right)\sqrt{.00421\left[\frac{1}{4}+\frac{1}{4}\right]}$$
$$=t(.99881,39).04588=3.2499(.04588)=.149\,.$$

Means that are greater than .149 apart are significantly different. Means that are less than .149 apart are not significantly different. Once again, we display the results visually. We order the sample means from smallest to largest and indicate groups of means that are not significantly different by underlining the group.

| Lab. | 4 | 7 | 5 | 6 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|
| Mean | 4.0964 | 4.2871 | 4.6906 | 4.7175 | 4.7919 | 4.8612 | 4.9031 |

Laboratories 4 and 7 are distinct from all other laboratories. Labs 5, 6, and 3 cannot be distinguished. Similarly, labs 6, 3, and 2 cannot be distinguished; however, lab 5 is significantly different from lab 2 and also lab 1. Labs 3, 2, and 1 cannot be distinguished, but lab 1 is significantly different from lab 6.

The Bonferroni simultaneous 95% confidence interval for, say, $\mu_2-\mu_5$ has endpoints

$$(4.8612-4.6906)\pm.149$$

which gives the interval (.021,.320). Transforming back to the original scale from the logarithmic scale, we are 95% confident that values for lab 2 average being between $e^{.021}=1.02$ and $e^{.320}=1.38$ times greater than the values for lab 5. Similar conclusions are drawn for the other twenty comparisons between pairs of means.

If we examine all 13 means, we have $\binom{13}{2}=78$ comparisons to make. The Bonferroni significant difference for $\alpha=.05$ is

$$BSD=t\left(1-\frac{.025}{78},39\right)\sqrt{.00421\left[\frac{1}{4}+\frac{1}{4}\right]}$$
$$=t(.9997,39).04588=3.7125(.04588)=.170.$$

Unlike the *LSD*, with more means to consider the *BSD* is larger. Now, means that are greater than .170 apart are significantly different. Means that are less than .170 apart are not significantly different. Again, we use a visual display, but with more means we list the ordered means vertically and use letters, rather than lines, to indicate groups that are not significantly different.

| Lab. | Mean | | | |
|---|---|---|---|---|
| 4 | 4.0964 | A | | |
| 9 | 4.2518 | A | B | |
| 7 | 4.2871 | | B | |
| 13 | 4.4610 | C | | |
| 5 | 4.6906 | D | | |
| 6 | 4.7175 | D | E | |
| 11 | 4.7176 | D | E | |
| 3 | 4.7919 | D | E | F |
| 2 | 4.8612 | | E | F |
| 12 | 4.8731 | | E | F |
| 10 | 4.9028 | | | F |
| 1 | 4.9031 | | | F |
| 8 | 4.9603 | | | F |

Here, for example, labs 4 and 9 are not significantly different, nor are labs 9 and 7 but 4 and 7 are different. Lab 13 is significantly different from all other labs. □

EXAMPLE 6.2.2. Consider again the six contrasts from Example 6.1.2 and Table 6.3. To perform the $\alpha = .05$ level Bonferroni adjustments on these six contrasts, once again divide the sums of squares in Table 6.3 by the *MSE* to get $F$ statistics but now compare the $F$ statistics to $F(.991\bar{6}, 1, 39) = 7.73$, where $.991\bar{6} = 1 - .05/6$. As given in Example 6.1.2, the $F$ statistics are

| Contrast | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|
| $F$ | 15.68 | 182.28 | 0.83 | 229.76 | 22.45 | 88.03 |

Comparing these to 7.73 shows that once again all of the contrasts are significantly different from zero except $C_3$, the comparison between the two labs in San Francisco. □

*Minitab commands*

Minitab can be used to obtain the $F$ and $t$ percentage points needed for Bonferroni's method. In this section we have used $t(.99881, 39)$, $t(.9997, 39)$, and $F(.991\bar{6}, 1, 39)$. To obtain these, use Minitab's inverse cumulative distribution function command.

```
MTB > invcdf .99881;
SUBC> t 39.
MTB > invcdf .9997;
SUBC> t 39.
MTB > invcdf .9916666;
SUBC> f 1 39.
```

## 6.3 Studentized range methods

Studentized range methods are generally used *only for comparing pairs of means in balanced analysis of variance problems*. They are not based on the analysis of variance $F$ test but on an alternative test of $\mu_1 = \mu_2 = \cdots = \mu_a$.

The *range* of a random sample is the difference between the largest observation and the smallest observation. For a known variance $\sigma^2$, the *range* of a random sample from a normal population has a distribution that can be worked out. This distribution depends on $\sigma^2$ and the number of observations

in the sample. It is only reasonable that the distribution depend on the number of observations because the difference between the largest and smallest observations ought to be larger in a sample of 75 observations than in a sample of 3 observations. Just by chance, we would expect the extreme observations to become more extreme in larger samples.

Knowing the distribution of the range is not very useful because the distribution depends on $\sigma^2$, which we do not know. To eliminate this problem, divide the range by an independent estimate of the standard deviation, say, $\hat{\sigma}$ having $r\hat{\sigma}^2/\sigma^2 \sim \chi^2(r)$. The distribution of this *studentized range* no longer depends on $\sigma^2$ but rather it depends on the degrees of freedom for the variance estimate. For a sample of $n$ observations and a variance estimate with $r$ degrees of freedom, the distribution of the studentized range is written as

$$Q(n,r).$$

Tables are given in Appendix B.5. The $\alpha$ percentile is denoted $Q(\alpha,n,r)$.

As discussed in Section 5.2, if $\mu_1 = \mu_2 = \cdots = \mu_a$ in a balanced ANOVA, the $\bar{y}_{i\cdot}$s form a random sample of size $a$ from a $N(\mu_1, \sigma^2/N)$ population. Looking at the range of this sample and dividing by the natural independent chi-squared estimate of the standard deviation leads to the statistic

$$Q = \frac{\max \bar{y}_{i\cdot} - \min \bar{y}_{i\cdot}}{\sqrt{MSE/N}}.$$

If the observed value of this studentized range statistic is consistent with its coming from a $Q(a,dfE)$ distribution, then the data are consistent with the null hypothesis of equal means $\mu_i$. If the $\mu_i$s are not all equal, the studentized range $Q$ tends to be larger than if the means were all equal; the difference between the largest and smallest observations will involve not only random variation but also the differences in the $\mu_i$s. Thus, for an $\alpha = .05$ level test, if the observed value of $Q$ is larger than $Q(.95,a,dfE)$, we reject the claim that the means are all equal.

The studentized range multiple comparison methods discussed in this section begin with this studentized range test.

### *6.3.1 Tukey's honest significant difference*

John Tukey's honest significant difference method is to reject the equality of a pair of means, say, $\mu_i$ and $\mu_j$ at the $\alpha = .05$ level, if

$$\frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|}{\sqrt{MSE/N}} > Q(.95,a,dfE).$$

Obviously, this test cannot be rejected for any pair of means unless the test based on the maximum and minimum sample means is also rejected. For an equivalent way of performing the test, reject equality of $\mu_i$ and $\mu_j$ if

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > Q(.95,a,dfE)\sqrt{MSE/N}.$$

With a fixed $\alpha$, the honest significant difference is

$$HSD \equiv Q(1-\alpha,a,dfE)\sqrt{MSE/N}.$$

For any pair of sample means with an absolute difference greater than the $HSD$, we conclude that the corresponding population means are significantly different. The $HSD$ is the number that an observed difference must be greater than in order for the population means to have an 'honestly' significant difference. The use of the word 'honest' is a reflection of the view that the *LSD* method allows 'too many' rejections.

Tukey's method can be extended to provide simultaneous $(1-\alpha)100\%$ confidence intervals for all differences between pairs of means. The interval for the difference $\mu_i - \mu_j$ has end points

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm HSD$$

where $HSD$ depends on $\alpha$. For $\alpha = .05$, we are 95% confident that the collection of all such intervals simultaneously contains all of the corresponding differences between pairs of population means.

EXAMPLE 6.3.1. For comparing the first 7 laboratories in Mandel's data with $\alpha = .05$, the honest significant difference is approximately

$$HSD = Q(.95, 7, 40)\sqrt{MSE/4} = 4.39\sqrt{.00421/4} = .142.$$

Here we have used $Q(.95, 7, 40)$ rather than the correct value $Q(.95, 7, 39)$ because the correct value was not available in the table used. Treatment means that are more than .142 apart are significantly different. Means that are less than .142 apart are not significantly different. Note that the $HSD$ value is similar to the corresponding $BSD$ value of .149; this frequently occurs. Once again, we display the results visually.

| Lab. | 4 | 7 | 5 | 6 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|
| Mean | 4.0964 | 4.2871 | 4.6906 | 4.7175 | 4.7919 | 4.8612 | 4.9031 |

These results are nearly the same as for the $BSD$ except that labs 6 and 2 are significantly different by the $HSD$ criterion.

The HSD simultaneous 95% confidence interval for, say, $\mu_2 - \mu_5$ has endpoints

$$(4.8612 - 4.6906) \pm .142$$

which gives the interval (.029, .313). Transforming back to the original scale from the logarithmic scale, we are 95% confident that values for lab 2 average being between $e^{.029} = 1.03$ and $e^{.313} = 1.37$ times greater than values for lab 5. Again, there are 20 more intervals to examine.

If we consider all 13 means, the honest significant difference is approximately

$$HSD = Q(.95, 13, 40)\sqrt{MSE/4} = 4.98\sqrt{.00421/4} = .162$$

Unlike the $LSD$, but like the $BSD$, with more means to consider the $HSD$ is larger. Now, means that are greater than .162 apart are significantly different. Means that are less than .162 apart are not significantly different. Again, we use a vertical display with letters to indicate groups that are not significantly different.

| Lab. | Mean | | | |
|---|---|---|---|---|
| 4 | 4.0964 | A | | |
| 9 | 4.2518 | A | B | |
| 7 | 4.2871 | | B | |
| 13 | 4.4610 | C | | |
| 5 | 4.6906 | D | | |
| 6 | 4.7175 | D | E | |
| 11 | 4.7176 | D | E | |
| 3 | 4.7919 | D | E | F |
| 2 | 4.8612 | G | E | F |
| 12 | 4.8731 | G | E | F |
| 10 | 4.9028 | G | | F |
| 1 | 4.9031 | G | | F |
| 8 | 4.9603 | G | | |

The results are similar to those for the corresponding $BSD$ of .170 except that labs 3 and 8 are now different. □

Table 6.4: *Comparison values for Newman–Keuls method as applied to Mandel's data*

| $r$ | $Q(.95,r,40)$ | $HSD$ | $r$ | $Q(.95,r,40)$ | $HSD$ |
|---|---|---|---|---|---|
| 13 | 4.98 | .162 | 7 | 4.39 | .142 |
| 12 | 4.90 | .159 | 6 | 4.23 | .137 |
| 11 | 4.82 | .156 | 5 | 4.04 | .131 |
| 10 | 4.74 | .154 | 4 | 3.79 | .123 |
| 9 | 4.64 | .151 | 3 | 3.44 | .112 |
| 8 | 4.52 | .147 | 2 | 2.86 | .093 |

### *6.3.2 Newman–Keuls multiple range method*

The Newman–Keuls multiple range method involves repeated use of the honest significant difference method with some minor adjustments. Multiple range methods are difficult to describe in general, so we simply demonstrate how they work.

EXAMPLE 6.3.2. To use the Newman–Keuls method for comparing the first 7 laboratories in Mandel's data, we need the $HSD$ value for comparing not only 7 laboratories, but also for comparing 6, 5, 4, 3, and 2 laboratories. Table 6.4 presents all the values needed, not only for comparing the first 7 labs, but also for comparing all 13 labs. Again, we approximate $Q(.95,r,39)$ with $Q(.95,r,40)$, so $HSD = Q(.95,r,40)\sqrt{MSE/4}$ where $\sqrt{MSE/4} = \sqrt{.00421/4} = .0324423$.

As before, the seven means are ordered from smallest to largest. The smallest mean, 4.0964, and the largest mean, 4.9031, are compared using the $r = 7$ value of $HSD$ from Table 6.4. These means are more than .142 apart so we go to the next stage.

At the second stage, the smallest mean, 4.0964, is compared with the second largest mean, 4.8612, and the second smallest mean, 4.2871, is compared to largest mean, 4.9031. These are groups of means that are 6 apart, so they are compared using the $HSD$ value for $r = 6$. Both differences in means are greater than .137, so we progress to the third stage.

In the third stage, the smallest mean, 4.0964, is compared to the third largest mean, 4.7919, the second smallest mean, 4.2871, is compared to the second largest mean, 4.8612, and the third smallest mean, 4.6906, is compared to the largest mean, 4.9031. These are groups of means that are 5 apart, so they are compared using the $HSD$ value for $r = 5$. All three differences in means are greater than .131, so we progress to the fourth stage and so on.

At any particular stage, means that are $r$ apart get compared using the $HSD$ value for comparing groups of $r$ means. The only exception to this rule is that *if at any given stage we conclude that certain means are not significantly different, then at later stages we never reconsider the possibility that they may contain significant differences*. The standard visual display is given below.

| Lab. | 4 | 7 | 5 | 6 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|
| Mean | 4.0964 | 4.2871 | 4.6906 | 4.7175 | 4.7919 | 4.8612 | 4.9031 |

All ordered means that were $r = 4$ apart were different. Of the means that were $r = 3$ apart, two groups were not significantly different. One of these consists of labs 5, 6, and 3, while the other group consists of labs 3, 2, and 1. For $r = 2$, we do not consider the possibility that there may be differences between labs 5, 6, and 3 or between labs 3, 2, and 1. We do consider possible differences between 4 and 7 and between 7 and 5.

If the mean for lab 6 was 4.7875, rather than its actual value 4.7175, the exception referred to in the previous paragraph would have come into play. In examining labs 5, 6, and 3, the difference between the largest and smallest of the three consecutive means 4.6906, 4.7875, and 4.7919 would still be less than the $HSD$ for $r = 3$ which is .112. Thus the three labs would still be considered not significantly different. The rule is that, since the three are not significantly different, we no longer

consider the possibility that any subset of the means could be different. If we allowed ourselves to compare the consecutive means 4.6906 and 4.7875 with $r = 2$, the appropriate $HSD$ value is .093 and the means for labs 5 and 6 would be considered significantly different. However, because the triple 4.6906, 4.7875, and 4.7919 are not significantly different, we never compare 4.6906 and 4.7875 directly.

The visual display for all 13 laboratories is given below.

| Lab. | Mean | | |
|---|---|---|---|
| 4 | 4.0964 | A | |
| 9 | 4.2518 | B | |
| 7 | 4.2871 | B | |
| 13 | 4.4610 | C | |
| 5 | 4.6906 | D | |
| 6 | 4.7175 | D | |
| 11 | 4.7176 | D | |
| 3 | 4.7919 | D | E |
| 2 | 4.8612 | F | E |
| 12 | 4.8731 | F | E |
| 10 | 4.9028 | F | E |
| 1 | 4.9031 | F | E |
| 8 | 4.9603 | F | |

□

## 6.4 Scheffé's method

Scheffé's method is valid for examining any and all contrasts simultaneously. *This method is primarily used with contrasts that were suggested by the data*. Scheffé's method should not be used for comparing pairs of means in a balanced ANOVA because the $HSD$ method has properties comparable to Scheffé's but is better for comparing pairs of means.

Scheffé's method is closely related to the analysis of variance $F$ test. Recalling the definition of the $MSTrts$, the analysis of variance $F$ test is rejected when

$$\frac{SSTrts/(a-1)}{MSE} > F(1-\alpha, a-1, dfE). \tag{6.4.1}$$

Recall from Section 5.4 that for any contrast $\sum_i \lambda_i \mu_i$,

$$SS\left(\sum_i \lambda_i \mu_i\right) \leq SSTrts. \tag{6.4.2}$$

It follows immediately that

$$\frac{SS\left(\sum_i \lambda_i \mu_i\right)/(a-1)}{MSE} \leq \frac{SSTrts/(a-1)}{MSE}.$$

Scheffé's method is to replace $SSTrts$ in (6.4.1) with $SS\left(\sum_i \lambda_i \mu_i\right)$ and to reject $H_0 : \sum_i \lambda_i \mu_i = 0$ if

$$\frac{SS\left(\sum_i \lambda_i \mu_i\right)/(a-1)}{MSE} > F(1-\alpha, a-1, dfE).$$

From (6.4.1) and (6.4.2), Scheffé's test cannot possibly be rejected unless the ANOVA test is rejected. This controls the experimentwise error rate for multiple tests. However, there always exists a contrast that contains all of the $SSTrts$, i.e., there is always a contrast that achieves equality in relation (6.4.2), so if the ANOVA test is rejected, there is always some contrast that can be rejected using Scheffé's method. This contrast may not be interesting but it exists, cf. Section 5.4.

Scheffé's method can be adapted to provide simultaneous $(1-\alpha)100\%$ confidence intervals for contrasts. These have the endpoints

$$\sum_{i=1}^{a}\lambda_i\bar{y}_{i\cdot} \pm \sqrt{(a-1)F(1-\alpha,a-1,dfE)}\,\mathrm{SE}\left(\sum_{i=1}^{a}\lambda_i\bar{y}_{i\cdot}\right).$$

EXAMPLE 6.4.1. Just for a change, we reexamine the electrical characteristic data of Chapter 5 rather than illustrating the methods with Mandel's data. The electrical characteristic data has $MSE = .56155$ with $dfE = 24$. We examined the orthogonal contrasts

$$C_1 \equiv (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 = \mu_1 - \mu_2,$$

$$C_2 \equiv (1/2)\mu_1 + (1/2)\mu_2 + (-1)\mu_3 + (0)\mu_4 = \frac{\mu_1+\mu_2}{2} - \mu_3,$$

and

$$C_3 \equiv (1/3)\mu_1 + (1/3)\mu_2 + (1/3)\mu_3 + (-1)\mu_4 = \frac{\mu_1+\mu_2+\mu_3}{3} - \mu_4.$$

The sums of squares for $C_1$, $C_2$, and $C_3$ are

$$SS(C_1) = 0.0178,\ SS(C_2) = 0.7738,\ \text{and } SS(C_3) = 10.0818.$$

These orthogonal contrasts were constructed because $C_3$ is easily interpretable and contains a large proportion of the available sum of squares for treatments; $SSTrts = 10.873$. The vast bulk of the treatment differences are due to the difference between sheet 4 and the average of the other sheets. The data suggest these contrasts, so it is not appropriate to ignore the selection process when testing whether the contrasts are 0. Scheffé's method compares

$$\frac{SS(C_3)/(a-1)}{MSE} = \frac{10.0818/3}{.56155} = 5.98$$

to an $F(3,24)$ distribution. $F(.999,3,24) = 7.55$ and $F(.99,3,24) = 4.72$, so there is very substantial evidence that sheet 4 differs from the average of the other sheets as evaluated using Scheffé's method. The similar computation for $C_1$ gives an $F$ of 0.01 and for $C_2$ an $F$ of 0.46. Both are less than 1, so neither is significant.

In our earlier consideration of these data, we also examined the orthogonal contrasts

$$C_5 \equiv (1)\mu_1 + (1)\mu_2 + (0)\mu_3 + (-2)\mu_4,$$

$C_1$, and

$$C_6 \equiv (1)\mu_1 + (1)\mu_2 + (-3)\mu_3 + (1)\mu_4.$$

The sums of squares for $C_5$, $C_1$, and $C_6$ are

$$SS(C_5) = 10.803,\ SS(C_1) = 0.018,\ \text{and } SS(C_6) = 0.052.$$

These contrasts were specifically constructed so that $S(C_5) \doteq SSTrts$. The only way to test a contrast that was constructed so as to contain all of the sums of squares treatments is to behave as if the contrast were the entire contribution from the treatments. Scheffé's method uses the test statistic

$$\frac{SS(C_5)/(a-1)}{MSE} = \frac{10.803/3}{.56155} = 6.41$$

and compares it to an $F(3,24)$ distribution. This is essentially the analysis of variance $F$ test.

The 95% Scheffé confidence interval for $C_3$ has endpoints

$$(1/3)16.4429 + (1/3)16.5143 + (1/3)16.0714 + (-1)14.9571 \pm \sqrt{3F(.95,3,24)}\sqrt{.56155\frac{(1/3)^2+(1/3)^2+(1/3)^2+(-1)^2}{7}}.$$

$F(.95,3,24) = 3.01$, so the endpoints reduce to $1.386 \pm .983$ and the interval is $(0.40, 2.37)$.

□

As with the *LSD* method, the overall $F$ test and thus Scheffé's method should be adapted to the contrasts of interest. For example, if we are considering only the first seven labs in Mandel's data, we would use an overall $F$ test with only six degrees of freedom in the numerator and Scheffé's method for examining contrasts among the seven labs uses 6 in place of $a - 1 = 13$.

## 6.5 Other methods

Other multiple comparison methods have been developed that are similar in spirit to the studentized range methods. Just as studentized range methods were developed for comparing pairs of means in balanced analysis of variance problems, these other methods were developed for examining other sets of contrasts in balanced ANOVA. Again, the methods are not based on the analysis of variance $F$ test but on alternative tests of $\mu_1 = \mu_2 = \cdots = \mu_a$. We will briefly discuss two of these methods: Ott's analysis of means method (AOM) and Dunnett's many-one $t$ statistics. In addition, we mention another studentized range method proposed by Duncan that can also be modified for application with AOM and Dunnett's method.

### *6.5.1 Ott's analysis of means method*

Ott (1967) introduced a graphical method called analysis of means for comparing each mean to the average of all the means. It is most often used in quality control work and the graphical method is closely related to control charts for means, cf. Shewhart (1931). Ott's work was founded upon earlier work that is referenced in his article. Nelson (1993) contains a brief, clear introduction, extensions, some tables, and references to other tables.

Balanced one-way ANOVA methods are founded on the fact that if $\mu_1 = \mu_2 = \cdots = \mu_a$, the $\bar{y}_{i\cdot}$s form a random sample of size $a$ from a $N(\mu_1, \sigma^2/N)$ population. We have already seen that the distribution of the studentized range is known when $\mu_1 = \mu_2 = \cdots = \mu_a$, so comparing the observed studentized range to the known distribution provides a test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$. This test was then modified to provide multiple comparison methods.

The AOM method is based on knowing the distribution of

$$\max_i \frac{|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|}{\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}$$

when the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ is true. The distribution depends on the number of treatments $a$ and the $dfE$. The $1-\alpha$ percentile of this distribution is often denoted $h(\alpha, a, dfE)$. An $\alpha$ level test of $H_0$ is rejected if

$$\max_i \frac{|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|}{\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})} > h(\alpha, a, dfE).$$

When this test is rejected, we can do multiple comparisons to identify which individual means are different from the overall average. A particular mean $\mu_i$ is considered different if

$$\frac{|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|}{\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})} > h(\alpha, a, dfE).$$

Clearly, if the overall test is not rejected, none of the individual means will be considered different.

The test given above is easily seen to be equivalent to the following: $\mu_i$ is considered different from the average of all the $\mu$s (or equivalently from the average of the other $\mu$s) if $\bar{y}_{i\cdot}$ is *not* between the values

$$\bar{y}_{\cdot\cdot} \pm h(\alpha, a, dfE)\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}).$$

This leads to a simple graphical procedure. Plot the pairs $(i, \bar{y}_{i\cdot})$. On this plot add horizontal lines at $\bar{y}_{\cdot\cdot} \pm h(\alpha, a, dfE)\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$. Any $\bar{y}_{i\cdot}$ that lies outside the horizontal lines indicates a $\mu_i$ that is different from the others. While it is not crucial, traditionally the graphical display also includes a center line at $\bar{y}_{\cdot\cdot}$. This graphical display is very similar to a control chart for means. The AOM is focused on testing whether one particular mean is different from the rest of the means. This may be particularly appropriate for quality control problems.

The primary detail that we have not yet covered is the exact formula for $\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$. To compute this, first note that

$$\begin{aligned} \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} &= \bar{y}_{i\cdot} - \frac{\bar{y}_{1\cdot} + \cdots + \bar{y}_{a\cdot}}{a} \\ &= \frac{a-1}{a}\bar{y}_{i\cdot} - \frac{1}{a}\sum_{k \neq i} \bar{y}_{k\cdot}. \end{aligned}$$

It follows that the standard error is

$$\begin{aligned} \mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) &= \sqrt{MSE\left[\left(\frac{a-1}{a}\right)^2 + (a-1)\left(\frac{1}{a}\right)^2\right] \Big/ N} \\ &= \sqrt{MSE\left[\frac{a-1}{aN}\right]}. \end{aligned}$$

In fact, this argument explicates exactly what AOM is examining. AOM is simultaneously testing whether the contrasts $[(a-1)/a]\mu_i - (1/a)\sum_{k \neq i}\mu_k$, $i = 1, \ldots, a$ are all equal to 0. Equivalently, we can multiply the contrasts by $a$ and think of the contrasts as being $(a-1)\mu_i - \sum_{k \neq i}\mu_k$, $i = 1, \ldots, a$. It is not difficult to see that these contrasts all equal 0 if and only if the $\mu_i$s are all equal.

A modification similar to the Newman–Keuls procedure can be used with AOM. The modification involves changing the value of $a$ in $h(\alpha, a, dfE)$. Order the values of $|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|$. When examining the largest value of $|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|$ compare it to $h(\alpha, a, dfE)\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$, when examining the second largest value of $|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|$, compare it to $h(\alpha, a-1, dfE)\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$, etc. To maintain consistency, if, say, the second largest value of $|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|$ is not greater than $h(\alpha, a-1, dfE)\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$, all of the smaller values of $|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|$ should also be considered nonsignificant. Note that $h(\alpha, 1, dfE) = 0$ for any $\alpha$, so that if all the other means are declared different, the mean with the smallest deviation from $\bar{y}_{\cdot\cdot}$ will also be declared different, assuming that the deviation is positive.

### *6.5.2 Dunnett's many-one t statistic method*

Dunnett's method is designed for situations in which there is a standard treatment (or placebo or control) and where interest lies in comparing each of the other treatments to the standard. Miller (1981) contains a thorough discussion along with references to the early work by Dunnett and Paulson.

Suppose that the standard treatment is $i = 1$. Dunnett's method is based on knowing the distribution of

$$\max_i \frac{|\bar{y}_{i\cdot} - \bar{y}_{1\cdot}|}{\mathrm{SE}(\bar{y}_{i\cdot} - \bar{y}_{1\cdot})}$$

when the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ is true. If we denote the $1 - \alpha$ percentile of the

distribution as $d(1-\alpha,a,dfE)$, an $\alpha$ level test of $H_0$ is rejected if

$$\max_i \frac{|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|}{\mathrm{SE}(\bar{y}_{i\cdot}-\bar{y}_{1\cdot})} > d(1-\alpha,a,dfE).$$

When the overall test is rejected, we can do multiple comparisons to identify which $\mu_i$s are different from $\mu_1$. A particular mean $\mu_i$ is considered different if

$$\frac{|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|}{\mathrm{SE}(\bar{y}_{i\cdot}-\bar{y}_{1\cdot})} > d(1-\alpha,a,dfE).$$

Clearly, if the overall test is not rejected, none of the individual means will be considered different. It is also clear that $\mu_1=\mu_2=\cdots=\mu_a$ if and only if $\mu_i-\mu_1=0$ for all $i>1$. The standard error is that for comparing two means, so

$$\mathrm{SE}(\bar{y}_{i\cdot}-\bar{y}_{1\cdot}) = \sqrt{MSE\left[\frac{1}{N}+\frac{1}{N}\right]}.$$

Simultaneous $(1-\alpha)100\%$ confidence intervals for the $\mu_i-\mu_1$s have endpoints

$$\bar{y}_{i\cdot}-\bar{y}_{1\cdot} \pm d(1-\alpha,a,dfE)\mathrm{SE}(\bar{y}_{i\cdot}-\bar{y}_{1\cdot}).$$

A modification similar to Newman–Keuls can be used with Dunnett's method. This modification orders the values of $|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|$ and adjusts the value of the $a$ parameter in $d(1-\alpha,a,dfE)$. When examining the largest value of $|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|$, use $d(1-\alpha,a,dfE)$, when examining the second largest value of $|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|$, use $d(1-\alpha,a-1,dfE)$, etc. Of course to maintain consistency, when it is determined that, say, the second largest value of $|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|$ is not greater than $d(1-\alpha,a-1,dfE)\mathrm{SE}(\bar{y}_{i\cdot}-\bar{y}_{1\cdot})$, all of the smaller values of $|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|$ should be considered as nonsignificant also.

### *6.5.3 Duncan's multiple range method*

Duncan has developed a multiple range procedure similar to that of Newman–Keuls. Newman–Keuls uses a series of tabled values $Q(1-\alpha,a,dfE)$, $Q(1-\alpha,a-1,dfE)$, ..., $Q(1-\alpha,2,dfE)$. Duncan's method simply changes the tabled values. Duncan uses $Q\big([1-\alpha]^{a-1},a,dfE\big)$, $Q\big([1-\alpha]^{a-2},a-1,dfE\big)$, ..., $Q(1-\alpha,2,dfE)$. See Miller (1981) for a discussion of the rationale behind these choices.

Using Duncan's value $Q\big([1-\alpha]^{a-1},a,dfE\big)$ to compare the largest and smallest means does not control the experimentwise error rate at $\alpha$. (It controls it at $1-[1-\alpha]^{a-1}$.) As a result, Duncan suggests performing the analysis of variance $F$ test first and proceeding only if the $F$ test indicates that there are differences among the means at level $\alpha$. Duncan's method is more likely to conclude that a pair of means is different than the Newman–Keuls method and less likely to establish a difference than the LSD method. Just as the Newman–Keuls approach can be used to modify the AOM and Dunnett's method, Duncan's idea can also be applied to the AOM and Dunnett's method.

## 6.6 Summary of multiple comparison procedures

In this section we review and compare the uses of the various multiple comparison procedures.

The most general procedures are the least significant difference, the Bonferroni, and the Scheffé methods. These can be used for arbitrary sets of preplanned contrasts. They are listed in order from least conservative (most likely to reject an individual null hypothesis) to most conservative (least likely to reject). Scheffé's method can also be used for examining contrasts suggested by the data. Bonferroni's method has the advantage that it can easily be applied to almost any multiple testing problem.

Table 6.5: *Rubber stress at five laboratories*

| Lab. | Sample size | Sample mean | Sample variance |
|---|---|---|---|
| 1 | 4 | 57.00 | 32.00 |
| 2 | 4 | 67.50 | 46.33 |
| 3 | 4 | 40.25 | 14.25 |
| 4 | 4 | 56.50 | 5.66 |
| 5 | 4 | 52.50 | 6.33 |

To compare all of the treatment groups in a balanced analysis of variance, we can use the least significant difference, the Duncan, the Newman–Keuls, the Bonferroni, and the Tukey methods. Again, these are (roughly) listed in the order from least conservative to most conservative. In some cases, for example when comparing Bonferroni and Tukey, an exact statement of which is more conservative is not possible.

To decide on a method, you need to decide on how conservative you want to be. If it is very important not to claim differences when there are none, you should be very conservative. If it is most important to identify differences that *may* exist, then you should choose less conservative methods.

Finally, we discussed two specialized methods for balanced ANOVA. The analysis of means provides for testing whether each group differs from all the other groups and Dunnett's method allows multiple testing of each group against a fixed standard group.

Many of the methods have corresponding methods for constructing multiple confidence intervals. Various computer programs execute these procedures. For example, newer versions of Minitab's 'oneway' command compute these for Fisher's LSD method, Tukey's method, and Dunnett's method.

## 6.7 Exercises

EXERCISE 6.7.1. Exercise 5.7.1 involved measurements from different laboratories on the stress at 600% elongation for a certain type of rubber. The summary statistics are repeated in Table 6.5. Ignoring any reservations you may have about the appropriateness of the analysis of variance model for these data, compare all pairs of laboratories using $\alpha = .10$ for the LSD, Bonferroni, Tukey, and Newman–Keuls methods. Give joint 95% confidence intervals using Tukey's method for all differences between pairs of labs.

EXERCISE 6.7.2. Use Scheffé's method with $\alpha = .01$ to test whether the contrast in Exercise 5.7.2d is zero.

EXERCISE 6.7.3. Use Bonferroni's method with an $\alpha$ near .01 to give simultaneous confidence intervals for the mean weight in each height group for Exercise 5.7.3.

EXERCISE 6.7.4. Use the LSD, Bonferroni, and Scheffé's methods to test whether the four orthogonal contrasts in Exercise 5.7.4 are zero. Use $\alpha = .05$.

EXERCISE 6.7.5. Exercise 5.7.5 contained data on stress measurements for four different types of rubber. Four observations were taken on each type of rubber; the means are repeated below

| Material | A | B | C | D |
|---|---|---|---|---|
| Mean | 26.4425 | 26.0225 | 23.5325 | 29.9600 |

and the sample variance of the 16 observations is 14.730793. Test for differences between all pairs of materials using $\alpha = .05$ for the LSD, Bonferroni, Tukey, and Newman–Keuls methods. Give 95% confidence intervals for the differences between all pairs of materials using the BSD method.

EXERCISE 6.7.6. In Exercise 5.7.6 on the stress of four types of rubber an outlier was noted in material B. Redo the multiple comparisons of the previous problem eliminating the outlier and using only the methods that are still applicable.

EXERCISE 6.7.7. In Exercise 5.7.7 on the peel-strength of different adhesive systems, parts (b) and (c) amount to doing LSD multiple comparisons for all pairs of systems. Compare the LSD results with the results obtained using the Tukey and Newman–Keuls methods with $\alpha = .01$.

EXERCISE 6.7.8. For the weight gain data of Exercise 5.7.8, use the LSD, Bonferroni, and Scheffé methods to test whether the following contrasts are zero: 1) the contrast that compares the two drugs and 2) the contrast that compares the control with the average of the two drugs. Pick an $\alpha$ level but clearly state the level chosen.

EXERCISE 6.7.9. For the Cushing's syndrome data of Exercise 5.7.9, use all appropriate methods to compare all pairwise differences among the three treatments. Pick an $\alpha$ level but clearly state the level chosen.

EXERCISE 6.7.10. Use Scheffé's method with $\alpha = .05$ and the data of Exercise 5.7.10 to test the significance of the contrast

| Age | 0.5 | 1.0 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 |
|---|---|---|---|---|---|---|---|
| Coeff. | −5 | −5 | 2 | 2 | 2 | 2 | 2 |

Chapter 7

# Simple linear and polynomial regression

This chapter examines data that come as pairs of numbers, say $(x, y)$, and the problem of fitting a line to them. More generally, it examines the problem of predicting one variable ($y$) from values of another variable ($x$). Consider for the moment the popular wisdom that people who read a lot tend to have large vocabularies and poor eyes. Thus reading causes both conditions: large vocabularies and poor eyes. If this is true, it may be possible to predict the size of someone's vocabulary from the condition of their eyes. Of course this does not mean that having poor eyes causes large vocabularies. Quite the contrary, if anything poor eyes probably keep people from reading and thus cause small vocabularies. Regression analysis is concerned with predictive ability, not with causation.

Section 7.1 of this chapter introduces an example along with many of the basic ideas and methods of simple linear regression. The next five sections go into the details of simple linear regression. Sections 7.7 and 7.8 deal with an idea closely related to simple linear regression: the correlation between two variables. Section 7.9 deals with methods for checking the assumptions made in simple linear regression. If the assumptions are violated, we need alternative methods of analysis. Section 7.10 presents methods for transforming the original data so that the assumptions become reasonable on the transformed data. *Sections 7.9 and 7.10 apply quite generally to analysis of variance and regression models*. They are not restricted to simple linear regression. Section 7.11 treats an alternative to transformations as a method for dealing with nonlinearity in the relationship between $y$ and $x$, namely fitting polynomials (parabolas, etc.) to the data. Section 7.12 explores the relationship between one-way analysis of variance and fitting polynomials.

## 7.1 An example

Data from *The Coleman Report* were reproduced in Mosteller and Tukey (1977). The data were collected from schools in the New England and Mid-Atlantic states of the USA. In this chapter we consider only two variables: $y$ – the mean verbal test score for sixth graders and $x$ – a composite measure of socioeconomic status. The data are presented in Table 7.1.

Figure 7.1 contains a scatter plot of the data. Note that there is a rough linear relationship. The

Table 7.1: *Coleman Report data*

| School | $y$ | $x$ | School | $y$ | $x$ |
|---|---|---|---|---|---|
| 1 | 37.01 | 7.20 | 11 | 23.30 | −12.86 |
| 2 | 26.51 | −11.71 | 12 | 35.20 | 0.92 |
| 3 | 36.51 | 12.32 | 13 | 34.90 | 4.77 |
| 4 | 40.70 | 14.28 | 14 | 33.10 | −0.96 |
| 5 | 37.10 | 6.31 | 15 | 22.70 | −16.04 |
| 6 | 33.90 | 6.16 | 16 | 39.70 | 10.62 |
| 7 | 41.80 | 12.70 | 17 | 31.80 | 2.66 |
| 8 | 33.40 | −0.17 | 18 | 31.70 | −10.99 |
| 9 | 41.01 | 9.85 | 19 | 43.10 | 15.03 |
| 10 | 37.20 | −0.05 | 20 | 41.01 | 12.77 |

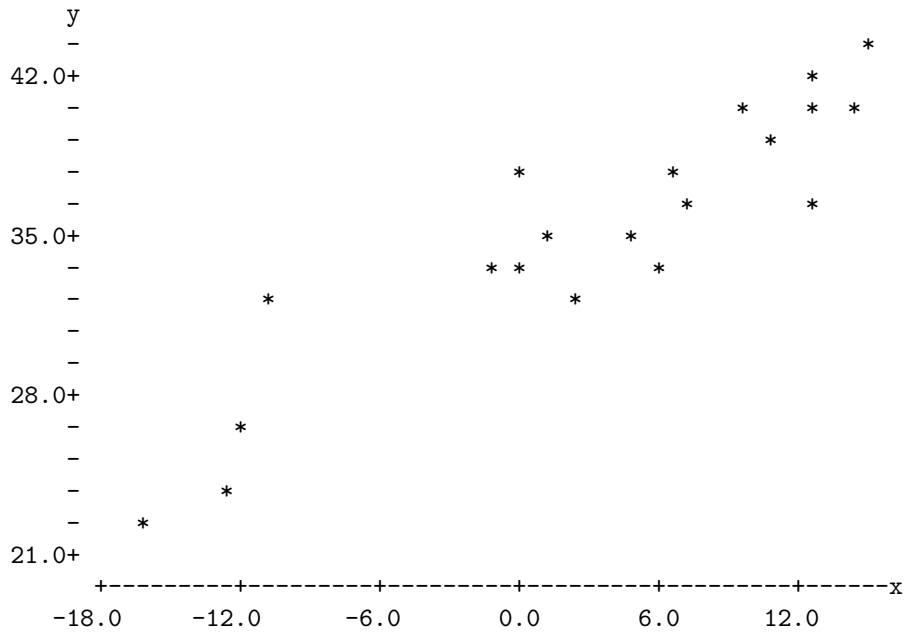



Figure 7.1: *Plot of y versus x.*

higher the composite socioeconomic status variable, the higher the mean verbal test score. However, there is a considerable amount of error in the relationship. By no means do the points lie exactly on a straight line.

We assume a basic linear relationship between the $y$s and $x$s, something like $y = \beta_0 + \beta_1 x$. Here $\beta_1$ is the slope of the line and $\beta_0$ is the intercept. Unfortunately, the observed $y$ values do not fit exactly on a line so $y = \beta_0 + \beta_1 x$ is only an approximation. We need to modify this equation to allow for the variability of the observations about the line. We do this by building a random error term into the linear relationship. Write the relationship as $y = \beta_0 + \beta_1 x + \varepsilon$, where $\varepsilon$ indicates the random error. In this model for the behavior of the data, $\varepsilon$ accounts for the deviations between the $y$ values we actually observe and the line $\beta_0 + \beta_1 x$ where we expect to observe any $y$ value that corresponds to $x$. As we are interested in predicting $y$ from known $x$ values, *we treat x as a known (nonrandom) variable*.

We assume that the relationship $y = \beta_0 + \beta_1 x + \varepsilon$ applies to all of our observations. For the current data, that means we assume this relationship holds for all of the 20 pairs of values in Table 7.1. This assumption is stated as *the simple linear regression model* for these data,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{7.1.1}$$

$i = 1, \ldots, 20$. For this model to be useful, we need to make some assumptions about the errors, the $\varepsilon_i$s. The standard assumption is that the

$$\varepsilon_i \text{s are independent } N(0, \sigma^2).$$

Given data for which these assumptions are reasonable, we can estimate the unknown parameters. Although we assume a linear relationship between the $y$s and $x$s, the model does not assume that we know the slope $\beta_1$ or the intercept $\beta_0$. Together these unknown parameters would tell us the exact nature of the linear relationship but both need to be estimated. We use the notation $\hat{\beta}_1$ and $\hat{\beta}_0$ to denote estimates of $\beta_1$ and $\beta_0$, respectively. To perform statistical inferences we also need to estimate the variance of the errors, $\sigma^2$. Note that $\sigma^2$ is also the variance of the $y$ observations because none of $\beta_0$, $\beta_1$, and $x$ are random.

Simple linear regression involves many assumptions. It assumes that the relationship between $y$ and $x$ is linear, it assumes that the errors are normally distributed, it assumes that the errors all have the same variance, it assumes that the errors are all independent, and it assumes that the errors all have mean 0. This last assumption is redundant. It turns out that the errors all have mean 0 if and only if the relationship between $y$ and $x$ is linear. As far as possible, we will want to verify (validate) that these assumptions are reasonable before we put much faith in the estimates and statistical inferences that can be obtained from simple linear regression. Section 7.9 deals with checking these assumptions.

Before getting into a detailed discussion of simple linear regression, we illustrate some highlights using the *Coleman Report* data. We need to fit model (7.1.1) to the data. A computer program typically yields parameter estimates, standard errors for the estimates, $t$ ratios for testing whether the parameters are zero, $P$ values for the tests, and an analysis of variance table. These results are often displayed in a fashion similar to that illustrated below.

| Predictor | $\hat{\beta}_k$ | $\mathrm{SE}(\hat{\beta}_k)$ | $t$ | $P$ |
|---|---|---|---|---|
| Constant | 33.3228 | 0.5280 | 63.11 | 0.000 |
| $x$ | 0.56033 | 0.05337 | 10.50 | 0.000 |

Analysis of Variance

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 1 | 552.68 | 552.68 | 110.23 | 0.000 |
| Error | 18 | 90.25 | 5.01 | | |
| Total | 19 | 642.92 | | | |

Much can be learned from these two tables of statistics. The estimated regression equation is

$$y = 33.3 + 0.560x.$$

This equation allows us to predict a value for $y$ when the value of $x$ is given. In particular, for these data an increase of one unit in socioeconomic status tends to increase mean verbal test scores by about .56 units. This is not to say that some program to increase socioeconomic statuses by one unit will increase mean verbal test scores by about .56 unit. The .56 *describes* the *current* data, *it does not imply a causal relationship*. If we want to predict the mean verbal test score for a school that is very similar to the ones in this study, this equation should give good predictions. If we want to predict the mean verbal test score for a school that is very different from the ones in this study, this equation is likely to give poor predictions. In fact, if we collect new data from schools with very different socioeconomic statuses, the data are not similar to these, so this fitted model would be highly questionable if applied to the new situation. Nevertheless, a simple linear regression model with a different intercept and slope might fit the new data well. Similarly, data collected after a successful program to raise socioeconomic statuses are unlikely to be similar to the data collected before such a program. The relationship between socioeconomic status and mean verbal test scores may be changed by such a program. In particular, the things causing both socioeconomic status and mean verbal test score may be changed in unknown ways by such a program. These are crucial points and bear repeating. *The regression equation describes an observed relationship between mean verbal test scores and socioeconomic status. It can be used to predict mean verbal test scores from socioeconomic status in similar situations. It does not imply that changing the socioeconomic status a fixed amount will cause the mean verbal test scores to change by a proportional amount*.

In simple linear regression, the reference distribution for statistical inferences is almost invariably $t(dfE)$ where $dfE$ is the degrees of freedom for error from the analysis of variance table. For these data, $dfE = 18$. We now consider some illustrations of statistical inferences.

From our standard theory of Chapter 3, the 95% confidence interval for $\beta_1$ has endpoints

$$\hat{\beta}_1 \pm t(.975, dfE)\mathrm{SE}\left(\hat{\beta}_1\right).$$

From a $t$ table, $t(.975,18) = 2.101$, so, using the tabled statistics, the endpoints are

$$.56033 \pm 2.101(.05337).$$

The confidence interval is $(.448,.672)$, so we are 95% confident that the slope $\beta_1$ is between .448 and .672.

The $t$ statistics for testing $H_0 : \beta_k = 0$ versus $H_A : \beta_k \neq 0$ are reported in the first table. For example, the test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ has

$$t_{obs} = \frac{0.56033}{.05337} = 10.50.$$

The significance level of the test is the $P$ value,

$$P = \Pr[|t| > 10.50] = .000.$$

The value .000 indicates a large amount of evidence that $\beta_1 \neq 0$. Note that if $\beta_1 = 0$, the linear relationship becomes $y = \beta_0 + \varepsilon$, so there is no relationship between $y$ and $x$, i.e., $y$ does not depend on $x$. The small $P$ value indicates that the slope is not zero and thus the variable $x$ helps to explain the variable $y$.

The primary value of the analysis of variance table is that it gives the degrees of freedom, the sum of squares, and the mean square for error. The mean squared error is the estimate of $\sigma^2$ and the sum of squares error and degrees of freedom for error are vital for comparing different regression models that we may choose to consider. Note that the sums of squares for regression and error add up to the sum of squares total and that the degrees of freedom for regression and error also add up to the degrees of freedom total.

The analysis of variance table gives an alternative but equivalent test for whether the $x$ variable helps to explain $y$. The alternative test of

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0$$

is based on

$$F = \frac{MSReg}{MSE} = \frac{552.68}{5.01} = 110.23.$$

Note that the value of this statistic is $110.23 = (10.50)^2$; the $F$ statistic is just the square of the corresponding $t$ statistic for testing $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The $F$ and $t$ tests are equivalent. In particular the $P$ values are identical. In this case, both are infinitesimal, zero to three decimal places. Our conclusion that $\beta_1 \neq 0$ means that the $x$ variable helps to explain the variation in the $y$ variable. In other words, it is possible to predict the verbal test scores for a school's sixth grade class from the socioeconomic measure. Of course, the fact that some predictive ability exists does not mean that the predictive ability is sufficient to be useful.

The *coefficient of determination*, $R^2$, measures the percentage of the total variability in $y$ that is explained by the $x$ variable. If this number is large, it suggests a substantial predictive ability. In our example

$$R^2 \equiv \frac{SSReg}{SSTot} = \frac{552.68}{642.92} = 86.0\%,$$

so 86.0% of the total variability is explained by the regression model. This is a large percentage, so it appears that the $x$ variable has substantial predictive power. However, *a large $R^2$ does not imply that the model is good in absolute terms*. It may be possible to show that this model does not fit the data adequately. In other words, while this model is explaining much of the variability, we may be able to establish that it is not explaining as much of the variability as it ought. (Example 7.9.2 involves a model with a high $R^2$ that is demonstrably inadequate.) Conversely, a model with a low $R^2$ value may be the perfect model but the data may simply have a great deal of variability. For example, if you have temperature measurements obtained by having someone walk outdoors and

guess the Celsius temperature and then use the true Fahrenheit temperatures as a predictor, the exact linear relationship between Celsius and Fahrenheit temperatures may make a line the ideal model. Nonetheless, the obvious inaccuracy involved in people guessing Celsius temperatures may cause a low $R^2$. Moreover, even a high $R^2$ of 86% may provide inadequate predictions for the purposes of the study, while in other situations an $R^2$ of, say, 14% may be perfectly adequate. It depends on the purpose of the study. Finally, it must be recognized that a large $R^2$ may be an unrepeatable artifact of a particular data set. *The coefficient of determination is a useful tool but it must be used with care. In particular, it is a much better measure of the predictive ability of a model than of the correctness of a model.*

Consider the problem of estimating the value of the line at $x = -16.04$. This value of $x$ is the minimum observed value for socioeconomic status, so it is somewhat dissimilar to the other $x$ values in the data. Its dissimilarity causes there to be substantial variability in estimating the regression line (mean value of $y$) at this point. The point on the line is $\beta_0 + \beta_1(-16.04)$ and the estimator is

$$\hat{\beta}_0 + \hat{\beta}_1 x = 33.32 + .560(-16.04) = 24.34.$$

For constructing 95% $t$ intervals, the percentile needed is $t(.975, 18) = 2.101$. The standard error for the estimate of the point on the line is usually available from computer programs; in this example it is 1.140. The 95% confidence interval for the point on the line $\beta_0 + \beta_1(-16.04)$ has endpoints

$$24.34 \pm 2.101(1.140)$$

which gives the interval $(21.9, 26.7)$. We are 95% confident that the population mean of the school-wise mean verbal test scores for New England and Mid-Atlantic sixth graders with a school socioeconomic measure of $-16.04$ is between 21.9 and 26.7.

The prediction $\hat{y}$ for a new observation with $x = -16.04$ is simply the estimated point on the line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(-16.04) = 24.34.$$

Prediction of a new observation is subject to more error than estimation of a point on the line. A new observation has the same variance as all other observations, so the prediction interval must account for this variance as well as for the variance of estimating the point on the line. The standard error for the prediction interval is computed as

$$\text{SE}(\textit{Prediction}) = \sqrt{MSE + SE(\textit{Line})^2}. \tag{7.1.2}$$

In this example,

$$\text{SE}(\textit{Prediction}) = \sqrt{5.01 + (1.140)^2} = 2.512.$$

The prediction interval endpoints are

$$24.34 \pm 2.101(2.512).$$

and the 95% prediction interval is $(19.1, 29.6)$. We are 95% confident that sixth graders's mean verbal test scores would be between 19.1 and 29.6 for a *different* New England or Mid-Atlantic school with a socioeconomic measure of $-16.04$. Note that the prediction interval is considerably wider than the corresponding confidence interval. Note also that this is just another special case of the prediction theory in Section 3.5. As such, these results are analogous to those obtained for the one sample, two sample, and one-way ANOVA data structures.

*Minitab commands*

The Minitab commands given below generate the table of estimates and the analysis of variance table. Column c1 contains the test scores $y$ and column c2 contains the composite socioeconomic

statuses $x$. The primary command is to regress c1 on 1 predictor variable, c2. This same command allows for more predictor variables and we will use that capability in this chapter as well as in the chapters on multiple regression. In our example, the subcommand 'predict $-16.04$' was used; this subcommand gives the estimate of the line (prediction) when $x = -16.04$, the standard error for the estimate of the line, the 95% confidence interval for the value of the line at $x = -16.04$, and the 95% prediction interval when $x = -16.04$.

```
MTB > name c1 'test' c2 'socio'
MTB > regress c1 on 1 c2;
SUBC> predict -16.04.
```

### 7.2 The simple linear regression model

In general, simple linear regression seeks to fit a line to pairs of numbers $(x,y)$ that are subject to error. These pairs of numbers may arise when there is a perfect linear relationship between $x$ and a variable $y_*$ but where $y_*$ cannot be measured without error. Our actual observations $y$ are then the sum of $y_*$ and the measurement error. Alternatively, we may sample a population of objects and take two measurements on each object. In this case, both elements of the pair $(x,y)$ are random. In simple linear regression we think of using the $x$ measurement to predict the $y$ measurement. While $x$ is actually random in this scenario, we use it as if it were fixed because we cannot predict $y$ until we have actually observed the $x$ value. We want to use the particular observed value of $x$ to predict $y$, so for our purposes $x$ is a fixed number. In any case, *the xs are always treated as fixed numbers in simple linear regression*.

The model for simple linear regression is a line with the addition of errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where $y$ is the variable of primary interest and $x$ is the predictor variable. Both the $y_i$s and the $x_i$s are observable, the $y_i$s are assumed to be random and the $x_i$s are assumed to be known fixed constants. The unknown constants (regression parameters) $\beta_0$ and $\beta_1$ are the intercept and the slope of the line, respectively. The $\varepsilon_i$s are unobservable errors that are assumed to be independent of each other with mean zero and the same variance, i.e.,

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

Typically the errors are also assumed to have normal distributions, i.e.,

$$\varepsilon_i \text{s independent } N(0, \sigma^2).$$

Sometimes the assumption of independence is replaced by the assumption that $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

Note that since $\beta_0$, $\beta_1$, and the $x_i$s are all assumed to be fixed constants,

$$\mathrm{E}(y_i) = \mathrm{E}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + \mathrm{E}(\varepsilon_i) = \beta_0 + \beta_1 x_i,$$

$$\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2,$$

and if the $\varepsilon_i$s are independent, the $y_i$s are independent.

### 7.3 Estimation of parameters

The unknown parameters in the simple linear regression model are the slope, $\beta_1$, the intercept, $\beta_0$, and the variance, $\sigma^2$. All of the estimates $\hat{\beta}_1$, $\hat{\beta}_0$, and $MSE$, can be computed from just six summary statistics

$$n, \quad \bar{x}_\cdot, \quad s_x^2, \quad \bar{y}_\cdot, \quad s_y^2, \quad \sum_{i=1}^{n} x_i y_i,$$

i.e., the sample size, the sample mean and variance of the $x_i$s, the sample mean and variance of the $y_i$s, and $\sum_{i=1}^n x_i y_i$. The only one of these that is any real work to obtain on a decent hand calculator is $\sum_{i=1}^n x_i y_i$. The standard estimates of the parameters are, respectively,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_\cdot) y_i}{\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2}$$

$$\hat{\beta}_0 = \bar{y}_\cdot - \hat{\beta}_1 \bar{x}_\cdot$$

and the *mean squared error*

$$\begin{aligned} MSE &= \frac{\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2} \\ &= \frac{1}{n-2}\left[\sum_{i=1}^n (y_i - \bar{y}_\cdot)^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2\right] \\ &= \frac{1}{n-2}\left[(n-1)s_y^2 - \hat{\beta}_1^2 (n-1) s_x^2\right]. \end{aligned}$$

The slope estimate $\hat{\beta}_1$ given above is the form that is most convenient for deriving its statistical properties. In this form it is just a linear combination of the $y_i$s. However, $\hat{\beta}_1$ is commonly written in a variety of ways to simplify various computations and, unfortunately for students, they are expected to recognize all of them. Observing that $0 = \sum_{i=1}^n (x_i - \bar{x}_\cdot)$ so that $0 = \sum_{i=1}^n (x_i - \bar{x}_\cdot)\bar{y}_\cdot$, we can also write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_\cdot)(y_i - \bar{y}_\cdot)}{\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} = \frac{s_{xy}}{s_x^2} = \frac{(\sum_{i=1}^n x_i y_i) - n\bar{x}_\cdot \bar{y}_\cdot}{(n-1)s_x^2}. \tag{7.3.1}$$

Here

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_\cdot)(y_i - \bar{y}_\cdot)$$

is the sample covariance between $x$ and $y$. The last equality on the right of equation (7.3.1) gives a form suitable for computing $\hat{\beta}_1$ from the summary statistics.

EXAMPLE 7.3.1. For the *Coleman Report* data,

$$n = 20, \quad \bar{x}_\cdot = 3.1405, \quad s_x^2 = 92.64798395,$$

$$\bar{y}_\cdot = 35.0825, \quad s_y^2 = 33.838125, \quad \sum_{i=1}^n x_i y_i = 3189.8793.$$

The estimates are

$$\hat{\beta}_1 = \frac{3189.8793 - 20(3.1405)(35.0825)}{(20-1)92.64798395} = .560325468,$$

$$\hat{\beta}_0 = 35.0825 - .560325468(3.1405) = 33.32279787$$

and

$$\begin{aligned} &MSE \\ &= \frac{1}{20-2}\left[(20-1)33.838125 - (.560325468)^2(20-1)92.64798395\right] \\ &= \frac{1}{18}\left[642.924375 - 552.6756109\right] \\ &= \frac{90.2487641}{18} = 5.01382. \end{aligned} \tag{7.3.2}$$

Up to round off error, these are the same results as tabled in Section 7.1. □

It is not clear that these estimates of $\beta_0$, $\beta_1$, and $\sigma^2$ are even reasonable. The estimate of the slope $\beta_1$ seems particularly unintuitive. However, from Proposition 7.3.2 below, the estimates are unbiased, so they are at least estimating what we claim that they estimate.

**Proposition 7.3.2.** $\mathrm{E}\left(\hat{\beta}_1\right) = \beta_1$, $\mathrm{E}\left(\hat{\beta}_0\right) = \beta_0$, and $\mathrm{E}(MSE) = \sigma^2$.

Proofs of the unbiasedness of the slope and intercept are given in the appendix to this chapter.

The parameter estimates are unbiased but that alone does not ensure that they are good estimates. These estimates are the best estimates available in several senses. We briefly mention these optimality properties but for a detailed discussion see Christensen (1987, chapter II). Assuming that the errors have independent normal distributions, all of the estimates have the smallest variance of any unbiased estimates. The regression parameters are also maximum likelihood estimates. Maximum likelihood estimates are those values of the parameters that are most likely to generate the data that were actually observed. Without assuming that the errors are normally distributed, the regression parameters have the smallest variance of any unbiased estimates that are linear functions of the $y$ observations. (Linear functions allow multiplying the $y_i$s by constants and adding terms together. Remember, the $x_i$s are constants, as are any functions of the $x_i$s.) Note that with this weaker assumption, i.e., giving up normality, we get a weaker result, minimum variance among only linear unbiased estimates instead of all unbiased estimates. The regression parameter estimates are also least squares estimates. Least squares estimates are choices of $\beta_0$ and $\beta_1$ that minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

*Under the standard assumptions, least squares estimates of the regression parameters are best (minimum variance) linear unbiased estimates (BLUEs), and for normally distributed data they are minimum variance unbiased estimates and maximum likelihood estimates.*

To draw statistical inferences about the regression parameters, we need standard errors for the estimates. To find the standard errors we need to know the variance of each estimate.

**Proposition 7.3.3.**

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x}_\cdot)^2} = \frac{\sigma^2}{(n-1)s_x^2}$$

and

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_\cdot^2}{\sum_{i=1}^{n} (x_i - \bar{x}_\cdot)^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_\cdot^2}{(n-1)s_x^2}\right].$$

The proof of this proposition is given in the appendix at the end of the chapter. Note that, except for the unknown parameter $\sigma^2$, the variances can be computed using the same six numbers we used to compute $\hat{\beta}_0$, $\hat{\beta}_1$, and $MSE$. Using $MSE$ to estimate $\sigma^2$ and taking square roots, we get the standard errors.

$$\mathrm{SE}\left(\hat{\beta}_1\right) = \sqrt{\frac{MSE}{(n-1)s_x^2}}$$

and

$$\mathrm{SE}\left(\hat{\beta}_0\right) = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}_\cdot^2}{(n-1)s_x^2}\right]}.$$

Table 7.2: *Analysis of variance*

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Intercept($\beta_0$) | 1 | $n\bar{y}_\cdot^2 \equiv C$ | $n\bar{y}_\cdot^2$ | |
| Regression($\beta_1$) | 1 | $\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2$ | $SSReg$ | $\frac{MSReg}{MSE}$ |
| Error | $n-2$ | $\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ | $SSE/(n-2)$ | |
| Total | $n$ | $\sum_{i=1}^n y_i^2$ | | |

EXAMPLE 7.3.4. For the *Coleman Report* data, using the numbers $n$, $\bar{x}_\cdot$, and $s_x^2$,

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{(20-1)92.64798395} = \frac{\sigma^2}{1760.311695}$$

and

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2\left[\frac{1}{20} + \frac{3.1405^2}{(20-1)92.64798395}\right] = \sigma^2\left[.055602837\right].$$

The $MSE$ is 5.014, so the standard errors are

$$\mathrm{SE}\left(\hat{\beta}_1\right) = \sqrt{\frac{5.014}{1760.311695}} = .05337$$

and

$$\mathrm{SE}\left(\hat{\beta}_0\right) = \sqrt{5.014\left[.055602837\right]} = .5280.$$

□

We always like to have estimates with small variances. The forms of the variances show how to achieve this. For example, the variance of $\hat{\beta}_1$ gets smaller when $n$ or $s_x^2$ gets larger. Thus, more observations (larger $n$) result in a smaller slope variance and more dispersed $x_i$ values (larger $s_x^2$) also result in a smaller slope variance. Of course all of this assumes that the simple linear regression model is correct.

## 7.4 The analysis of variance table

A standard tool in regression analysis is the construction of an analysis of variance table. The best form is given in Table 7.2. In this form there is one degree of freedom for every observation, cf. the total line, and the sum of squares total is the sum of all of the squared observations. The degrees of freedom and sums of squares for intercept, regression, and error can be added to obtain the degrees of freedom and sums of squares total. We see that one degree of freedom is used to estimate the intercept, one is used for the slope, and the rest are used to estimate the variance.

The more commonly used form for the analysis of variance table is given as Table 7.3. It eliminates the line for the intercept and corrects the total line so that the degrees of freedom and sums of squares still add up.

These two forms for the analysis of variance table are analogous to the two different forms discussed in Section 5.2 for the one-way ANOVA analysis of variance table.

EXAMPLE 7.4.1. Consider again the *Coleman Report* data. The analysis of variance table was given in Section 7.1; Table 7.4 illustrates the necessary computations. Most of the computations were made earlier in equation (7.3.2) during the process of obtaining the $MSE$ and all are based on the usual six numbers, $n$, $\bar{x}_\cdot$, $s_x^2$, $\bar{y}_\cdot$, $s_y^2$, and $\sum x_i y_i$. More directly, the computations depend on $n$, $\hat{\beta}_1$, $s_x^2$,

Table 7.3: *Analysis of variance*

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Regression($\beta_1$) | 1 | $\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2$ | $SSReg$ | $\frac{MSReg}{MSE}$ |
| Error | $n-2$ | $\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ | $SSE/(n-2)$ | |
| Total | $n-1$ | $\sum_{i=1}^n (y_i - \bar{y}_\cdot)^2$ | | |

Table 7.4: *Analysis of variance*

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Regression($\beta_1$) | 1 | $.560325^2(20-1)92.64798$ | 552.6756109 | $\frac{552.68}{5.014}$ |
| Error | $20-2$ | 90.2487641 | $90.2487641/18$ | |
| Total | $20-1$ | $(20-1)33.838125$ | | |

and $s_y^2$. The corrected version of $SSTot$ is $(n-1)s_y^2$. Note that the $SSE$ is obtained as $SSTot - SSReg$. The correction factor $C$ in Table 7.2 is $20(35.0825)^2$ but it is not used in these computations for Table 7.4. □

## 7.5 Inferential procedures

The general theory of Chapter 3 applies to inferences about regression parameters. The theory requires 1) a parameter ($Par$), 2) an estimate ($Est$) of the parameter, 3) the standard error of the estimate (SE($Est$)) and 4) a known (tabled) distribution for

$$\frac{Est - Par}{\mathrm{SE}(Est)}$$

that is symmetric about 0. The computations for most of the applications considered in this section were illustrated in Section 7.1 for the *Coleman Report* data.

Consider inferences about the slope *parameter* $\beta_1$. The estimate $\hat{\beta}_1$ and the standard error of $\hat{\beta}_1$ are as given in Section 7.3. The appropriate reference distribution is

$$\frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}\left(\hat{\beta}_1\right)} \sim t(n-2).$$

Using standard methods, the 99% confidence interval for $\beta_1$ has endpoints

$$\hat{\beta}_1 \pm t(.995, n-2)\,\mathrm{SE}\left(\hat{\beta}_1\right).$$

An $\alpha = .05$ test of, say, $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ rejects $H_0$ if

$$\frac{|\hat{\beta}_1 - 0|}{\mathrm{SE}\left(\hat{\beta}_1\right)} > t(.975, n-2).$$

An $\alpha = .05$ test of $H_0 : \beta_1 \geq 1$ versus $H_A : \beta_1 < 1$ rejects $H_0$ if

$$\frac{\hat{\beta}_1 - 1}{\mathrm{SE}\left(\hat{\beta}_1\right)} < -t(.95, n-2).$$

For inferences about the intercept parameter $\beta_0$, the estimate $\hat{\beta}_0$, and the standard error of $\hat{\beta}_0$ are as given in Section 7.3. The appropriate reference distribution is

$$\frac{\hat{\beta}_0 - \beta_0}{\mathrm{SE}\left(\hat{\beta}_0\right)} \sim t(n-2).$$

A 95% confidence interval for $\beta_0$ has endpoints

$$\hat{\beta}_0 \pm t(.975, n-2)\,\mathrm{SE}\left(\hat{\beta}_0\right).$$

An $\alpha = .01$ test of $H_0 : \beta_0 = 0$ versus $H_A : \beta_0 \neq 0$ rejects $H_0$ if

$$\frac{|\hat{\beta}_0 - 0|}{\mathrm{SE}\left(\hat{\beta}_0\right)} > t(.995, n-2).$$

An $\alpha = .05$ test of $H_0 : \beta_0 \leq 0$ versus $H_A : \beta_0 > 0$ rejects $H_0$ if

$$\frac{\hat{\beta}_0 - 0}{\mathrm{SE}\left(\hat{\beta}_0\right)} > t(.95, n-2).$$

Typically inferences about $\beta_0$ are not of substantial interest. $\beta_0$ is the intercept, it is the value of the line when $x = 0$. Typically, the line is only an approximation to the behavior of the $(x, y)$ pairs in the neighborhood of the observed data. This approximation is only valid in the neighborhood of the observed data. If we have not collected data near $x = 0$, the intercept is describing behavior of the line outside the range of valid approximation.

We can also draw inferences about a point on the line $y = \beta_0 + \beta_1 x$. For any fixed point $x$, $\beta_0 + \beta_1 x$ has an estimate

$$\hat{y} \equiv \hat{\beta}_0 + \hat{\beta}_1 x.$$

To get a standard error for $\hat{y}$, we first need its variance. As shown in the appendix to this chapter, the variance of $\hat{y}$ is

$$\mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x}.)^2}{(n-1)s_x^2}\right], \tag{7.5.1}$$

so the standard error of $\hat{y}$ is

$$\mathrm{SE}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) = \sqrt{MSE\left[\frac{1}{n} + \frac{(x - \bar{x}.)^2}{(n-1)s_x^2}\right]}. \tag{7.5.2}$$

The appropriate distribution for inferences about the point $\beta_0 + \beta_1 x$ is

$$\frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) - (\beta_0 + \beta_1 x)}{\mathrm{SE}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right)} \sim t(n-2).$$

Using standard methods, the 99% confidence interval for $(\beta_0 + \beta_1 x)$ has endpoints

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \pm t(.995, n-2)\,\mathrm{SE}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right).$$

We typically prefer to have small standard errors. Even when $\sigma^2$, and thus $MSE$, is large, from equation (7.5.2) we see that the standard error of $\hat{y}$ will be small when the number of observations

$n$ is large, when the $x_i$ values are well spread out, i.e., $s_x^2$ is large, and when $x$ is close to $\bar{x}_\cdot$. In other words, the line can be estimated efficiently in the neighborhood of $\bar{x}_\cdot$ by collecting a lot of data. Unfortunately, if we try to estimate the line far from where we collected the data, the standard error of the estimate gets large. The standard error gets larger as $x$ gets farther away from the center of the data, $\bar{x}_\cdot$, because the term $(x-\bar{x}_\cdot)^2$ gets larger. This effect is standardized by the original observations; the term in question is $(x-\bar{x}_\cdot)^2/(n-1)s_x^2$, so $(x-\bar{x}_\cdot)^2$ must be large relative to $(n-1)s_x^2$ before a problem develops. In other words, the distance between $x$ and $\bar{x}_\cdot$ must be several times the standard deviation $s_x$ before a problem develops. Nonetheless, large standard errors occur when we try to estimate the line far from where we collected the data. Moreover, the regression line is often just an approximation that holds in the neighborhood of where the data were collected. This approximation may be invalid for data points far from the original data. So, in addition to the problem of having large standard errors, estimates far from the neighborhood of the original data may be totally invalid.

Estimating a point on the line is distinct from prediction of a new observation for a given $x$ value. Ideally, the prediction would be the true point on the line for the value $x$. However, the true line is an unknown quantity, so our prediction is the estimated point on the line at $x$. The distinction between prediction and estimating a point on the line arises because a new observation is subject to variability about the line. In making a prediction we must account for the variability of the new observation even when the line is known, as well as account for the variability associated with our need to estimate the line. The new observation is assumed to be independent of the past data, so the variance of the prediction is $\sigma^2$ (the variance of the new observation) plus the variance of the estimate of the line as given in (7.5.1). The standard error replaces $\sigma^2$ with $MSE$ and takes the square root, i.e.,

$$\mathrm{SE}(\mathit{Prediction}) = \sqrt{MSE\left[1+\frac{1}{n}+\frac{(x-\bar{x}_\cdot)^2}{(n-1)s_x^2}\right]}.$$

Note that this is the same as the formula given in equation (7.1.2). Prediction intervals follow in the usual way. For example, the 99% prediction interval associated with $x$ has endpoints

$$(\hat{y}) \pm t(.995, n-2)\,\mathrm{SE}(\mathit{Prediction}).$$

As discussed earlier, estimation of points on the line should be restricted to $x$ values in the neighborhood of the original data. For similar reasons, predictions should also be made only in the neighborhood of the original data. While it is possible, by collecting a lot of data, to estimate the line well even when the variance $\sigma^2$ is large, it is not always possible to get good prediction intervals. Prediction intervals are subject to the variability of both the observations and the estimate of the line. The variability of the observations cannot be eliminated or reduced. If this variability is too large, we may get prediction intervals that are too large to be useful. If the simple linear regression model is the 'truth', there is nothing to be done, i.e., no way to improve the prediction intervals. If the simple linear regression model is only an approximation to the true process, a more sophisticated model may give a better approximation and produce better prediction intervals.

## 7.6 An alternative model

For some purposes, it is more convenient to work with an alternative to the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The alternative model is

$$y_i = \beta_{*0} + \beta_1 (x_i - \bar{x}_\cdot) + \varepsilon_i$$

where we have adjusted the predictor variable for its mean. The key difference between the parameters in the two models is that

$$\beta_0 = \beta_{*0} - \beta_1 \bar{x}_\cdot.$$

In fact, this is the basis for our formula for estimating $\beta_0$. The new parameter $\beta_{*0}$ has a very simple estimate, $\hat{\beta}_{*0} \equiv \bar{y}_\cdot$. It then follows that

$$\hat{\beta}_0 = \bar{y}_\cdot - \hat{\beta}_1 \bar{x}_\cdot.$$

The reason that this model is useful is because the predictor variable $x_i - \bar{x}_\cdot$ has the property $\sum_{i=1}^n (x_i - \bar{x}_\cdot) = 0$. This property leads to the simple estimate of $\beta_{*0}$ but also to the fact that $\bar{y}_\cdot$ and $\hat{\beta}_1$ are independent. Independence simplifies the computation of variances for regression line estimates. We will not go further into these claims at this point but the results follow trivially from the matrix approach to regression that will be treated in later chapters.

The key point about the alternative model is that it is equivalent to the original model. The $\beta_1$ parameters are the same, as are their estimates and standard errors. The models give the same predictions, the same ANOVA table $F$ test, and the same $R^2$. Even the intercept parameters are equivalent, i.e., they are related in a precise fashion so that knowing about the intercept in either model yields equivalent information about the intercept in the other model.

## 7.7 Correlation

The correlation coefficient is a measure of the linear relationship between two variables. The population correlation coefficient, usually denoted $\rho$, was discussed in Chapter 1. The sample correlation is defined as

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x}_\cdot)(y_i - \bar{y}_\cdot)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2 \sum_{i=1}^n (y_i - \bar{y}_\cdot)^2}}.$$

The sample correlation coefficient is related to the estimated slope. From equation (7.3.1) it is easily seen that

$$r = \hat{\beta}_1 \frac{s_x}{s_y}.$$

EXAMPLE 7.7.1. *Simulated data with various correlations*
Figures 7.2 through 7.5 contain plots of 25 correlated observations. These are presented so the reader can get some feeling for the meaning of various sample correlation values. The caption of each plot gives the sample correlation $r$ and also the population correlation $\rho$. The population correlation is only useful in that it provides some feeling for the amount of sampling variation to be found in $r$ based on samples of 25 from (jointly) normally distributed data. □

A commonly used statistic in regression analysis is the coefficient of determination,

$$R^2 \equiv \frac{SSReg}{SSTot}.$$

This is the percentage of the total variation in the dependent variable that is explained by the regression. For simple linear regression,

$$R^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2}{\sum_{i=1}^n (y_i - \bar{y}_\cdot)^2} = \hat{\beta}_1^2 \frac{s_x^2}{s_y^2} = r^2.$$

In later chapters we will consider regression problems with more than one predictor variable. For such problems $R^2$ does not equal $r^2$. In fact, with more than one predictor, there are several $r^2$s that one could compute. It is not clear which of these one would want to compare to $R^2$.

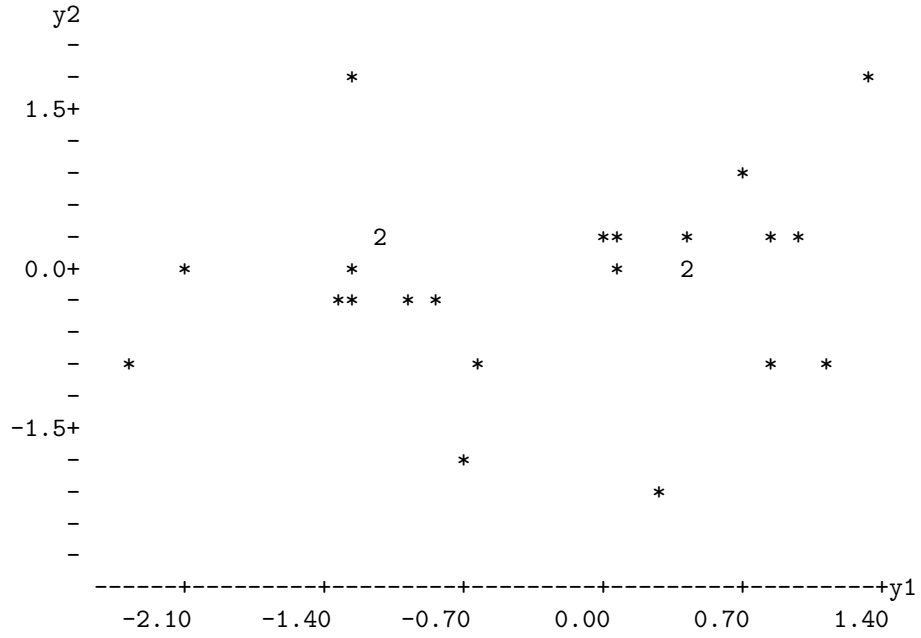



Figure 7.2: *Correlation plot,* $\rho = 0.000$*,* $r = 0.144$*.*

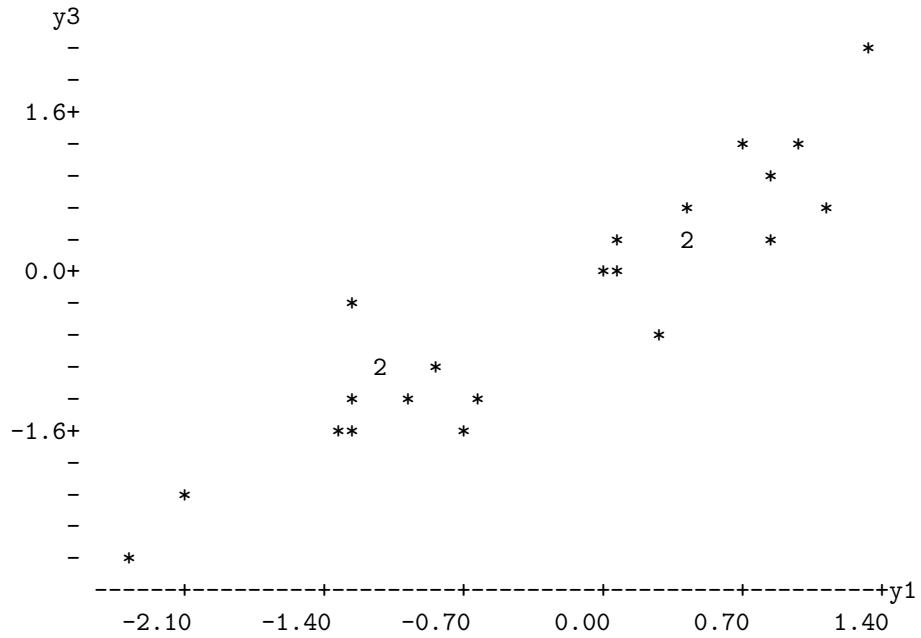



Figure 7.3: *Correlation plot,* $\rho = 0.894$*,* $r = 0.929$*.*

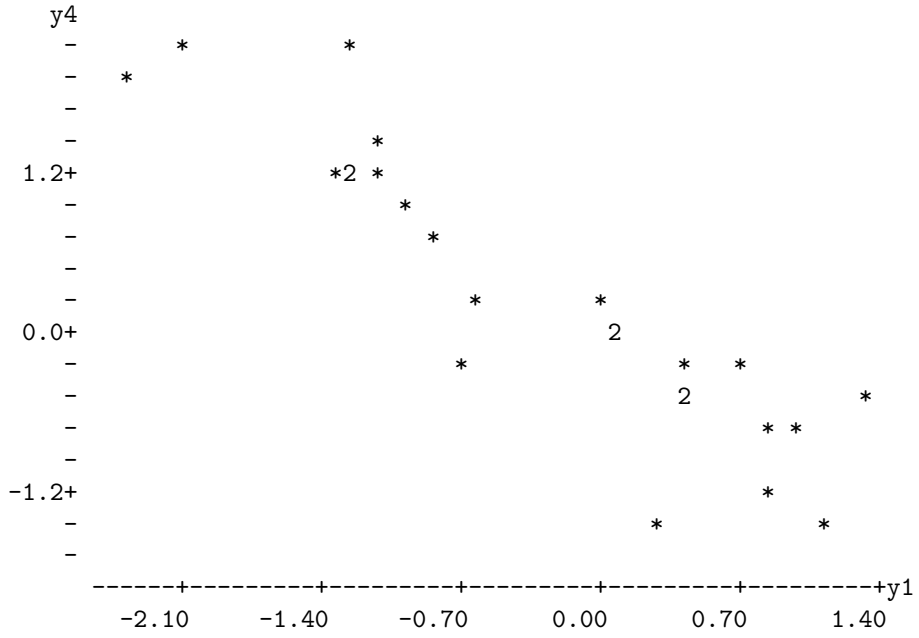



Figure 7.4: *Correlation plot,* $\rho = -0.894$*,* $r = -0.929$*.*

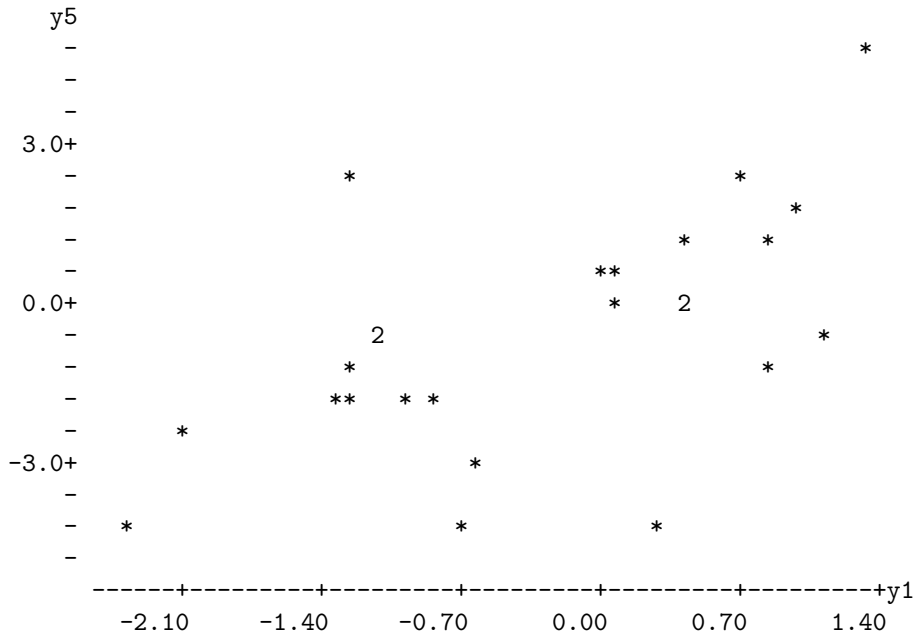



Figure 7.5: *Correlation plot,* $\rho = 0.447$*,* $r = 0.593$*.*

## 7.8 Recognizing randomness: simulated data with zero correlation

Just as it is important to be able to look at a plot and tell when the $x$ and $y$ variables are related, it is important to be able to look at a plot and tell that two variables are unrelated. In other words, we need to be able to identify plots that only display random variation. This skill is of particular importance in Section 7.9 where we use plots to evaluate the assumptions made in simple linear regression. To check the assumptions of the regression model, we use plots that should display only random variation when the assumptions are true. Any systematic pattern in the model checking plots indicates a problem with our assumed regression model.

EXAMPLE 7.8.1. *Simulated data with zero correlation*
We now examine data on six uncorrelated variables, C10 through C15. Figures 7.6 through 7.14 contain various plots of the variables. Since all the variable pairs have zero correlation, i.e., $\rho = 0$, any 'patterns' that are recognizable in these plots are due entirely to random variation. In particular, note that there is no real pattern in Figure 7.13.

The point of this example is to familiarize the reader with the appearance of random plots. The reader should try to identify systematic patterns in these plots, remembering that there are none. This suggests that in the model checking plots that appear later, any systematic pattern of interest should be more pronounced than anything that can be detected in Figures 7.6 through 7.15.

Below are the sample correlations $r$ for each pair of variables. Although $\rho = 0$, none of the $r$ values is zero and some of them are quite far from 0.

Sample correlations

| | C10 | C11 | C12 | C13 | C14 | C15 |
|---|---|---|---|---|---|---|
| C10 | 1.000 | | | | | |
| C11 | 0.005 | 1.000 | | | | |
| C12 | −0.145 | −0.209 | 1.000 | | | |
| C13 | −0.162 | −0.416 | 0.488 | 1.000 | | |
| C14 | −0.034 | −0.038 | −0.265 | 0.003 | 1.000 | |
| C15 | −0.218 | −0.202 | 0.310 | 0.114 | 0.134 | 1.000 |

□

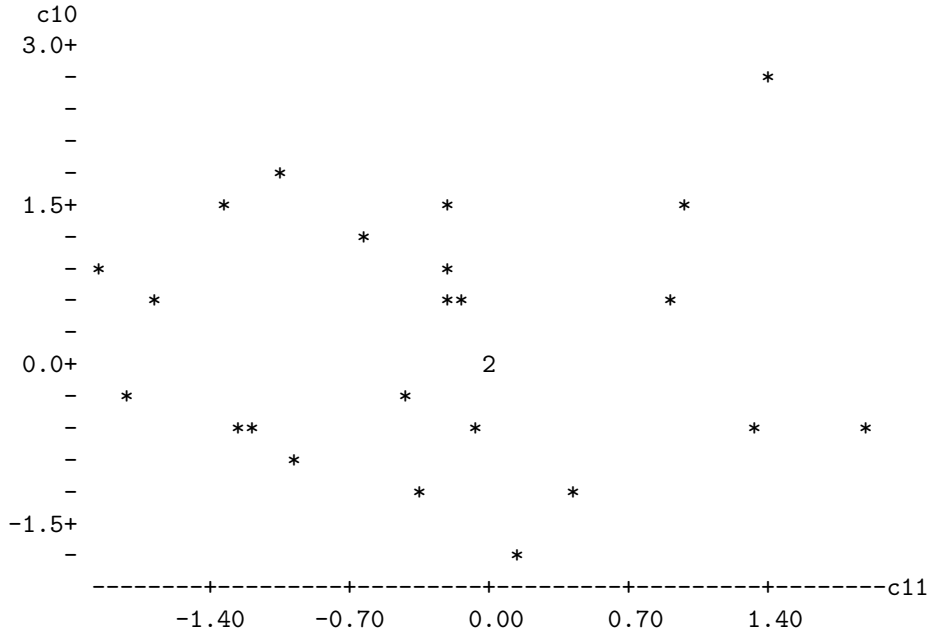



Figure 7.6: *Plot of data with* $\rho = 0$.

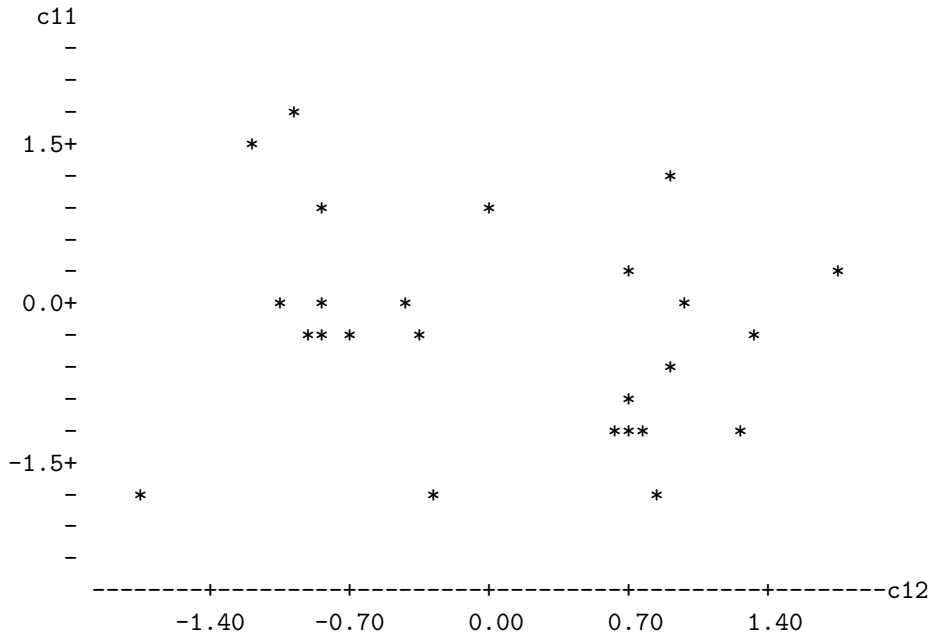



Figure 7.7: *Plot of data with* $\rho = 0$.

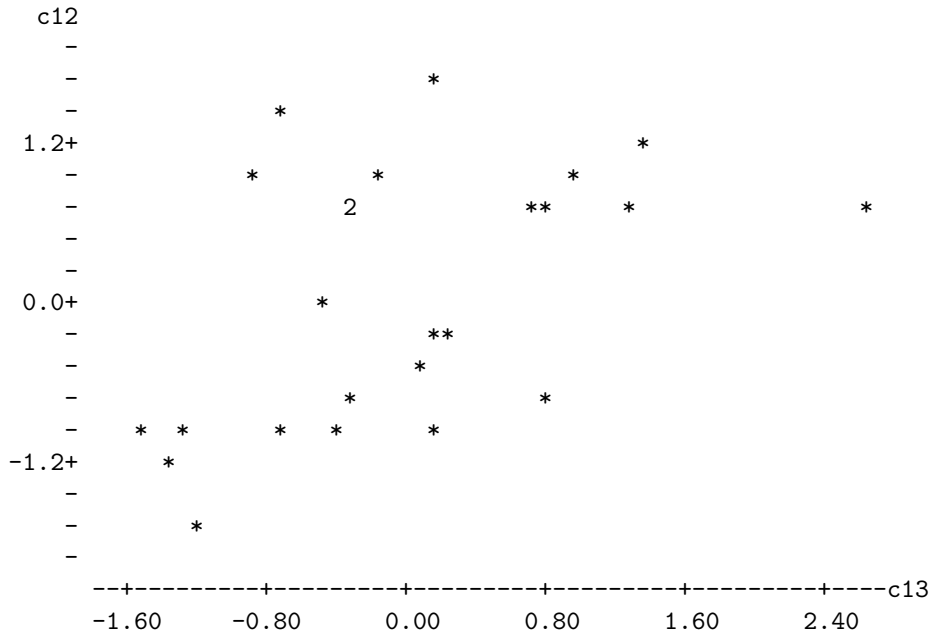



Figure 7.8: *Plot of data with* $\rho = 0$.

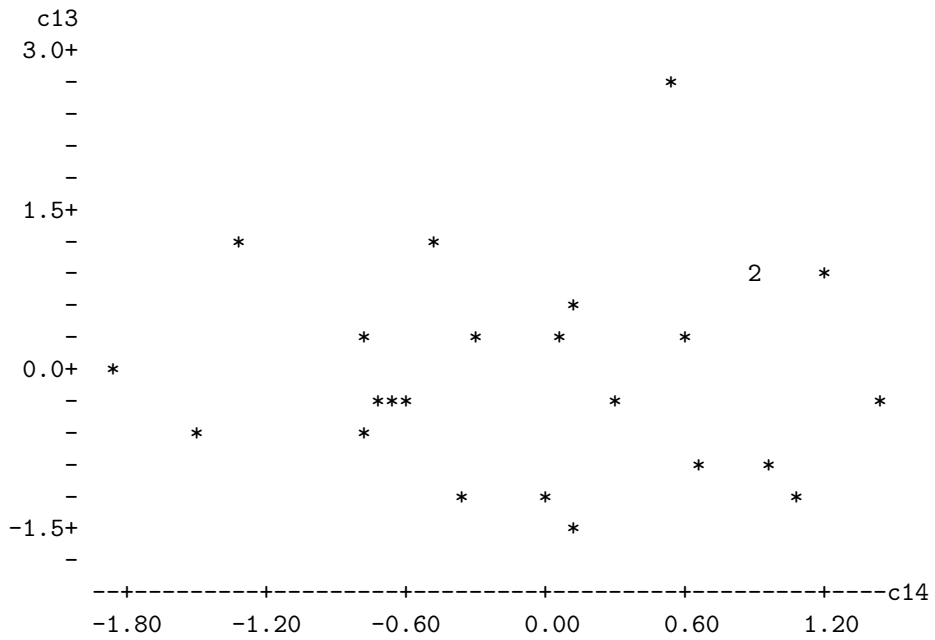



Figure 7.9: *Plot of data with* $\rho = 0$.

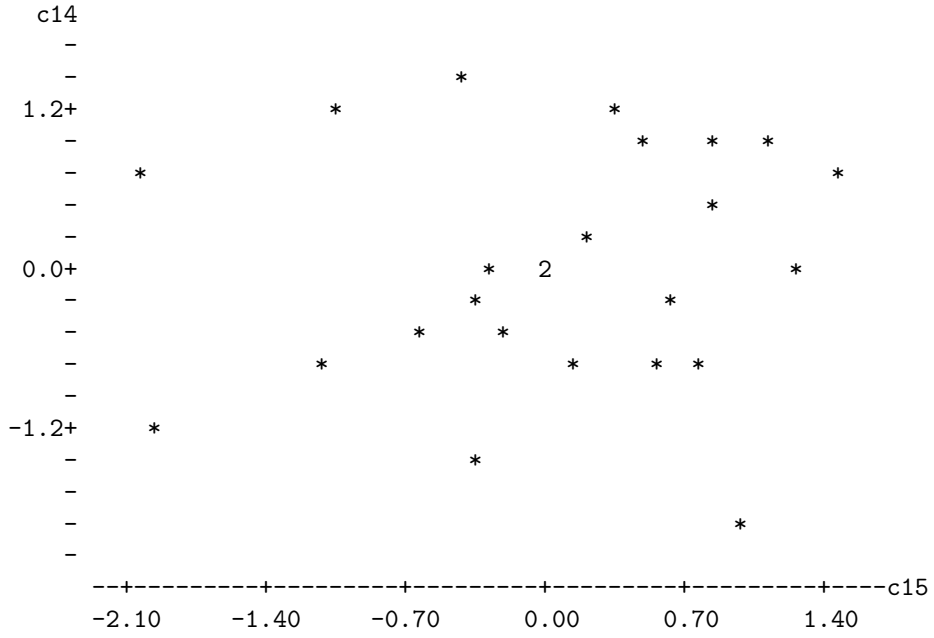



Figure 7.10: *Plot of data with* $\rho = 0$.

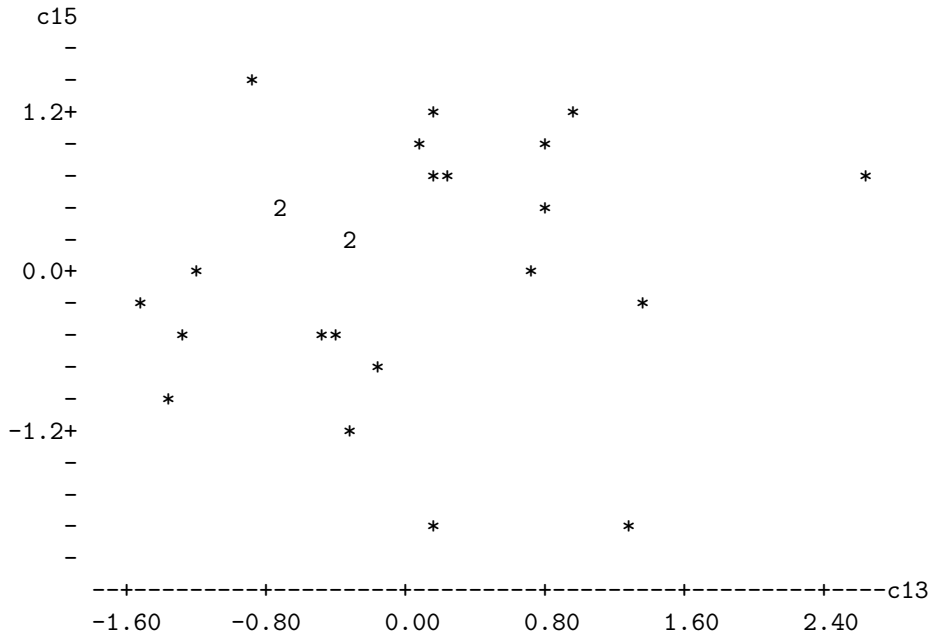



Figure 7.11: *Plot of data with* $\rho = 0$.

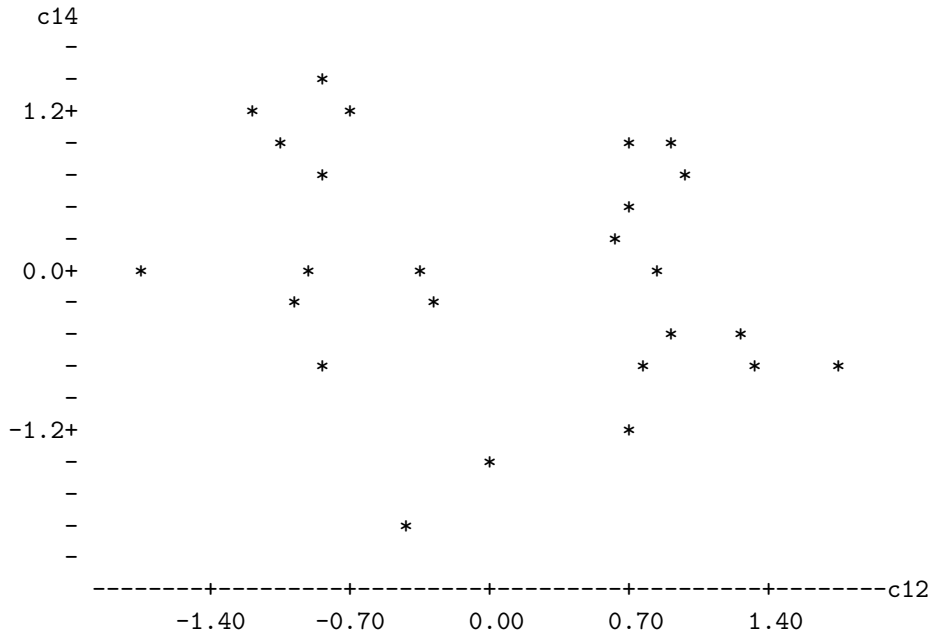



Figure 7.12: *Plot of data with* $\rho = 0$.

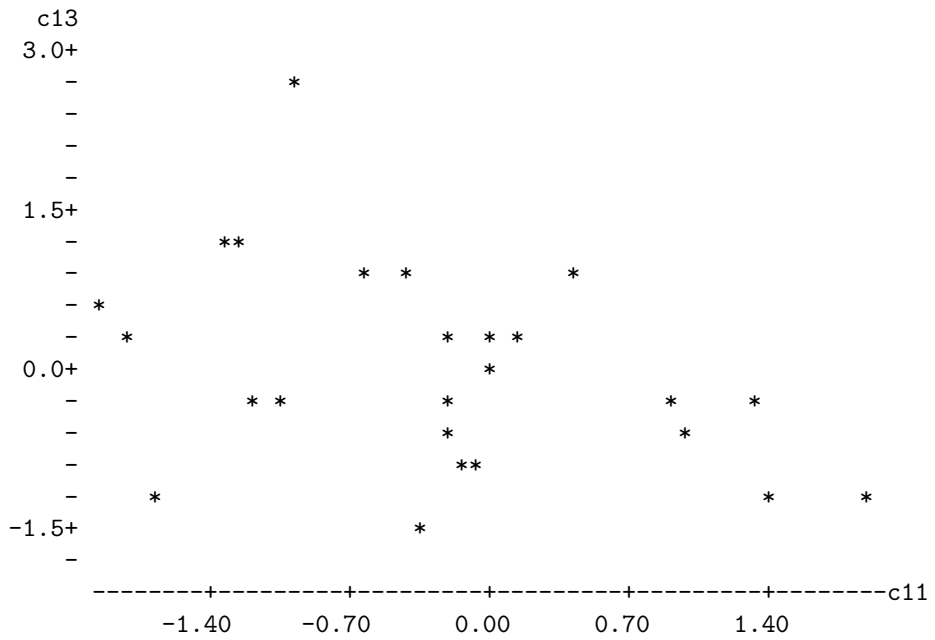



Figure 7.13: *Plot of data with* $\rho = 0$.

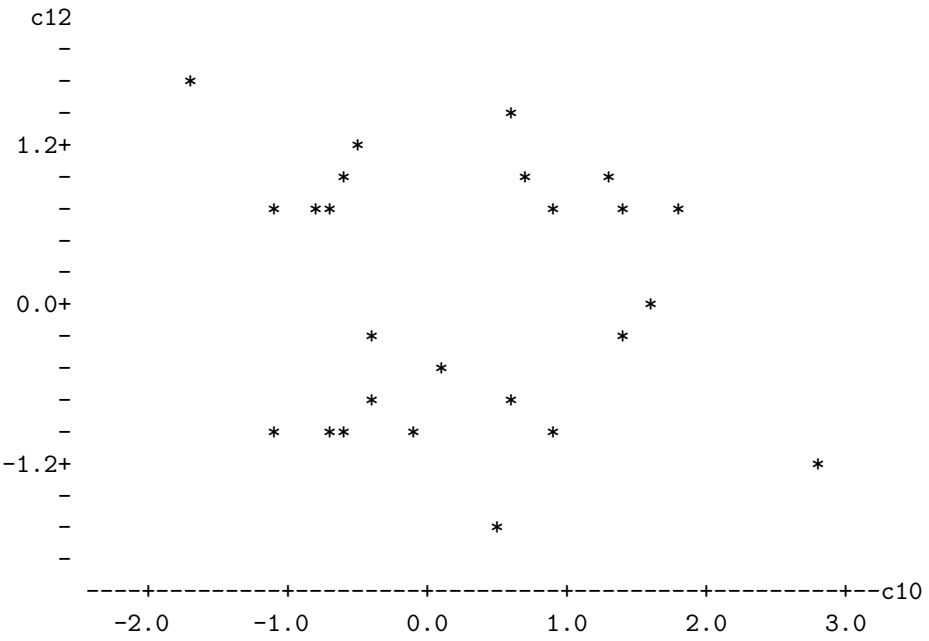



Figure 7.14: *Plot of data with* $\rho = 0$.

*Minitab commands*

The plots and sample correlations in this section were obtained with the Minitab commands given below.

```
MTB > random 25 c10-c15;
SUBC> normal 0 1.
MTB > plot c10 c11
MTB > plot c11 c12
MTB > plot c12 c13
MTB > plot c13 c14
MTB > plot c14 c15
MTB > plot c15 c13
MTB > plot c14 c12
MTB > plot c13 c11
MTB > plot c12 c10
MTB > note     OBTAIN SAMPLE CORRELATION MATRIX
MTB > corr c10-c15
```

## 7.9 Checking assumptions: residual analysis

The assumptions involved in regression can all be thought of in terms of the errors. The assumptions are that

1. the $\varepsilon_i$s are independent,
2. $\mathrm{E}(\varepsilon_i) = 0$ for all $i$,
3. $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all $i$,
4. the $\varepsilon_i$s are normally distributed.

To have faith in our analysis, we need to validate these assumptions as far as possible. These are *assumptions* and cannot be validated completely, but we can try to detect gross violations of the assumptions.

The first assumption, that the $\varepsilon_i$s are independent, is the most difficult to validate. If the observations are taken at regular time intervals, they may lack independence and standard time series methods may be useful in the analysis. We will not consider this further, the interested reader can consult the time series literature, e.g., Shumway (1988). In general, we rely on the data analyst to think hard about whether there are reasons for the data to lack independence.

The second assumption is that $\mathrm{E}(\varepsilon_i) = 0$. This is violated when we have the wrong regression model. The simple linear regression model with $\mathrm{E}(\varepsilon_i) = 0$ specifies that

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i.$$

If we fit this model when it is incorrect, we will not have errors with $\mathrm{E}(\varepsilon_i) = 0$. Having the wrong model is called *lack of fit*.

The last two assumptions are that the errors all have some common variance $\sigma^2$ and that they are normally distributed. The term *homoscedasticity* refers to having a constant (homogeneous) variance. The term *heteroscedasticity* refers to having nonconstant (heterogeneous) variances.

In checking the error assumptions, we are hampered by the fact that the errors are not observable; we must estimate them. The model involves

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

or equivalently,

$$y_i - \beta_0 - \beta_1 x_i = \varepsilon_i.$$

We can estimate $\varepsilon_i$ with the *residual*

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Actually, I prefer to call this predicting the error rather than estimating it. *One estimates fixed unknown parameters and predicts unobserved random variables*.

Previously, we used residuals to check assumptions in one-way analysis of variance. The discussion here is similar but more extensive. The methods presented here also apply to ANOVA models, but, especially in balanced ANOVA, many of the issues are not as crucial. Note also that, as in one-way ANOVA, the *SSE* for simple linear regression is precisely the sum of the squared residuals.

Two of the error assumptions are independence and homoscedasticity of the variances. Unfortunately, the residuals are neither independent nor do they have the same variance. The residuals all involve the random variables $\hat{\beta}_0$ and $\hat{\beta}_1$, so they are not independent. Moreover, the $i$th residual involves $\hat{\beta}_0 + \hat{\beta}_1 x_i$, the variance of which depends on $(x_i - \bar{x}_\cdot)$. Thus the variance of $\hat{\varepsilon}_i$ depends on $x_i$. There is little we can do about the lack of independence except hope that it does not cause severe problems. On the other hand, we can adjust for the differences in variances. The variance of a residual is

$$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$$

where $h_i$ is the *leverage* of the $i$th case. Leverages are discussed a bit later in this section and more extensively in relation to multiple regression.

Given the variance of a residual, we can obtain a standard error for it,

$$\mathrm{SE}(\hat{\varepsilon}_i) = \sqrt{MSE(1 - h_i)}.$$

We can now adjust the residuals so they all have a variance of about 1; these *standardized residuals* are

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{MSE(1 - h_i)}}.$$

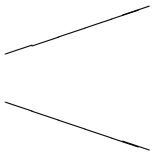
The main tool used in checking assumptions is plotting the residuals or, more commonly, the standardized residuals. If the assumptions are correct, plots of the standardized residuals versus any variable should look random. If the variable plotted against the $r_i$s is continuous with no major gaps, the plots should look similar to the plots given in the previous section. In analysis of variance problems, we often plot the residuals against indicators of the treatment groups, so the discrete nature of the number of groups keeps the plots from looking like those of the previous section. The single most popular diagnostic plot is probably the plot of the standardized residuals against the predicted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

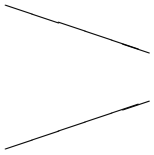
however the $r_i$s can be plotted against any variable that provides a value associated with each case.

Violations of the error assumptions are indicated by any systematic pattern in the residuals. This could be, for example, a pattern of increased variability as the predicted values increase, or some curved pattern in the residuals, or any change in the variability of the residuals.

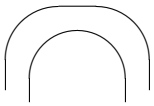
A residual plot that displays an increasing variance looks roughly like a horn opening to the right.



A residual plot indicating a decreasing variance is a horn opening to the left.



Plots that display curved shapes typically indicate lack of fit. One example of a curve is given below.



EXAMPLE 7.9.1. *Coleman Report data*

Figures 7.15 through 7.17 contain standardized residual plots for the *Coleman Report* Data. Figure 7.15 is a plot against the predicted values; Figure 7.16 is a plot against the sole predictor variable $x$. The shapes of these two plots are identical. This always occurs in simple linear regression because the predictions $\hat{y}$ are a linear function of the one predictor $x$. The one caveat to the claim of identical shapes is that the plots may be reversed. If the estimated slope is negative, the largest $x$ values correspond to the smallest $\hat{y}$ values. Figures 7.15 and 7.16 look like random patterns but it should be noted that if the smallest standardized residual were dropped (the small one on the right), the plot might suggest decreasing variability. The normal plot of the standardized residuals in Figure 7.17 does not look too bad. □

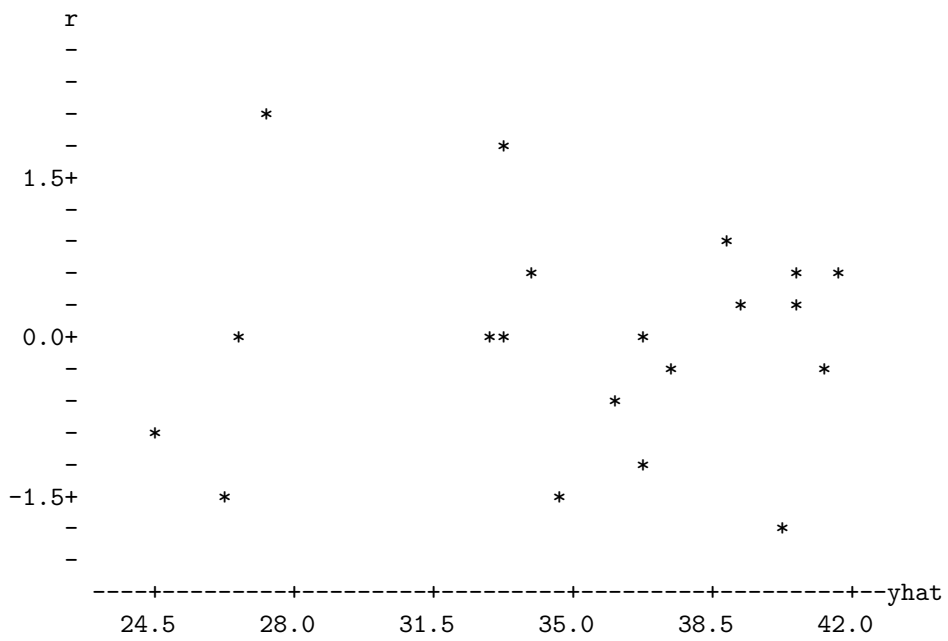



Figure 7.15: *Plot of the standardized residuals r versus ŷ,* Coleman Report.

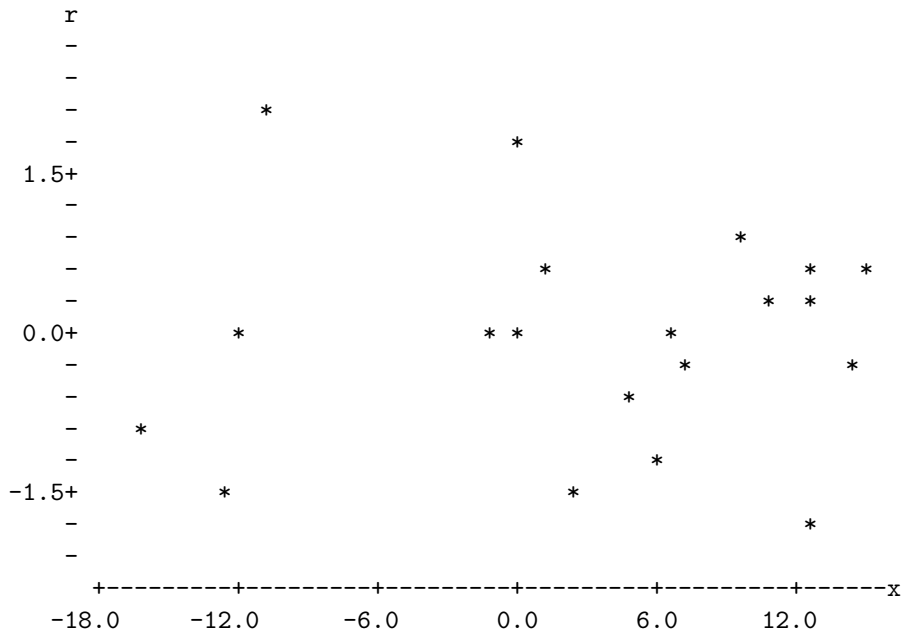



Figure 7.16: *Plot of the standardized residuals r versus x,* Coleman Report.

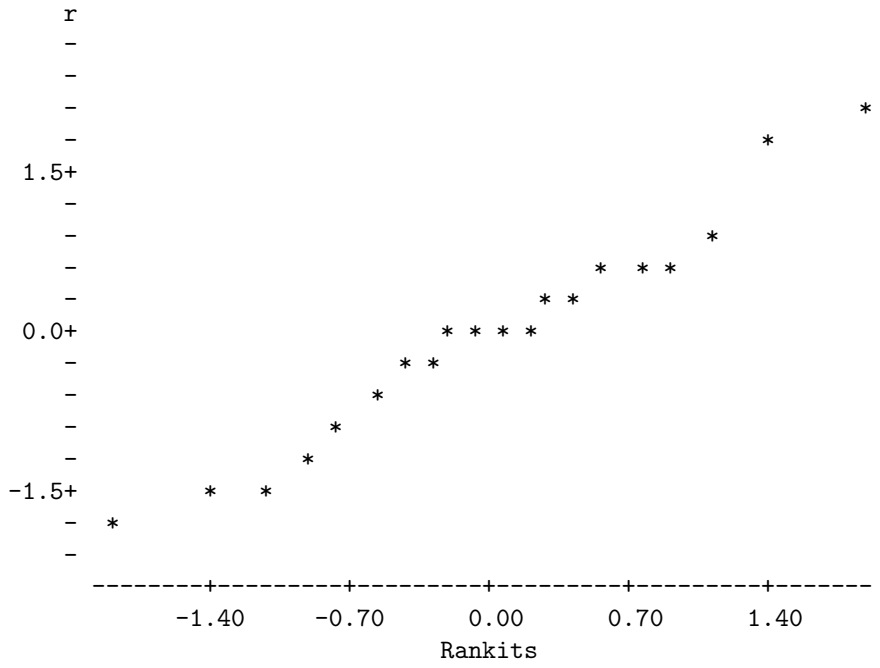



Figure 7.17: *Normal plot,* Coleman Report, $W' = .966$.

*Minitab commands*

We now illustrate the Minitab commands necessary for the analysis in Example 7.9.1. The subtlest thing going on here is in the command 'regress c1 on 1 c2 c11 c12'. The number 1 indicates that there is one predictor variable $x$ and the first column (c2) after the number 1 is taken to be that predictor. The command recognizes columns c11 and c12 as not being predictors; in fact, the command puts the standardized residuals $r_i$ in the first column (c11) listed after the predictor and the predicted values $\hat{y}_i$ in the second column (c12) listed after

Table 7.5: *Hooker data*

| Case | Temperature | Pressure | Case | Temperature | Pressure |
|---|---|---|---|---|---|
| 1 | 180.6 | 15.376 | 17 | 191.1 | 19.490 |
| 2 | 181.0 | 15.919 | 18 | 191.4 | 19.758 |
| 3 | 181.9 | 16.106 | 19 | 193.4 | 20.480 |
| 4 | 181.9 | 15.928 | 20 | 193.6 | 20.212 |
| 5 | 182.4 | 16.235 | 21 | 195.6 | 21.605 |
| 6 | 183.2 | 16.385 | 22 | 196.3 | 21.654 |
| 7 | 184.1 | 16.959 | 23 | 196.4 | 21.928 |
| 8 | 184.1 | 16.817 | 24 | 197.0 | 21.892 |
| 9 | 184.6 | 16.881 | 25 | 199.5 | 23.030 |
| 10 | 185.6 | 17.062 | 26 | 200.1 | 23.369 |
| 11 | 185.7 | 17.267 | 27 | 200.6 | 23.726 |
| 12 | 186.0 | 17.221 | 28 | 202.5 | 24.697 |
| 13 | 188.5 | 18.507 | 29 | 208.4 | 27.972 |
| 14 | 188.8 | 18.356 | 30 | 210.2 | 28.559 |
| 15 | 189.5 | 18.869 | 31 | 210.8 | 29.211 |
| 16 | 190.6 | 19.386 | | | |

the predictor variable. Actually, Minitab will let you write the command as 'regress c1 on 1 c2 put standardized resids in c11 put predicted values in c12'. Minitab just ignores all the words in this command other than 'regress'.

```
MTB > names c1 'test' c2 'socio'
MTB > regress c1 on 1 c2 c11 c12
MTB > names c11 'r' c12 'yhat'
MTB > note     PLOT STD. RESIDS AGAINST PRED. VALUES
MTB > plot c11 c12
MTB > note     PLOT STD. RESIDS AGAINST x
MTB > plot c11 c2
MTB > note     COMPUTE NORMAL SCORES (RANKITS) FOR THE
MTB > note     STANDARDIZED RESIDUALS
MTB > nscores c11 c10
MTB > note     MAKE NORMAL PLOT
MTB > plot c11 c10
MTB > note     COMPUTE W' STATISTIC
MTB > corr c11 c10
MTB > note     CORR PRINTS OUT A NUMBER LIKE .978
MTB > let k1=.978**2
MTB > print k1
```

*Another example*

EXAMPLE 7.9.2. *Hooker data*
Forbes (1857) reported data on the relationship between atmospheric pressure and the boiling point of water that were collected in the Himalaya mountains by Joseph Hooker. Weisberg (1985, p. 28) presented a subset of 31 observations that are reproduced in Table 7.5.

A scatter plot of the data is given in Figure 7.18. The data appear to fit a line very closely. The usual summary tables are given below.

| Predictor | $\hat{\beta}_k$ | SE($\hat{\beta}_k$) | $t$ | $P$ |
|---|---|---|---|---|
| Constant | −64.413 | 1.429 | −45.07 | 0.000 |
| Temperature | 0.440282 | 0.007444 | 59.14 | 0.000 |

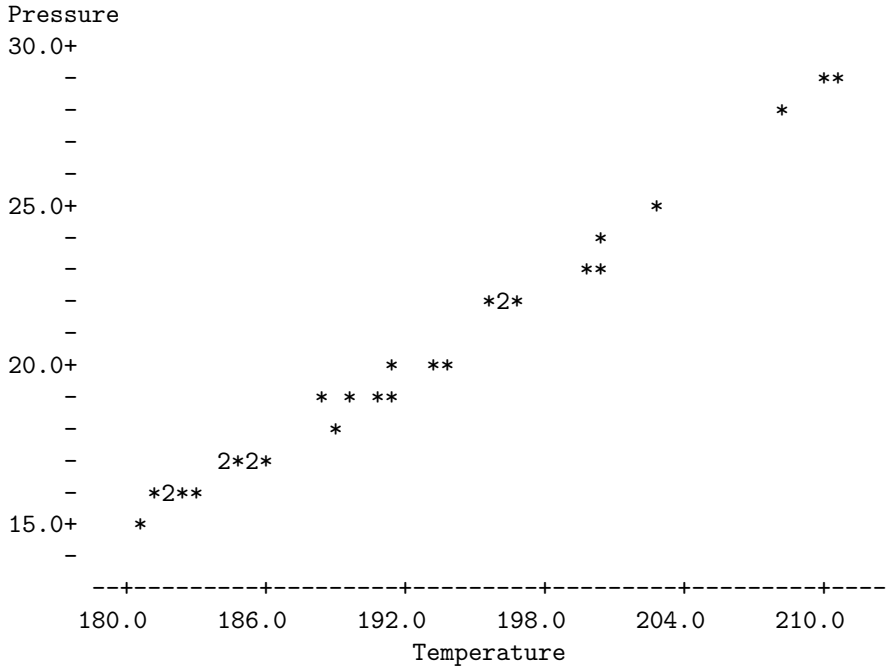



Figure 7.18: *Scatter plot of Hooker data.*

Analysis of variance

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 1 | 444.17 | 444.17 | 3497.89 | 0.000 |
| Error | 29 | 3.68 | 0.13 | | |
| Total | 30 | 447.85 | | | |

The coefficient of determination is an exceptionally large

$$R^2 = \frac{444.17}{447.85} = 99.2\%.$$

The plot of residuals versus predicted values is given in Figure 7.19. A pattern is very clear; the residuals form something like a parabola. In spite of a very large $R^2$ and a scatter plot that looks very linear, the residual plot shows that a lack of fit obviously exists. After seeing the residual plot, you can go back to the scatter plot and detect suggestions of nonlinearity. The simple linear regression model is clearly inadequate, so we do not bother presenting a normal plot. In the next two sections, we will examine ways of dealing with this lack of fit.

□

*Outliers*

Outliers are bizarre data points. They are points that do not seem to fit with the other observations in a data set. We can characterize bizarre points as having either bizarre $x$ values or bizarre $y$ values. There are two valuable tools for identifying outliers.

*Leverages* are values between 0 and 1 that measure how bizarre an $x$ value is relative to the other $x$ values in the data. *A leverage near 1 is a very bizarre point*. Leverages that are small are similar to the other data. The sum of all the leverages in a simple linear regression is always 2, thus the average leverage is $2/n$. Points with leverages larger that $4/n$ or $6/n$ are often considered high

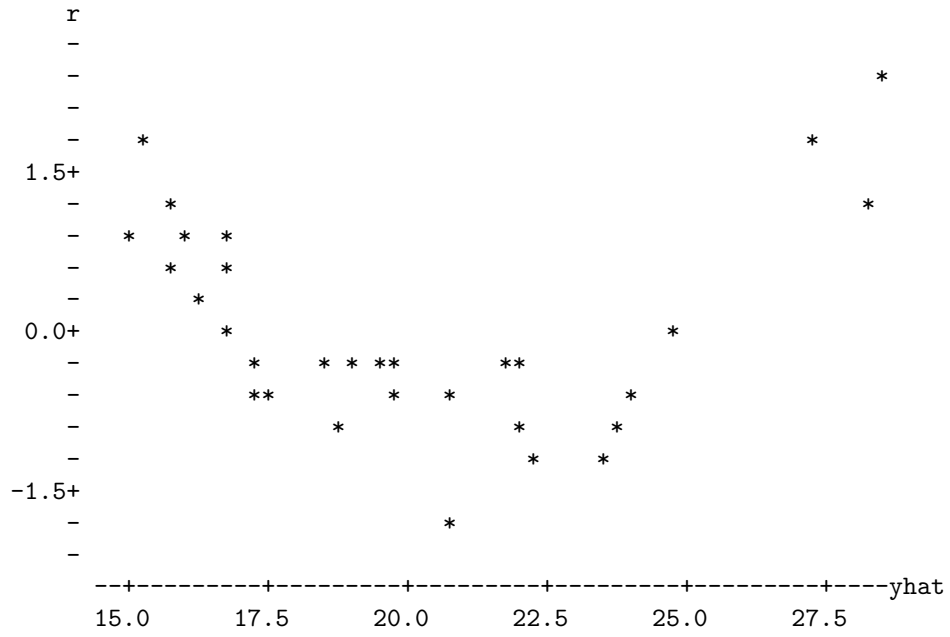



Figure 7.19: *Standardized residuals versus predicted values for Hooker data.*

leverage points. The concept of leverage will be discussed in more detail when we discuss multiple regression.

Outliers in the $y$ values can be detected from the *standardized deleted residuals*. Standardized deleted residuals are just standardized residuals, but the residual for the $i$th case is computed from a regression that does not include the $i$th case. For example, the third deleted residual is

$$\hat{\varepsilon}_{[3]} = y_3 - \hat{\beta}_{0[3]} - \hat{\beta}_{1[3]} x_3$$

where the estimates $\hat{\beta}_{0[3]}$ and $\hat{\beta}_{1[3]}$ are computed from a regression in which case 3 has been dropped from the data. The third standardized deleted residual is simply the third deleted residual divided by its standard error. The standardized deleted residuals really contain the same information as the standardized residuals; the largest standardized deleted residuals are also the largest standardized residuals. The main virtue of the standardized deleted residuals is that they can be compared to a $t(n-3)$ distribution to test whether they could reasonably have occurred when the model is true. The degrees of freedom in the test are $n-3$ because the simple linear regression model was fitted without the $i$th case so there are only $n-1$ data points in the fit and $(n-1)-2$ degrees of freedom for error.

If one compares the *largest* absolute standardized deleted residual to a $t$ distribution, one is essentially testing whether *every* case is an outlier. Thus a total of $n$ tests are being performed and the overall error rate from an individual $\alpha$ level test may be as high as $n\alpha$. For $n = 20$ and $\alpha = .05$, $n\alpha = 1$, so we can reasonably expect to 'find an outlier' in the *Coleman Report* data even when none exists. Obviously, one needs a more stringent requirement for declaring a case to be an outlier. The criterion for declaring a case to be an outlier should be something like significance at the $.05/n$ level rather than at the .05 level. This is just the Bonferroni adjustment discussed in the previous chapter. *Often the standardized deleted residuals are simply called $t$ residuals and denoted $t_i$.*

EXAMPLE 7.9.3. The leverages and standardized deleted residuals are given in Table 7.6 for the *Coleman Report* data with one predictor. Compared to the leverage rule of thumb $4/n = 4/20 = .2$, only case 15 has a noticeably large leverage. None of the cases is above the $6/n$ rule of thumb.

Table 7.6: *Outlier diagnostics for the* Coleman Report *data*

| Case | Leverages | Std. del. residuals | Case | Leverages | Std. del. residuals |
|---|---|---|---|---|---|
| 1 | 0.059362 | −0.15546 | 11 | 0.195438 | −1.44426 |
| 2 | 0.175283 | −0.12019 | 12 | 0.052801 | 0.61394 |
| 3 | 0.097868 | −1.86339 | 13 | 0.051508 | −0.49168 |
| 4 | 0.120492 | −0.28961 | 14 | 0.059552 | 0.14111 |
| 5 | 0.055707 | 0.10792 | 15 | 0.258992 | −0.84143 |
| 6 | 0.055179 | −1.35054 | 16 | 0.081780 | 0.19341 |
| 7 | 0.101914 | 0.63059 | 17 | 0.050131 | −1.41912 |
| 8 | 0.056226 | 0.07706 | 18 | 0.163429 | 2.52294 |
| 9 | 0.075574 | 1.00744 | 19 | 0.130304 | 0.63836 |
| 10 | 0.055783 | 1.92501 | 20 | 0.102677 | 0.24410 |

In simple linear regression, one does not really need to evaluate the leverages directly because the necessary information about bizarre $x$ values is readily available from the $x, y$ plot. In multiple regression with three or more predictor variables, leverages are vital because no one scatter plot can give the information on bizarre $x$ values. In the scatter plot of the *Coleman Report* data, Figure 7.1, there are no outrageous $x$ values, although there is a noticeable gap between the smallest four values and the rest. From Table 7.1 we see that the cases with the smallest $x$ values are 2, 11, 15, and 18. These cases also have the highest leverages reported in Table 7.6. The next two highest leverages are for cases 4 and 19; these have the largest $x$ values.

For an overall $\alpha = .05$ level test of the deleted residuals, the tabled value needed is

$$t\left(1 - \frac{.05}{2(20)}, 17\right) = 3.54\,.$$

None of the standardized deleted residuals approach this, so there is no evidence of any unaccountably bizarre $y$ values.

A handy way to identify cases with large leverages, residuals, standardized residuals, or standardized deleted residuals is with an index plot. This is simply a plot of the value against the case number as in Figure 7.20 for leverages. In this version of the plot, the symbol plotted is the last digit of the case number. □

*Minitab commands*

We now illustrate the Minitab commands necessary for obtaining the leverages and standardized deleted residuals for the *Coleman Report* data. Both sets of values are obtained by using subcommands of the regress command. The 'hi' subcommand gives leverages, while the 'tresid' subcommand gives standardized deleted residuals. The last command gives the index plot for leverages.

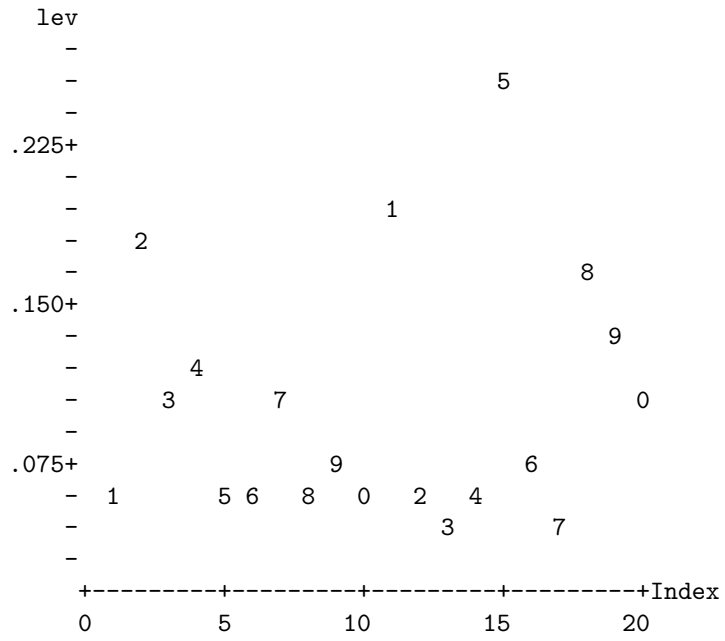



Figure 7.20: *Index plot of leverages for the* Coleman Report *data.*

```
MTB > names c1 'test' c2 'socio'
MTB > regress c1 on 1 c2;
SUBC> hi c13;
SUBC> tresid c14.
MTB > tsplot c13
```

*Effects of high leverage*

EXAMPLE 7.9.4. Figure 7.21 contains some data along with their least squares estimated line. The four points on the left form a perfect line with slope 1 and intercept 0. There is one high leverage point far away to the right. The actual data are given below along with their leverages.

| Case | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $y$ | 1 | 2 | 3 | 4 | $-3$ |
| $x$ | 1 | 2 | 3 | 4 | 20 |
| Leverage | .30 | .26 | .24 | .22 | .98 |

The case with $x = 20$ is an extremely high leverage point; it has a leverage of nearly 1. The estimated regression line is forced to go very nearly through this high leverage point. In fact, this plot has two clusters of points that are very far apart, so a rough approximation to the estimated line is the line that goes through the mean $x$ and $y$ values for each of the two clusters. This example has one cluster of four cases on the left of the plot and another cluster consisting solely of the one case on the right. The average values for the four cases on the left give the point $(\bar{x}, \bar{y}) = (2.5, 2.5)$. The one case on the right is $(20, -3)$. A little algebra shows the line through these two points to be $\hat{y} = 3.286 - 0.314x$. The estimated line using least squares turns out to be $\hat{y} = 3.128 - 0.288x$, which is not too different. The least squares line goes through the two points $(2.5, 2.408)$ and $(20, -2.632)$, so the least squares line is a little lower at $x = 2.5$ and a little higher at $x = 20$.

Obviously, the single point on the right of Figure 7.21 dominates the estimated straight line. For example, if the point on the right was $(20, 15)$, the estimated line would go roughly through

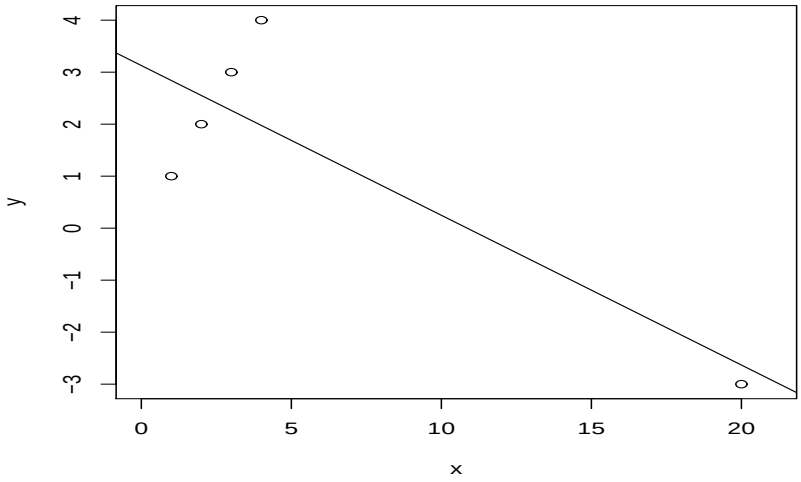



Figure 7.21: *Plot of y versus x.*

this point and $(2.5, 2.5)$. Substantially changing the $y$ value at $x = 20$ always gives an extremely different estimated line than the ones we just considered. Wherever the point on the right is, the estimated line follows it. This happens regardless of the fact that the four cases on the left follow a *perfect* straight line with slope 1 and intercept 0. The behavior of the four points on the left is almost irrelevant to the fitted line when there is a high leverage point on the right. They have an effect on the quality of the rough two-point approximation to the actual estimated line but their overall effect is small.

To summarize what can be learned from Figure 7.21, we have a reasonable idea about what happens to $y$ for $x$ values near the range 1 to 4 and we have some idea of what happens when $x$ is 20 but, barring outside information, we have not the slightest idea what happens to $y$ when $x$ is between 4 and 20. Fitting a line to the complete data suggests that we know something about the behavior of $y$ for any value of $x$ between 1 and 20. This is just silly! We would be better off to analyze the two clusters of points separately and to admit that learning about $y$ when $x$ is between 4 and 20 requires us to obtain data on $y$ when $x$ is between 4 and 20. In this example, the two separate statistical analyses are trivial. The cluster on the left follows a perfect line so we simply report that line. The cluster on the right is a single point so we report the point. □

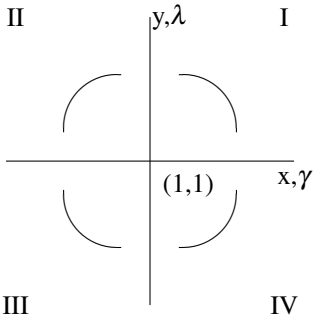



Figure 7.22: *The circle of $x, y$ transformations.*

## 7.10 Transformations

If the residuals show a problem with lack of fit, heteroscedasticity, or nonnormality, one way to deal with the problem is to try transforming the $y_i$s. Typically, this only works well when $y_{max}/y_{min}$ is reasonably large. The use of transformations is often a matter of trial and error. Various transformations are tried and the one that gives the best fitting model is used. In this context, the best fitting model should have residual plots indicating that the model assumptions are reasonably valid. The first approach to transforming the data should be to consider transformations that are suggested by any theory associated with the data collection. Another approach to choosing a transformation is to try a variance stabilizing transformation. These were discussed in Section 2.5 and are repeated below for data $y_i$ with $\mathrm{E}(y_i) = \mu_i$ and $\mathrm{Var}(y_i) = \sigma_i^2$.

Variance stabilizing transformations

| Data | Distribution | Mean, variance relationship | Transformation |
|---|---|---|---|
| Count | Poisson | $\mu_i \propto \sigma_i^2$ | $\sqrt{y_i}$ |
| Amount | Gamma | $\mu_i \propto \sigma_i$ | $\log(y_i)$ |
| Proportion | Binomial$/N$ | $\frac{\mu_i(1-\mu_i)}{N} \propto \sigma_i^2$ | $\sin^{-1}\left(\sqrt{y_i}\right)$ |

Whenever the data have the indicated mean, variance relationship, the corresponding variance stabilizing transformation should work reasonably well.

The shape of an $x, y$ plot can also suggest possible transformations to straighten it out. We consider power transformations of both $y$ and $x$, thus $y$ is transformed into, say, $y^{\lambda}$ and $x$ is transformed into $x^{\gamma}$. Note that $\lambda = 1$ and $\gamma = 1$ indicate no transformation. As we will justify later, we treat $\lambda = 0$ and $\gamma = 0$ as log transformations.

Figure 7.22 indicates the kinds of transformations appropriate for some different shapes of $x, y$ curves. For example, if the $x, y$ curve is similar to that in quadrant I, i.e., the $y$ values decrease as $x$ increases and the curve opens to the lower left, appropriate transformations involve increasing $\lambda$ or increasing $\gamma$ or both. Here we refer to increasing $\lambda$ and $\gamma$ relative to the no transformation values of $\lambda = 1$ and $\gamma = 1$. In particular, Figure 7.23 gives an $x, y$ plot for part of a cosine curve that is shaped like the curve in quadrant I. Figure 7.24 is a plot of the numbers after $x$ has been transformed into $x^{1.5}$ and $y$ has been transformed into $y^{1.5}$. Note that the curve in Figure 7.24 is much straighter than the curve in Figure 7.23. If the $x, y$ curve increases and opens to the lower right such as those in quadrant II, appropriate transformations involve increasing $\lambda$ or decreasing $\gamma$ or both. An $x, y$ curve

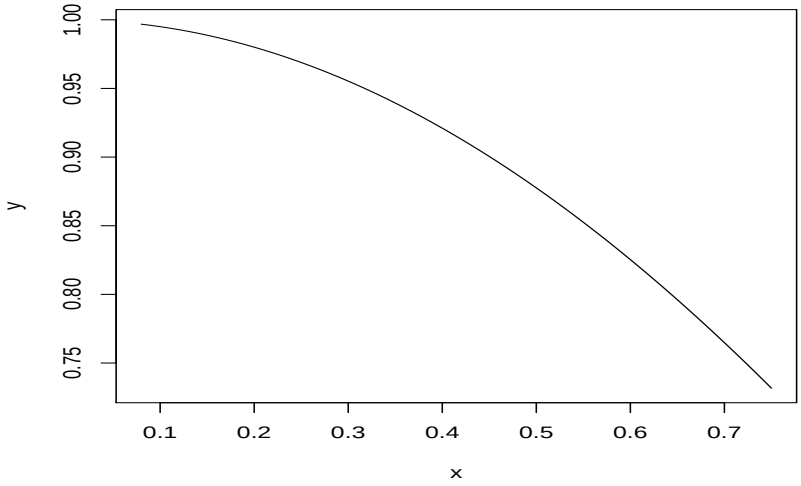



Figure 7.23: *Curved $x, y$ plot.*

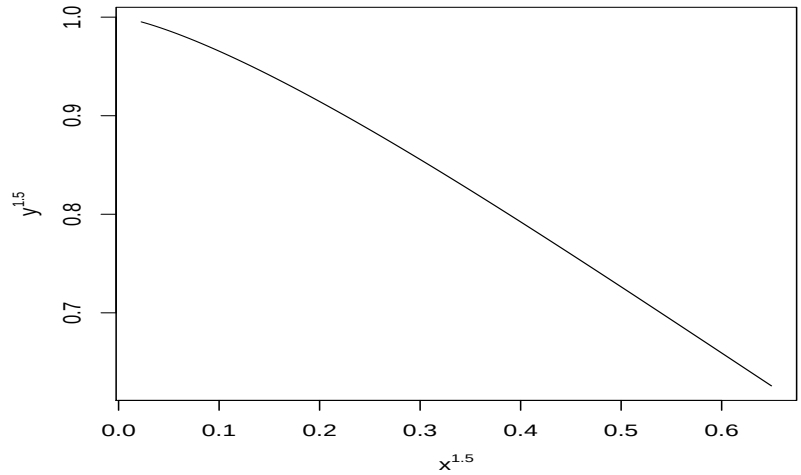



Figure 7.24: *Plot of $x^{1.5}, y^{1.5}$.*

similar to that in quadrant III suggests decreasing $\lambda$ or decreasing $\gamma$ or both. The graph given in Figure 7.22 is often referred to as *the circle of $x, y$ transformations*.

We established in the previous section that the Hooker data does not fit a straight line and that the scatter plot in Figure 7.18 increases with a slight tendency to open to the upper left. This is the same shaped curve as in quadrant IV of Figure 7.22. The circle of $x, y$ transformations suggests that to straighten the curve, we should try transformations with decreased values of $\lambda$ or increased values of $\gamma$ or both. Thus we might try transforming $y$ into $y^{1/2}$, $y^{1/4}$, $\log(y)$, or $y^{-1}$. Similarly, we might try transforming $x$ into $x^{1.5}$ or $x^2$.

To get a preliminary idea of how well various transformations work, we should do a series of

plots. We might begin by examining the four plots in which $y^{1/2}$, $y^{1/4}$, $\log(y)$, and $y^{-1}$ are plotted against $x$. We might then plot $y$ against both $x^{1.5}$ and $x^2$. We should also plot all possibilities involving one of $y^{1/2}$, $y^{1/4}$, $\log(y)$, and $y^{-1}$ plotted against one of $x^{1.5}$ and $x^2$ and we may need to consider other choices of $\lambda$ and $\gamma$. For the Hooker data, looking at these plots would probably only allow us to eliminate the worst transformations. Recall that Figure 7.18 looks remarkably straight and it is only after fitting a simple linear regression model and examining residuals that the lack of fit (the curvature of the $x, y$ plot) becomes apparent. Evaluating the transformations would require fitting a simple linear regression for every pair of transformed variables that has a plot that looks reasonably straight.

Observe that many of the power transformations considered here break down with values of $y$ that are negative. For example, it is difficult to take square roots and logs of negative numbers. Fortunately, data are often positive or at least nonnegative. Measured amounts, counts and proportions are almost always nonnegative. When problems arise, a small constant is often added to all cases so that they all become positive. Of course, it is unclear what constant should be added.

Obviously, the circle of transformations, just like the variance stabilizing transformations, provides only suggestions on how to transform the data. The process of choosing a particular transformation remains one of trial and error. We begin with reasonable candidates and examine how well these transformations agree with the simple linear regression model. When we find a transformation that agrees well with the assumptions of simple linear regression, we proceed to analyze the data. Obviously, an alternative to transforming the data is to change the model. In the next section we consider a new class of models that incorporate transformations of the $x$ variable. In the remainder of this section, we focus on a systematic method for choosing a transformation of $y$.

### *7.10.1 Box–Cox transformations*

We now consider a systematic method, introduced by Box and Cox (1964), for choosing a power transformation. Consider the family of power transformations, say, $y_i^\lambda$. This includes the square root transformation as the special case $\lambda = 1/2$ and other interesting transformations such as the reciprocal transformation $y_i^{-1}$. By making a minor adjustment, we can bring log transformations into the power family. Consider the transformations

$$y_i^{(\lambda)} = \begin{cases} \left(y_i^\lambda - 1\right)/\lambda & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases}.$$

For any fixed $\lambda \neq 0$, the transformation $y_i^{(\lambda)}$ is equivalent to $y_i^\lambda$, because the difference between the two transformations consists of subtracting a constant and dividing by a constant. In other words, fitting the model

$$y_i^\lambda = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is equivalent to fitting the model

$$y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

although the regression parameters in the two models have slightly different meanings. While the transformation $\left(y_i^\lambda - 1\right)/\lambda$ is undefined for $\lambda = 0$, as $\lambda$ approaches 0, $\left(y_i^\lambda - 1\right)/\lambda$ approaches $\log(y_i)$, so the log transformation fits in naturally.

Unfortunately, the results of fitting models to $y_i^{(\lambda)}$ with different values of $\lambda$ are not directly comparable. Thus it is difficult to decide which transformation in the family to use. This problem is easily evaded (cf. Cook and Weisberg, 1982) by further modifying the family of transformations so that the results of fitting with different $\lambda$s are comparable. Let $\tilde{y}$ be the geometric mean of the $y_i$s, i.e.,

$$\tilde{y} = \left[\prod_{i=1}^n y_i\right]^{1/n} = \exp\left[\frac{1}{n}\sum_{i=1}^n \log(y_i)\right]$$

Table 7.7: *Choice of power transformation*

| $\lambda$ | 1/2 | 1/3 | 1/4 | 0 | −1/4 | −1/2 |
|---|---|---|---|---|---|---|
| $SSE(\lambda)$ | 1.21 | 0.87 | 0.78 | 0.79 | 1.21 | 1.98 |

and define the family of transformations

$$z_i^{(\lambda)} = \begin{cases} \left[y_i^{\lambda} - 1\right] / \left[\lambda \tilde{y}^{\lambda-1}\right] & \lambda \neq 0 \\ \tilde{y}\log(y_i) & \lambda = 0 \end{cases}.$$

The results of fitting the model

$$z_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \varepsilon_i$$

can be summarized via $SSE(\lambda)$. These values are directly comparable for different values of $\lambda$. The choice of $\lambda$ that yields the smallest $SSE(\lambda)$ is the best fitting model. (It maximizes the likelihood with respect to $\lambda$.) Actually, *this method of choosing a transformation works for any ANOVA or regression model*.

Box and Draper (1987, p. 290) discuss finding a confidence interval for the transformation parameter $\lambda$. An approximate $(1-\alpha)100\%$ confidence interval consists of all $\lambda$ values that satisfy

$$\log SSE(\lambda) - \log SSE(\hat{\lambda}) \leq \chi^2(1-\alpha, 1)/dfE$$

where $\hat{\lambda}$ is the value of $\lambda$ that minimizes $SSE(\lambda)$. When $y_{max}/y_{min}$ is not large, the interval tends to be wide.

EXAMPLE 7.10.1. *Hooker data*

In the previous section, we found that Hooker's data on atmospheric pressure and boiling points displayed a lack of fit when regressing pressure on temperature. We now consider using power transformations to eliminate the lack of fit.

Table 7.7 contains $SSE(\lambda)$ values for some reasonable choices of $\lambda$. Assuming that $SSE(\lambda)$ is a very smooth (convex) function of $\lambda$, the best $\lambda$ value is probably between 0 and 1/4. If the curve being minimized is very flat between 0 and 1/4, there is a possibility that the minimizing value is between 1/4 and 1/3. One could pick more $\lambda$ values and compute more $SSE(\lambda)$s but I have a bias towards simple transformations. (They are easier to sell to clients.)

The log transformation of $\lambda = 0$ is simple (certainly simpler than the fourth root) and $\lambda = 0$ is near the optimum, so we will consider it further. We now use the simple log transformation, rather than adjusting for the geometric mean. The usual summary tables are given below.

| Predictor | $\hat{\beta}_k$ | $\mathrm{SE}(\hat{\beta}_k)$ | $t$ | $P$ |
|---|---|---|---|---|
| Constant | −1.02214 | 0.03365 | −30.38 | 0.000 |
| Temp. | 0.0208698 | 0.0001753 | 119.08 | 0.000 |

Analysis of variance: log Hooker data

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 1 | 0.99798 | 0.99798 | 14180.91 | 0.000 |
| Error | 29 | 0.00204 | 0.00007 | | |
| Total | 30 | 1.00002 | | | |

The coefficient of determination is again extremely high, $R^2 = 99.8\%$. The plot of the standardized residuals versus the predicted values is given in Figure 7.25. There is no obvious lack of fit or inconstancy of variances. Figure 7.26 contains a normal plot of the standardized residuals. The

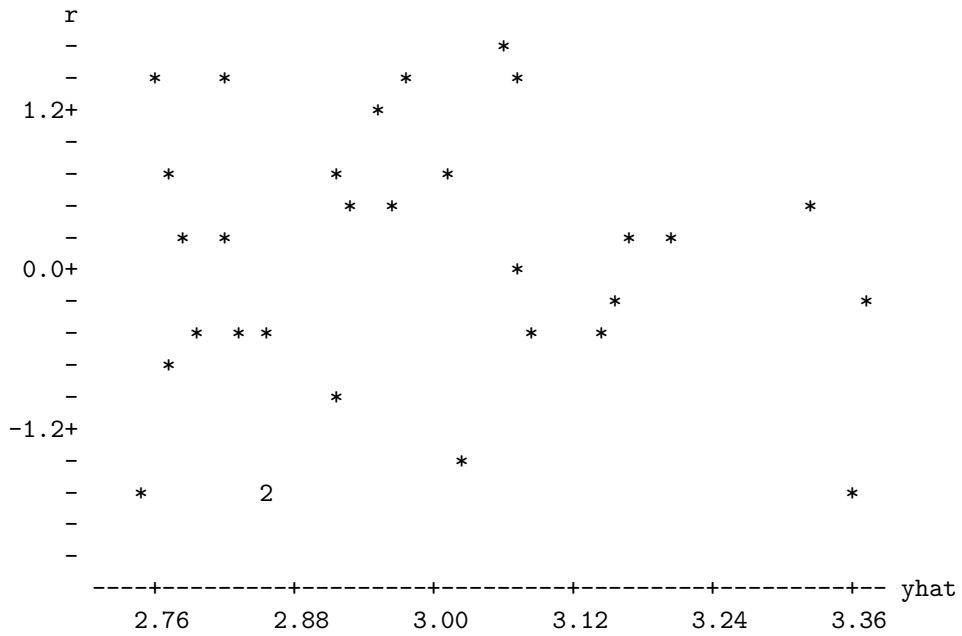



Figure 7.25: *Standardized residuals versus predicted values, logs of Hooker data.*

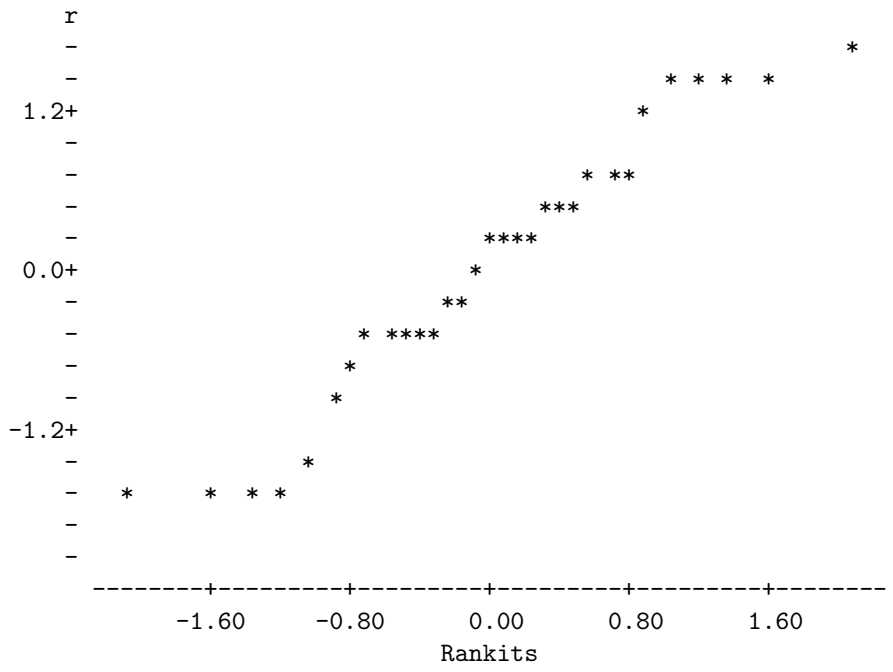



Figure 7.26: *Normal plot for logs of Hooker data,* $W' = 0.960$.

normal plot is not horrible but it is not wonderful either. There is a pronounced shoulder at the bottom and perhaps even an S shape.

If we are interested in the mean value of log pressure for a temperature of 205°F, the estimate is $3.2562 = -1.02214 + .0208698(205)$ with a standard error of 0.00276 and a 95% confidence interval of $(3.2505, 3.2618)$. In the original units, the estimate is $e^{3.2562} = 25.95$ and the confidence interval becomes $(e^{3.2505}, e^{3.2618})$ or $(25.80, 26.10)$. The point prediction for a new log observation at 205°F has the same value as the point estimate and has a 95% prediction interval of $(3.2381, 3.2742)$. In the original units, the prediction is again 25.95 and the prediction interval becomes $(e^{3.2381}, e^{3.2742})$ or $(25.49, 26.42)$. □

One way to test whether a transformation is needed is to use a *constructed variable* as introduced by Atkinson (1973). Let

$$w_i = y_i [\log(y_i/\tilde{y}) - 1]$$

and fit the multiple regression model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i.$$

Multiple regression gives results similar to those for simple linear regression; typical output includes estimates of the $\beta$s, standard errors, $t$ statistics, and an ANOVA table. A test of $H_0 : \beta_2 = 0$ gives an approximate test that no transformation is needed. The test is performed using the standard methods of Chapter 3. Details are illustrated in the example below and discussed in the chapters on multiple regression. In addition, the estimate $\hat{\beta}_2$ provides, indirectly, an estimate of $\lambda$,

$$\hat{\lambda} = 1 - \hat{\beta}_2.$$

Frequently, this is not a very good estimate of $\lambda$ but it gives an idea of where to begin a search for good $\lambda$s.

EXAMPLE 7.10.2. *Hooker data*
Performing the multiple regression of pressure on both temperature and the constructed variable $w$ gives the following results.

| Predictor | $\hat{\beta}_k$ | SE($\hat{\beta}_k$) | $t$ | $P$ |
|---|---|---|---|---|
| Constant | −43.426 | 2.074 | −20.94 | 0.000 |
| Temperature | 0.411816 | 0.004301 | 95.75 | 0.000 |
| $w$ | 0.80252 | 0.07534 | 10.65 | 0.000 |

The $t$ statistic is $10.65 = .80252/.07534$ for testing that the regression coefficient of the constructed variable is 0. The $P$ value is 0.000, which strongly indicates the need for a transformation. The estimate of $\lambda$ is

$$\hat{\lambda} = 1 - \hat{\beta}_2 = 1 - 0.80 = .2,$$

which is consistent with what we learned from Table 7.7. From Table 7.7 we suspected that the best transformation would be between 0 and .25. Of course this estimate of $\lambda$ is quite crude, finding the 'best' transformation requires a more extensive version of Table 7.7. I limited the choices of $\lambda$ in Table 7.7 because I was unwilling to consider transformations that I did not consider simple. □

*Computational techniques*

Below are given Minitab commands for performing the Box–Cox transformations and the constructed variable test. To perform multiple regression using the 'regress' command, you need to specify the number of predictor variables, in this case 2.

```
MTB > name c1 'temp' c2 'press'
MTB > note     CONSTRUCT THE GEOMETRIC MEAN
MTB > let c9 = loge(c2)
MTB > mean c9 k1
   MEAN    =      2.9804
MTB > let k2 = expo(k1)
MTB > note     PRINT THE GEOMETRIC MEAN
MTB > print k2
K2       19.6960
MTB > note     CONSTRUCT THE z VARIABLES
MTB > note     FOR DIFFERENT LAMBDAS
MTB > let c20=(c2**.5-1)/(.5*k2**(.5-1))
MTB > let c21=(c2**.25-1)/(.25*k2**(.25-1))
MTB > let c22=(c2**.333333-1)/(.333333*k2**(.333333-1))
MTB > let c23=loge(c2)*k2
MTB > let c24=(c2**(-.5) - 1)/(-.5*k2**(-1.5))
MTB > let c25=(c2**(-.25) -1)/(-.25*k2**(-1.25))
MTB > note     REGRESS z FOR LAMBDA = .5 ON c1
MTB > regress c20 on 1 c1
MTB > note     4 MORE REGRESSIONS ARE NECESSARY
MTB > note
MTB > note     CONSTRUCT THE VARIABLE w
MTB > let c3=c2*(c9-k1-1)
MTB > note     PERFORM THE MULTIPLE REGRESSION ON x AND THE
MTB > note     CONSTRUCTED VARIABLE w
MTB > regress c2 on 2 c1 c3
```

*Transforming the predictor variable*

Weisberg (1985, p. 156) suggests applying a log transformation to the predictor variable $x$ whenever $x_{max}/x_{min}$ is larger than 10 or so. There is also a procedure, originally due to Box and Tidwell (1962), that is akin to the constructed variable test but that is used for checking the need to transform $x$. As presented by Weisberg, this procedure consists of fitting the original model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

to obtain $\hat{\beta}_1$ and then fitting the model

$$y_i = \eta_0 + \eta_1 x_i + \eta_2 x_i \log(x_i) + \varepsilon_i.$$

Here, $x_i \log(x_i)$ is just an additional predictor variable that we compute from the values of $x_i$. The test of $H_0 : \eta_2 = 0$ is a test for whether a transformation of $x$ is needed. If $\eta_2 \neq 0$, transforming $x$ into $x^\gamma$ is suggested where a rough estimate of $\gamma$ is

$$\hat{\gamma} = \frac{\hat{\eta}_2}{\hat{\beta}_1} + 1$$

and $\gamma = 0$ is viewed as the log transformation. Typically, only $\gamma$ values between about $-2$ and 2 are considered useable. Of course none of this is going to make any sense if $x$ takes on negative values, and if $x_{max}/x_{min}$ is not large, computational problems may occur when trying to fit a model that contains both $x_i$ and $x_i \log(x_i)$.

## 7.11 Polynomial regression

With Hooker's data, the simple linear regression of pressure on temperature showed a lack of fit. In the previous section, we used a power transformation in an attempt to eliminate the lack of fit. In this section we introduce an alternative method, a special case of multiple regression called *polynomial regression*. With a single predictor variable $x$, we can try to eliminate lack of fit by fitting larger models. In particular, we can fit the *quadratic* (parabolic) model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

We could also try a *cubic* model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i,$$

the *quartic* model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i,$$

or even higher degree polynomials. If we view our purpose as finding good, easily interpretable approximate models for the data, *high degree polynomials can behave poorly*. As we will see later, the process of fitting the observed data can cause high degree polynomials to give very erratic results in areas very near the observed data. A good approximate model should work well, not only at the observed data, but also near it. Thus, we should focus on low degree polynomials.

EXAMPLE 7.11.1. We again examine Hooker's data. Computer programs give output for polynomial regression that is very similar to that for simple linear regression. Typical summary tables for fitting the quadratic model are given below.

| Predictor | $\hat{\beta}_k$ | $\mathrm{SE}(\hat{\beta}_k)$ | $t$ | $P$ |
|---|---|---|---|---|
| Constant | 88.02 | 13.93 | 6.32 | 0.000 |
| Temp. | −1.1295 | 0.1434 | −7.88 | 0.000 |
| Temp squared | 0.0040330 | 0.0003682 | 10.95 | 0.000 |

Analysis of variance

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 2 | 447.15 | 223.58 | 8984.23 | 0.000 |
| Error | 28 | 0.70 | 0.02 | | |
| Total | 30 | 447.85 | | | |

The $MSE$, regression parameter estimates, and standard errors are used in the usual way. The $t$ statistics and $P$ values are for the two-sided tests of whether the corresponding $\beta$ parameters are 0. The $t$ statistic for $\beta_2$ is $-7.88$, which is highly significant, so the quadratic model accounts for a significant amount of the lack of fit displayed by the simple linear regression model. (It is not clear yet that the quadratic accounts for all of the lack of fit.)

Usually, *the only interesting test for a regression coefficient is the one for the highest term in the polynomial*. In particular, it usually makes little sense to have a quadratic (second degree) model that does not include a first degree term, so there is little point in testing $\beta_1 = 0$. One reason for this is that simple linear transformations of the predictor variable change the roles of lower order terms. For example, the Hooker data uses temperature measured in Fahrenheit as a predictor variable. While it is not actually the case, suppose the quadratic model for the Hooker data was consistent with $\beta_1 = 0$. If we then changed to measuring temperature in Celsius, we would be unlikely to have a new quadratic model that is still consistent with $\beta_1 = 0$. When there is a quadratic term in the model, a linear term based on Fahrenheit measurements has a completely different meaning than a linear term based on Celsius measurements. On the other hand, the Fahrenheit and Celsius quadratic

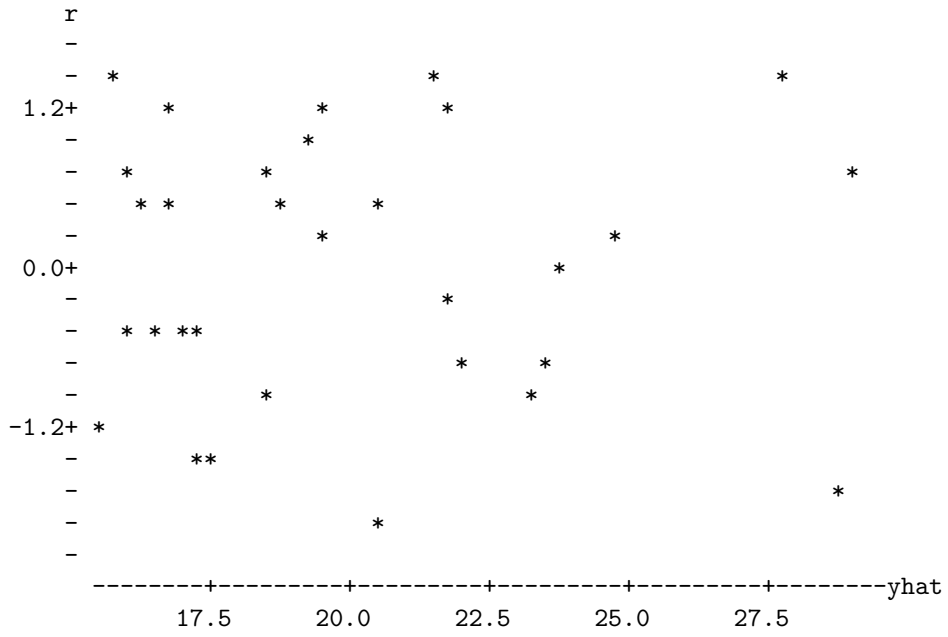



Figure 7.27: *Standardized residuals versus predicted values, quadratic model.*

models that include linear terms and intercepts are equivalent, just as the simple linear regressions based on Fahrenheit and Celsius are equivalent.

We will not discuss the ANOVA table in detail, but note that with two predictors, $x$ and $x^2$, there are 2 degrees of freedom for regression. In general, if we fit a polynomial of degree $a$, there will be $a$ degrees of freedom for regression, one degree of freedom for every term other than the intercept. Correspondingly, when fitting a polynomial of degree $a$, there are $n-a-1$ degrees of freedom for error. *The ANOVA table $F$ statistic provides a test of whether the quadratic model explains the data better than the model with only an intercept*.

The coefficient of determination is computed and interpreted as before. It is the *SSReg* divided by the *SSTot*, so it measures the amount of the total variability that is explained by the predictor variables temperature and temperature squared. For these data, $R^2 = 99.8\%$, which is an increase from 99.2% for the simple linear regression model. It is not appropriate to compare the $R^2$ for this model to the $R^2$ from the log transformed model of the previous section because they are computed from data that use different scales.

The standardized residual plots are given in Figures 7.27 and 7.28. The plot against the predicted values looks good, just as it did for the transformed data examined in the previous section. The normal plot for this model has a shoulder at the top but it looks much better than the normal plot for the simple linear regression on the log transformed data.

If we are interested in the mean value of pressure for a temperature of 205°F, the quadratic model estimate is (up to a little of round off error)

$$\hat{y} = 25.95 = 88.02 - 1.1295\,(205) + .004033\,(205)^2.$$

The standard error (as reported by the computer program) is 0.0528 and a 95% confidence interval is $(25.84, 26.06)$. This compares to a point estimate of 25.95 and a 95% confidence interval of $(25.80, 26.10)$ obtained in the previous section from regressing the log of pressure on temperature. The quadratic model *prediction* for a new observation at 205°F is again 25.95 with a 95% prediction interval of $(25.61, 26.29)$. The corresponding prediction interval from the log transformed data

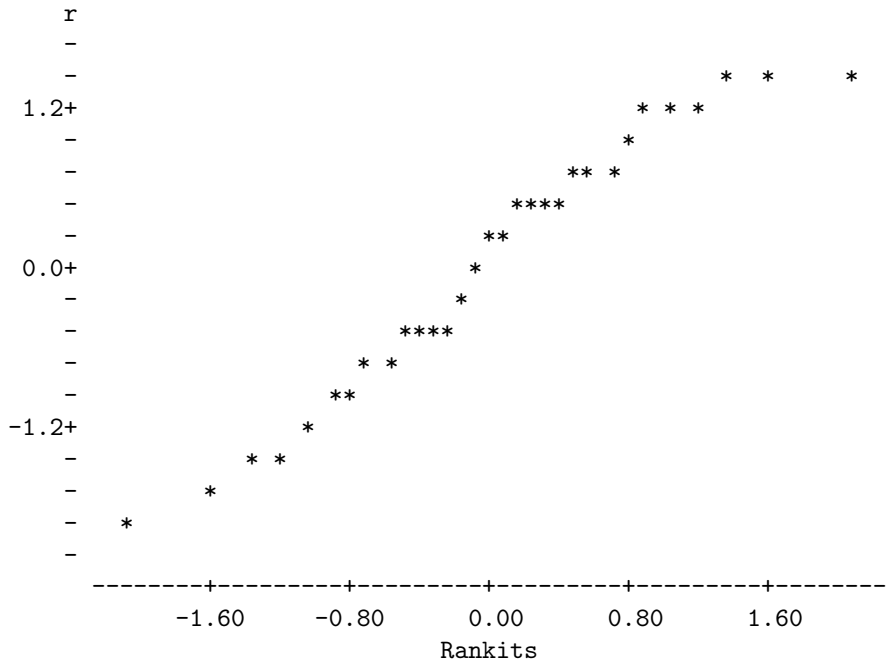



Figure 7.28: *Normal plot for quadratic model,* $W' = 0.966$.

is $(25.49, 26.42)$. In this example, the results of the two methods for dealing with lack of fit are qualitatively very similar, at least at 205°F.

Finally, we tried fitting a cubic model to these data. The cubic model suffers from substantial numerical instability. (Some computer programs object to fitting it.) This may be related to the fact that the $R^2$ is so high. The $\beta_3$ coefficient does not seem to be significantly different from 0, so considering the good residual plots, the quadratic model seems adequate. (One easy way to improve numerical stability is to adjust the predictor variables for their mean as in Section 7.6. In other words, one builds a polynomial using powers of the predictor variable $x_i - \bar{x}_\cdot$.) □

EXAMPLE 7.11.2. We now present a simple example that illustrates two points: that leverages depend on the model and that high order polynomials can fit the data in very strange ways. The data for the example are given below.

| Case | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $y$ | 0.445 | 1.206 | 0.100 | −2.198 | 0.536 | 0.329 | −0.689 |
| $x$ | 0.0 | 0.5 | 1.0 | 10.0 | 19.0 | 19.5 | 20.0 |

I selected the $x$ values. The $y$ values are a sample of size 7 from a $N(0,1)$ distribution. Note that with seven distinct $x$ values, we can fit a polynomial of degree 6.

The data are plotted in Figure 7.29. Just by chance (honest folks), I observed a very small $y$ value at $x = 10$, so the data appear to follow a parabola that opens up. The small $y$ value at $x = 10$ totally dominates the impression given by Figure 7.29 If the $y$ value at $x = 10$ had been near 3 rather than near $-2$, the data would appear to be a parabola that opens down. If the $y$ value had been between 0 and 1, the data would appear to fit a line with a slightly negative slope. When thinking about fitting a parabola, the case with $x = 10$ is an extremely high leverage point.

Depending on the $y$ value at $x = 10$, the data suggest a parabola opening up, a parabola opening down, or that we do not need a parabola to explain the data. Regardless of the $y$ value observed at $x = 10$, the fitted parabola must go nearly through the point $(10, y)$. On the other hand, if we think

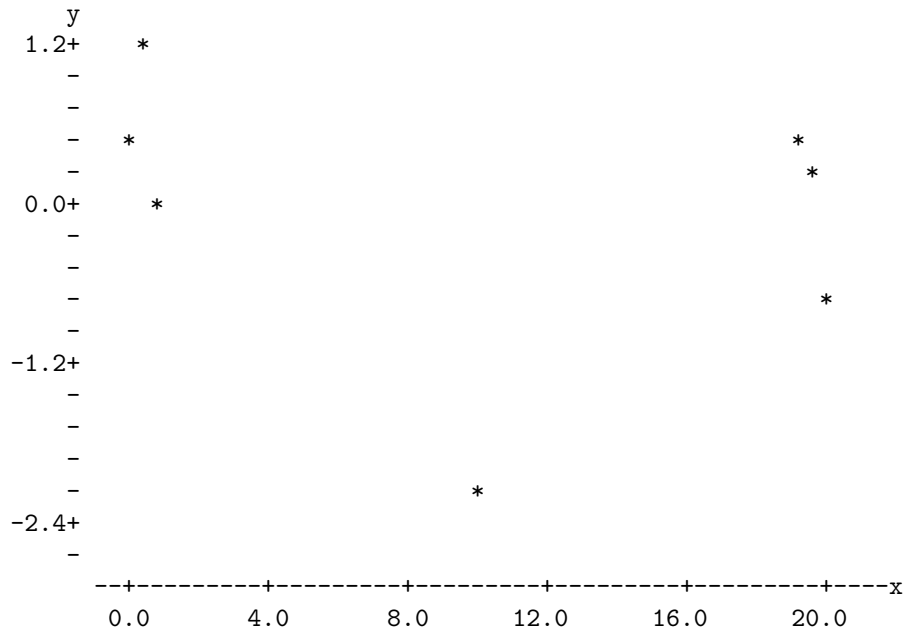



Figure 7.29: *Plot of y versus x.*

Table 7.8: *Leverages*

| | Model | | | | | |
|---|---|---|---|---|---|---|
| $x$ | Linear | Quadratic | Cubic | Quartic | Quintic | Hexic |
| 0.0 | 0.33 | 0.40 | 0.64 | 0.87 | 0.94 | 1.00 |
| 0.5 | 0.31 | 0.33 | 0.33 | 0.34 | 0.67 | 1.00 |
| 1.0 | 0.29 | 0.29 | 0.55 | 0.80 | 0.89 | 1.00 |
| 10.0 | 0.14 | 0.96 | 0.96 | 1.00 | 1.00 | 1.00 |
| 19.0 | 0.29 | 0.29 | 0.55 | 0.80 | 0.89 | 1.00 |
| 19.5 | 0.31 | 0.33 | 0.33 | 0.34 | 0.67 | 1.00 |
| 20.0 | 0.33 | 0.40 | 0.64 | 0.87 | 0.94 | 1.00 |

only about fitting a line to these data, the small $y$ value at $x = 10$ has much less effect. In fitting a line, the value $y = -2.198$ will look unusually small (it will have a very noticeable standardized residual), but it will not force the fitted line to go nearly through the point $(10, -2.198)$.

Table 7.8 gives the leverages for all of the polynomial models that can be fitted to these data. Note that there are no large leverages for the simple linear regression model (the linear polynomial). For the quadratic (parabolic) model, all of the leverages are reasonably small except the leverage of .96 at $x = 10$ which very nearly equals 1. Thus, in the quadratic model, the value of $y$ at $x = 10$ dominates the fitted polynomial. The cubic model has extremely high leverage at $x = 10$, but the leverages are also beginning to get large at $x = 0, 1, 19, 20$. For the quartic model, the leverage at $x = 10$ is 1 to two decimal places; the leverages for $x = 0, 1, 19, 20$ are also nearly 1. The same pattern continues with the quintic model but the leverages at $x = 0.5, 19.5$ are also becoming large. Finally, with the sixth degree (hexic) polynomial, all of the leverages are exactly one. This indicates that the sixth degree polynomial has to go through every data point exactly and thus every data point is extremely influential on the estimate of the sixth degree polynomial.

As we fit larger polynomials, we get more high leverage cases (and more numerical instability). *The estimated polynomials must go very nearly through all high leverage cases. To accomplish this*

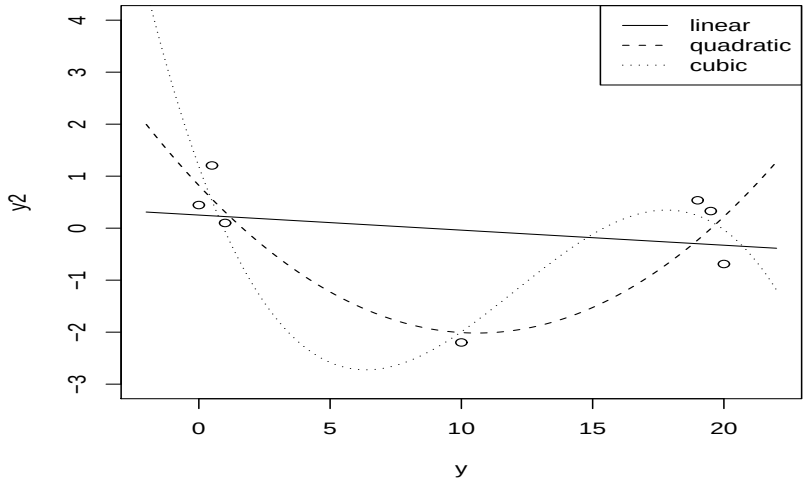



Figure 7.30: *Plots of linear (solid), quadratic (long dashes), and cubic (short dashes) regression curves.*

*the estimated polynomials may get very strange*. Below we give all of the fitted polynomials for these data.

| Model | | | Estimated polynomial |
|---|---|---|---|
| Linear | $\hat{y}$ | = | $0.252 - 0.029x$ |
| Quadratic | $\hat{y}$ | = | $0.822 - 0.536x + 0.0253x^2$ |
| Cubic | $\hat{y}$ | = | $1.188 - 1.395x + 0.1487x^2 - 0.0041x^3$ |
| Quartic | $\hat{y}$ | = | $0.713 - 0.141x - 0.1540x^2 + 0.0199x^3$ $- 0.00060x^4$ |
| Quintic | $\hat{y}$ | = | $0.623 + 1.144x - 1.7196x^2 + 0.3011x^3$ $- 0.01778x^4 + 0.000344x^5$ |
| Hexic | $\hat{y}$ | = | $0.445 + 3.936x - 5.4316x^2 + 1.2626x^3$ $- 0.11735x^4 + 0.004876x^5$ $- 0.00007554x^6$ |

Figures 7.30 and 7.31 contain graphs of these estimated polynomials.

Figure 7.30 contains the estimated linear, quadratic, and cubic polynomials. The linear and quadratic curves fit about as one would expect from looking at the scatter plot Figure 7.29. For $x$ values near the range 0 to 20, we could use these curves to predict $y$ values and get reasonable, if not necessarily good, results. One could not say the same for the estimated cubic polynomial. The cubic curve takes on $\hat{y}$ values near $-3$ for some $x$ values that are near 6. The $y$ values in the data are between about $-2$ and 1.2; nothing in the data suggests that $y$ values near $-3$ are likely to occur. Such predicted values are entirely the product of fitting a cubic polynomial. If we really knew that a cubic polynomial was correct for these data, the estimated polynomial would be perfectly appropriate. But most often we use polynomials to approximate the behavior of the data and for these data the cubic polynomial gives a poor approximation.

Figure 7.31 gives the estimated quartic, quintic, and hexic polynomials. Note that the scale on the $y$ axis has changed drastically from Figure 7.30. Qualitatively, the fitted polynomials behave like the cubic except their behavior is even worse. These polynomials do very strange things everywhere except near the observed data.

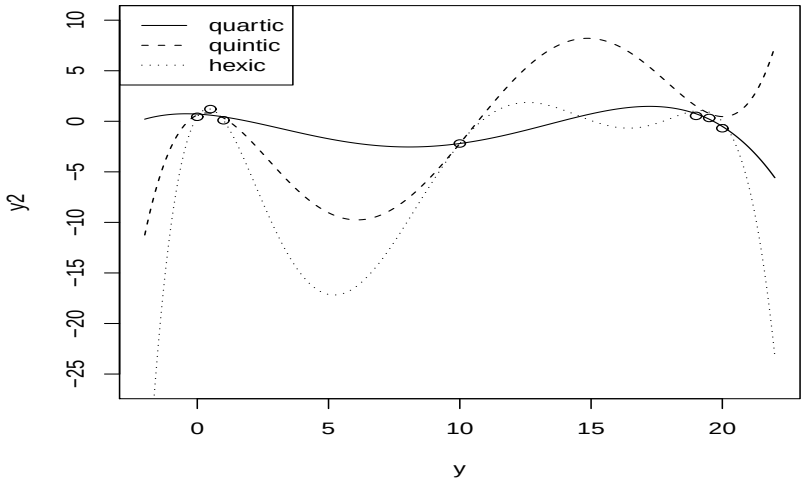



Figure 7.31: *Plots of quartic (solid), quintic (long dashes), and hexic (short dashes) regression curves.*

Another phenomenon that sometimes occurs when fitting large models to data is that the mean squared error gets unnaturally small. Table 7.9 gives the analysis of variance tables for all of the polynomial models. Our original data were a sample from a $N(0,1)$ distribution. The data were constructed with no regression structure so the best estimate of the variance comes from the total line and is $7.353/6 = 1.2255$. This value is a reasonable estimate of the true value 1. The *MSE* from the simple linear regression model also provides a reasonable estimate of $\sigma^2 = 1$. The larger models do not work as well. Most have variance estimates near .5, while the hexic model does not even allow an estimate of $\sigma^2$ because it fits every data point perfectly. By fitting models that are too large one can often make the *MSE* artificially small. For example, the quartic model has a *MSE* of .306 and an $F$ statistic of 5.51; if it were not for the small value of $dfE$, such an $F$ value would be highly significant. *If you find a large model that has an unnaturally small MSE with a reasonable number of degrees of freedom, everything can appear to be significant even though nothing you look at is really significant*.

Just as the mean squared error often gets unnaturally small when fitting large models, $R^2$ gets unnaturally large. As we have seen, there can be no possible reason to use a larger model than the quadratic with its $R^2$ of .71, but the cubic, quartic, quintic, and hexic models have $R^2$s of .78, .92, .93, and 1, respectively. □

Table 7.9: *Analysis of variance tables*

Simple linear regression

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 1 | 0.457 | 0.457 | 0.33 | 0.59 |
| Error | 5 | 6.896 | 1.379 | | |
| Total | 6 | 7.353 | | | |

Quadratic model

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 2 | 5.185 | 2.593 | 4.78 | 0.09 |
| Error | 4 | 2.168 | 0.542 | | |
| Total | 6 | 7.353 | | | |

Cubic model

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 3 | 5.735 | 1.912 | 3.55 | 0.16 |
| Error | 3 | 1.618 | 0.539 | | |
| Total | 6 | 7.353 | | | |

Quartic model

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 4 | 6.741 | 1.685 | 5.51 | 0.16 |
| Error | 2 | 0.612 | 0.306 | | |
| Total | 6 | 7.353 | | | |

Quintic model

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 5 | 6.856 | 1.371 | 2.76 | 0.43 |
| Error | 1 | 0.497 | 0.497 | | |
| Total | 6 | 7.353 | | | |

Hexic model

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 6 | 7.353 | 1.2255 | — | — |
| Error | 0 | 0.000 | — | | |
| Total | 6 | 7.353 | | | |

*Minitab commands*

Below we illustrate Minitab commands for fitting quadratic, cubic, and quartic models. These include the prediction subcommand used with the quadratic model for $x = 205$. Note that the prediction subcommand requires us to enter both the value of $x$ and the value of $x^2$ when using the quadratic model.

```
MTB > names c1 'y' c2 'x'
MTB > note     FIT QUADRATIC MODEL
MTB > let c22=c2**2
MTB > regress c2 on 2 c2 c22;
SUBC> pred 205 42025.
MTB > note     FIT CUBIC MODEL
MTB > let c23=c2**3
MTB > regress c1 on 3 c2 c22 c23
MTB > note     FIT QUARTIC MODEL
MTB > let c24=c2**4
MTB > regress c1 on 4 c2 c22-c24
```

## 7.12 Polynomial regression and one-way ANOVA

The main reason for introducing polynomial regression at this point is to exploit its relationships with analysis of variance. In some analysis of variance problems, the treatment groups are determined by quantitative levels of a factor. For example, one might take observations on the depth of hole made by a drill press in a given amount of time with 20, 30, or 40 pounds of downward thrust applied. The groups are determined by the quantitative levels, 20, 30, and 40. In such a situation we could fit a one-way analysis of variance with three groups, or we could fit a simple linear regression model. Simple linear regression is appropriate because all the data come as pairs. The pairs are $(x_i, y_{ij})$, where $x_i$ is the numerical level of thrust and $y_{ij}$ is the depth of the hole on the $j$th trial with $x_i$ pounds of downward thrust. Not only can we fit a simple linear regression, but we can fit polynomials to the data. In this example, we could fit no polynomial above second degree (quadratic), because three points determine a parabola and we only have three distinct $x$ values. If we ran the experiment with 20, 25, 30, 35, and 40 pounds of thrust, we could fit at most a fourth degree (quartic) polynomial because five points determine a fourth degree polynomial and we would only have five $x$ values.

*In general, some number $a$ of distinct $x$ values allows fitting of an $a-1$ degree polynomial.* Moreover, *fitting the $a-1$ degree polynomial is equivalent to fitting the one-way ANOVA with groups defined by the $a$ different $x$ values*. However, as discussed in the previous section, fitting high degree polynomials is often a very questionable procedure. The problem is not with how the model fits the observed data but with the suggestions that a high degree polynomial makes about the behavior of the process for $x$ values other than those observed. In the example with 20, 25, 30, 35, and 40 pounds of thrust, the quartic polynomial will fit as well as the one-way ANOVA model but the quartic polynomial may have to do some very weird things in the areas between the observed $x$ values. Of course, the ANOVA model gives no indications of behavior for $x$ values other than those that were observed. When performing regression, we usually like to have some smooth fitting model giving predictions that, in some sense, interpolate between the observed data points. High degree polynomials often fail to achieve this goal.

EXAMPLE 7.12.1. Beineke and Suddarth (1979) and Devore (1991, p. 380) consider data on roof supports involving trusses that use light gauge metal connector plates. Their dependent variable is an axial stiffness index (ASI) measured in kips per inch. The predictor variable is the length of the light gauge metal connector plates. The data are given in Table 7.10 in a format consistent with performing a regression analysis on the data.

Table 7.10: *Axial stiffness index data*

| Plate | ASI | Plate | ASI | Plate | ASI | Plate | ASI | Plate | ASI |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 309.2 | 6 | 402.1 | 8 | 392.4 | 10 | 346.7 | 12 | 407.4 |
| 4 | 409.5 | 6 | 347.2 | 8 | 366.2 | 10 | 452.9 | 12 | 441.8 |
| 4 | 311.0 | 6 | 361.0 | 8 | 351.0 | 10 | 461.4 | 12 | 419.9 |
| 4 | 326.5 | 6 | 404.5 | 8 | 357.1 | 10 | 433.1 | 12 | 410.7 |
| 4 | 316.8 | 6 | 331.0 | 8 | 409.9 | 10 | 410.6 | 12 | 473.4 |
| 4 | 349.8 | 6 | 348.9 | 8 | 367.3 | 10 | 384.2 | 12 | 441.2 |
| 4 | 309.7 | 6 | 381.7 | 8 | 382.0 | 10 | 362.6 | 12 | 465.8 |

Table 7.11: *ASI summary statistics*

| Plate | $N$ | $\bar{y}_{i\cdot}$ | $s_i^2$ | $s_i$ |
|---|---|---|---|---|
| 4 | 7 | 333.2143 | 1338.6981 | 36.59 |
| 6 | 7 | 368.0571 | 816.3629 | 28.57 |
| 8 | 7 | 375.1286 | 433.7990 | 20.83 |
| 10 | 7 | 407.3571 | 1981.1229 | 44.51 |
| 12 | 7 | 437.1714 | 675.8557 | 26.00 |

The data could also be considered as an analysis of variance with plate lengths being different treatments and with seven observations on each treatment. Table 7.11 gives the usual summary statistics for a one-way ANOVA.

Viewed as regression data, we might think of fitting a simple linear regression model

$$y_h = \beta_0 + \beta_1 x_h + \varepsilon_h,$$

$h = 1, \ldots, 35$. Note that while $h$ varies from 1 to 35, there are only five distinct values of $x_h$ that occur in the data. As an analysis of variance, we usually use two subscripts to identify an observation: one to identify the treatment group and one to identify the observation within the group. The ANOVA model would often be written as

$$y_{ij} = \mu_i + \varepsilon_{ij} \tag{7.12.1}$$

where $i = 1,2,3,4,5$ and $j = 1, \ldots, 7$. We can also rewrite the regression model using the two subscripts $i$ and $j$ in place of $h$,

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij},$$

where $i = 1,2,3,4,5$ and $j = 1, \ldots, 7$. Note that all of these models account for exactly 35 observations.

Figure 7.32 contains a scatter plot of the data. With multiple observations at each $x$ value, the regression is really only fitted to the mean of the $y$ values at each $x$ value. The means of the $y$s are plotted against the $x$ values in Figure 7.33. The overall trend of the data is easier to evaluate in this plot than in the full scatter plot. We see an overall increasing trend which is very nearly linear except for a slight anomaly with 6 inch plates. We need to establish if these visual effects are real or just random variation. We would also like to establish whether there is a simple regression model that is appropriate for any trend that may exist. With only five distinct $x$ values, we can fit at most a quartic (fourth degree) polynomial, say,

$$y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_{ij}, \tag{7.12.2}$$

so a simple model should be something smaller than a quartic, i.e., either a cubic, quadratic, or a linear polynomial.

Table 7.12 contains ANOVA tables for fitting the linear, quadratic, cubic, and quartic polynomial regressions and for fitting the one-way ANOVA model. From our earlier discussion, the $F$ test in the simple linear regression ANOVA table strongly suggests that there is an overall trend in the

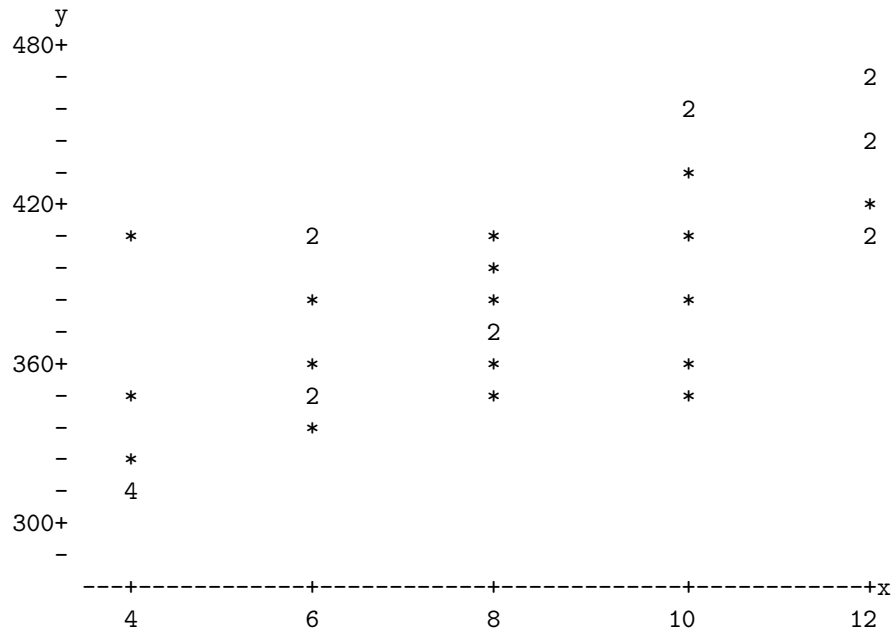



Figure 7.32: *ASI data versus plate length.*

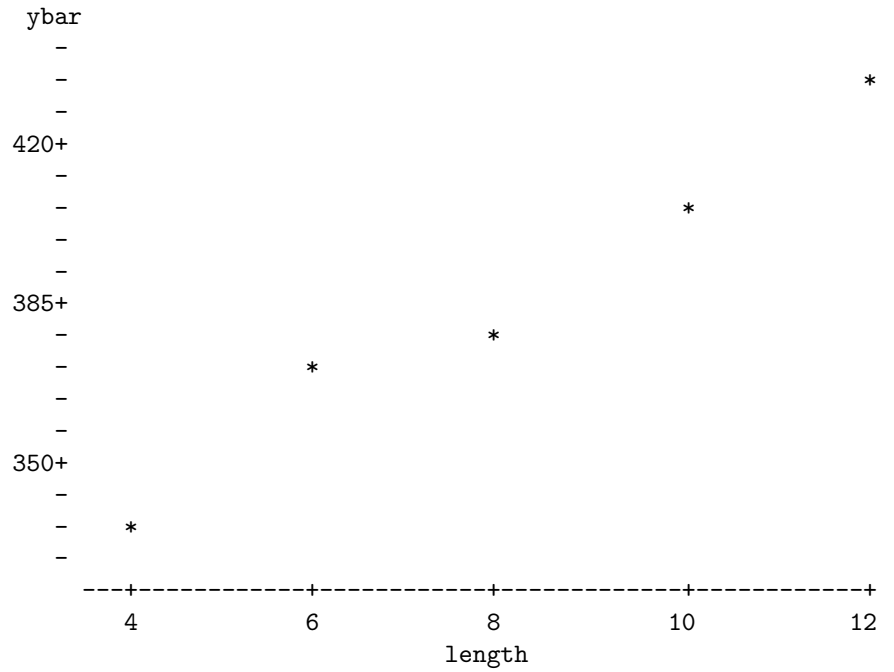



Figure 7.33: *ASI means versus plate length.*

Table 7.12: *Analysis of variance tables for ASI data*

Analysis of variance: simple linear regression

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 1 | 42780 | 42780 | 43.19 | 0.000 |
| Error | 33 | 32687 | 991 | | |
| Total | 34 | 75468 | | | |

Analysis of variance: quadratic polynomial

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 2 | 42894 | 21447 | 21.07 | 0.000 |
| Error | 32 | 32573 | 1018 | | |
| Total | 34 | 75468 | | | |

Analysis of variance: cubic polynomial

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 3 | 43345 | 14448 | 13.94 | 0.000 |
| Error | 31 | 32123 | 1036 | | |
| Total | 34 | 75468 | | | |

Analysis of variance: quartic polynomial

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 4 | 43993 | 10998 | 10.48 | 0.000 |
| Error | 30 | 31475 | 1049 | | |
| Total | 34 | 75468 | | | |

Analysis of variance: one-way ANOVA

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Trts(plates) | 4 | 43993 | 10998 | 10.48 | 0.000 |
| Error | 30 | 31475 | 1049 | | |
| Total | 34 | 75468 | | | |

data. From Figure 7.33 we see that this trend must be increasing, i.e., as lengths go up, by and large the ASI readings go up. ANOVA tables for higher degree polynomial models have been discussed briefly in the previous section but for now the key point to recognize is that *the ANOVA table for the quartic polynomial is identical to the ANOVA table for the one-way analysis of variance*. This occurs because models (7.12.1) and (7.12.2) are equivalent.

The first question of interest is whether a quartic polynomial is needed or whether a cubic model would be adequate. This is easily evaluated from the table of estimates and standard errors for the quartic fit. For computational reasons, the results reported are for a polynomial involving powers of $x - \bar{x}_\cdot$ rather than powers of $x$, cf. Section 7.6. This has *no* effect on our subsequent discussion, see Exercise 7.13.15.

| Predictor | $\hat{\beta}_k$ | $\mathrm{SE}(\hat{\beta}_k)$ | $t$ | $P$ |
|---|---|---|---|---|
| Constant | 375.13 | 12.24 | 30.64 | 0.000 |
| $(x-\bar{x}_\cdot)$ | 8.768 | 5.816 | 1.51 | 0.142 |
| $(x-\bar{x}_\cdot)^2$ | 3.983 | 4.795 | 0.83 | 0.413 |
| $(x-\bar{x}_\cdot)^3$ | 0.2641 | 0.4033 | 0.65 | 0.517 |
| $(x-\bar{x}_\cdot)^4$ | −0.2096 | 0.2667 | −0.79 | 0.438 |

There is little evidence ($P = .438$) that $\beta_4 \neq 0$, so a cubic polynomial seems to be an adequate explanation of the data.

The table of estimates given above is inappropriate for evaluating $\beta_3$ in the cubic model (even the cubic model based on $x - \bar{x}_\cdot$). To evaluate $\beta_3$, we need to fit the cubic model. If we then decide that a parabola is an adequate model, evaluating $\beta_2$ in the parabola requires one to fit the quadratic model. *In general, regression estimates are* only *valid for the model fitted. A new model requires new estimates*.

In Section 5.5, we discussed comparing models as a way of arriving at the $F$ test in a one-way analysis of variance. Comparing a submodel against a larger model to determine the adequacy of the submodel is a key method in regression analysis. Recall that model comparisons are based on the difference between the sums of squares for error of the submodel and the sums of squares for error of the larger model. Obviously, the simple linear regression model is a submodel of the quadratic model which is a submodel of the cubic model, which is a submodel of the quartic model, and we have seen that the quartic model is equivalent to the one-way ANOVA model. Given below are the degrees of freedom and sums of squares for error for the four polynomial regression models and the model with only an intercept $\beta_0$ (grand mean). (See Section 5.5 for discussion of the grand mean model.) The differences in sums of squares error for adjacent models are also given; the differences in degrees of freedom error are just 1.

Model comparisons

| Model | $dfE$ | $SSE$ | Difference |
|---|---|---|---|
| Intercept | 34 | 75468 | —- |
| Linear | 33 | 32687 | 42780 |
| Quadratic | 32 | 32573 | 114 |
| Cubic | 31 | 32123 | 450 |
| Quartic | 30 | 31475 | 648 |

Note that the $dfE$ and $SSE$ for the intercept model are those from the corrected Total lines in the ANOVAs of Table 7.12. The $dfE$s and $SSE$s for the other models also come from Table 7.12.

To test the quartic model against the cubic model we take

$$F = \frac{648/1}{31475/30} = .62.$$

This is just the square of the $t$ statistic for testing $\beta_4 = 0$ in the quartic model. The reference distribution for the $F$ statistic is $F(1,30)$ and the $P$ value is .44, as it was for the $t$ test.

If we decide that we do not need the quartic term, we can test whether we need the cubic term. We can test the quadratic model against the cubic model with

$$F = \frac{450/1}{32123/31} = 0.434.$$

The reference distribution is $F(1,31)$. This test is equivalent to the $t$ test of $\beta_3 = 0$ *in the cubic model. The t test of* $\beta_3 = 0$ *in the quartic model is inappropriate*. An alternative to this $F$ test can also be used. The denominator of this test is $32123/31$, the mean squared error from the cubic model. If we accepted the cubic model only after testing the quartic model, the result of the quartic test is open to question and thus the estimate of $\sigma^2$ from the cubic model, i.e., the $MSE$ from the cubic model, is open to question. It might be better just to use the estimate of $\sigma^2$ from the quartic model, which is the mean squared error from the one-way ANOVA. If we do this, the test statistic for the cubic term becomes

$$F = \frac{450/1}{31475/30} = 0.429.$$

The reference distribution for the alternative test is $F(1,30)$. In this example the two $F$ tests give essentially the same answers. This should, by definition, almost always be the case. If, for example, one test were significant at .05 and the other were not, they are both likely to have $P$ values near .05 and the fact that one is a bit larger than .05 and the other is a bit smaller than .05 should not be a cause for concern.

If we decide that neither the quartic nor the cubic terms are important, we can test whether we need the quadratic term. Testing the quadratic model against the simple linear model gives

$$F = \frac{114/1}{32573/32} = 0.112$$

which is compared to an $F(1,32)$ distribution. This test is equivalent to the $t$ test of $\beta_2 = 0$ in the quadratic model. Again, an alternative test can also be used. The denominator of this test is $32573/32$, the mean squared error from the quadratic model. If we accepted the quadratic model only after testing the cubic and quartic models, this estimate of $\sigma^2$ may be biased and it might be better to use the estimate of $\sigma^2$ from the quartic model, i.e., the one-way ANOVA model. If we do this, the test statistic for the quadratic term becomes

$$F = \frac{114/1}{31475/30} = 0.109$$

and the reference distribution is $F(1,30)$.

If we decide that we need none of the higher order terms, we can test whether we need the linear term. Testing the intercept model against the simple linear model gives

$$F = \frac{42780/1}{32687/33} = 43.190.$$

This is just the test for zero slope *in the simple linear model*. Again, the alternative test can be used. The denominator of this test is the mean squared error from the linear model, $32687/33$. If we accepted the linear model only after testing the higher order models, it may be better to use the mean squared error from the one-way ANOVA model. The alternative $F$ test for the linear term has

$$F = \frac{42780/1}{31475/30} = 40.775.$$

*The model comparison tests just discussed can be reconstructed from contrasts in the one-way ANOVA*. Below are given some simple contrasts that correspond to the differences in sums of squares error for the model comparisons.

| | Orthogonal polynomial contrasts | | | | |
|---|---|---|---|---|---|
| Plate | Linear | Quadratic | Cubic | Quartic | $\bar{y}_{i\cdot}$ |
| 4 | −2 | 2 | −1 | 1 | 333.2143 |
| 6 | −1 | −1 | 2 | −4 | 368.0571 |
| 8 | 0 | −2 | 0 | 6 | 375.1286 |
| 10 | 1 | −1 | −2 | −4 | 407.3571 |
| 12 | 2 | 2 | 1 | 1 | 437.1714 |
| *Est* | 247.2142 | 15.1000 | 25.3571 | −80.4995 | |
| *SS* | 42780.4 | 114.0 | 450.1 | 648.0 | |

Recall that the estimate of, say, the linear contrast is

$$(-2)(333.2143) + (-1)(368.0571) + (0)(375.1286) + (1)(407.3571) + (2)(437.1714) = 247.2142$$

and that with seven observations on each plate length, the sum of squares for the linear contrast is

$$SS(\text{linear}) = \frac{(247.2142)^2}{[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2]/7} = 42780.4.$$

Table 7.13: *Analysis of variance for ASI data*

| Source | *df* | *SS* | *MS* | *F* | *P* |
|---|---|---|---|---|---|
| Treatments | 4 | 43993 | 10998 | 10.48 | 0.000 |
| (linear) | (1) | (42780) | (42780) | (40.78) | |
| (quadratic) | (1) | ( 114) | ( 114) | (0.11) | |
| (cubic) | (1) | ( 450) | ( 450) | (0.43) | |
| (quartic) | (1) | ( 648) | ( 648) | (0.62) | |
| Error | 30 | 31475 | 1049 | | |
| Total | 34 | 75468 | | | |

This is precisely the difference in error sums of squares between the intercept and straight line models. Similar results hold for the other contrasts.

These contrasts are called *orthogonal polynomial contrasts* because they are orthogonal in balanced ANOVAs and reproduce the sums of squares for comparing different polynomial regression models. We leave it to the reader to verify that the contrasts are orthogonal, cf. Section 5.4, but recall that with orthogonal contrasts we have the identity

$$SSTrts = 43992.5 = 42780.4 + 114.0 + 450.1 + 648.0.$$

Table 7.13 contains an expanded analysis of variance table for the one-way ANOVA that incorporates the information from the contrasts. Using the orthogonal polynomial contrasts allows us to make all of the model comparisons by using simple analysis of variance computations rather than fitting polynomial regression models.

From Table 7.13, the $P$ value of .000 indicates strong evidence that the five groups are different, i.e., there is strong evidence for the quartic polynomial. The results from the contrasts are so clear that we did not bother to report $P$ values for them. There is a huge effect for the linear contrast. The other three $F$ statistics are all much less than 1, so there is no evidence of the need for a quartic, cubic, or quadratic polynomial. As far as we can tell, a line fits the data just fine. For completeness, some residual plots are presented as Figures 7.34 through 7.38. Note that the normal plot for the simple linear regression in Figure 7.35 is less than admirable, while the normal plot for the one-way ANOVA in Figure 7.38 is only slightly better. It appears that one should not put great faith in the normality assumption. □

The linear, quadratic, cubic, and quartic contrasts for the ASI data are simple only because the ANOVA is balanced and the treatment groups are equally spaced. The treatments occur at 4, 6, 8, 10, and 12 inches. Thus the treatments occur at intervals of 2 inches. If the treatments were at irregular intervals or if the group sample sizes were unequal, orthogonal linear, quadratic, cubic, and quartic contrasts still exist, but they are difficult to find. With either unequal spacings or unequal numbers, it is easier just to do the appropriate regressions. *With a balanced ANOVA and regularly spaced intervals, the orthogonal polynomial contrasts can be determined from the number of treatment groups and thus they can be tabled*. Such a table is given in Appendix B.4 for linear, quadratic, and cubic contrasts.

Comparing one of the reduced polynomial models against the one-way ANOVA model is often referred to as a test of *lack of fit*. This is especially true when the reduced model is the simple linear regression model. In these tests, the degrees of freedom, sums of squares, and mean squares used in the numerator of the tests are all described as being for *lack of fit*. The denominator of the test is based on the error from the one-way ANOVA. The mean square, sum of squares, and degrees of freedom for error in the one-way ANOVA are often referred to as the mean square, sum of squares, and degrees of freedom for *pure error*. This lack of fit test can be performed *whenever* the data contain multiple observations at *any* $x$ values. Often the appropriate unbalanced one-way ANOVA includes treatments with only one observation on them. These treatments do not provide an estimate of $\sigma^2$, so they simply play no role in obtaining the mean square for pure error.

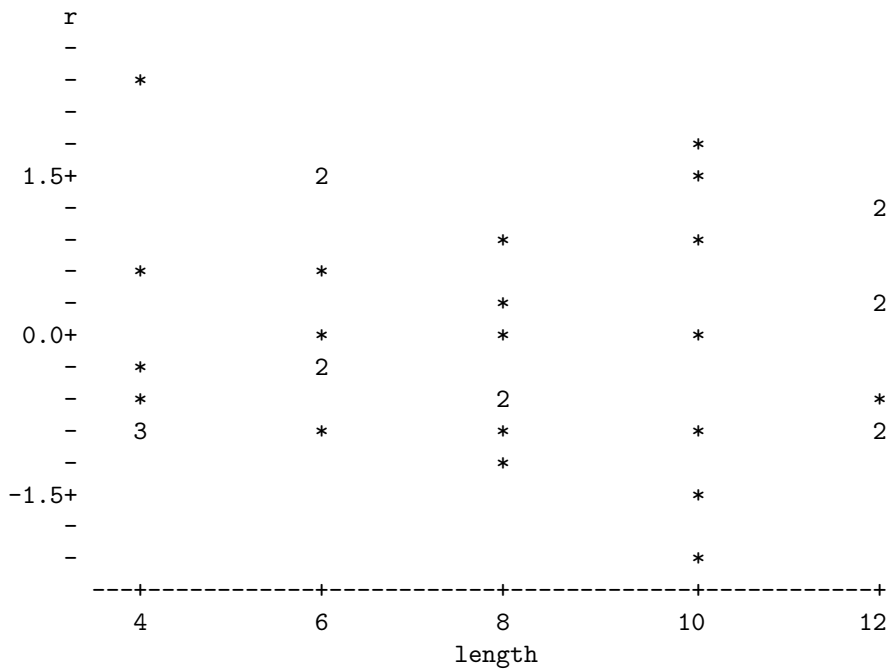



Figure 7.34: *ASI SLR standardized residuals versus plate length.*

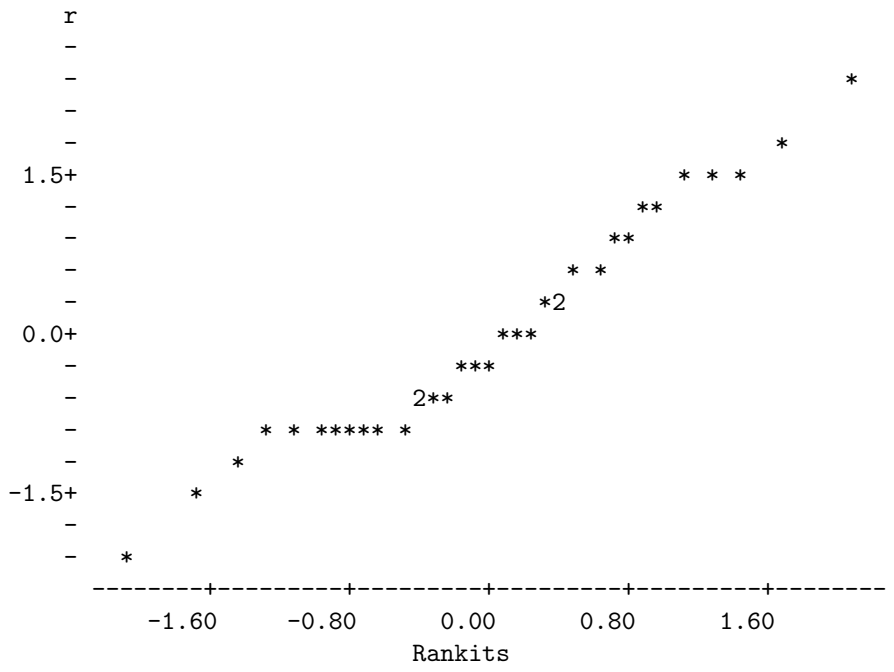



Figure 7.35: *ASI SLR standardized residuals normal plot,* $W' = .960$.

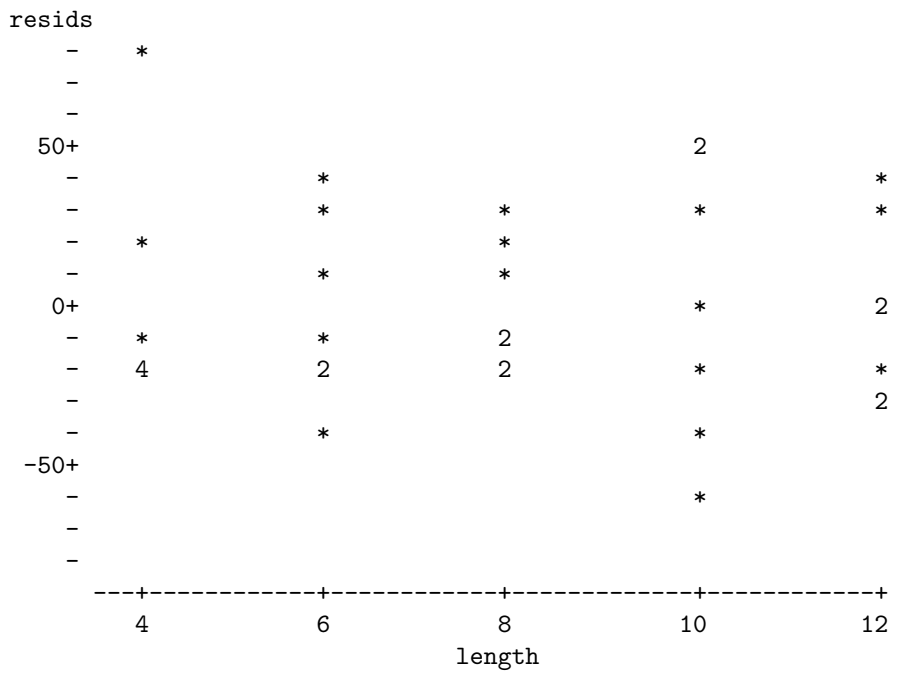



Figure 7.36: *ASI ANOVA residuals versus plate length.*

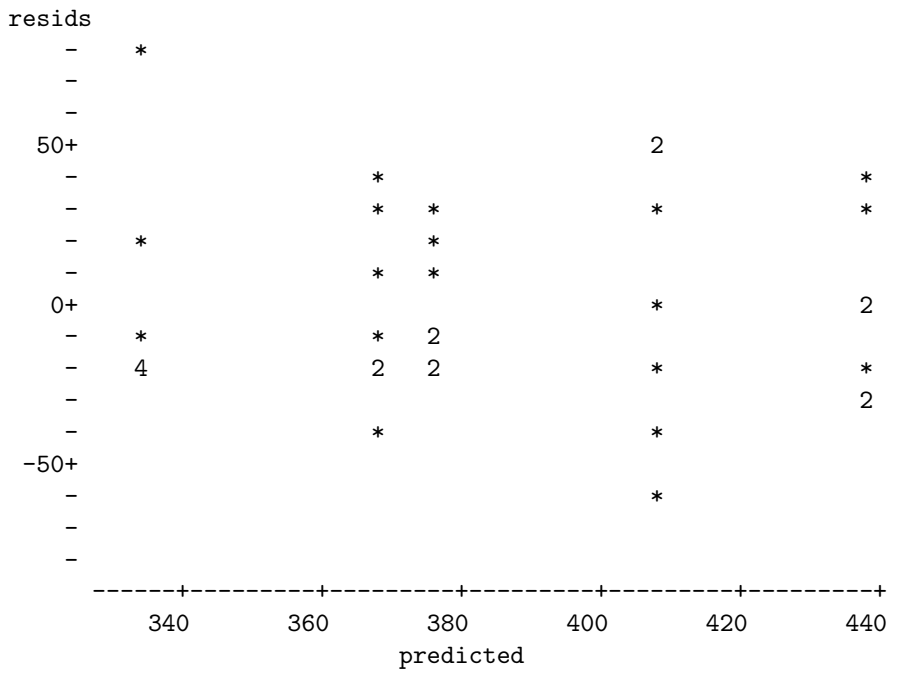



Figure 7.37: *ASI ANOVA residuals versus predicted values.*

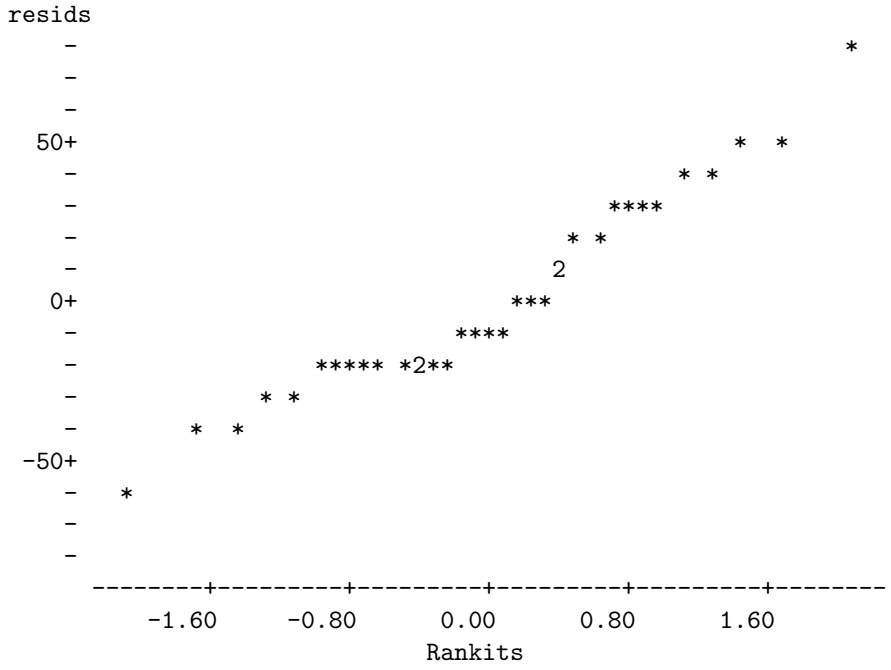



Figure 7.38: *ASI ANOVA residuals normal plot,* $W' = .966$.

For testing lack of fit in the simple linear regression model with the ASI data, the numerator sum of squares can be obtained either by differencing the sums of squares for error in the simple linear regression model and the one-way ANOVA model or by adding up the sums of squares for the quadratic, cubic, and quartic contrasts. Here the sum of squares for lack of fit is $32687 - 31475 = 1212 = 114 + 450 + 648$ and the degrees of freedom for lack of fit are $33 - 30 = 3$. The mean square for lack of fit is $1212/3 = 404$. The pure error comes from the one-way ANOVA table. The lack of fit test $F$ statistic for the simple linear regression model is

$$F = \frac{404}{1049} = .39$$

which is less than 1, so there is no evidence of a lack of fit in the simple linear regression model. If an $\alpha = .05$ test were desired, the test statistic would be compared to $F(.95, 3, 30)$.

*Appendix: simple linear regression proofs*

PROOF OF UNBIASEDNESS FOR THE REGRESSION ESTIMATES.
To begin, The $\beta$s and $x_i$s are all fixed numbers so

$$\mathrm{E}(y_i) = \mathrm{E}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + \mathrm{E}(\varepsilon_i) = \beta_0 + \beta_1 x_i \, .$$

Also note that $\sum_{i=1}^{n} (x_i - \bar{x}_\cdot) = 0$, so $\sum_{i=1}^{n} (x_i - \bar{x}_\cdot)\bar{x}_\cdot = 0$. It follows that

$$\sum_{i=1}^{n} (x_i - \bar{x}_\cdot)^2 = \sum_{i=1}^{n} (x_i - \bar{x}_\cdot) x_i - \sum_{i=1}^{n} (x_i - \bar{x}_\cdot)\bar{x}_\cdot = \sum_{i=1}^{n} (x_i - \bar{x}_\cdot) x_i \, .$$

Now consider the slope estimate.

$$\mathrm{E}\left(\hat{\beta}_1\right) \quad = \quad \mathrm{E}\left( \frac{\sum_{i=1}^{n} (x_i - \bar{x}_\cdot) y_i}{\sum_{i=1}^{n} (x_i - \bar{x}_\cdot)^2} \right)$$

$$\begin{aligned}
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x}.)\,\mathrm{E}(y_i)}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2} \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x}.)\,(\beta_0+\beta_1 x_i)}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2} \\
&= \beta_0\frac{\sum_{i=1}^{n}(x_i-\bar{x}.)}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2}+\beta_1\frac{\sum_{i=1}^{n}(x_i-\bar{x}.)\,x_i}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2} \\
&= \beta_0\frac{0}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2}+\beta_1\frac{\sum_{i=1}^{n}(x_i-\bar{x}.)^2}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2} \\
&= \beta_1
\end{aligned}$$

The proof for the intercept goes as follows:

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_0\right) &= \mathrm{E}\left(\bar{y}.-\hat{\beta}_1\bar{x}.\right) \\
&= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)-\mathrm{E}\left(\hat{\beta}_1\right)\bar{x}. \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(y_i)-\beta_1\bar{x}. \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i)-\beta_1\bar{x}. \\
&= \beta_0+\beta_1\frac{1}{n}\sum_{i=1}^{n}(x_i)-\beta_1\bar{x}. \\
&= \beta_0+\beta_1\bar{x}.-\beta_1\bar{x}. \\
&= \beta_0\,.
\end{aligned}$$

PROOF OF VARIANCE FORMULAE.
To begin,

$$\mathrm{Var}(y_i)=\mathrm{Var}(\beta_0+\beta_1 x_i+\varepsilon_i)=\mathrm{Var}(\varepsilon_i)=\sigma^2.$$

Now consider the slope estimate. Recall that the $y_i$s are independent.

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_1\right) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x}.)\,y_i}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2}\right) \\
&= \frac{1}{\left[\sum_{i=1}^{n}(x_i-\bar{x}.)^2\right]^2}\mathrm{Var}\left(\sum_{i=1}^{n}(x_i-\bar{x}.)\,y_i\right) \\
&= \frac{1}{\left[\sum_{i=1}^{n}(x_i-\bar{x}.)^2\right]^2}\sum_{i=1}^{n}(x_i-\bar{x}.)^2\,\mathrm{Var}(y_i) \\
&= \frac{1}{\left[\sum_{i=1}^{n}(x_i-\bar{x}.)^2\right]^2}\sum_{i=1}^{n}(x_i-\bar{x}.)^2\,\sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x}.)^2} \\
&= \frac{\sigma^2}{(n-1)s_x^2}.
\end{aligned}$$

Rather than establishing the variance of $\hat{\beta}_0$ directly, we find $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)$ for an arbitrary value $x$. The variance of $\hat{\beta}_0$ is the special case with $x = 0$. A key result is that $\bar{y}_\cdot$ and $\hat{\beta}_1$ are independent. This was discussed in relation to the alternative regression model of Section 7.6. The independence of these estimates is based on the errors having independent normal distributions with the same variance. More generally, if the errors have the same variance and zero covariance, we still get $\mathrm{Cov}(\bar{y}_\cdot, \hat{\beta}_1) = 0$, see Exercise 7.13.14.

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) &= \mathrm{Var}\left(\bar{y}_\cdot - \hat{\beta}_1 \bar{x}_\cdot + \hat{\beta}_1 x\right) \\
&= \mathrm{Var}\left(\bar{y}_\cdot + \hat{\beta}_1 (x - \bar{x}_\cdot)\right) \\
&= \mathrm{Var}(\bar{y}_\cdot) + \mathrm{Var}\left(\hat{\beta}_1\right)(x - \bar{x}_\cdot)^2 - 2(x - \bar{x}_\cdot)\mathrm{Cov}\left(\bar{y}_\cdot, \hat{\beta}_1\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(y_i) + \mathrm{Var}\left(\hat{\beta}_1\right)(x - \bar{x}_\cdot)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 + \frac{\sigma^2 (x - \bar{x}_\cdot)^2}{(n-1)s_x^2} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x}_\cdot)^2}{(n-1)s_x^2}\right].
\end{aligned}$$

In particular, when $x = 0$ we get

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_\cdot^2}{(n-1)s_x^2}\right].$$

Table 7.14: *Age and maintenance costs of truck tractors*

| Age | Cost | Age | Cost | Age | Cost |
|---|---|---|---|---|---|
| 0.5 | 163 | 4.0 | 495 | 5.0 | 890 |
| 0.5 | 182 | 4.0 | 723 | 5.0 | 1522 |
| 1.0 | 978 | 4.0 | 681 | 5.0 | 1194 |
| 1.0 | 466 | 4.5 | 619 | 5.5 | 987 |
| 1.0 | 549 | 4.5 | 1049 | 6.0 | 764 |
| | | 4.5 | 1033 | 6.0 | 1373 |

Table 7.15: *Angle between the plane of the equator and the plane of rotation about the sun*

| Year | Angle | Year | Angle | Year | Angle | Year | Angle |
|---|---|---|---|---|---|---|---|
| −229 | $51.33\bar{3}$ | 880 | 35.000 | 1500 | 28.400 | 1600 | 31.000 |
| −139 | $51.33\bar{3}$ | 1070 | 34.000 | 1500 | $29.26\bar{6}$ | 1656 | $29.03\bar{3}$ |
| 140 | $51.16\bar{6}$ | 1300 | 32.000 | 1570 | $29.91\bar{6}$ | 1672 | 28.900 |
| 390 | 30.000 | 1460 | 30.000 | 1570 | 31.500 | 1738 | $28.33\bar{3}$ |

## 7.13 Exercises

EXERCISE 7.13.1. Draper and Smith (1966, p. 41) considered data on the relationship between the age of truck tractors (in years) and the cost (in dollars) of maintaining them over a six month period. The data are given in Table 7.14. Plot cost versus age and fit a regression of cost on age. Give 95% confidence intervals for the slope and intercept. Give a 99% confidence interval for the mean cost of maintaining tractors that are 2.5 years old. Give a 99% prediction interval for the cost of maintaining a particular tractor that is 2.5 years old.

Reviewing the plot of the data, how much faith should be placed in these estimates for tractors that are 2.5 years old?

EXERCISE 7.13.2. Stigler (1986, p. 6) reported data from Cassini (1740) on the angle between the plane of the equator and the plane of the earth's revolution about the sun. The data are given in Table 7.15. The years −229 and −139 indicate 230 B.C. and 140 B.C. respectively. The angles are listed as the minutes above 23 degrees.

Plot the data. Are there any obvious outliers? If outliers exist, compare the fit of the line with and without the outliers. In particular, compare the different 95% confidence intervals for the slope and intercept.

EXERCISE 7.13.3. Mulrow et al. (1988) presented data on the calibration of a differential scanning calorimeter. The melting temperatures of mercury and naphthalene are known to be 234.16 and 353.24 Kelvin, respectively. The data are given in Table 7.16. Plot the data. Fit a simple linear regression $y = \beta_0 + \beta_1 x + \varepsilon$ to the data. Under ideal conditions, the simple linear regression should have $\beta_0 = 0$ and $\beta_1 = 1$; test whether these hypotheses are true using $\alpha = .05$. Give a 95% confidence interval for the population mean of observations taken on this calorimeter for which the true melting point is 250. Give a 95% prediction interval for a new observation taken on this calorimeter for which the true melting point is 250.

Is there any way to check whether it is appropriate to use a line in modeling the relationship between $x$ and $y$? If so, do so.

EXERCISE 7.13.4. Using the complete data of Exercise 7.13.2, test the need for a transformation of the simple linear regression model. Repeat the test after eliminating any outliers. Compare the results.

Table 7.16: *Melting temperatures*

| Chemical | $x$ | $y$ |
|---|---|---|
| Naphthalene | 353.24 | 354.62 |
| | 353.24 | 354.26 |
| | 353.24 | 354.29 |
| | 353.24 | 354.38 |
| Mercury | 234.16 | 234.45 |
| | 234.16 | 234.06 |
| | 234.16 | 234.61 |
| | 234.16 | 234.48 |

Table 7.17: *IQs and achievement scores*

| IQ | Achiev. | IQ | Achiev. | IQ | Achiev. | IQ | Achiev. | IQ | Achiev. |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 49 | 105 | 50 | 134 | 78 | 107 | 43 | 122 | 66 |
| 117 | 47 | 89 | 72 | 125 | 39 | 121 | 75 | 130 | 63 |
| 98 | 69 | 96 | 45 | 140 | 66 | 90 | 40 | 116 | 43 |
| 87 | 47 | 105 | 47 | 137 | 69 | 132 | 80 | 101 | 44 |
| 106 | 45 | 95 | 46 | 142 | 68 | 116 | 55 | 92 | 50 |
| 134 | 55 | 126 | 67 | 130 | 71 | 137 | 73 | 120 | 60 |
| 77 | 72 | 111 | 66 | 92 | 31 | 113 | 48 | 80 | 31 |
| 107 | 59 | 121 | 59 | 125 | 53 | 110 | 41 | 117 | 55 |
| 125 | 27 | 106 | 49 | 120 | 64 | 114 | 29 | 93 | 50 |

EXERCISE 7.13.5. Dixon and Massey (1969) presented data on the relationship between IQ scores and results on an achievement test in a general science course. Table 7.17 contains a subset of the data. Fit the simple linear regression model of achievement on IQ and the quadratic model of achievement on IQ and IQ squared. Evaluate both models and decide which is the best.

EXERCISE 7.13.6. Snedecor and Cochran (1967, Section 6.18) presented data obtained in 1942 from South Dakota on the relationship between the size of farms (in acres) and the number of acres planted in corn. The data are given in Table 7.18.

Plot the data. Fit a simple linear regression to the data. Examine the residuals and discuss what you find. Test the need for a power transformation. Is it reasonable to examine the square root or log transformations? If so, do so.

EXERCISE 7.13.7. In Exercises 5.7.2 and 7.13.6 we considered data on the relationship between farm sizes and the acreage in corn. Fit the linear, quadratic, cubic, and quartic polynomial models to the logs of the acreages in corn. Find the model that fits best. Check the assumptions for this model.

Table 7.18: *Acreage in corn for different farm acreages*

| Farm | Corn | Farm | Corn | Farm | Corn |
|---|---|---|---|---|---|
| $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
| 80 | 25 | 160 | 45 | 320 | 110 |
| 80 | 10 | 160 | 40 | 320 | 30 |
| 80 | 20 | 240 | 65 | 320 | 55 |
| 80 | 32 | 240 | 80 | 320 | 60 |
| 80 | 20 | 240 | 65 | 400 | 75 |
| 160 | 60 | 240 | 85 | 400 | 35 |
| 160 | 35 | 240 | 30 | 400 | 140 |
| 160 | 20 | 320 | 70 | 400 | 90 |
| | | | | 400 | 110 |

Table 7.19: *Weights for various heights*

| Ht. | Wt. | Ht. | Wt. |
|---|---|---|---|
| 65 | 120 | 63 | 110 |
| 65 | 140 | 63 | 135 |
| 65 | 130 | 63 | 120 |
| 65 | 135 | 72 | 170 |
| 66 | 150 | 72 | 185 |
| 66 | 135 | 72 | 160 |

Compute the sums of squares for the following contrasts using the means of the logs of the corn acreages:

| | Farm acreages | | | | |
|---|---|---|---|---|---|
| Contrast | 80 | 160 | 240 | 320 | 400 |
| Linear | $-2$ | $-1$ | 0 | 1 | 2 |
| Quadratic | 2 | $-1$ | $-2$ | $-1$ | 2 |
| Cubic | $-1$ | 2 | 0 | $-2$ | 1 |
| Quartic | 1 | $-4$ | 6 | $-4$ | 1 |
| Means | 2.9957 | 3.6282 | 4.1149 | 4.0904 | 4.4030 |

Compare the contrast sums of squares to the polynomial model fitting procedure.

EXERCISE 7.13.8. Repeat Exercise 7.13.6 but instead of using the number of acres of corn as the dependent variable, use the proportion of acreage in corn as the dependent variable. Compare the results to those given earlier.

EXERCISE 7.13.9. In Exercises 7.13.1 and 5.7.10, we performed a simple linear regression and a one-way ANOVA on the data of Table 7.14. Test for lack of fit, i.e., whether the simple linear regression is an adequate reduced model as compared to the one-way ANOVA model.

EXERCISE 7.13.10. The analysis of variance in Exercise 5.7.3 was based on the height and weight data given in Table 7.19. Fit a simple linear regression of weight on height for these data and check the assumptions. Give a 99% confidence interval for the mean weight of people with a 72 inch height and compare it to the interval from Exercise 5.7.3. Test the lack of fit of the simple linear regression model compared to the larger one-way ANOVA model.

EXERCISE 7.13.11. Jensen (1977) and Weisberg (1985, p. 101) considered data on the outside diameter of crank pins that were produced in an industrial process. The diameters of batches of crank pins were measured on various days; if the industrial process is 'under control' the diameters should not depend on the day they were measured. A subset of the data is given in Table 7.20 in a format consistent with performing a regression analysis on the data. The diameters of the crank pins are actually $.742 + y_{ij}10^{-5}$ inches, where the $y_{ij}$s are reported in Table 7.20. Perform an analysis of variance and polynomial regressions on the data. Give the lack of fit test for the simple linear regression.

EXERCISE 7.13.12. Exercise 7.13.3 involves the calibration of a measuring instrument. Often, calibration curves are used in reverse, i.e., we would use the calorimeter to measure a melting point $y$ and use the regression equation to give a point estimate of $x$. If a new substance has a measured melting point of 300 Kelvin, using the simple linear regression model what is the estimate of the true melting point? Use a prediction interval to determine whether the measured melting point of $y = 300$ is consistent with the true melting point being $x = 300$. Is an observed value of 300 consistent with a true value of 310?

Table 7.20: *Jensen's crank pin data*

| Days | Diameters | Days | Diameters | Days | Diameters | Days | Diameters |
|---|---|---|---|---|---|---|---|
| 4 | 93 | 10 | 93 | 16 | 82 | 22 | 90 |
| 4 | 100 | 10 | 88 | 16 | 72 | 22 | 92 |
| 4 | 88 | 10 | 87 | 16 | 80 | 22 | 82 |
| 4 | 85 | 10 | 87 | 16 | 72 | 22 | 77 |
| 4 | 89 | 10 | 87 | 16 | 89 | 22 | 89 |

EXERCISE 7.13.13. Working-Hotelling confidence bands are a method for getting confidence intervals for every point on a line with a guaranteed simultaneous coverage. The method is essentially the same as Scheffé's method for simultaneous confidence intervals discussed in Section 6.4. For estimating the point on the line at a value $x$, the endpoints of the $(1-\alpha)100\%$ simultaneous confidence intervals are

$$(\hat{\beta}_0+\hat{\beta}_1 x) \pm \sqrt{2F(1-\alpha,2,dfE)}\,\mathrm{SE}(\hat{\beta}_0+\hat{\beta}_1 x).$$

Using the *Coleman Report* data of Table 7.1, find 95% simultaneous confidence intervals for the values $x = -17,-6,0,6,17$. Plot the estimated regression line and sketch the Working-Hotelling confidence bands. We are 95% confident that the entire line $\beta_0+\beta_1 x$ lies between the confidence bands. Compute the regular confidence intervals for $x = -17,-6,0,6,17$ and compare them to the results of the Working-Hotelling procedure.

EXERCISE 7.13.14. Use part (4) of Proposition 1.2.11 to show that $\mathrm{Cov}(\bar{y}_\cdot,\hat{\beta}_1) = 0$ whenever $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all $i$ and $\mathrm{Cov}(\varepsilon_i,\varepsilon_j) = 0$ for all $i \neq j$. Hint: write out $\bar{y}_\cdot$ and $\hat{\beta}_1$ in terms of the $y_i$s.

EXERCISE 7.13.15. Using the axial stiffness index data of Table 7.10, fit linear, quadratic, cubic, and quartic polynomial regression models using powers of $x$, the plate length, and using powers of $x-\bar{x}_\cdot$, the plate length minus the average plate length. Compare the results of the two procedures. If your computer program will not fit some of the models, report on that in addition to comparing results for the models you could fit.

Chapter 8

# The analysis of count data

For the most part, this book concerns itself with measurement data and the corresponding analyses based on normal distributions. In this chapter we consider data that consist of counts. We begin in Section 8.1 by examining a set of data on the number of females admitted into graduate school at the University of California, Berkeley. A key feature of these data is that only two outcomes are possible: admittance or rejection. Data with only two outcomes are referred to as *binary (or dichotomous) data*. Often the two outcomes are referred to generically as success and failure. In Section 8.2, we expand our discussion by comparing two sets of dichotomous data; we compare Berkeley graduate admission rates for females and males. Section 8.3 examines *polytomous data*, i.e., count data in which there are more than two possible outcomes. For example, numbers of Swedish females born in the various months of the year involve counts for 12 possible outcomes. Section 8.4 examines comparisons between two samples of polytomous data, e.g., comparing the numbers of females and males that are born in the different months of the year. Section 8.5 looks at comparisons among more than two samples of polytomous data. The penultimate section considers a method of reducing large tables of counts that involve several samples of polytomous data into smaller more interpretable tables. The final section deals with a count data analogue of simple linear regression.

Sections 8.1 and 8.2 involve analogues of Chapters 2 and 4 that are appropriate for dichotomous data. The basic analyses in these sections simply involve new applications of the ideas in Chapter 3. Analyzing polytomous data requires techniques that are different from the methods of Chapter 3. Sections 8.3, 8.4, and 8.5 are polytomous data analogues of Chapters 2, 4, and 5. Everitt (1977) and Fienberg (1980) give more detailed introductions to the analysis of count data. Sophisticated analyses of count data frequently use analogues of ANOVA and regression called log-linear models. Christensen (1990b) provides an intermediate level account of log-linear models.

## 8.1 One binomial sample

The few distributions that are most commonly used in statistics arise naturally. The normal distribution arises for measurement data because the variability in the data often results from the mean of a large number of small errors and the central limit theorem indicates that such means tend to be normally distributed.

The binomial distribution arises naturally with count data because of its simplicity. Consider a number of trials, say $n$, each a success or failure. If each trial is independent of the other trials and if the probability of obtaining a success is the same for every trial, then the random number of successes has a binomial distribution. *The beauty of discrete data is that the probability models can often be justified solely by how the data were collected. This does not happen with measurement data*. The binomial distribution depends on two parameters, $n$, the number of independent trials, and the constant probability of success, say $p$. Typically, we know the value of $n$, while $p$ is the unknown parameter of interest. Binomial distributions were examined in Section 1.4.

Bickel et al. (1975) report data on admissions to graduate school at the University of California,

Berkeley. The numbers of females that were admitted and rejected are given below along with the total number of applicants.

Graduate admissions at Berkeley

| | Admitted | Rejected | Total |
|---|---|---|---|
| Female | 557 | 1278 | 1835 |

It seems reasonable to view the 1835 females as a random sample from a population of potential female applicants. We are interested in the probability $p$ that a female applicant is admitted to graduate school. A natural estimate of the parameter $p$ is the proportion of females that were actually admitted, thus our estimate of the parameter is

$$\hat{p} = \frac{557}{1835} = .30354.$$

We have a parameter of interest, $p$, and an estimate of that parameter, $\hat{p}$. If we can identify a standard error and an appropriate distribution, we can use the methods of Chapter 3 to perform statistical inferences.

The key to finding a standard error is to find the variance of the estimate. As we will see later,

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}. \tag{8.1.1}$$

To estimate the standard deviation of $\hat{p}$, we simply use $\hat{p}$ to estimate $p$ in (8.1.1) and take the square root. Thus the standard error is

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.30354(1-.30354)}{1835}} = .01073.$$

The final requirement for using the results of Chapter 3 is to find an appropriate reference distribution for

$$\frac{\hat{p}-p}{\mathrm{SE}(\hat{p})}.$$

We can think of each trial as scoring either a 1, if the trial is a success, or a 0, if the trial is a failure. With this convention $\hat{p}$, the proportion of successes, is really the average of the 0–1 scores and since $\hat{p}$ is an average we can apply the central limit theorem. (In fact, $\mathrm{SE}(\hat{p})$ is very nearly $s/\sqrt{n}$, where $s$ is computed from the 0–1 scores.) The central limit theorem simply states that for a large number of trials $n$, the distribution of $\hat{p}$ is approximately normal with a population mean that is the population mean of $\hat{p}$ and a population variance that is the population variance of $\hat{p}$. We have already given the variance of $\hat{p}$ and we will see later that $\mathrm{E}(\hat{p}) = p$. Thus for large $n$ we have the approximation

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

The variance is unknown but by the law of large numbers it is approximately equal to our estimate of it, $\hat{p}(1-\hat{p})/n$. Standardizing the normal distribution (cf. Exercise 1.6.2) gives the approximation

$$\frac{\hat{p}-p}{\mathrm{SE}(\hat{p})} \sim N(0,1). \tag{8.1.2}$$

This distribution requires a sample size that is large enough for both the central limit theorem approximation and the law of large numbers approximation to be reasonably valid. For values of $p$ that are not too close to 0 or 1, the approximation works reasonably well with sample sizes as small as 20.

We now have $Par = p$, $Est = \hat{p}$, $\mathrm{SE}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ and the distribution in (8.1.2). As in Chapter 3, a 95% confidence interval for $p$ has limits

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Here $1.96 = z(.975) = t(.975, \infty)$. Recall that a $(1-\alpha)100\%$ confidence interval requires the $(1-\alpha/2)$ percentile of the distribution. For the female admissions data, the limits are

$$.30354 \pm 1.96(.01073)$$

which gives the interval $(.28, .32)$. We are 95% confident that the population proportion of females admitted to Berkeley's graduate school is between .28 and .32. (As is often the case, it is not exactly clear what population these data relate to.)

We can also perform, say, an $\alpha = .01$ test of the null hypothesis $H_0 : p = 1/3$ versus the alternative $H_A : p \neq 1/3$. The test rejects $H_0$ if

$$\frac{\hat{p} - 1/3}{\mathrm{SE}(\hat{p})} > 2.58$$

or if

$$\frac{\hat{p} - 1/3}{\mathrm{SE}(\hat{p})} < -2.58.$$

Here $2.58 = z(.995) = t(.975, \infty)$. An $\alpha$ level two-sided test requires the $(1 - \frac{\alpha}{2})100\%$ point of the distribution. The Berkeley data yield the test statistic

$$\frac{.30354 - .33333}{.01073} = -2.78$$

which is smaller than $-2.58$, so we reject the null hypothesis of $p = 1/3$ with $\alpha = .01$. In other words, we can reject, with strong assurance, the claim that one third of female applicants are admitted to graduate school at Berkeley. Since the test statistic is negative, we have evidence that the true proportion is less than one third. The test as constructed here is equivalent to checking whether $p = 1/3$ is within a 99% confidence interval.

There is an alternative, slightly different, way of performing tests such as $H_0 : p = 1/3$ versus $H_A : p \neq 1/3$. The difference involves using a different standard error. The variance of the estimate $\hat{p}$ is $p(1-p)/n$. In obtaining a standard error, we estimated $p$ with $\hat{p}$ and took the square root of the estimated variance. Recalling that tests are performed *assuming that the null hypothesis is true*, it makes sense in the testing problem to use the assumption $p = 1/3$ in computing a standard error for $\hat{p}$. Thus an alternative standard error for $\hat{p}$ in this testing problem is

$$\sqrt{\frac{1}{3}\left(1 - \frac{1}{3}\right) \Big/ 1835} = .01100.$$

The test statistic now becomes

$$\frac{.30354 - .33333}{.01100} = -2.71.$$

Obviously, since the test statistic is slightly different, one could get slightly different answers for tests using the two different standard errors. Moreover, the results of this test will not always agree with a corresponding confidence interval for $p$ because this test uses a different standard error than the confidence interval.

We should remember that the $N(0,1)$ distribution being used for the test is only a large sample approximation. (In fact, all of our results are only approximations.) The difference between the two standard errors is often minor compared to the level of approximation inherent in using the standard normal as a reference distribution. In any case, whether we ascribe the differences to the standard errors or to the quality of the normal approximations, the exact behavior of the two test statistics can be quite different when the sample size is small. Moreover, *when p is near 0 or 1, the sample sizes must be quite large to get a good normal approximation*.

The main theoretical results for a single binomial sample are establishing that $\hat{p}$ is a reasonable

Table 8.1: *Graduate admissions at Berkeley*

| | Admitted | Rejected | Total |
|---|---|---|---|
| Females | 557 | 1278 | 1835 |
| Males | 1198 | 1493 | 2691 |

estimate of $p$ and that the variance formula given earlier is correct. The data are $y \sim \text{Bin}(n, p)$. As seen in Section 1.4, $\text{E}(y) = np$ and $\text{Var}(y) = np(1-p)$. The estimate of $p$ is $\hat{p} = y/n$. The estimate is unbiased because

$$\text{E}(\hat{p}) = \text{E}(y/n) = \text{E}(y)/n = np/n = p.$$

The variance of the estimate is

$$\text{Var}(\hat{p}) = \text{Var}(y/n) = \text{Var}(y)/n^2 = np(1-p)/n^2 = p(1-p)/n.$$

### *8.1.1 The sign test*

We now consider an alternative analysis for paired comparisons based on the binomial distribution. Consider Burt's data on IQs of identical twins raised apart from Exercise 4.5.7 and Table 4.9. The earlier discussion of paired comparisons involved assuming and validating the normal distribution for the differences in IQs between twins. In the current discussion, we make the same assumptions as before except we replace the normality assumption with the weaker assumption that the distribution of the differences is symmetric. In the earlier discussion, we would test $H_0 : \mu_1 - \mu_2 = 0$. In the current discussion, we test whether there is a $50:50$ chance that $y_1$, the IQ for the foster parent raised twin, is larger than $y_2$, the IQ for the genetic parent raised twin. In other words, we test whether $\Pr(y_1 - y_2 > 0) = .5$. We have a sample of $n = 27$ pairs of twins. If $\Pr(y_1 - y_2 > 0) = .5$, the number of pairs with $y_1 - y_2 > 0$ has a $\text{Bin}(27, .5)$ distribution. From Table 4.9, 13 of the 27 pairs have larger IQs for the foster parent raised child. (These are the differences with a positive sign, hence the name sign test.) The proportion is $\hat{p} = 13/27 = .481$. The test statistic is

$$\frac{.481 - .5}{\sqrt{.5(1-.5)/27}} = -.20$$

which is nowhere near significant.

A similar method could be used to test, say, whether there is a $50:50$ chance that $y_1$ is at least 3 IQ points greater than $y_2$. This hypothesis translates into $\Pr(y_1 - y_2 \geq 3) = .5$. The test is then based on the number of differences that are 3 or more.

The point of the sign test is the weakening of the assumption of normality. If the normality assumption is appropriate, the $t$ test of Section 4.1 is more powerful. When the normality assumption is not appropriate, some modification like the sign test should be used. In this book, the usual approach is to check the normality assumption and, if necessary, to transform the data to make the normality assumption reasonable. For a more detailed introduction to *nonparametric* methods such as the sign test, see, for example, Conover (1971).

## 8.2 Two independent binomial samples

In this section we compare two independent binomial samples. Consider again the Berkeley admissions data. Table 8.1 contains data on admissions and rejections for the 1835 females considered in Section 8.1 along with data on 2691 males. We assume that the sample of females is independent of the sample of males. Throughout, we refer to the females as the first sample and the males as the second sample.

We consider being admitted to graduate school a 'success'. Assuming that the females are a

binomial sample, they have a sample size of $n_1 = 1835$ and some probability of success, say, $p_1$. The observed proportion of female successes is

$$\hat{p}_1 = \frac{557}{1835} = .30354.$$

Treating the males as a binomial sample, the sample size is $n_2 = 2691$ and the probability of success is, say, $p_2$. The observed proportion of male successes is

$$\hat{p}_2 = \frac{1198}{2691} = .44519.$$

Our interest is in comparing the success rate of females and males. The appropriate parameter is the difference in proportions,

$$Par = p_1 - p_2.$$

The natural estimate of this parameter is

$$Est = \hat{p}_1 - \hat{p}_2 = .30354 - .44519 = -.14165.$$

With independent samples, we can find the variance of the estimate and thus the standard error. Since the females are independent of the males,

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2).$$

Using the variance formula in equation (8.1.1),

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \tag{8.2.1}$$

Estimating $p_1$ and $p_2$ and taking the square root gives the standard error,

$$\begin{aligned} \mathrm{SE}(\hat{p}_1 - \hat{p}_2) &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= \sqrt{\frac{.30354(1-.30354)}{1835} + \frac{.44519(1-.44519)}{2691}} \\ &= .01439. \end{aligned}$$

For large sample sizes $n_1$ and $n_2$, both $\hat{p}_1$ and $\hat{p}_2$ have approximate normal distributions and they are independent, so $\hat{p}_1 - \hat{p}_2$ has an approximate normal distribution and the appropriate reference distribution is approximately

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\mathrm{SE}(\hat{p}_1 - \hat{p}_2)} \sim N(0,1).$$

We now have all the requirements for applying the results of Chapter 3. A 95% confidence interval for $p_1 - p_2$ has endpoints

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where the value $1.96 = z(.975)$ is obtained from the $N(0,1)$ distribution. For comparing the female and male admissions, the 95% confidence interval for the population difference in proportions has endpoints

$$-.14165 \pm 1.96(.01439).$$

The interval is $(-.17, -.11)$. Thus we are 95% confident that the proportion of women being admitted to graduate school at Berkeley is between .11 and .17 less than that for men.

To test $H_0 : p_1 = p_2$, or equivalently $H_0 : p_1 - p_2 = 0$, against $H_A : p_1 - p_2 \neq 0$, reject an $\alpha = .10$ test if

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\mathrm{SE}(\hat{p}_1 - \hat{p}_2)} > 1.645$$

or if

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\mathrm{SE}(\hat{p}_1 - \hat{p}_2)} < -1.645.$$

Again, the value 1.645 is obtained from the $N(0,1) \equiv t(\infty)$ distribution. With the Berkeley data, the observed value of the test statistic is

$$\frac{-.14165 - 0}{.01439} = -9.84.$$

This is far smaller than $-1.645$, so the test rejects the null hypothesis of equal proportions at the .10 level. The test statistic is negative, so there is evidence that the proportion of women admitted to graduate school is lower than the proportion of men.

Once again, an alternative standard error is often used in testing problems. The test assumes that the null hypothesis is true and under the null hypothesis $p_1 = p_2$, so in constructing a standard error for the test statistic it makes sense to pool the data into one estimate of this common proportion. The pooled estimate is a weighted average of the individual estimates,

$$\begin{aligned} \hat{p}_* &= \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \\ &= \frac{1835(.30354) + 2691(.44519)}{1835 + 2691} \\ &= \frac{557 + 1198}{1835 + 2691} \\ &= .38776\,. \end{aligned}$$

Using $\hat{p}_*$ to estimate both $p_1$ and $p_2$ in equation (8.2.1) and taking the square root gives the alternative standard error

$$\begin{aligned} \mathrm{SE}(\hat{p}_1 - \hat{p}_2) &= \sqrt{\frac{\hat{p}_*(1-\hat{p}_*)}{n_1} + \frac{\hat{p}_*(1-\hat{p}_*)}{n_2}} \\ &= \sqrt{\hat{p}_*(1-\hat{p}_*)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \\ &= \sqrt{.38776(1-.38776)\left[\frac{1}{1835} + \frac{1}{2691}\right]} \\ &= .01475 \end{aligned}$$

The alternative test statistic is

$$\frac{-.14165 - 0}{.01475} = -9.60.$$

Again, the two test statistics are slightly different but the difference should be minor compared to the level of approximation involved in using the normal distribution.

A final note. Before you conclude that the data in Table 8.1 provide evidence of sex discrimination, you should realize that females tend to apply to different graduate programs than males. A more careful examination of the complete Berkeley data shows that the difference observed here results from females applying more frequently than males to highly restrictive programs, cf. Christensen (1990b, p. 96).

Table 8.2: *Swedish female births by month*

| Month | Females | $\hat{p}$ | Probability | $E$ | $(O-E)/\sqrt{E}$ |
|---|---|---|---|---|---|
| January | 3537 | .083 | 1/12 | 3549.25 | −0.20562 |
| February | 3407 | .080 | 1/12 | 3549.25 | −2.38772 |
| March | 3866 | .091 | 1/12 | 3549.25 | 5.31678 |
| April | 3711 | .087 | 1/12 | 3549.25 | 2.71504 |
| May | 3775 | .087 | 1/12 | 3549.25 | 3.78930 |
| June | 3665 | .086 | 1/12 | 3549.25 | 1.94291 |
| July | 3621 | .085 | 1/12 | 3549.25 | 1.20435 |
| August | 3596 | .084 | 1/12 | 3549.25 | 0.78472 |
| September | 3491 | .082 | 1/12 | 3549.25 | −0.97775 |
| October | 3391 | .080 | 1/12 | 3549.25 | −2.65629 |
| November | 3160 | .074 | 1/12 | 3549.25 | −6.53372 |
| December | 3371 | .079 | 1/12 | 3549.25 | −2.99200 |
| Total | 42591 | 1 | 1 | 42591.00 | |

## 8.3 One multinomial sample

In this section we investigate the analysis a single polytomous variable, i.e., a count variable with more than two possible outcomes. In particular, we assume that the data are a sample from a *multinomial* distribution, cf. Section 1.5. The multinomial distribution is a generalization of the binomial that allows more than two outcomes. We assume that each trial gives one of, say, $q$ possible outcomes. Each trial must be independent and the probability of each outcome must be the same for every trial. The multinomial distribution gives probabilities for the number of trials that fall into each of the possible outcome categories. The binomial distribution is a special case of the multinomial distribution in which $q = 2$.

The first two columns of Table 8.2 give months and numbers of Swedish females born in each month. The data are from Cramér (1946) who did not name the months. We assume that the data begin in January.

With polytomous data such as those listed in Table 8.2, there is no one parameter of primary interest. One might be concerned with the proportions of births in January, or December, or in any of the twelve months. With no one parameter of interest, the methods of Chapter 3 do not apply. Column 3 of Table 8.2 gives the observed proportions of births for each month. These are simply the monthly births divided by the total births for the year. Note that the proportion of births in March seems high and the proportion of births in November seems low.

A simplistic, yet interesting, hypothesis is that the proportion of births is the same for every month. To test this null hypothesis, we compare the number of observed births to the number of births we would expect to see if the hypothesis were true. The number of births we expect to see in any month is just the probability of having a birth in that month times the total number of births. The equal probabilities are given in column 4 of Table 8.2 and the expected values are given in column 5. The entries in column 5 are labeled $E$ for expected value and are computed as $(1/12)42591 = 3549.25$. *It cannot be overemphasized that the expectations are computed under the assumption that the null hypothesis is true*.

Comparing observed values with expected values can be tricky. Suppose an observed value is 2145 and the expected value is 2149. The two numbers are off by 4; the observed value is pretty close to the expected. Now suppose the observed value is 1 and the expected value is 5. Again the two numbers are off by 4 but now the difference between observed and expected seems quite substantial. A difference of 4 means something very different depending on how large both numbers are. To account for this phenomenon, we standardized the difference between observed and expected counts. We do this by dividing the difference by the square root of the expected count. Thus, when

we compare observed counts with expected counts we look at

$$\frac{O-E}{\sqrt{E}} \tag{8.3.1}$$

where $O$ stands for the observed count and $E$ stands for the expected count. The values in (8.3.1) are called *Pearson residuals*, after Karl Pearson.

The Pearson residuals for the Swedish female births are given in column 6 of Table 8.2. As noted earlier, the two largest deviations from the assumption of equal probabilities occur for March and November. Reasonably large deviations also occur for May and to a lesser extent December, April, October, and February. In general, *the Pearson residuals can be compared to observations from a $N(0,1)$ distribution to evaluate whether a residual is large*. For example, the residuals for March and November are 5.3 and $-6.5$. These are not values one is likely to observe from a $N(0,1)$ distribution; they provide strong evidence that birth rates in March are really larger than $1/12$ and that birth rates in November are really smaller than $1/12$.

Births seem to peak in March and they, more or less, gradually decline until November. After November, birth rates are still low but gradually increase until February. In March birth rates increase markedly. Birth rates are low in the fall and lower in the winter; they jump in March and remain relatively high, though decreasing, until September. This analysis could be performed using the monthly proportions of column 2 but the results are clearer using the residuals.

A statistic for testing whether the null hypothesis of equal proportions is reasonable can be obtained by squaring the residuals and adding them together. This statistic is known as *Pearson's* $\chi^2$ (chi-squared) statistic and is computed as

$$X^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}.$$

For the female Swedish births,

$$X^2 = 121.24.$$

Note that small values of $X^2$ indicate observed values that are similar to the expected values, so small values of $X^2$ are consistent with the null hypothesis. Large values of $X^2$ occur whenever one or more observed values are far from the expected values. To perform a test, we need some idea of how large $X^2$ could reasonably be when the null hypothesis is true. It can be shown that for a problem such as this with 1) a fixed number of cells $q$, here $q = 12$, with 2) a null hypothesis consisting of known probabilities such as those given in column 4 of Table 8.2, and with 3) large sample sizes for each cell, the null distribution of $X^2$ is approximately

$$X^2 \sim \chi^2(q-1).$$

The degrees of freedom are only $q-1$ because the $\hat{p}$s *must* add up to 1. Thus, if we know $q-1=11$ of the proportions, we can figure out the last one. Only $q-1$ of the cells are really free to vary. From Appendix B.2, the 99.5th percentile of a $\chi^2(11)$ distribution is $\chi^2(.995,11) = 26.76$. The observed $X^2$ value of 121.24 is much larger than this, so the observed value of $X^2$ could not reasonably come from a $\chi^2(11)$ distribution. In particular, an $\alpha = .005$ test of the null hypothesis is rejected easily, so the $P$ value for the test is 'much' less than .005. It follows that there is overwhelming evidence that the proportion of female Swedish births is not the same for all months.

In this example, our null hypothesis was that the probability of a female birth was the same in every month. A more reasonable hypothesis might be that the probability of a female birth is the same on every day. The months have different numbers of days so under this null hypothesis they have different probabilities. For example, assuming a 365 day year, the probability of a female birth in January is $31/365$ which is somewhat larger than $1/12$. Exercise 8.8.4 involves testing this alternative null hypothesis.

We can use results from Section 8.1 to help in the analysis of multinomial data. If we consider

Table 8.3: *Swedish births: monthly observations ($O_{ij}$s) and monthly proportions by sex*

| | Observations | | | Proportions | |
|---|---|---|---|---|---|
| Month | Female | Male | Total | Female | Male |
| January | 3537 | 3743 | 7280 | .083 | .082 |
| February | 3407 | 3550 | 6957 | .080 | .078 |
| March | 3866 | 4017 | 7883 | .091 | .088 |
| April | 3711 | 4173 | 7884 | .087 | .091 |
| May | 3775 | 4117 | 7892 | .089 | .090 |
| June | 3665 | 3944 | 7609 | .086 | .086 |
| July | 3621 | 3964 | 7585 | .085 | .087 |
| August | 3596 | 3797 | 7393 | .084 | .083 |
| September | 3491 | 3712 | 7203 | .082 | .081 |
| October | 3391 | 3512 | 6903 | .080 | .077 |
| November | 3160 | 3392 | 6552 | .074 | .074 |
| December | 3371 | 3761 | 7132 | .079 | .082 |
| Total | 42591 | 45682 | 88273 | 1.000 | 1.000 |

only the month of December, we can view each trial as a success if the birth is in December and a failure otherwise. Writing the probability of a birth in December as $p_{12}$, from Table 8.2 the estimate of $p_{12}$ is

$$\hat{p}_{12} = \frac{3371}{42591} = .07915$$

with standard error

$$\mathrm{SE}(\hat{p}_{12}) = \sqrt{\frac{.07915(1-.07915)}{42591}} = .00131$$

and a 95% confidence interval has endpoints

$$.07915 \pm 1.96(.00131).$$

The interval reduces to $(.077, .082)$. Tests for monthly proportions can be performed in a similar fashion. Bonferroni adjustments can be made to all tests and confidence intervals to control the experimentwise error rate for multiple tests or intervals, cf. Section 6.2.

## 8.4 Two independent multinomial samples

Table 8.3 gives monthly births for Swedish females and males along with various marginal totals. We wish to determine whether monthly birth rates differ for females and males. Denote the females as population 1 and the males as population 2. Thus we have a sample of 42591 females and, by assumption, an independent sample of 45682 males.

In fact, it is more likely that there is actually only one sample here, one consisting of 88273 births. It is more likely that the births have been divided into 24 categories depending on sex and birth month. Such data can be treated as two independent samples with (virtually) no loss of generality. The interpretation of results for two independent samples is considerably simpler than the interpretation necessary for one sample cross-classified by both sex and month. Thus we discuss such data as though they are independent samples. The alternative interpretation involves a multinomial sample with the probabilities for month and sex pairs all being independent.

The number of births in month $i$ for sex $j$ is denoted $O_{ij}$, where $i = 1, \ldots, 12$ and $j = 1, 2$. Thus, for example, the number of males born in December is $O_{12,2} = 3761$. Let $O_{i\cdot}$ be the total for month $i$, $O_{\cdot j}$ be the total for sex $j$, and $O_{\cdot\cdot}$ be the total over all months and sexes. For example, May has $O_{5\cdot} = 7892$, males have $O_{\cdot 2} = 45682$, and the grand total is $O_{\cdot\cdot} = 88273$.

Our interest now is in whether the population proportion of births for each month is the same for females as for males. We no longer make any assumption about what these proportions are, our null hypothesis is simply that the proportions are the same in each month. Again, we wish to compare

Table 8.4: *Estimated expected Swedish births by month ($\hat{E}_{ij}$s) and pooled proportions*

| | Expectations | | | Pooled |
|---|---|---|---|---|
| Month | Female | Male | Total | proportions |
| January | 3512.54 | 3767.46 | 7280 | .082 |
| February | 3356.70 | 3600.30 | 6957 | .079 |
| March | 3803.48 | 4079.52 | 7883 | .089 |
| April | 3803.97 | 4080.03 | 7884 | .089 |
| May | 3807.83 | 4084.17 | 7892 | .089 |
| June | 3671.28 | 3937.72 | 7609 | .086 |
| July | 3659.70 | 3925.30 | 7585 | .086 |
| August | 3567.06 | 3825.94 | 7393 | .084 |
| September | 3475.39 | 3727.61 | 7203 | .082 |
| October | 3330.64 | 3572.36 | 6903 | .078 |
| November | 3161.29 | 3390.71 | 6552 | .074 |
| December | 3441.13 | 3690.87 | 7132 | .081 |
| Total | 42591.00 | 45682.00 | 88273 | 1.000 |

the observed values, the $O_{ij}$s with expected values, but now, since we do not have hypothesized proportions for any month, we must estimate the expected values.

Under the null hypothesis that the proportions are the same for females and males, it makes sense to pool the male and female data to get an estimate of the proportion of births in each month. Using the column of monthly totals in Table 8.3, the estimated proportion for January is the January total divided by the total for the year, i.e.,

$$\hat{p}_1^0 = \frac{7280}{88273} = .0824714.$$

In general, for month $i$ we have

$$\hat{p}_i^0 = \frac{O_{i\cdot}}{O_{\cdot\cdot}}$$

where the superscript of 0 is used to indicate that these proportions are estimated under the null hypothesis of identical monthly rates for males and females. The estimate of the expected number of females born in January is just the number of females born in the year times the estimated probability of a birth in January,

$$\hat{E}_{11} = 42591(.0824714) = 3512.54.$$

The expected number of males born in January is the number of males born in the year times the estimated probability of a birth in January,

$$\hat{E}_{12} = 45682(.0824714) = 3767.46.$$

In general,

$$\hat{E}_{ij} = O_{\cdot j}\,\hat{p}_i^0 = O_{\cdot j}\frac{O_{i\cdot}}{O_{\cdot\cdot}} = \frac{O_{i\cdot}O_{\cdot j}}{O_{\cdot\cdot}}.$$

Again, *the estimated expected values are computed assuming that the proportions of births are the same for females and males in every month, i.e., assuming that the null hypothesis is true*. The estimated expected values under the null hypothesis are given in Table 8.4. Note that the totals for each month and for each sex remain unchanged.

The estimated expected values are compared to the observations using Pearson residuals, just as in Section 8.3. The Pearson residuals are

$$\tilde{r}_{ij} \equiv \frac{O_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}}}.$$

Table 8.5: *Pearson residuals for Swedish birth months, ($\tilde{r}_{ij}$s)*

| Month | Female | Male |
|---|---|---|
| January | 0.41271 | −0.39849 |
| February | 0.86826 | −0.83837 |
| March | 1.01369 | −0.97880 |
| April | −1.50731 | 1.45542 |
| May | −0.53195 | 0.51364 |
| June | −0.10365 | 0.10008 |
| July | −0.63972 | 0.61770 |
| August | 0.48452 | −0.46785 |
| September | 0.26481 | −0.25570 |
| October | 1.04587 | −1.00987 |
| November | −0.02288 | 0.02209 |
| December | −1.19554 | 1.15438 |

A more apt name for the Pearson residuals in this context may be *crude standardized residuals*. It is the standardization here that is crude and not the residuals. The standardization in the Pearson residuals ignores the fact that $\hat{E}$ is itself an estimate. Better, but considerably more complicated, standardized residuals can be defined for count data, cf. Christensen (1990b, Section IV.9). For the Swedish birth data, the Pearson residuals are given in Table 8.5. Note that when compared to a $N(0,1)$ distribution, none of the residuals is very large; all are smaller than 1.51 in absolute value.

As in Section 8.3, the sum of the squared Pearson residuals gives Pearson's $\chi^2$ statistic for testing the null hypothesis of no differences between females and males. Pearson's test statistic is

$$X^2 = \sum_{ij} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}.$$

For the Swedish birth data, computing the statistic from the 24 cells in Table 8.5 gives

$$X^2 = 14.9858.$$

For a formal test, $X^2$ is compared to a $\chi^2$ distribution. The appropriate number of degrees of freedom for the $\chi^2$ test is the number of cells in the table adjusted to account for all the parameters we have estimated as well as the constraint that the sex totals sum to the grand total. There are $12 \times 2$ cells but only $12-1$ free months and only $2-1$ free sex totals. The appropriate distribution is $\chi^2((12-1)(2-1)) = \chi^2(11)$. *The degrees of freedom are the number of data rows in Table 8.3 minus 1 times the number of data columns in Table 8.3 minus 1.* The 90th percentile of a $\chi^2(11)$ distribution is $\chi^2(.9,11) = 17.28$, so the observed test statistic $X^2 = 14.9858$ could reasonably come from a $\chi^2(11)$ distribution. In particular, the test is not significant at the .10 level. Moreover, $\chi^2(.75,11) = 13.70$, so the test has a $P$ value between .25 and .10. There is no evidence of any differences in the monthly birth rates for males and females.

Another way to evaluate the null hypothesis is by comparing the observed monthly birth proportions by sex. These observed proportions are given in Table 8.3. If the populations of females and males have the same proportions of births in each month, the observed proportions of births in each month should be similar (except for sampling variation). One can compare the numbers directly in Table 8.3 or one can make a visual display of the observed proportions as in Figure 8.1.

The methods just discussed apply equally well to the binomial data of Table 8.1. Applying the $X^2$ test given here to the data of Table 8.1 gives

$$X^2 = 92.2.$$

The statistic $X^2$ is equivalent to the test statistic given in Section 8.2 using the pooled estimate $\hat{p}_*$ to compute the standard error. The test statistic in Section 8.2 is $-9.60$, and if we square this we get

$$(-9.60)^2 = 92.2 = X^2.$$

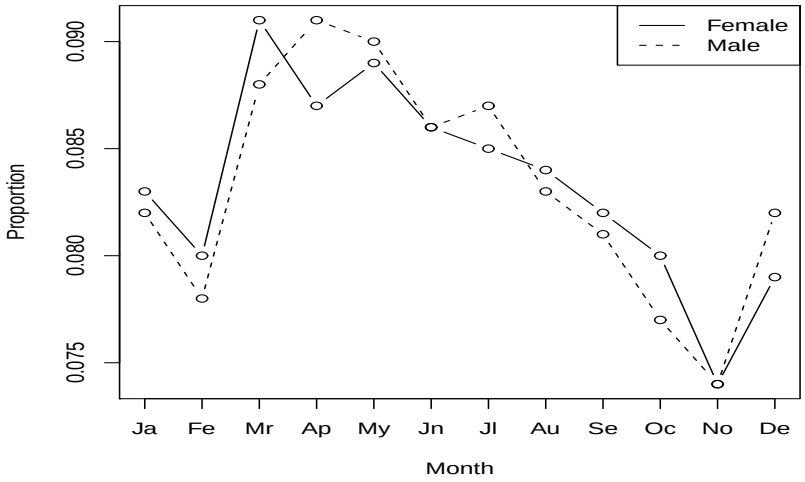



Figure 8.1: *Monthly Swedish birth proportions by sex: solid line, female; dashed line, male.*

The $-9.60$ is compared to a $N(0,1)$, while the 92.2 is compared to a $\chi^2(1)$ because Table 8.1 has 2 rows and 2 columns. A $\chi^2(1)$ distribution is obtained by squaring a $N(0,1)$ distribution, so $P$ values are identical and critical values are equivalent.

*Minitab commands*

Minitab commands for generating the analysis of Swedish birth rates are given below. Column c1 contains the observations, the $O_{ij}$s. Column c2 contains indices from 1 to 12 indicating the month of each observation and c3 contains indices for the two sexes. The subcommand 'colpercents' provides the proportions discussed in the analysis. The subcommand 'chisquare 3' gives the observations, estimated expected values, and Pearson residuals along with the Pearson test statistic.

```
MTB > read 'swede2.dat' c1 c2 c3
MTB > table c2 c3;
SUBC> frequencies c1;
SUBC> colpercents;
SUBC> chisquare 3.
```

## 8.5 Several independent multinomial samples

The methods of Section 8.4 extend easily to dealing with more than two samples. Consider the data in Table 8.6 that was extracted from Lazerwitz (1961). The data involve samples from three religious groups and consist of numbers of people in various occupational groups. The occupations are labeled A, professions; B, owners, managers, and officials; C, clerical and sales; and D, skilled. The three religious groups are Protestant, Roman Catholic, and Jewish. This is a subset of a larger collection of data that includes many more religious and occupational groups. The fact that we are restricting ourselves to a subset of a larger data set has no effect on the analysis. As discussed in Section 8.4, the analysis of these data is essentially identical regardless of whether the data come from one sample of 1926 individuals cross-classified by religion and occupation, or four independent samples of sizes 348, 477, 411, and 690 taken from the occupational groups, or three independent samples of sizes 1135, 648, and 143 taken from the religious groups. We choose to view the data as independent

Table 8.6: *Religion and occupations*

| | Occupation | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | Total |
| Protestant | 210 | 277 | 254 | 394 | 1135 |
| Roman Catholic | 102 | 140 | 127 | 279 | 648 |
| Jewish | 36 | 60 | 30 | 17 | 143 |
| Total | 348 | 477 | 411 | 690 | 1926 |

samples from the three religious groups. The data in Table 8.6 constitutes a $3 \times 4$ table because, excluding the totals, the table has 3 rows and 4 columns.

We again test whether the populations are the same. In other words, the null hypothesis is that the probability of falling into any occupational group is identical for members of the various religions. Under this null hypothesis, it makes sense to pool the data from the three religions to obtain estimates of the common probabilities. For example, under the null hypothesis of identical populations, the estimate of the probability that a person is a professional is

$$\hat{p}_1^0 = \frac{348}{1926} = .180685.$$

For skilled workers the estimated probability is

$$\hat{p}_4^0 = \frac{690}{1926} = .358255.$$

Denote the observations as $O_{ij}$ with $i$ identifying a religious group and $j$ indicating occupation. We use a dot to signify summing over a subscript. Thus the total for religious group $i$ is

$$O_{i\cdot} = \sum_j O_{ij},$$

the total for occupational group $j$ is

$$O_{\cdot j} = \sum_i O_{ij},$$

and

$$O_{\cdot\cdot} = \sum_{ij} O_{ij}$$

is the grand total. Recall that the null hypothesis is that the probability of being in an occupation group is the same for each of the three populations. Pooling information over religions, we have

$$\hat{p}_j^0 = \frac{O_{\cdot j}}{O_{\cdot\cdot}}$$

as the estimate of the probability that someone in the study is in occupational group $j$. *This estimate is only appropriate when the null hypothesis is true*.

The estimated expected count under the null hypothesis for a particular occupation and religion is obtained by multiplying the number of people sampled in that religion by the probability of the occupation. For example, the estimated expected count under the null hypothesis for Jewish professionals is

$$\hat{E}_{31} = 143(.180685) = 25.84.$$

Similarly, the estimated expected count for Roman Catholic skilled workers is

$$\hat{E}_{24} = 648(.358255) = 232.15.$$

Table 8.7: *Estimated expected counts ($\hat{E}_{ij}$s)*

| Religion | A | B | C | D | Total |
|---|---|---|---|---|---|
| Protestant | 205.08 | 281.10 | 242.20 | 406.62 | 1135 |
| Roman Catholic | 117.08 | 160.49 | 138.28 | 232.15 | 648 |
| Jewish | 25.84 | 35.42 | 30.52 | 51.23 | 143 |
| Total | 348.00 | 477.00 | 411.00 | 690.00 | 1926 |

Table 8.8: *Residuals ($\tilde{r}_{ij}$s)*

| Religion | A | B | C | D |
|---|---|---|---|---|
| Protestant | 0.34 | −0.24 | 0.76 | −0.63 |
| Roman Catholic | −1.39 | −1.62 | −0.96 | 3.07 |
| Jewish | 2.00 | 4.13 | −0.09 | −4.78 |

In general,

$$\hat{E}_{ij} = O_{i\cdot}\hat{p}_j^0 = O_{i\cdot}\frac{O_{\cdot j}}{O_{\cdot\cdot}} = \frac{O_{i\cdot}O_{\cdot j}}{O_{\cdot\cdot}}.$$

Again, *the estimated expected values are computed assuming that the null hypothesis is true*. The expected values for all occupations and religions are given in Table 8.7.

The estimated expected values are compared to the observations using Pearson residuals. The Pearson residuals are

$$\tilde{r}_{ij} = \frac{O_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}}}.$$

These crude standardized residuals are given in Table 8.8 for all occupations and religions. The largest negative residual is $-4.78$ for Jewish people with occupation D. This indicates that Jewish people were substantially underrepresented among skilled workers relative to the other two religious groups. On the other hand, Roman Catholics were substantially overrepresented among skilled workers, with a positive residual of 3.07. The other large residual in the table is 4.13 for Jewish people in group B. Thus Jewish people were more highly represented among owners, managers, and officials than the other religious groups. Only one other residual is even moderately large, the 2.00 indicating a high level of Jewish people in the professions. The main feature of these data seems to be that the Jewish group was different from the other two. A substantial difference appears in every occupational group except clerical and sales.

As in Sections 8.3 and 8.4, the sum of the squared Pearson residuals gives Pearson's $\chi^2$ statistic for testing the null hypothesis that the three populations are the same. Pearson's test statistic is

$$X^2 = \sum_{ij}\frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}.$$

Summing the squares of the values in Table 8.8 gives

$$X^2 = 60.0.$$

The appropriate number of degrees of freedom for the $\chi^2$ test is the number of data rows in Table 8.6 minus 1 times the number of data columns in Table 8.6 minus 1. Thus the appropriate reference distribution is $\chi^2((3-1)(4-1)) = \chi^2(6)$. The 99.5th percentile of a $\chi^2(6)$ distribution is $\chi^2(.995,6) = 18.55$ so the observed statistic $X^2 = 60.0$ could not reasonably come from a $\chi^2(6)$ distribution. In particular, the test is significant at the .005 level, clearly indicating that the proportions of people in the different occupation groups differ with religious group.

As in the previous section, we can informally evaluate the null hypothesis by examining the

Table 8.9: *Observed proportions by religion*

| | Occupation | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | Total |
| Protestant | .185 | .244 | .224 | .347 | 1.00 |
| Roman Catholic | .157 | .216 | .196 | .431 | 1.00 |
| Jewish | .252 | .420 | .210 | .119 | 1.00 |

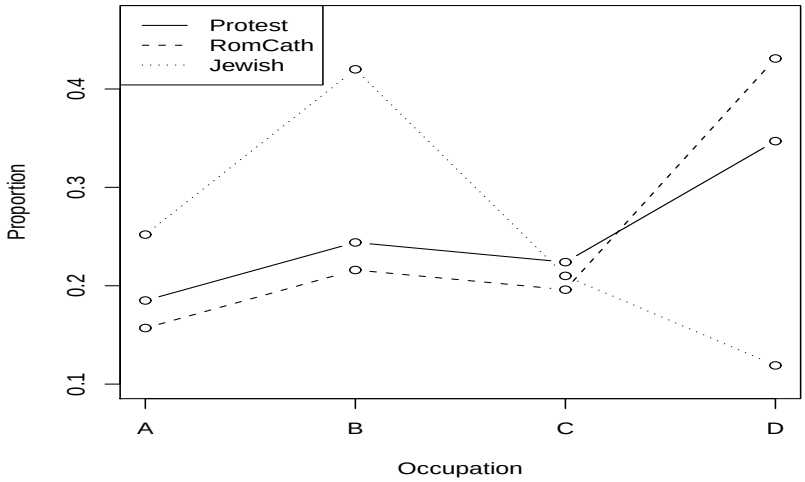



Figure 8.2 *Occupational proportions by religion: solid – protestant, long dashes – catholic, short dashes – jewish.*

observed proportions for each religious group. The observed proportions are given in Table 8.9. Under the null hypothesis, the observed proportions in each occupation category should be the same for all the religions (up to sampling variability). Figure 8.2 displays the observed proportions graphically. The Jewish group is obviously very different from the other two groups in occupations B and D and is very similar in occupation C. The Jewish proportion seems somewhat different for occupation A. The Protestant and Roman Catholic groups seem similar except that the Protestants are a bit underrepresented in occupation D and therefore are overrepresented in the other three categories. (Remember that the four proportions for each religion must add up to one, so being underrepresented in one category forces an overrepresentation in one or more other categories.)

## 8.6 Lancaster–Irwin partitioning

Lancaster–Irwin partitioning is a method for breaking a table of count data into smaller tables. When used to its maximum extent, partitioning is similar in spirit to looking at contrasts in analysis of variance. The basic idea is that a table of counts can be broken into two component tables, a reduced table and a collapsed table. Table 8.10 illustrates such a partition for the data of Table 8.6. In the reduced table, the row for the Jewish group has been eliminated, leaving a subset of the original table. In the collapsed table, the two rows in the reduced table, Protestant and Roman Catholic, have been collapsed into a single row.

In Lancaster–Irwin partitioning, we pick a group of either rows or columns, say rows. The reduced table involves all of the columns but only the chosen subgroup of rows. The collapsed table involves all of the columns and all of the rows *not* in the chosen subgroup, along with a row that

Table 8.10: *A Lancaster–Irwin partition of Table 8.6*

Reduced table

| Religion | A | B | C | D | Total |
|---|---|---|---|---|---|
| Protestant | 210 | 277 | 254 | 394 | 1135 |
| Roman Catholic | 102 | 140 | 127 | 279 | 648 |
| Total | 312 | 417 | 381 | 673 | 1783 |

Collapsed table

| Religion | A | B | C | D | Total |
|---|---|---|---|---|---|
| Prot. & R.C. | 312 | 417 | 381 | 673 | 1783 |
| Jewish | 36 | 60 | 30 | 17 | 143 |
| Total | 348 | 477 | 411 | 690 | 1926 |

combines (collapses) all of the subgroup rows into a single row. In Table 8.10 the chosen subgroup of rows contains the Protestants and Roman Catholics. The reduced table involves all occupational groups but only the Protestants and Roman Catholics. In the collapsed table the occupational groups are unaffected but the Protestants and Roman Catholics are combined into a single row. The other rows remain the same; in this case the other rows consist only of the Jewish row. As alluded to above, rather than picking a group of rows to form the partitioning, we could select a group of columns.

Lancaster–Irwin partitioning is by no means a unique process. There are as many ways to partition a table as there are ways to pick a group of rows or columns. In Table 8.10 we made a particular selection based on the residual analysis of these data from the previous section. The main feature we discovered in the residual analysis was that the Jewish group seemed to be different from the other two groups. Thus it seemed to be of interest to compare the Jewish group with a combination of the others and then to investigate what differences there might be among the other religious groups. The partitioning of Table 8.10 addresses precisely these questions.

Tables 8.11 and 8.12 provide statistics for the analysis of the reduced table and collapsed table. The reduced table simply reconfirms our previous conclusions. The $X^2$ value of 12.3 indicates substantial evidence of a difference between Protestants and Roman Catholics. The percentage point $\chi^2(.995,3) = 12.84$ indicates that the $P$ value for the test is a bit greater than .005. The residuals indicate that the difference was due almost entirely to the fact that Roman Catholics have relatively higher representation among skilled workers. (Or equivalently, that Protestants have relatively lower representation among skilled workers.) Overrepresentation of Roman Catholics among skilled workers forces their underrepresentation among other occupational groups but the level of underrepresentation in the other groups was approximately constant as indicated by the approximately equal residuals for Roman Catholics in the other three occupation groups. We will see later that for Roman Catholics in the other three occupation groups, their distribution among those groups was almost the same as those for Protestants. This reinforces the interpretation that the difference was due almost entirely to the difference in the skilled group.

The conclusions that can be reached from the collapsed table are also similar to those drawn in the previous section. The $X^2$ value of 47.5 on 3 degrees of freedom indicates overwhelming evidence that the Jewish group was different from the combined Protestant–Roman Catholic group. The residuals can be used to isolate the sources of the differences. The two groups differed in proportions of skilled workers and proportions of owners, managers, and officials. There was a substantial difference in the proportions of professionals. There was almost no difference in the proportion of clerical and sales workers between the Jewish group and the others.

The $X^2$ value computed for Table 8.6 was 60.0. The $X^2$ value for the collapsed table is 47.5 and the $X^2$ value for the reduced table is 12.3. Note that $60.0 \doteq 59.8 = 47.5 + 12.3$. It is not by chance that the sum of the $X^2$ values for the collapsed and reduced tables is approximately equal to the $X^2$ value for the original table. In fact, this relationship is a primary reason for using the Lancaster–

Table 8.11: *Reduced table*

| | Observations | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | Total |
| Protestant | 210 | 277 | 254 | 394 | 1135 |
| Roman Catholic | 102 | 140 | 127 | 279 | 648 |
| Total | 312 | 417 | 381 | 673 | 1783 |

| | Estimated expected counts | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | Total |
| Protestant | 198.61 | 265.45 | 242.53 | 428.41 | 1135 |
| Roman Catholic | 113.39 | 151.55 | 138.47 | 244.59 | 648 |
| Total | 312.00 | 417.00 | 381.00 | 673.00 | 1783 |

| | Pearson residuals | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | |
| Protestant | 0.81 | 0.71 | 0.74 | −1.66 | |
| Roman Catholic | −1.07 | −0.94 | −0.97 | 2.20 | |

$X^2 = 12.3, df = 3$

Table 8.12: *Collapsed table*

| | Observations | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | Total |
| Prot. & R.C. | 312 | 417 | 381 | 673 | 1783 |
| Jewish | 36 | 60 | 30 | 17 | 143 |
| Total | 348 | 477 | 411 | 690 | 1926 |

| | Estimated expected counts | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | Total |
| Prot. & R.C. | 322.16 | 441.58 | 380.48 | 638.77 | 1783 |
| Jewish | 25.84 | 35.42 | 30.52 | 51.23 | 143 |
| Total | 348.00 | 477.00 | 411.00 | 690.00 | 1926 |

| | Pearson residuals | | | | |
|---|---|---|---|---|---|
| Religion | A | B | C | D | |
| Prot. & R.C. | −0.57 | −1.17 | 0.03 | 1.35 | |
| Jewish | 2.00 | 4.13 | −0.09 | −4.78 | |

$X^2 = 47.5, df = 3$

Irwin partitioning method. The approximate equality $60.0 \doteq 59.8 = 47.5 + 12.3$ indicates that the vast bulk of the differences between the three religious groups is due to the collapsed table, i.e., the difference between the Jewish group and the other two. Roughly 80% (47.5/60) of the original $X^2$ value is due to the difference between the Jewish group and the others. Of course the $X^2$ value 12.2 for the reduced table is still large enough to strongly suggest differences between Protestants and Roman Catholics.

Not all data will yield an approximation as close as $60.0 \doteq 59.8 = 47.5 + 12.3$ for the partitioning. The fact that we have an approximate equality rather than an exact equality is due to our choice of the test statistic $X^2$. Pearson's statistic is simple and intuitive; it compares observed values with expected values and standardizes by the size of the expected value. An alternative test statistic also exists called the likelihood ratio test statistic. The motivation behind the likelihood ratio test statistic is not as transparent as that behind Pearson's statistic, so we will not discuss the likelihood ratio test statistic in any detail. However, one advantage of the likelihood ratio test statistic is that the sum of its values for the reduced table and collapsed table gives *exactly* the likelihood ratio test statistic for the original table. For more discussion of the likelihood ratio test statistic, see Christensen (1990b, chapter II).

Table 8.13:

| | Observations | | | |
|---|---|---|---|---|
| Religion | A | B | C | Total |
| Protestant | 210 | 277 | 254 | 741 |
| Roman Catholic | 102 | 140 | 127 | 369 |
| Total | 312 | 417 | 381 | 1110 |

| | Estimated expected counts | | | |
|---|---|---|---|---|
| Religion | A | B | C | Total |
| Protestant | 208.28 | 278.38 | 254.34 | 741 |
| Roman Catholic | 103.72 | 138.62 | 126.66 | 369 |
| Total | 312.00 | 417.00 | 381.00 | 1110 |

| | Pearson residuals | | | |
|---|---|---|---|---|
| Religion | A | B | C | |
| Protestant | 0.12 | −0.08 | 0.00 | |
| Roman Catholic | −0.17 | 0.12 | 0.03 | |

$X^2 = .065$, $df = 2$

*Further partitioning*

We began this section with the $3 \times 4$ data of Table 8.6 that has 6 degrees of freedom for its $X^2$ test. We partitioned the data into two $2 \times 4$ tables, each with 3 degrees of freedom. We can continue to use the Lancaster–Irwin method to partition the reduced and collapsed tables given in Table 8.10. The process of partitioning previously partitioned tables can be continued until the original table is broken into a collection of $2 \times 2$ tables. Each $2 \times 2$ table has one degree of freedom for its chi-squared test, so partitioning provides a way of breaking a large table into one degree of freedom components. This is similar in spirit to looking at contrasts in analysis of variance. Contrasts break the sum of squares for treatments into one degree of freedom components.

What we have been calling the reduced table involves all four occupational groups along with the two religious groups Protestant and Roman Catholic. The table was given in both Table 8.10 and Table 8.11. We now consider this table further. It was discussed earlier that the difference between Protestants and Roman Catholics can be ascribed almost entirely to the difference in the proportion of skilled workers in the two groups. To explore this we choose a new partition based on a group of *columns* that includes all occupations other than the skilled workers. Thus we get the 'reduced' table in Table 8.13 with occupations A, B, and C and the 'collapsed' table in Table 8.14 with occupation D compared to the accumulation of the other three.

Table 8.13 allows us to examine the proportions of Protestants and Catholics in the occupational groups A, B, and C. We are not investigating whether Catholics were more or less likely than Protestants to enter these occupational groups; we are examining their distribution *within* the groups. The analysis is based only on those individuals *that were in this collection of three occupational groups*. The $X^2$ value is exceptionally small, only .065. There is no evidence of any difference between Protestants and Catholics for these three occupational groups.

Table 8.13 is a $2 \times 3$ table. We could partition it again into two $2 \times 2$ tables but there is little point in doing so. We have already established that there is no evidence of differences.

Table 8.14 has the three occupational groups A, B, and C collapsed into a single group. This table allows us to investigate whether Catholics were more or less likely than Protestants to enter this group of three occupations. The $X^2$ value is a substantial 12.2 on one degree of freedom, so we can tentatively conclude that there was a difference between Protestants and Catholics. From the residuals, we see that *among people in the four occupational groups*, Catholics were more likely than Protestants to be in the skilled group and less likely to be in the other three.

Table 8.14 is a $2 \times 2$ table so no further partitioning is possible. Note again that the $X^2$ of 12.3

Table 8.14:

| | Observations | | |
|---|---|---|---|
| Religion | A & B & C | D | Total |
| Protestant | 741 | 394 | 1135 |
| Roman Catholic | 369 | 279 | 648 |
| Total | 1110 | 673 | 1783 |

| | Estimated expected counts | | |
|---|---|---|---|
| Religion | A & B & C | D | Total |
| Protestant | 706.59 | 428.41 | 1135 |
| Roman Catholic | 403.41 | 244.59 | 648 |
| Total | 1110.00 | 673.00 | 1783 |

| | Pearson residuals | | |
|---|---|---|---|
| Religion | A & B & C | D | |
| Protestant | 1.29 | −1.66 | |
| Roman Catholic | −1.71 | 2.20 | |

$X^2 = 12.2, df = 1$

Table 8.15:

| | Observations | | |
|---|---|---|---|
| Religion | A & B & D | C | Total |
| Prot. & R.C. | 1402 | 381 | 1783 |
| Jewish | 113 | 30 | 143 |
| Total | 1515 | 411 | 1926 |

| | Estimated expected counts | | |
|---|---|---|---|
| Religion | A & B & D | C | Total |
| Prot. & R.C. | 1402.52 | 380.48 | 1783 |
| Jewish | 112.48 | 30.52 | 143 |
| Total | 1515.00 | 411.00 | 1926 |

| | Pearson residuals | | |
|---|---|---|---|
| Religion | A & B & D | C | |
| Prot. & R.C. | −0.00 | 0.03 | |
| Jewish | 0.04 | −0.09 | |

$X^2 = .01, df = 1$

from Table 8.11 is approximately equal to the sum of the .065 from Table 8.13 and the 12.2 from Table 8.14.

Finally, we consider additional partitioning of the collapsed table given in Tables 8.10 and 8.12. It was noticed earlier that the Jewish group seemed to differ from Protestants and Catholics in every occupational group except C, clerical and sales. Thus we choose a partitioning that isolates group C. Table 8.15 gives a collapsed table that compares C to the combination of groups A, B, and D. Table 8.16 gives a reduced table that involves only occupational groups A, B, and D.

Table 8.15 demonstrates no difference between the Jewish group and the combined Protestant–Catholic group. Thus the proportion of people in clerical and sales was the same for the Jewish group as for the combined Protestant and Roman Catholic group. Any differences between the Jewish and Protestant–Catholic groups must be in the proportions of people *within* the three occupational groups A, B, and D.

Table 8.16 demonstrates major differences between occupations A, B, and D for the Jewish group and the combined Protestant–Catholic group. As seen earlier and reconfirmed here, skilled workers had much lower representation among the Jewish group, while professionals and especially owners, managers, and officials had much higher representation among the Jewish group.

Table 8.16:

| | Observations | | | |
|---|---|---|---|---|
| Religion | A | B | D | Total |
| Prot. & R.C. | 312 | 417 | 673 | 1402 |
| Jewish | 36 | 60 | 17 | 113 |
| Total | 348 | 477 | 690 | 1515 |

| | Estimated expected counts | | | |
|---|---|---|---|---|
| Religion | A | B | D | Total |
| Prot. & R.C. | 322.04 | 441.42 | 638.53 | 1402 |
| Jewish | 25.96 | 35.58 | 51.47 | 113 |
| Total | 348.00 | 477.00 | 690.00 | 1515 |

| | Pearson residuals | | | |
|---|---|---|---|---|
| Religion | A | B | D | |
| Prot. & R.C. | −0.59 | −1.16 | 1.36 | |
| Jewish | 1.97 | 4.09 | −4.80 | |

$X^2 = 47.2, df = 2$

Table 8.17:

| | Observations | | |
|---|---|---|---|
| Religion | B | D | Total |
| Prot. & R.C. | 417 | 673 | 1090 |
| Jewish | 60 | 17 | 77 |
| Total | 477 | 690 | 1167 |

| | Estimated expected counts | | |
|---|---|---|---|
| Religion | B | D | Total |
| Prot. & R.C. | 445.53 | 644.47 | 1090 |
| Jewish | 31.47 | 45.53 | 77 |
| Total | 477.00 | 690.00 | 1167 |

| | Pearson residuals | | |
|---|---|---|---|
| Religion | B | D | |
| Prot. & R.C. | −1.35 | 1.12 | |
| Jewish | 5.08 | −4.23 | |

$X^2 = 46.8, df = 1$

Table 8.16 can be further partitioned into Tables 8.17 and 8.18. Table 8.17 is a reduced $2 \times 2$ table that considers the difference between the Jewish group and others with respect to occupational groups B and D. Table 8.18 is a $2 \times 2$ collapsed table that compares occupational group A with the combination of groups B and D.

Table 8.17 shows a major difference between occupational groups B and D. Table 8.18 may or may not show a difference between group A and the combination of groups B and D. The $X^2$ values are 46.8 and 5.45 respectively. The question is whether an $X^2$ value of 5.45 is suggestive of a difference between religious groups when we have examined the data in order to choose the partitions of Table 8.6. Note that the two $X^2$ values sum to 52.25, whereas the $X^2$ value for Table 8.16, from which they were constructed, is only 47.2. The approximate equality is a very rough approximation. Nonetheless, we see from the relative sizes of the two $X^2$ values that the majority of the difference between the Jewish group and the other religious groups was in the proportion of owners, managers, and officials as compared to the proportion of skilled workers.

Ultimately, we have partitioned Table 8.6 into Tables 8.13, 8.14, 8.15, 8.17, and 8.18. These are all $2 \times 2$ tables except for Table 8.13. We could also have partitioned Table 8.13 into two $2 \times 2$ tables but we chose to leave it because it showed so little evidence of any difference between

Table 8.18:

| | Observations | | |
|---|---|---|---|
| Religion | A | B & D | Total |
| Prot. & R.C. | 312 | 1090 | 1402 |
| Jewish | 36 | 77 | 113 |
| Total | 348 | 1167 | 1515 |

| | Estimated expected counts | | |
|---|---|---|---|
| Religion | A | B & D | Total |
| Prot. & R.C. | 322.04 | 1079.96 | 1402 |
| Jewish | 25.96 | 87.04 | 113 |
| Total | 348.00 | 1167.00 | 1515 |

| | Pearson residuals | | |
|---|---|---|---|
| Religion | A | B & D | |
| Prot. & R.C. | −0.56 | 0.30 | |
| Jewish | 1.97 | −1.08 | |

$X^2 = 5.45$, $df = 1$

Protestants and Roman Catholics for the three occupational groups considered. The $X^2$ value of 60.0 for Table 8.6 was approximately partitioned into $X^2$ values of .065, 12.2, .01, 46.8, and 5.45 respectively. Except for the .065 from Table 8.13, each of these values is computed from a $2 \times 2$ table, so each has 1 degree of freedom. The .065 is computed from a $2 \times 3$ table, so it has 2 degrees of freedom. The sum of the five $X^2$ values is 64.5 which is roughly equal to the 60.0 from Table 8.6.

The five $X^2$ values can all be used in testing. Not only does such testing involve the usual problems associated with multiple testing but we even let the data suggest the partitions. It is inappropriate to compare these $X^2$ values to their usual $\chi^2$ percentage points to obtain tests. A simple way to adjust for both the multiple testing and the data dredging (letting the data suggest partitions) is to compare all $X^2$ values to the percentage points appropriate for Table 8.6. For example, the $\alpha = .05$ test for Table 8.6 uses the critical value $\chi^2(.95, 6) = 12.58$. By this standard, Table 8.17 with $X^2 = 46.8$ shows a significant difference between religious groups and Table 8.14 with $X^2 = 12.2$ nearly shows a significant difference between religious groups. The value of $X^2 = 5.45$ for Table 8.18 gives no evidence of a difference based on this criterion even though such a value would be highly suggestive if we could compare it to a $\chi^2(1)$ distribution. This method is similar in spirit to Scheffé's method from Section 6.4 and suffers from the same extreme conservatism.

## 8.7 Logistic regression

Logistic regression is a method of modeling the relationships between probabilities and predictor variables. We begin with an example.

EXAMPLE 8.7.1. Woodward et al. (1941) reported data on 120 mice divided into 12 groups of 10. The mice in each group were exposed to a specific dose of chloracetic acid and the observations consist of the number in each group that lived and died. Doses were measured in grams of acid per kilogram of body weight. The data are given in Table 8.19, along with the proportions of mice who died at each dose. We could analyze these data using the methods discussed earlier in this chapter; we have samples from twelve populations and we could test to see if the populations are the same. In addition though, we can try to model the relationship between dose level and the probability of dying. If we can model the probability of dying as a function of dose, we can make predictions about the probability of dying for any dose levels that are similar to those in the original data. □

Logistic regression as applied to this example is somewhat like fitting a simple linear regression to one-way ANOVA data as discussed in Section 7.12. In Section 7.12 we considered data on the

Table 8.19: *Lethality of chloracetic acid*

| Dose | Group | Died | Survived | Total | $\hat{p}_i$ |
|---|---|---|---|---|---|
| .0794 | 1 | 1 | 9 | 10 | .1 |
| .1000 | 2 | 2 | 8 | 10 | .2 |
| .1259 | 3 | 1 | 9 | 10 | .1 |
| .1413 | 4 | 0 | 10 | 10 | .0 |
| .1500 | 5 | 1 | 9 | 10 | .1 |
| .1588 | 6 | 2 | 8 | 10 | .2 |
| .1778 | 7 | 4 | 6 | 10 | .4 |
| .1995 | 8 | 6 | 4 | 10 | .6 |
| .2239 | 9 | 4 | 6 | 10 | .4 |
| .2512 | 10 | 5 | 5 | 10 | .5 |
| .2818 | 11 | 5 | 5 | 10 | .5 |
| .3162 | 12 | 8 | 2 | 10 | .8 |

ASI indices given in Table 7.10. These data have seven observations on each of five plate lengths. The data can be analyzed as either a one-way ANOVA or as a simple linear regression, and in Section 7.12 we examined relationships between the two approaches. In particular, we mentioned that the estimated regression line could be obtained by fitting a line to the sample means for the five groups. The analysis of the lethality data takes a similar approach. Instead of fitting a line to sample means, we perform a regression on the observed proportions. Unfortunately, a standard regression is inappropriate because the observed proportions do not have constant variance. For $i = 1, \ldots, q$, $\hat{p}_i$ is the observed proportion from $N_i$ binomial trials, so as discussed in Section 8.1, $\text{Var}(\hat{p}_i) = p_i(1-p_i)/N_i$. One approach is to use the variance stabilizing transformation from Sections 2.3 and 7.10 on the $\hat{p}_i$s and then apply standard regression methods. As alluded to in Section 2.3, there are better methods available and this section briefly introduces some of them.

We begin with a reasonably simple analysis of the chloracetic acid data. This analysis involves not only a transformation of the $\hat{p}_i$s but incorporating weights into the simple linear regression procedure. Weighted regression is a method for dealing with nonconstant variances in the observations. If the variances are not constant, *observations with large variances should be given relatively little weight, while observations with small variances are given increased weight*. The details of weighted regression are discussed in Section 15.7. The discussion given there requires one to know the material in Chapter 13 and the first five sections of Chapter 15, but considerable insight can be obtained from Examples 15.7.1 and 15.7.2. These examples merely require the background from Section 7.12.

In weighted regression for binomial data we take the observations on the dependent variable as

$$\log[\hat{p}_i/(1-\hat{p}_i)].$$

We then fit the model

$$\log[\hat{p}_i/(1-\hat{p}_i)] = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with weights

$$w_i = N_i \hat{p}_i (1-\hat{p}_i).$$

The regression estimates from this method minimize the weighted sum of squares

$$\sum_{i=1}^{q} w_i \left(\log[\hat{p}_i/(1-\hat{p}_i)] - \beta_0 - \beta_1 x_i\right)^2.$$

There are a couple of serious drawbacks to this procedure. First, *the weights are really only appropriate if all the samples sizes $N_i$ are large*. The weights rely on large sample variance formulae and the law of large numbers. Second, *the values* $\log[\hat{p}_i/(1-\hat{p}_i)]$ *are not always defined*. If we have an observed proportion with $\hat{p}_i = 0$ or 1, $\log[\hat{p}_i/(1-\hat{p}_i)]$ is undefined. Either we are trying to take

the log of zero or we are trying to divide by zero. With $\hat{p}_i = y_i/N_i$, so that $y_i$ is the number of 'successes,' this problem occurs whenever $y_i$ equals 0 or $N_i$. The problem is often dealt with by adding or subtracting a small number to $y_i$. Generally, *the size of the small number should be chosen to be small in relation to the size of* $N_i$. In many applications, *all* of the $N_i$s are 1. *In any case with* $N_i = 1$, $\hat{p}_i$ *is always either 0 or 1, so* $\log[\hat{p}_i/(1-\hat{p}_i)]$ *is always undefined*. These drawbacks are not as severe with another method of analysis that we will examine later.

EXAMPLE 8.7.1 CONTINUED. We now return to the chloracetic acid data. In this example $N_i = 10$ for all $i$, so the sample sizes are all reasonably large. For dose $x = .1413$, the number of deaths was 0, so the observed proportion was zero. We handle this problem by treating the observed count as .5, so the observed proportion is taken as .5/10 = .05. A computer program for regression analysis will typically give output such as the following tables.

Raw parameter table

| Predictor | $\hat{\beta}_k$ | SE($\hat{\beta}_k$) | $t$ | $P$ |
|---|---|---|---|---|
| Constant | −3.1886 | 0.5914 | −5.39 | 0.000 |
| Dose | 13.181 | 2.779 | 4.74 | 0.000 |

Analysis of variance: weighted simple linear regression

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Regression | 1 | 15.282 | 15.282 | 22.50 | 0.000 |
| Error | 10 | 6.791 | 0.679 | | |
| Total | 11 | 22.074 | | | |

*The estimates of the regression parameters are appropriate but everything involving variances in these tables is wrong!* The problem is that with binomial data the variance depends solely on the probability and we have already accounted for the variance in defining the weights. Thus there is no separate parameter $\sigma^2$ to deal with but standard regression output is designed to adjust for such a parameter. To obtain appropriate standard errors, we need to divide the reported standard errors by $\sqrt{MSE}$. The adjusted table is given below.

Adjusted parameter table

| Predictor | $\hat{\beta}_k$ | SE($\hat{\beta}_k$) | $t$ | $P$ |
|---|---|---|---|---|
| Constant | −3.1886 | 0.7177 | −4.44 | 0.000 |
| Dose | 13.181 | 3.373 | 3.91 | 0.000 |

The table provides clear evidence of the need for both parameters. To predict the probability of death for rats given a dose $x$, the predicted probability $\hat{p}$ satisfies

$$\log[\hat{p}/(1-\hat{p})] = \hat{\beta}_0 + \hat{\beta}_1 x = -3.1886 + 13.181x.$$

Solving for $\hat{p}$ gives

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x)} = \frac{\exp(-3.1886 + 13.181x)}{1+\exp(-3.1886 + 13.181x)}.$$

For example, if $x = .3$, $-3.1886 + 13.181(.3) = .7657$ and $\hat{p} = e^{.7657}/(1+e^{.7657}) = .68$.

The only interest in the ANOVA table is in the error line. As we have seen, $\sqrt{MSE}$ is needed to adjust the standard errors. In addition, the $SSE$ provides a lack of fit test similar in spirit to that discussed in Section 7.12. To test for lack of fit compare $SSE$ to a $\chi^2(dfE)$ distribution. Large values of $SSE$ indicate lack of fit. In this example $SSE = 6.791$, which is smaller than $dfE = 10$, so the $\chi^2$ test gives no evidence of lack of fit. A line seems to fit these data adequately. □

*Minitab commands*

In Minitab let c1 contain the doses, c2 contain the number of deaths, and c3 contain the number of trials (10 in each case). The commands for this analysis are given below.

```
MTB > let c5=c2/c3
MTB > let c6=1-c5
MTB > let c7=loge(c5/c6)
MTB > let c8=c3*c5*c6
MTB > regress c7 on 1 c1;
SUBC> weights c8.
```

*The logistic model and maximum likelihood*

When we have a one-way ANOVA with treatments that are quantitative levels of some factor, we can fit either the one-way ANOVA model

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

or the simple linear regression model

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}.$$

We can think of the regression as a model for the $\mu_i$s, i.e.,

$$\mu_i = \beta_0 + \beta_1 x_i.$$

Logistic regression uses a very similar idea. The binomial situation here has 'observations' $\hat{p}_i = y_i/N_i$ where $y_i \sim \text{Bin}(N_i, p_i)$, $i = 1, \ldots, q$. In logistic regression, we model the parameters $p_i$. In particular, the model is

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 x_i. \tag{8.7.1}$$

The question is then how to fit this model. The weighted regression approach was discussed earlier. The weighted regression estimates are the values of $\beta_0$ and $\beta_1$ that minimize the function

$$\sum_{i=1}^{q} w_i \left(\log[\hat{p}_i/(1-\hat{p}_i)] - \beta_0 - \beta_1 x_i\right)^2.$$

An alternative method for estimating the parameters is to maximize something called the likelihood function.

Recall from Section 1.4 that the probability function for an individual binomial, say, $y_i \sim \text{Bin}(N_i, p_i)$ is

$$\Pr(y_i = r_i) = \binom{N_i}{r_i} p_i^{r_i} (1-p_i)^{N_i - r_i}.$$

We are dealing with *q independent* binomials, so probabilities for the entire collection of random variables are obtained by multiplying the probabilities for the individual events.

One of the things that students initially find confusing about statistical theory is that we often use the same symbols for random variables and for observations from those random variables. I am about to do the same thing. I want to write down the probability of the data that we actually saw. If we saw $y_i$, the probability of seeing that is

$$\binom{N_i}{y_i} p_i^{y_i} (1-p_i)^{N_i - y_i}.$$

If all together we saw $y_1, \ldots, y_q$, the probability of obtaining all those values is the product of the individual probabilities, i.e.,

$$\prod_{i=1}^{q} \binom{N_i}{y_i} p_i^{y_i} (1-p_i)^{N_i - y_i}. \tag{8.7.2}$$

This probability of getting the observed data is called the *likelihood function*. In the likelihood function we know all of the $N_i$s and $y_i$s but we do not know the $p_i$s. Thus the likelihood is a function of the $p_i$s. It is not too difficult to show that the maximum value of the likelihood function is obtained by taking $p_i = \hat{p}_i = y_i/N_i$ for all $i$. The observed proportions $\hat{p}_i$ are the values of the parameters that maximize the probability of getting the observed data. We say that such values are *maximum likelihood estimates (mles)* of the parameters $p_i$.

The model (8.7.1) specifies the $p_i$s in terms of $\beta_0$ and $\beta_1$. We can solve (8.7.1) for $p_i$ by writing

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{8.7.3}$$

If we now substitute this formula for $p_i$ into equation (8.7.2) we get the likelihood as a function of $\beta_0$ and $\beta_1$. The maximum likelihood estimates of $\beta_0$ and $\beta_1$ are simply the values of $\beta_0$ and $\beta_1$ that maximize the likelihood function. Equations (8.7.1) and (8.7.3) are equivalent ways of writing the model. Equation (8.7.3) is actually the logistic regression model and equation (8.7.1) is the corresponding *logit* model.

Computer programs are available for finding maximum likelihood estimates. Such programs typically give standard errors that are valid for large samples. If the large sample approximations are appropriate, the parameters, estimates, and standard errors can be used as in Chapter 3 with a $N(0,1)$ reference distribution. For the approximations to be valid, it is typically enough that the total number of trials in the entire data be large; the individual sample sizes $N_i$ need not be large.

Maximum likelihood theory also provides a test of lack of fit similar to the weighted regression $\chi^2$ test using the *SSE*. In maximum likelihood theory the test examines the value of the likelihood (8.7.2) when using the mles of $\beta_0$ and $\beta_1$ in equation (8.7.3) to determine the $p_i$s, and compares that value to the likelihood when using the observed proportions $\hat{p}_i$ as the $p_i$s. Using the observed proportions involves less structure so the likelihood value will be greater using them. The lack of fit test statistic is $-2$ times the log of the ratio of the likelihood using the estimated $\beta_k$s to the likelihood using the $\hat{p}_i$s. This test statistic is properly called the (generalized) likelihood ratio test statistic but is often simply called the *deviance*. (The likelihood ratio test was also mentioned in the previous section.) The deviance is compared to a $\chi^2(q-2)$ distribution where $q$ is the number of independent binomials and 2 is the number of regression parameters in the logistic model. Unlike the standard errors for the $\beta_i$s, *all the sample sizes $N_i$ must be large for the lack of fit test to be valid!*

EXAMPLE 8.7.2. Maximum likelihood for the chloracetic acid data gives the following results.

| Predictor | $\hat{\beta}_k$ | SE($\hat{\beta}_k$) | $t$ | $P$ |
|---|---|---|---|---|
| Constant | $-3.570$ | 0.7040 | $-5.07$ | 0.000 |
| Dose | 14.64 | 3.326 | 4.40 | 0.000 |

These are similar to the weighted regression results. The deviance of the maximum likelihood fit is 10.254 with $12 - 2 = 10$ degrees of freedom for the lack of fit test. The sample sizes are all reasonably large, so a $\chi^2$ test is appropriate. The test statistic is approximately equal to the degrees of freedom, so a test would not be rejected. A simple line seems to fit the data adequately. The maximum likelihood results were obtained using the computer program GLIM. □

We will not analyze more sophisticated count data in this book but we should mention that *both the maximum likelihood methods and the weighted regression methods extend to much more general models*, such as those treated in the remainder of the book. Both methods work when there are many predictors, so we can perform multiple logistic regression which is similar in spirit to multiple

Table 8.20: *French convictions*

| Year | Convictions | Accusations |
|---|---|---|
| 1825 | 4594 | 7234 |
| 1826 | 4348 | 6988 |
| 1827 | 4236 | 6929 |
| 1828 | 4551 | 7396 |
| 1829 | 4475 | 7373 |
| 1830 | 4130 | 6962 |

regression as treated in Chapters 13, 14, and 15. By modifying the matrix approach to ANOVA problems discussed in Section 16.5, the methods introduced here can be applied to models that are structured like analysis of variance and even analysis of covariance. Christensen (1990b) contains a more complete discussion of logistic regression and logit models. It also contains references to additional work.

## 8.8 Exercises

EXERCISE 8.8.1. Reiss et al. (1975) and Fienberg (1980) reported that 29 of 52 virgin female undergraduate university students who used a birth control clinic thought that extramarital sex is not always wrong. Give a 99% confidence interval for the population proportion of virgin undergraduate university females who use a birth control clinic and think that extramarital sex is not always wrong.

In addition, 67 of 90 virgin females who did not use the clinic thought that extramarital sex is not always wrong. Give a 99% confidence interval for the difference in proportions between the two groups and give a .05 level test that there is no difference.

EXERCISE 8.8.2. Pauling (1971) reports data on the incidence of colds among French skiers who where given either ascorbic acid or a placebo. Of 139 people given ascorbic acid, 17 developed colds. Of 140 people given the placebo, 31 developed colds. Do these data suggest that the proportion of people who get colds differs depending on whether they are given ascorbic acid?

EXERCISE 8.8.3. Quetelet (1842) and Stigler (1986, p. 175) report data on conviction rates in the French Courts of Assize (Law Courts) from 1825 to 1830. The data are given in Table 8.20. Test whether the conviction rate is the same for each year. Use $\alpha = .05$. (Hint: Table 8.20 is written in a nonstandard form. You need to modify it before applying the methods of this chapter.) If there are differences in conviction rates, use residuals to explore these differences.

EXERCISE 8.8.4. Use the data in Table 8.2 to test whether the probability of a birth in each month is the number of days in the month divided by 365. Thus the null probability for January is $31/365$ and the null probability for February is $28/365$.

EXERCISE 8.8.5. Snedecor and Cochran (1967) report data from an unpublished report by E. W. Lindstrom. The data concern the results of cross-breeding two types of corn (maize). In 1301 crosses of two types of plants, 773 green, 231 golden, 238 green-golden, and 59 golden-green-striped plants were obtained. If the inheritance of these properties is particularly simple, Mendelian genetics suggests that the probabilities for the four types of corn may be $9/16$, $3/16$, $3/16$, and $1/16$, respectively. Test whether these probabilities are appropriate. If they are inappropriate, identify the problem.

EXERCISE 8.8.6. In France in 1827, 6929 people were accused in the courts of assize and 4236

Table 8.21: *Occupation and religion*

| Religion | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| White Baptist | 43 | 78 | 64 | 135 | 135 | 57 | 86 | 114 |
| Black Baptist | 9 | 2 | 9 | 23 | 47 | 77 | 18 | 41 |
| Methodist | 73 | 80 | 80 | 117 | 102 | 58 | 66 | 153 |
| Lutheran | 23 | 36 | 43 | 59 | 46 | 26 | 49 | 46 |
| Presbyterian | 35 | 54 | 38 | 46 | 19 | 22 | 11 | 46 |
| Episcopalian | 27 | 27 | 20 | 14 | 7 | 5 | 2 | 15 |

Table 8.22: *Heights and chest circumferences*

| | Heights | | | | | |
|---|---|---|---|---|---|---|
| Chest | 64–65 | 66–67 | 68–69 | 70–71 | 71–73 | Total |
| 39 | 142 | 442 | 341 | 117 | 20 | 1062 |
| 40 | 118 | 337 | 436 | 153 | 38 | 1082 |
| Total | 260 | 779 | 777 | 270 | 58 | 2144 |

were convicted. In 1828, 7396 people were accused and 4551 were convicted. Give a 95% confidence interval for the proportion of people convicted in 1827. At the .01 level, test the null hypothesis that the conviction rate in 1827 was greater than or equal to $2/3$. Does the result of the test depend on the choice of standard error? Give a 95% confidence interval for the difference in conviction rates between the two years. Test the hypothesis of no difference in conviction rates using $\alpha = .05$ and both standard errors.

EXERCISE 8.8.7. Table 8.21 contains additional data from Lazerwitz (1961). These consist of a breakdown of the Protestants in Table 8.6 but with the addition of four more occupational categories. The additional categories are E, semiskilled; F, unskilled; G, farmers; H, no occupation. Analyze the data with an emphasis on partitioning the table.

EXERCISE 8.8.8. Stigler (1986, p. 208) reports data from the *Edinburgh Medical and Surgical Journal* (1817) on the relationship between heights and chest circumferences for Scottish militia men. Measurements were made in inches. We concern ourselves with two groups of men, those with 39 inch chests and those with 40 inch chests. The data are given in Table 8.22. Test whether the distribution of heights is the same for these two groups.

EXERCISE 8.8.9. Use weighted least squares to fit a logistic model to the data of Table 8.20 that relates probability of conviction to year. Is there evidence of a trend in the conviction rates over time? Is there evidence for a lack of fit?

EXERCISE 8.8.10. Is it reasonable to fit a logistic regression to the data of Table 8.22? Why or why not? Explain what such a model would be doing. Whether reasonable or not, fitting such a model can be done. Use weighted least squares to fit a logistic model and discuss the results. Is there evidence for a lack of fit?

Chapter 9

# Basic experimental designs

In this chapter we examine the three most basic experimental designs: completely randomized designs (CRDs), randomized complete block (RCB) designs, and Latin square designs. Completely randomized designs are the simplest of these and have been used previously without having been named. Also, we have previously performed an analysis for a randomized complete block design.

The basic object of experimental design is to construct an experiment that allows for a valid estimate of $\sigma^2$, the variance of the observations. Obtaining a valid *estimate of error* requires appropriate replication of the experiment. Having one observation on each treatment is not sufficient. All three of the basic designs considered in this chapter allow for a valid estimate of the variance.

A second important consideration is to construct a design that yields a small variance. A smaller variance leads to sharper statistical inferences, i.e., narrower confidence intervals and more powerful tests. A fundamental tool for reducing variability is *blocking*. The basic idea is to examine the treatments on homogeneous experimental material. With four drug treatments and observations on eight animals, a valid estimate of the error can be obtained by randomly assigning each of the drugs to two animals. *If the treatments are assigned completely at random to the experimental units* (animals), *the design is a completely randomized design*. Generally, a smaller variance for treatment comparisons is obtained when the eight animals consist of two litters of four siblings and each treatment is applied to one randomly selected animal from each litter. With each treatment applied in every litter, all comparisons among treatments can be performed *within* each litter. Having at least two litters is necessary to get a valid estimate of the variance of the comparisons. *Randomized complete block designs: 1) identify blocks of homogeneous experimental material (units) and 2) randomly assign each treatment to an experimental unit within each block*. The blocks are complete in the sense that each block contains all of the treatments.

*Latin squares use two forms of blocking at once*. For example, if we suspect that birth order within the litter might also have an important effect on our results, we continue to take observations on each treatment within every litter, but we also want to have each treatment observed in every birth order. This is accomplished by having four litters with treatments arranged in a Latin square design.

Another fundamental concept in experimental design is the idea that the experimenter has the ability to randomly assign the treatments to the experimental material available. This is not always the case.

EXAMPLE 9.0.1. In Chapter 5, we considered two examples of one-way analysis of variance; neither were designed experiments. For the suicide ages, a designed experiment would require that we take a group of people who we know will commit suicide and randomly assign one of the ethnic groups, non-Hispanic Caucasian, Hispanic, or Native American, to the people. Obviously a difficult task. With the electrical characteristic data, rather than having ceramic sheets divided into strips, a designed experiment would require starting with different pieces of ceramic material and randomly assigning the pieces to have come from a particular ceramic strip. □

Random assignment of treatments to experimental units allows one to infer causation from a

designed experiment. If treatments are *randomly* assigned to experimental units, then the only systematic differences between the units are the treatments. Barring an unfortuitous randomization, such differences must be caused by the treatments because they cannot be caused by anything else. However, as discussed below, the 'treatments' may be more involved than the experimenter realizes.

Random assignment of treatments does not mean haphazard assignment. Haphazard assignment is subject to the (unconscious) biases of the person making the assignments. Random assignment uses a reliable table of random numbers or a reliable computer program to generate random numbers. It then uses these numbers to assign treatments. For example, suppose we have four experimental units labeled $u_1$, $u_2$, $u_3$, and $u_4$ and four treatments labeled A, B, C, and D. Given a program or table that provides random numbers between 0 and 1 (i.e., random samples from a uniform(0,1) distribution), we associate numbers between 0 and .25 with treatment A, numbers between .25 and .50 with treatment B, numbers between .50 and .75 with treatment C, and numbers between .75 and 1 with treatment D. The first random number selected determines the treatment for $u_1$. If the first number is .6321, treatment C is assigned to $u_1$ because $.50 < .6321 < .75$. If the second random number is .4279, $u_2$ gets treatment B because $.25 < .4279 < .50$. If the third random number is .2714, $u_3$ would get treatment B, but we have already assigned treatment B to $u_2$, so we throw out the third number. If the fourth number is .9153, $u_3$ is assigned treatment D. Only one unit and one treatment are left, so $u_4$ gets treatment A. Any reasonable rule (decided ahead of time) can be used to make the assignment if a random number hits a boundary, e.g., if a random number comes up, say, .2500.

In cases such as those discussed previously, i.e., in *observational studies* where the treatments are not randomly assigned to experimental units, it is much more difficult to infer causation. If we find differences, there are differences in the corresponding populations, but it does not follow that the differences are caused by the labels given to the populations. If the average suicide age is lower for Native Americans, we know only that the phenomenon exists, we do not know what aspects of being Native American cause the phenomenon. Perhaps low socioeconomic status causes early suicides and Native Americans are over represented in the low socioeconomic strata. We don't know and it will be difficult to ever know using the only possible data on such matters, data that come from observational studies.

One also needs to realize that the treatments in an experiment may not be what the experimenter thinks they are. Suppose you want to test whether artificial sweeteners made with a new chemical cause cancer. You get some rats, randomly divide them into a treatment group and a control. You inject the treatment rats with a solution of the sweetener combined with another (supposedly benign) chemical. You leave the control rats alone. For simplicity you keep the treatment rats in one cage and the control rats in another cage. Eventually, you find an increased risk of cancer among the treatment rats as compared to the control rats. You can reasonably conclude that the treatments caused the increased cancer rate. Unfortunately, you do not really know whether the sweetener or the supposedly benign chemical or the combination of the two caused the cancer. In fact, you do not really know that it was the chemicals that caused the cancer. Perhaps the process of injecting the rats caused the cancer or perhaps something about the environment in the treatment rats's cage caused the cancer. A treatment consists of *all the ways* in which a group is treated differently from other groups. It is crucially important to *treat all experimental units as similarly as possible so that (as nearly as possible) the only differences between the units are the agents that were meant to be investigated*.

Ideas of blocking can also be useful in observational studies. While one cannot really create blocks in observational studies, one can adjust for important groupings.

EXAMPLE 9.0.2. If we wish to study whether cocaine users are more paranoid than other people, we may decide that it is important to block on socioeconomic status. This is appropriate if the underlying level of paranoia in the population differs by socioeconomic status. Conducting an experiment in this setting is difficult. Given groups of people of various socioeconomic statuses, it is

a rare researcher who has the luxury of deciding which subjects will ingest cocaine and which will not. □

The seminal work on experimental design was written by Fisher (1935). It is still well worth reading. My favorite source on the ideas of experimentation is Cox (1958). The books by Cochran and Cox (1957) and Kempthorne (1952) are classics. Cochran and Cox is more applied. Kempthorne is more theoretical. There is a huge literature in both journal articles and books on the general subject of designing experiments. The article by Coleman and Montgomery (1993) is interesting in that it tries to formalize many aspects of planning experiments that are often poorly specified.

## 9.1 Completely randomized designs

In a completely randomized design, a group of experimental units are available and the experimenter randomly assigns treatments to the experimental units. The data consist of a group of observations on each treatment. These groups of observations are subjected to a one-way analysis of variance.

EXAMPLE 9.1.1. In Example 6.0.1, we considered data from Mandel (1972) on the elasticity measurements of natural rubber made by 13 laboratories. While Mandel did not discuss how the data were obtained, it could well have been the result of a completely randomized design. For a CRD, we would need 52 pieces of the type of rubber involved. These should be randomly divided into 13 groups (using a table of random numbers or random numbers generated by a reliable computer program). The first group of samples is then sent to the first lab, the second group to the second lab, etc. For a CRD, it is important that a sample is not sent to a lab because the sample somehow seems appropriate for that particular lab.

Personally, I would also be inclined to send the four samples to a given lab at different times. If the four samples are sent at the same time, they might be analyzed by the same person, on the same machines, at the same time. Samples sent at different times might be treated differently. If samples are treated differently at different times, this additional source of variation should be included in any predictive conclusions we wish to make about the labs.

When samples sent at different times are treated differently, sending a batch of four samples at the same time constitutes *subsampling*. There are two sources of variation to deal with: variation from time to time and variation within a given time. The values from four samples at a given time help reduce the effect on treatment comparisons due to variability at a given time, but samples analyzed at different times are still *required* to obtain a valid estimate of the error. In fact, with subsampling, a perfectly valid analysis can be based on the means of the four subsamples. In our example, such an analysis gives only one 'observation' at each time, so the need for sending samples at more than one time is obvious. If the four samples were sent at the same time, there would be no replication, hence no estimate of error. Subsection 12.4.1 and Christensen (1987, section XI.4) discuss subsampling in more detail. □

## 9.2 Randomized complete block designs

In a randomized complete block design the experimenter obtains (constructs) blocks of homogeneous material that contain as many experimental units as there are treatments. The experimenter then randomly assigns a different treatment to each of the units in the block. The random assignments are performed independently for each block. The advantage of this procedure is that treatment comparisons are subject only to the variability within the blocks. Block to block variation is eliminated in the analysis. In a completely randomized design applied to the same experimental material, the treatment comparisons would be subject to both the within block and the between block variability.

The key to a good blocking design is in obtaining blocks that have little within block variability.

Table 9.1: *Spectrometer data*

| | Block | | | Trt. |
|---|---|---|---|---|
| Treatment | 1 | 2 | 3 | means |
| New-clean | 0.9331 | 0.8664 | 0.8711 | $0.8902\bar{0}$ |
| New-soiled | 0.9214 | 0.8729 | 0.8627 | $0.8856\bar{6}$ |
| Used-clean | 0.8472 | 0.7948 | 0.7810 | $0.8076\bar{6}$ |
| Used-soiled | 0.8417 | 0.8035 | 0.8099 | $0.8183\bar{6}$ |
| Block means | 0.885850 | 0.834400 | 0.831175 | 0.850475 |

Often this requires that the blocks be relatively small. A difficulty with RCB designs is that the blocks must be large enough to allow all the treatments to be applied within each block. This can be a serious problem if there is a substantial number of treatments or if maintaining homogeneity within blocks requires the blocks to be very small. If the treatments cannot all be fitted into each block, we need some sort of *incomplete block* design. Such designs will be considered in Chapters 16 and 17.

The analysis of a randomized complete block design is a two-way ANOVA without replication or interaction. The analysis is illustrated below and discussed in general in the following subsection.

EXAMPLE 9.2.1. Inman, Ledolter, Lenth, and Niemi (1992) studied the performance of an optical emission spectrometer. Table 9.1 gives some of their data on the percentage of manganese (Mn) in a sample. The data were collected using a sharp counterelectrode tip with the sample to be analyzed partially covered by a boron nitride disk. Data were collected under three temperature conditions. Upon fixing a temperature, the sample percentage of Mn was measured using 1) a new boron nitride disk with light passing through a clean window (new-clean), 2) a new boron nitride disk with light passing through a soiled window (new-soiled), 3) a used boron nitride disk with light passing through a clean window (used-clean), and 4) a used boron nitride disk with light passing through a soiled window (used-soiled). The four conditions, new-clean, new-soiled, used-clean, used-soiled are the treatments. The temperature was then changed and data were again collected for each of the four treatments. A block is always made up of experimental units that are homogeneous. The temperature conditions were held constant while observations were taken on the four treatments so the temperature levels identify blocks.

In analyzing a one-way ANOVA, the analysis of variance table is of little direct importance. For a randomized complete block design the analysis of variance table is crucial. Before we can proceed with any analysis of the treatments, we need an estimate of the variance $\sigma^2$. In one-way ANOVA, the *MSE* is simply a pooled estimate obtained from group sample variances. In a RCB design, the replications of the experiment occur in different blocks and the effect of these blocks must be taken into account. In particular, the three observations on each treatment do not form a random sample from a population, so it is inappropriate to compute the sample variance within each treatment group and it is totally inappropriate to pool such variance estimates. Instead, we expand the analysis of variance table by accounting for both treatments and blocks and estimate $\sigma^2$ with the leftover sum of squares.

The sum of squares total (corrected for the grand mean) is computed just as in one-way ANOVA; it is the sample variance of all 12 observations multiplied by the degrees of freedom $12-1$, e.g.,

$$SSTot = (12-1)s_y^2 = (11).002277797 = .025055762.$$

The mean square and sum of squares for treatments are also computed as in one-way ANOVA. Using Table 9.1, the treatment means are averages of 3 observations and the sample variance of the treatment means is .001893343, so

$$MSTrts = 3(.001893343) = .005680028\,.$$

There are 4 treatments, so the sum of squares is the mean square multiplied by the degrees of

Table 9.2: *Analysis of variance for spectrometer data*

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Trts | 3 | 0.0170401 | 0.0056800 | 70.05 | 0.000 |
| Blocks | 2 | 0.0075291 | 0.0037646 | 46.43 | 0.000 |
| Error | 6 | 0.0004865 | 0.0000811 | | |
| Total | 11 | 0.0250558 | | | |

freedom $(4-1)$,

$$SSTrts = (4-1).005680028 = .017040083\,.$$

The mean square and sum of squares for blocks are also computed as if they were treatments in a one-way ANOVA. The block means are averages of 4 observations and the sample variance of the block means is .000941143, so

$$MSBlocks = 4(.000941143) = .003764572\,.$$

There are 3 blocks, so the sum of squares is the mean square times the degrees of freedom $(3-1)$,

$$SSBlocks = (3-1).003764572 = .007529145\,.$$

The sum of squares error is obtained by subtraction,

$$\begin{aligned} SSE &= SSTot - SSTrts - SSBlocks \\ &= .025055762 - .017040083 - .007529145 \\ &= .000486534\,. \end{aligned}$$

Similarly,

$$\begin{aligned} dfE &= dfTot - dfTrts - dfBlocks \\ &= 11 - 3 - 2 \\ &= 6\,. \end{aligned}$$

The estimate of $\sigma^2$ is

$$MSE = \frac{SSE}{dfE} = \frac{.000486534}{6} = .000081089\,.$$

Given the definitions of $SSE$ and $dfE$, it is tautological that the sums of squares for treatments, blocks, and error add up to the sum of squares total, and similarly for the degrees of freedom.

All of the calculations are summarized in the analysis of variance table, Table 9.2. Table 9.2 also gives the analysis of variance $F$ test for the null hypothesis that the effects are the same for each treatment. By definition the $F$ statistic is $MSTrts/MSE$ and in this example it is huge, 70.05. The $P$ value is infinitesimal, so there is clear evidence that the 4 treatments do not behave the same.

Table 9.2 also contains an $F$ test for blocks. In a true blocking experiment, there is not much interest in testing whether block means are different. After all, one *chooses* the blocks so that they have different means. Nonetheless, the $F$ statistic $MSBlks/MSE$ is of some interest because it indicates how effective the blocking was, i.e., it indicates how much the variability was reduced by blocking. For this example, $MSBlks$ is 46 times larger than $MSE$, indicating that blocking was definitely worthwhile. In the model for RCB designs presented in the following subsection, there is no reason not to test for blocks, but some models used for RCBs do not allow a test for blocks. Regardless of the particular model, the analysis of treatments works in the same way.

Now that we have an estimate of the variance, we can proceed with the more interesting questions about how the treatments differ. We begin by examining pairwise differences. The multiple comparison methods of Chapter 6 all apply in the usual way after adjusting for the difference in

Table 9.3: *Contrasts for the spectrometer data*

| | Contrast labels | | | Trt. |
|---|---|---|---|---|
| Treatment | D | W | DW | means |
| New-clean | 1 | 1 | 1 | $0.8902\bar{0}$ |
| New-soiled | 1 | −1 | −1 | $0.8856\bar{6}$ |
| Used-clean | −1 | 1 | −1 | $0.8076\bar{6}$ |
| Used-soiled | −1 | −1 | 1 | $0.8183\bar{6}$ |
| *Est* | .149833 | −.006167 | .015233 | |
| *SS* | .0168375 | .0000285 | .0001740 | |

$dfE$. For example, there are 4 treatment means, each based on 3 observations, and the $MSE$ is 0.0000811 with 6 degrees of freedom, so for $\alpha = .05$ the honest significant difference is

$$HSD = Q(.95,4,6)\sqrt{0.0000811/3} = 4.90(.00519936) = .02548\,.$$

The differences between the treatments are illustrated below.

| Treatment | Used-clean | Used-soiled | New-soiled | New-clean |
|---|---|---|---|---|
| Mean | $0.8076\bar{6}$ | $0.8183\bar{6}$ | $0.8856\bar{6}$ | $0.8902\bar{0}$ |

We have no evidence of an effect due to the condition of the window when considering used boron nitride disks. The higher yields occur for soiled windows, but they are not significantly different. We also have no evidence of an effect due to the condition of the window for new boron nitride disks. The higher yields occur for clean windows but again the difference is not significant. Evidence does exist that the two means for used disks are different (less than) the two means for new disks.

The structure of the treatments suggests particular orthogonal contrasts that are of interest. Contrast coefficients, estimated contrasts, and sums of squares for the contrasts are given in Table 9.3.

The contrast labeled D looks at the difference in disks by averaging over windows. This involves averaging the two means for new disks, say, $\mu_{NC}$ and $\mu_{NS}$, and contrasting this average with the average of the two means for used disks, say, $\mu_{UC}$ and $\mu_{US}$. The contrast examining the difference in disks averaging over windows is $(\mu_{NC}+\mu_{NS})/2-(\mu_{UC}+\mu_{US})/2$ or

$$\frac{1}{2}\mu_{NC}+\frac{1}{2}\mu_{NS}-\frac{1}{2}\mu_{UC}-\frac{1}{2}\mu_{US}.$$

As discussed earlier, multiplying a contrast by a constant does not really change the contrast, so to eliminate the fractional multiplications and make the contrast a bit easier to work with, we multiply this contrast by 2 and make it

$$\mu_{NC}+\mu_{NS}-\mu_{UC}-\mu_{US}.$$

This is the contrast D reported in Table 9.3. Recall that the estimate of the D contrast is computed as

$$\hat{D} = (1)0.8902\bar{0}+(1)0.8856\bar{6}+(-1)0.8076\bar{6}+(-1)0.8183\bar{6} = .149833$$

and the sum of squares is computed as

$$SS(D) = \frac{(.149833)^2}{[1^2+1^2+(-1)^2+(-1)^2]/3} = .0168375.$$

We define the contrast W similarly; it looks at the difference in windows by averaging over disks. Again, we multiplied the averages by 2 to simplify the contrast.

Contrast DW looks at the *interaction* between disks and windows, i.e., how the difference between disks changes as we go from a clean window to a soiled window. The difference between new

and used disks with a clean window is $(\mu_{NC} - \mu_{UC})$ and the difference between new and used disks with a soiled window is $(\mu_{NS} - \mu_{US})$. The change in the disk difference between clean and soiled windows is $(\mu_{NC} - \mu_{UC}) - (\mu_{NS} - \mu_{US})$, or equivalently

$$\mu_{NC} - \mu_{NS} - \mu_{UC} + \mu_{US}.$$

This is the DW contrast. Note that the DW contrast coefficients in Table 9.3 can be obtained by multiplying the corresponding D and W contrast coefficients. *This procedure for obtaining interaction contrast coefficients by multiplying main effect contrast coefficients works quite generally*. Note that the DW contrast can also be obtained by looking at $(\mu_{NC} - \mu_{NS}) - (\mu_{UC} - \mu_{US})$, which is the change in the difference between clean and soiled windows as we go from new disks to used disks.

The analysis of variance is balanced; there are three observations on each treatment and four observations on each block. Thus, using the definition of orthogonality for a balanced one-way ANOVA, the treatment contrasts are orthogonal. It follows, both numerically and theoretically, that

$$SS(D) + SS(W) + SS(DW) = SSTrts.$$

From Table 9.3 we see that the vast majority of the sum of squares for treatments is due to the difference between disks averaged over windows (the D contrast). In particular, $SS(D)/SSTrts = .0168375/.0170401$, so 99% of the sum of squares for treatments is due to the D contrast. We also see that there is relatively little effect due to windows averaged over disks (the W contrast) and little effect due to the change in the disk differences due to windows (the DW contrast). In particular, the unadjusted $F$ statistic for testing whether the DW contrast is zero is

$$F = \frac{SS(DW)}{MSE} = \frac{.0001740}{0.0000811} = 2.15$$

which has a $P$ value of .193. Recall that a contrast has one degree of freedom, so $SS(DW) = MS(DW)$ in constructing the $F$ statistic. Similarly, the test for contrast W has $F = .35$ and $P = .575$.

While there is no statistical evidence for the existence of an interaction in the example, this does not prove that interaction does not exist. For example, if $MSE$ had turned out to be one-third of its actual value, the $F$ test for interaction would have been significant. When interactions exist, it is important to explore their nature. We now discuss some methods and ideas for examining interactions. This discussion is merely a precursor of the more extensive examination of interaction in Chapter 11. The difference between clean and soiled windows for new disks is .004533 and the difference for used disks is $-.010700$. These effects are actually in different directions! In one case, clean windows give higher readings and in the other case clean windows give lower readings. This seems to indicate that the effect of windows changes depending on the type of disk used, but in this example the $MSE$ is large enough that the difference can reasonably be ascribed to random variation. In other words, this change in effect is not statistically significant because the interaction contrast is not statistically significant.

The contrast examining the different windows averaged over disks (the W contrast) was insignificant. However, if the DW interaction existed, the windows would still have a demonstrable effect on yields. The windows would have an effect because the disks behave differently for clean windows than for soiled windows. Additionally, the large effect for D would be of less interest if interaction were present because D is obtained by averaging over windows even though we would know that the disk effect depends on the window used.

Figure 9.1 contains a plot of the treatment means. There are two curves, one for new disks and another for used disks. The differences between disks is indicated by the separation between the two lines. The differences in the windows are indicated visually by the slopes of the new and used disk lines. If the effect of windows was the same regardless of disk condition, these slopes would be the same and the line segments would be parallel *up to sampling error*. With these data the lines are reasonably parallel. When interaction exists, the plot indicates the nature of the interaction.

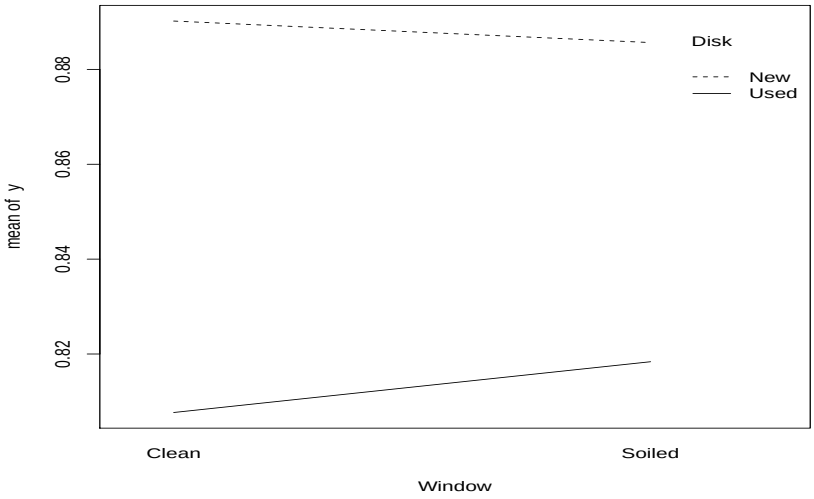



Figure 9.1: *Disk–window interaction plot for spectrometer data.*

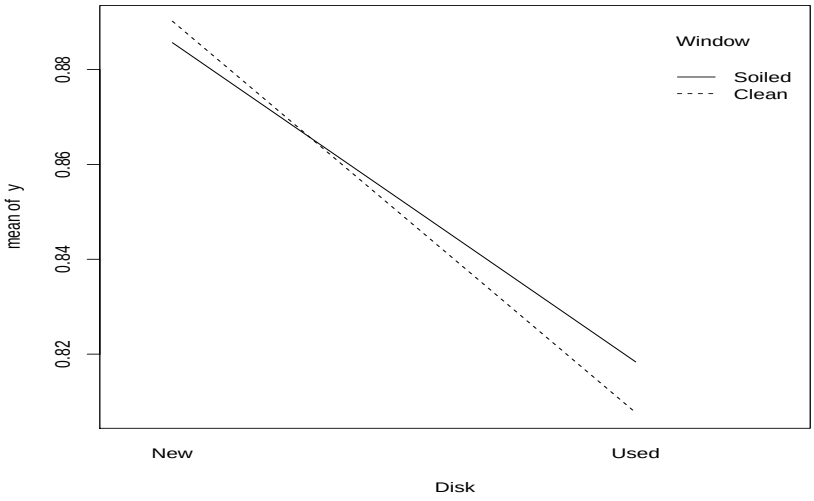



Figure 9.2: *Disk–window interaction plot for spectrometer data.*

Rather than describing the interaction as a difference in how the windows react to disks, we can describe the interaction in terms of how the effect of disk changes with type of window. As mentioned earlier, the two approaches are equivalent. Figure 9.2 contains another plot of the treatment means. There are two curves, one for clean windows and another for soiled windows. In this plot, the slopes indicate the differences due to disks and the separation of the lines indicates the differences due to windows. If there is no interaction, the curves should be parallel *up to sampling variation*. In this example, the curves intersect, suggesting that the curves are not parallel. However, after considering the level of sampling error, there is no evidence that the curves are not parallel.

Residual plots for the data are given in Figures 9.3 through 9.6. The residuals now must adjust

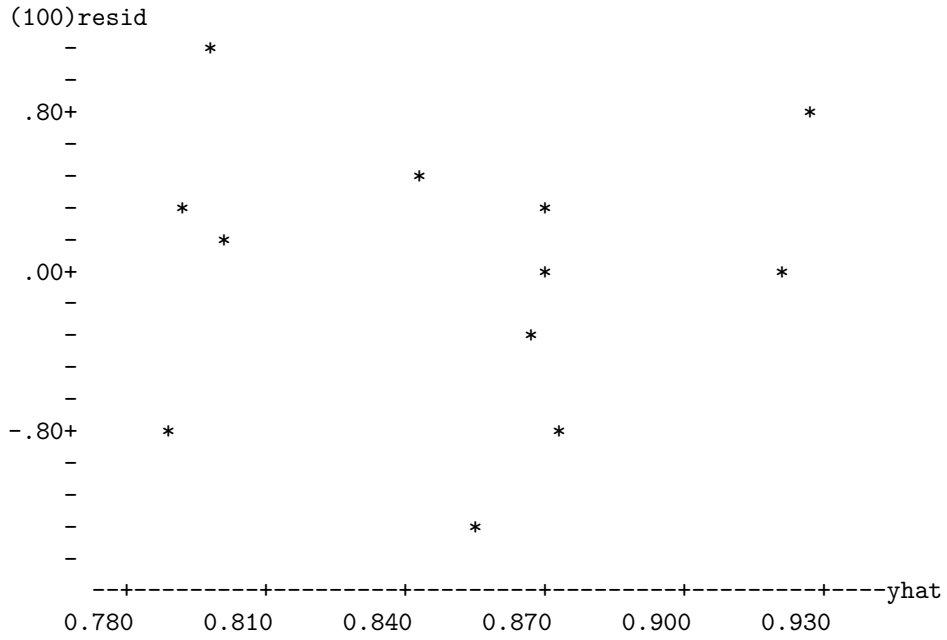



Figure 9.3: *Plot of residuals versus predicted values, spectrometer data.*

for both the treatments and the blocks. The residual for an observation $y_{ij}$ with treatment $i$ in block $j$ is defined as

$$\hat{\varepsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot},$$

where $\bar{y}_{i\cdot}$ is the $i$th treatment mean, $\bar{y}_{\cdot j}$ is the $j$th block mean, and $\bar{y}_{\cdot\cdot}$ is the grand mean of all 12 observations. By subtracting out both the treatment and block means, we have over adjusted for the overall level of the numbers, so the grand mean must be added back in. In balanced analysis of variance problems all the residuals have the same variance, so it is not necessary to standardize the residuals.

Figure 9.3 is a plot of the residuals versus the predicted values. The predicted values are

$$\hat{y}_{ij} = \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}.$$

Figure 9.4 plots the residuals versus indicators of the treatments. While the plot looks something like a bow tie, I am not overly concerned. Figure 9.5 contains a plot of the residuals versus indicators of blocks. The residuals look pretty good. From Figure 9.6, the residuals look reasonably normal. In the normal plot there are 12 residuals but the analysis has only 6 degrees of freedom for error. If you want to do a $W'$ test for normality, you might use a sample size of 12 and compare the value $W' = .970$ to $W'(\alpha, 12)$, but it may be appropriate to use the $dfE$ as the sample size for the test and use $W'(\alpha, 6)$. □

*Minitab commands*

The following Minitab commands generate the analysis of variance. Column c1 contains the spectrometer data, while column c2 contains integers 1 through 4 indicating the appropriate treatment, and c3 contains integers 1 through 3 that indicate the block. The predicted values are given by the 'fits' subcommand.

```
MTB > names c1 'y' c2 'Trts' c3 'Blks'
MTB > anova c1 = c2 c3;
```

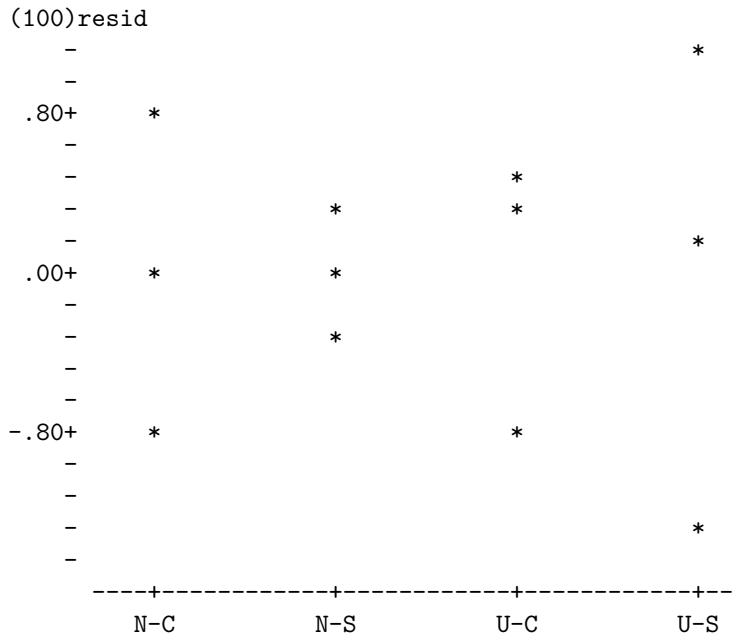



Figure 9.4: *Plot of residuals versus treatment groups, spectrometer data.*

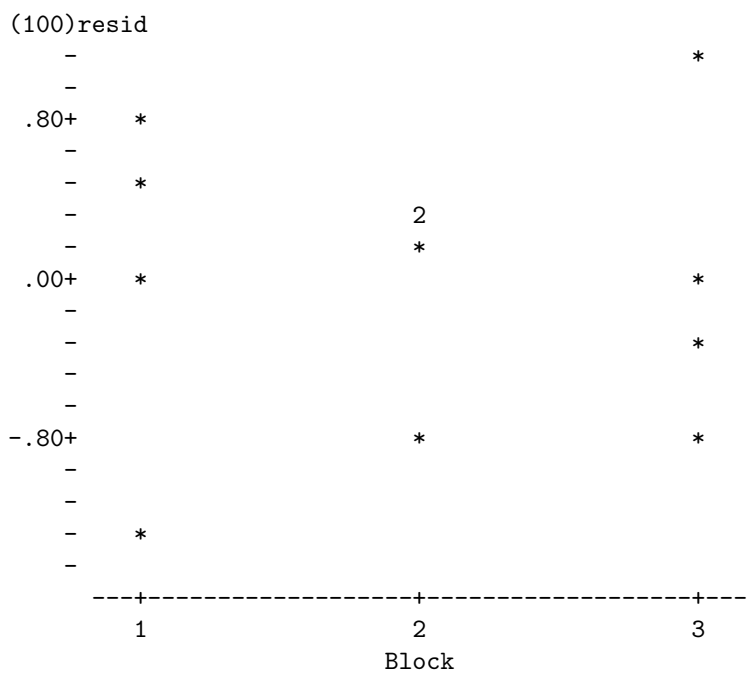



Figure 9.5: *Plot of residuals versus blocks, spectrometer data.*

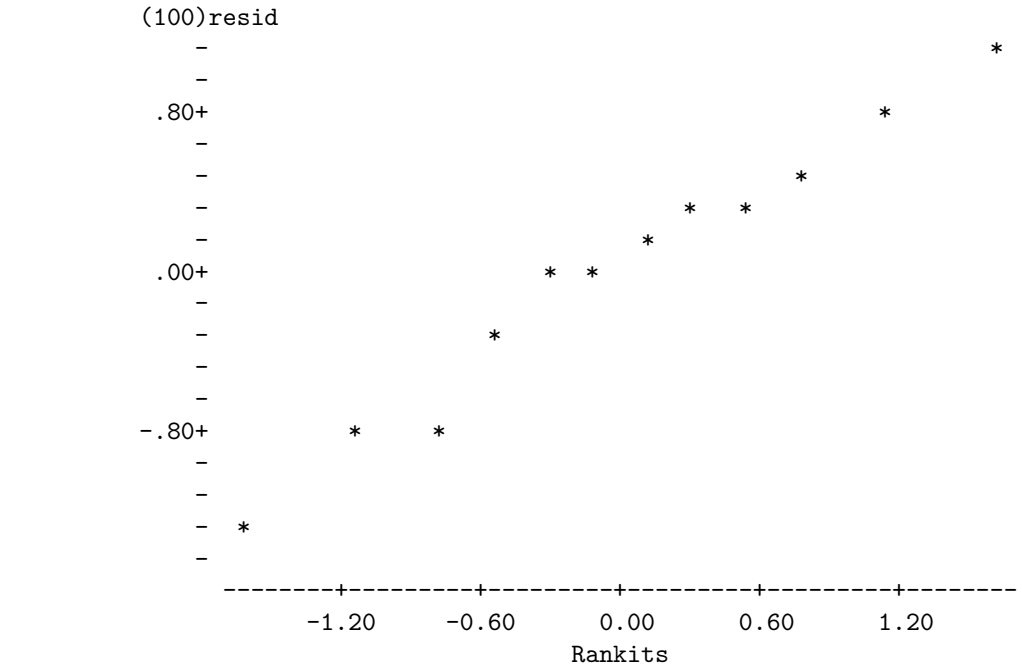



Figure 9.6: *Normal plot of residuals, spectrometer data,* $W' = 0.970$.

```
SUBC> means c2 c3;
SUBC> resid c10;
SUBC> fits  c11.
```

*Balanced two-way analysis of variance*

The model for a randomized complete block design is a two-way analysis of variance,

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij}\text{s independent } N(0,\sigma^2), \tag{9.2.1}$$

$i = 1,\ldots,a,\ j = 1,\ldots,b$. There are $b$ blocks with $a$ treatments observed within each block. The parameter $\mu$ is viewed as a grand mean, $\alpha_i$ is an unknown fixed effect for the $i$th treatment, and $\beta_j$ is an unknown fixed effect for the $j$th block. The necessary summary statistics are the sample variance of all $ab$ observations and the means for each treatment and block. It is frequently convenient to display the data as follows.

| Treatment | Block $j$ | | | | Trt. |
|---|---|---|---|---|---|
| $i$ | 1 | 2 | $\cdots$ | $b$ | means |
| 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1b}$ | $\bar{y}_{1\cdot}$ |
| 2 | $y_{21}$ | $y_{22}$ | $\cdots$ | $y_{2b}$ | $\bar{y}_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $a$ | $y_{a1}$ | $y_{a2}$ | $\cdots$ | $y_{ab}$ | $\bar{y}_{a\cdot}$ |
| Blk. means | $\bar{y}_{\cdot 1}$ | $\bar{y}_{\cdot 2}$ | $\cdots$ | $\bar{y}_{\cdot b}$ | $\bar{y}_{\cdot\cdot}$ |

The predicted values from this model are

$$\hat{y}_{ij} \equiv \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$$

Table 9.4: *Analysis of variance*

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Trts($\alpha$) | $a-1$ | $b\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$ | $SS(\alpha)/(a-1)$ | $\frac{MS(\alpha)}{MSE}$ |
| Blks($\beta$) | $b-1$ | $a\sum_{j=1}^{b}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$ | $SS(\beta)/(b-1)$ | $\frac{MS(\beta)}{MSE}$ |
| Error | $(a-1)(b-1)$ | $\sum_{i,j}\hat{\varepsilon}_{ij}^2$ | $SSE/dfE$ | |
| Total | $ab-1$ | $\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{\cdot\cdot})^2$ | | |

and the residuals are

$$\hat{\varepsilon}_{ij} \equiv y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}\,.$$

The computations involved in estimating $\sigma^2$ can be summarized in an analysis of variance table. The commonly used form for the analysis of variance table is given in Table 9.4. The degrees of freedom and sums of squares for treatments, blocks, and error add up to the degrees of freedom and sums of squares total (corrected for the grand mean). Note that the mean square for treatments is just the sample variance of the $\bar{y}_{i\cdot}$s times $b$ and that the mean square for blocks is just the sample variance of the $\bar{y}_{\cdot j}$s times $a$.

The $F$ statistic $MS(\alpha)/MSE$ is the ratio of the mean square treatments to the mean square error. It is used to test whether there are treatment effects, i.e., it is used to test

$$H_0 : \alpha_1 = \cdots = \alpha_a.$$

Note that if all the $\alpha_i$s are equal, we cannot *distinguish* between the effects of different treatments. In other words, we cannot isolate anything that can be identified as the effect of a treatment. $H_0$ does not imply that the treatments have no effect, it implies that they have the same effect. Generally, the effect of a treatment (an $\alpha_i$) is impossible to isolate (or estimate) because we cannot distinguish it from the overall effect of running the experiment ($\mu$) or indeed from any effect common to every block. The same thing is true of the block effects $\beta_j$; they cannot be isolated from $\mu$ or common effects of the treatments. What we *can* isolate are comparative differences in the effects of treatments (and blocks). *The $F$ statistic provides a test of whether there are differences in the treatment effects and not whether any treatment effects exist*. The only way you can test whether treatment effects exist is to redefine what you mean by treatment effects, so that they only exist when they are different.

The treatments are dealt with exactly as in a one-way ANOVA. For known $\lambda_i$*s that sum to zero*, a contrast in the treatment effects is

$$Par = \sum_{i=1}^{a} \lambda_i \alpha_i$$

with

$$Est = \sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot}\,.$$

Each treatment mean is the average of $b$ observations, so

$$\mathrm{SE}\left(\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot}\right) = \sqrt{MSE \sum_{i=1}^{a} \lambda_i^2/b}.$$

The reference distribution is $t(dfE)$. The sum of squares for the contrast is

$$SS\left(\sum_{i=1}^{a} \lambda_i \alpha_i\right) = \left[\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot}\right]^2 \Big/ \left[\sum_{i=1}^{a} \lambda_i^2/b\right].$$

If of interest, similar results hold for the block effects. For known $\xi_j$s that add to zero,

$$Par = \sum_{j=1}^{b} \xi_j \beta_j$$

with

$$Est = \sum_{j=1}^{b} \xi_j \bar{y}_{\cdot j} .$$

The block means are the average of $a$ observations, so

$$\mathrm{SE}\left( \sum_{j=1}^{b} \xi_j \bar{y}_{\cdot j} \right) = \sqrt{MSE \sum_{j=1}^{b} \xi_j^2 / a}.$$

The reference distribution is $t(dfE)$. The sum of squares for the contrast is

$$SS\left( \sum_{j=1}^{b} \xi_j \beta_j \right) = \left[ \sum_{j=1}^{b} \xi_j \bar{y}_{\cdot j} \right]^2 \Bigg/ \left[ \sum_{j=1}^{b} \xi_j^2 / a \right].$$

The $F$ statistic $MS(\beta)/MSE$ is the ratio of the mean square blocks to the mean square error. It is used to test whether there are block effects, i.e., it is used to test

$$H_0 : \beta_1 = \cdots = \beta_b.$$

Again, if all the $\beta_i$s are equal we cannot distinguish between the effects of different blocks, so the $F$ statistic provides a test of whether we can isolate comparative differences in the block effects. As discussed earlier, some models for RCB designs do not allow testing for block effects, but in any case, the ratio $MS(\beta)/MSE$ is of interest in that large values indicate that blocking was a worthwhile exercise.

The theoretical basis for this analysis of model (9.2.1) is precisely as in the balanced one-way ANOVA. Consider the analysis of treatment effects. (The analysis for block effects is similar.) The only thing random about a $y_{ij}$ is the corresponding $\varepsilon_{ij}$. The $\varepsilon_{ij}$s are independent, so the $y_{ij}$s are independent. If follows that the $\bar{y}_{i\cdot}$s are independent because they are computed from distinct groups of observations. Since $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, obviously $\bar{y}_{i\cdot} = \mu + \alpha_i + \bar{\beta}_\cdot + \bar{\varepsilon}_{i\cdot}$. Using Proposition 1.2.11,

$$\mathrm{E}(\bar{y}_{i\cdot}) = \mu + \alpha_i + \bar{\beta}_\cdot + \mathrm{E}(\bar{\varepsilon}_{i\cdot}) = \mu + \alpha_i + \bar{\beta}_\cdot$$

and

$$\mathrm{Var}(\bar{y}_{i\cdot}) = \mathrm{Var}(\bar{\varepsilon}_{i\cdot}) = \frac{\sigma^2}{b}.$$

More directly, the equalities follow because $\bar{\varepsilon}_{i\cdot}$ is the sample mean from a random sample of $b$ variables each with population mean 0 and population variance $\sigma^2$. If the errors are normally distributed, the $\bar{y}_{i\cdot}$s are independent $N(\mu + \alpha_i + \bar{\beta}_\cdot, \sigma^2/b)$ random variables. In any case, if $b$ is reasonably large, the normal distribution holds approximately because of the central limit theorem.

As in a balanced one-way ANOVA, the estimated contrast $\sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot}$ is an unbiased estimate of the parameter

$$\mathrm{E}\left( \sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} \right) = \sum_{i=1}^{a} \lambda_i (\mu + \alpha_i + \bar{\beta}_\cdot) = \sum_{i=1}^{a} \lambda_i \alpha_i .$$

This follows because $\sum_{i=1}^{a} \lambda_i = 0$. A derivation that is identical to the derivation used for a balanced one-way ANOVA gives

$$\mathrm{Var}\left( \sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot} \right) = \sigma^2 \frac{\sum_{i=1}^{a} \lambda_i^2}{b},$$

from which the standard error follows. The reference distribution for tests and confidence intervals relies on the fact that for normal errors,

$$\frac{SSE}{\sigma^2} \sim \chi^2(dfE)$$

with $MSE$ independent of the $\bar{y}_{i\cdot}$s and the $\bar{y}_{\cdot j}$s.

Also as in a balanced one-way ANOVA, the $\bar{y}_{i\cdot}$s are a random sample from a normal population with variance $\sigma^2/b$ if and only if the $\bar{y}_{i\cdot}$s have the same means, i.e., if and only if all the $\alpha_i$s are equal. When the $\alpha_i$s are all equal, the sample variance of the $\bar{y}_{i\cdot}$s is an estimate of $\sigma^2/b$ and the $MSTrts$ is an estimate of $\sigma^2$. In general, $MSTrts$ estimates

$$\mathrm{E}(MSTrts) = \sigma^2 + \frac{b}{a-1}\sum_{i-1}^{a}(\alpha_i - \bar{\alpha}_\cdot)^2$$

which is much larger than $\sigma^2$ if the treatment effects are very different relative to the size of $\sigma^2$ or if the number of blocks $b$ is large. Moreover, the structure of the treatment means implies that all of the multiple comparison methods of Chapter 6 can continue to be applied.

Proposition 1.2.11 also shows that the predicted values have

$$\mathrm{E}(\hat{y}_{ij}) = \mu + \alpha_i + \beta_j = \mathrm{E}(y_{ij}),$$

that the residuals have

$$\mathrm{E}(\hat{\varepsilon}_{ij}) = 0,$$

and that the $MSE$ is unbiased for $\sigma^2$, i.e.,

$$\mathrm{E}(MSE) = \sigma^2.$$

(Showing the last of these is much the most complicated.)

*Paired comparisons*

An interesting special case of complete block data is paired comparison data as discussed in Section 4.1. In paired comparison data, there are two treatments to contrast and each pair constitutes a complete block.

EXAMPLE 9.2.2. *Shewhart's hardness data*

In Section 4.1, we examined Shewhart's data on hardness of two items that were welded together. In this case, it is impossible to group arbitrary formless pairs of parts and then randomly assign a part to be either part 1 or part 2, so the data do not actually come from an RCB experiment. Nonetheless, the two-way ANOVA model remains reasonable with pairs playing the role of blocks.

The data were given in Section 4.1 along with the means for each of the two parts. The two-way ANOVA analysis also requires the mean for each pair of parts. The analysis of variance table for the blocking analysis is given in Table 9.5. In comparing the blocking analysis to the paired comparison analysis given earlier, allowance for round-off errors must be made. The $MSE$ is exactly half the value of $s_d^2 = 17.77165$ given in Section 4.1. The two-way ANOVA $t$ test for differences between the two parts has

$$t_{obs} = \frac{47.552 - 34.889}{\sqrt{8.8858\,[(1/27) + (1/27)]}} = 15.61\,.$$

This is exactly the same $t$ statistic as used in Section 4.1. The reference distribution is $t(26)$, again exactly the same. The analysis of variance $F$ statistic is just the square of the $t_{obs}$ and gives equivalent results for two-sided tests. Confidence intervals for the difference in means are also exactly the same in the blocking analysis and the paired comparison analysis. The one real difference between this analysis and the analysis of Section 4.1 is that this analysis provides an indication of whether pairing was worthwhile. □

Table 9.5: *Analysis of variance for hardness data*

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Pairs(Blocks) | 26 | 634.94 | 24.42 | 2.75 | 0.006 |
| Parts(Trts) | 1 | 2164.73 | 2164.73 | 243.62 | 0.000 |
| Error | 26 | 231.03 | 8.89 | | |
| Total | 53 | 3030.71 | | | |

Table 9.6: *Mangold root data*

| | Columns | | | | | Row |
|---|---|---|---|---|---|---|
| Rows | 1 | 2 | 3 | 4 | 5 | means |
| 1 | D(376) | E(371) | C(355) | B(356) | A(335) | 358.6 |
| 2 | B(316) | D(338) | E(336) | A(356) | C(332) | 335.6 |
| 3 | C(326) | A(326) | B(335) | D(343) | E(330) | 332.0 |
| 4 | E(317) | B(343) | A(330) | C(327) | D(336) | 330.6 |
| 5 | A(321) | C(332) | D(317) | E(318) | B(306) | 318.8 |
| Col. means | 331.2 | 342.0 | 334.6 | 340.0 | 327.8 | 335.12 |
| | | | | | | |
| Treatments | A | B | C | D | E | |
| Trt. means | 333.6 | 331.2 | 334.4 | 342.0 | 334.4 | |

## 9.3 Latin square designs

Latin square designs involve two simultaneous but distinct definitions of blocks. The treatments are arranged so that every treatment is observed in every block for both kinds of blocks.

EXAMPLE 9.3.1. Mercer and Hall (1911) and Fisher (1925, section 49) consider data on the weights of mangold roots. They used a Latin square design with 5 rows, columns, and treatments. The rectangular field on which the experiment was run was divided into five rows and five columns. This created 25 plots, arranged in a square, on which to apply the treatments A, B, C, D, and E. Each row of the square was viewed as a block, so every treatment was applied in every row. The unique feature of Latin square designs is that there is a second set of blocks. Every column was also considered a block, so every treatment was also applied in every column. The data are given in Table 9.6, arranged by rows and columns with the treatment given in the appropriate place and the observed root weight given in parentheses. The table also contains the means for rows, columns, and treatments. In each case, the mean is the average of 5 observations.

The analysis of variance table is constructed like that for a randomized complete block design except that now both rows and columns play roles similar to blocks. The sum of squares total (corrected for the grand mean) is computed just as in one-way ANOVA, the sample variance of all 25 observations is computed and multiplied by $25-1$, i.e.,

$$SSTot = (25-1)s_y^2 = (24)292.77\bar{6} = 7026.6.$$

The mean square and sum of squares for treatments are also computed as in one-way ANOVA. The treatment means are averages of 5 observations and the sample variance of the treatment means is 16.512, so

$$MSTrts = 5(16.512) = 82.56\,.$$

There are 5 treatments, so treatments have $5-1$ degrees of freedom and the sum of squares is the mean square times $(5-1)$ ,

$$SSTrts = (5-1)82.56 = 330.24\,.$$

The mean square and sum of squares for columns are also computed as if they were treatments

Table 9.7: *Analysis of variance for mangold root data*

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Trts | 4 | 330.2 | 82.6 | 0.56 | .696 |
| Columns | 4 | 701.8 | 175.5 | 1.20 | .360 |
| Rows | 4 | 4240.2 | 1060.1 | 7.25 | .003 |
| Error | 12 | 1754.3 | 146.2 | | |
| Total | 24 | 7026.6 | | | |

in a one-way ANOVA. The column means are averages of 5 observations and the sample variance of the column means is 35.092, so

$$MSCols = 5(35.092) = 175.46\,.$$

There are 5 columns, so the sum of squares is the mean square times $(5-1)$,

$$SSCols = (5-1)175.46 = 701.84\,.$$

The mean square and sum of squares for rows are again computed as if they were treatments in a one-way ANOVA. The row means are the average of 5 observations and the sample variance of the row means is 212.012, so

$$MSRows = 5(212.012) = 1060.06\,.$$

There are 5 rows, so the sum of squares is the mean square times $(5-1)$,

$$SSRows = (5-1)1060.06 = 4240.24\,.$$

The sum of squares error is obtained by subtraction,

$$\begin{aligned} SSE &= SSTot - SSTrts - SSCols - SSRows \\ &= 7026.6 - 330.2 - 701.8 - 4240.2 \\ &= 1754.3\,. \end{aligned}$$

Similarly,

$$\begin{aligned} dfE &= dfTot - dfTrts - dfCols - dfRows \\ &= 24 - 4 - 4 - 4 \\ &= 12\,. \end{aligned}$$

The estimate of $\sigma^2$ is

$$MSE = \frac{SSE}{dfE} = \frac{1754.3}{12} = 146.2\,.$$

All of the calculations are summarized in the analysis of variance table, Table 9.7. Table 9.7 also gives the analysis of variance $F$ test for the null hypothesis that the effects are the same for every treatment. The $F$ statistic $MSTrts/MSE$ is very small, 0.56, so there is no evidence that the treatments behave differently. Blocking on columns was not very effective as evidenced by the $F$ statistic of 1.20, but blocking on rows was very effective, $F = 7.25$.

Many experimenters are less than thrilled when told that there is no evidence for their treatments having any differential effects. Inspection of the treatment means given in Table 9.6 leads to the obvious conclusion that most of the differences are due to the fact that treatment $D$ is much larger than the others, so we look at this a bit more. (Besides, this gives us an excuse to look at a contrast

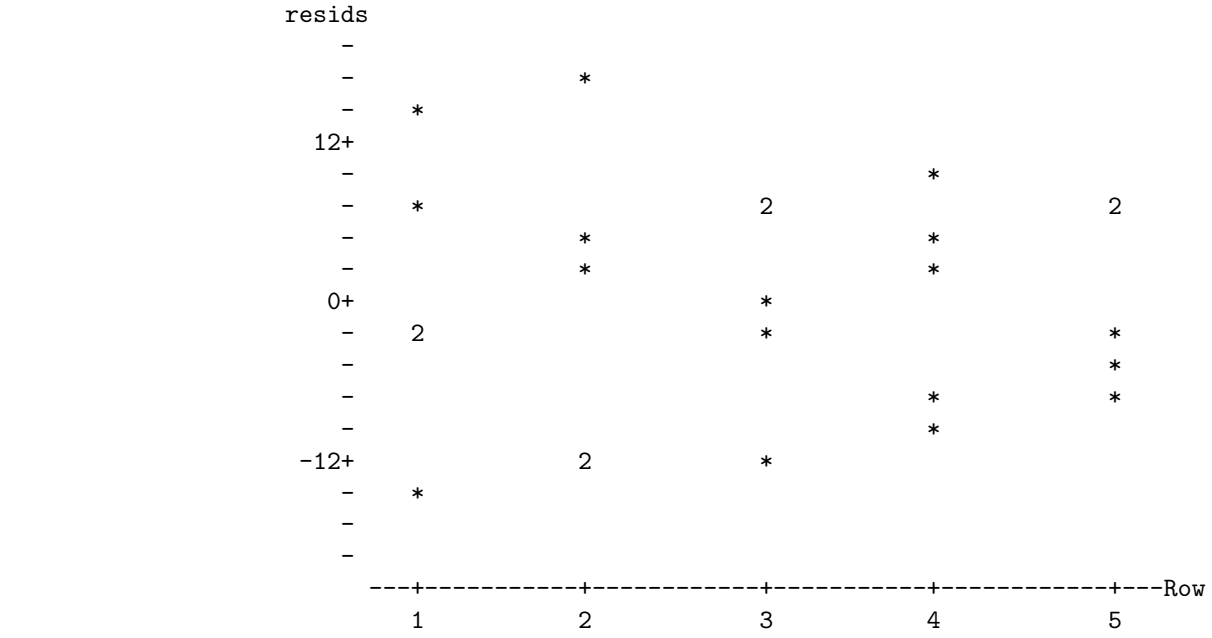



Figure 9.7: *Plot of residuals versus rows.*

in a Latin square design.) If we construct the sum of squares for a contrast that compares $D$ with the other means, say,

$$\mu_A + \mu_B + \mu_C - (4)\mu_D + \mu_E,$$

we get a sum of squares that contains the vast majority of the treatment sum of squares, i.e.,

$$SS(D \text{ vs. others}) = \frac{[333.6 + 331.2 + 334.4 - (4)342.0 + 334.40]^2}{[1+1+1+(-4)^2+1]/5} = 295.84\,.$$

However, the $F$ ratio for the contrast is quite small,

$$F = \frac{295.84}{146.2} = 2.02\,.$$

This is too small to provide any evidence for a difference between $D$ and the average of the other treatments, *even if we had not let the data suggest the contrast*. If this $F$ test cannot be rejected at even the unadjusted .05 level, there is no point in examining any multiple comparison methods to see if they will detect a difference, they will not.

Standard residual plots are given in Figures 9.7 through 9.10. They look quite good. □

*Computing techniques*

The following Minitab commands will give the sums of squares, means, and residuals necessary for the analysis. Here c1 is a column containing the mangold root yields, c2 has values from 1 to 5 indicating the row, c3 has values from 1 to 5 indicating the column, and c4 has values from 1 to 5 indicating the treatment.

```
MTB > names c1 'y' c2 'Rows' c3 'Cols' c4 'Trts'
MTB > ancova c1 = c2 c3 c4;
SUBC> means c2 c3 c4;
SUBC> resid c11.
```

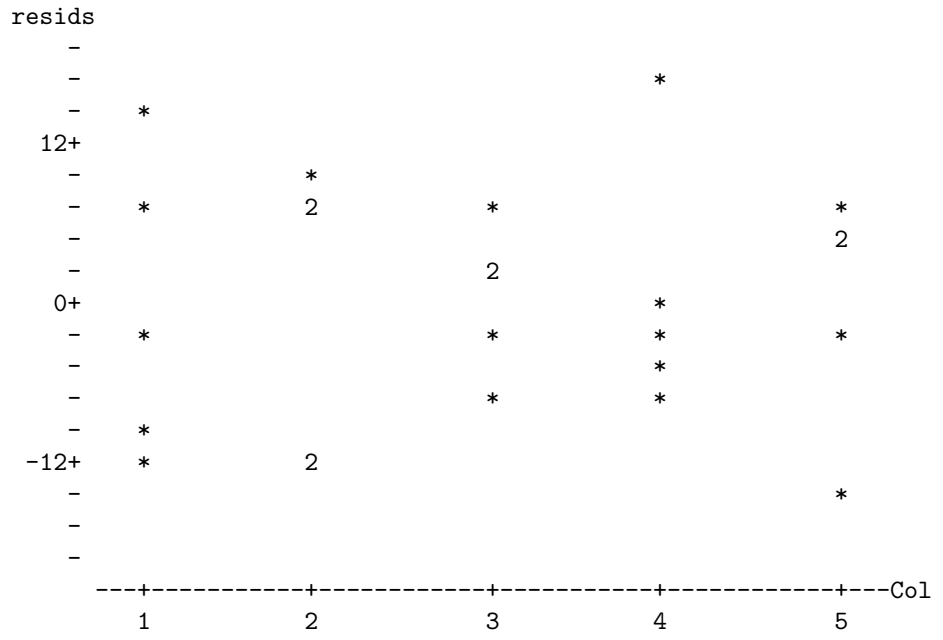



Figure 9.8: *Plot of residuals versus columns.*

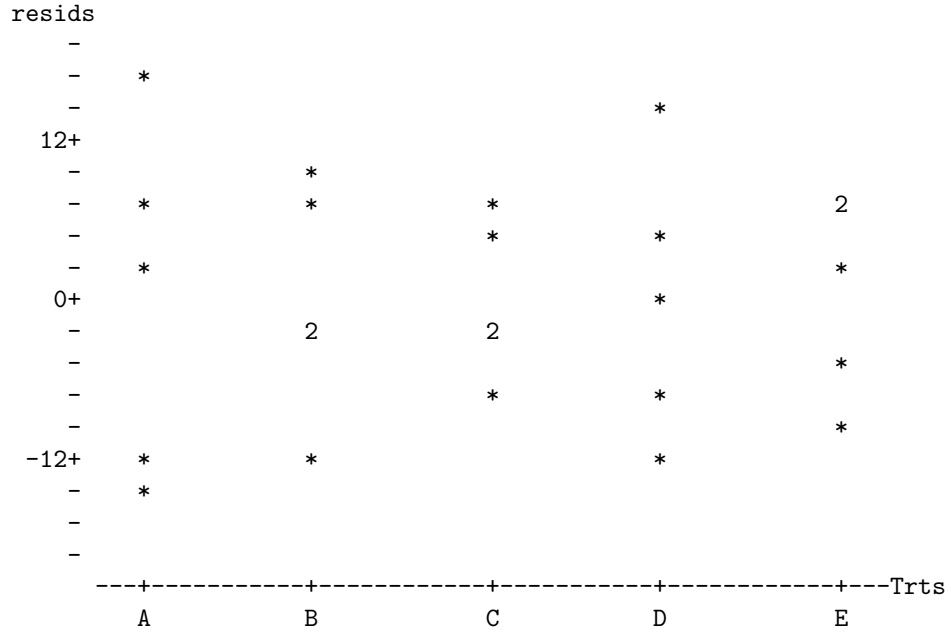



Figure 9.9: *Plot of residuals versus treatment groups.*

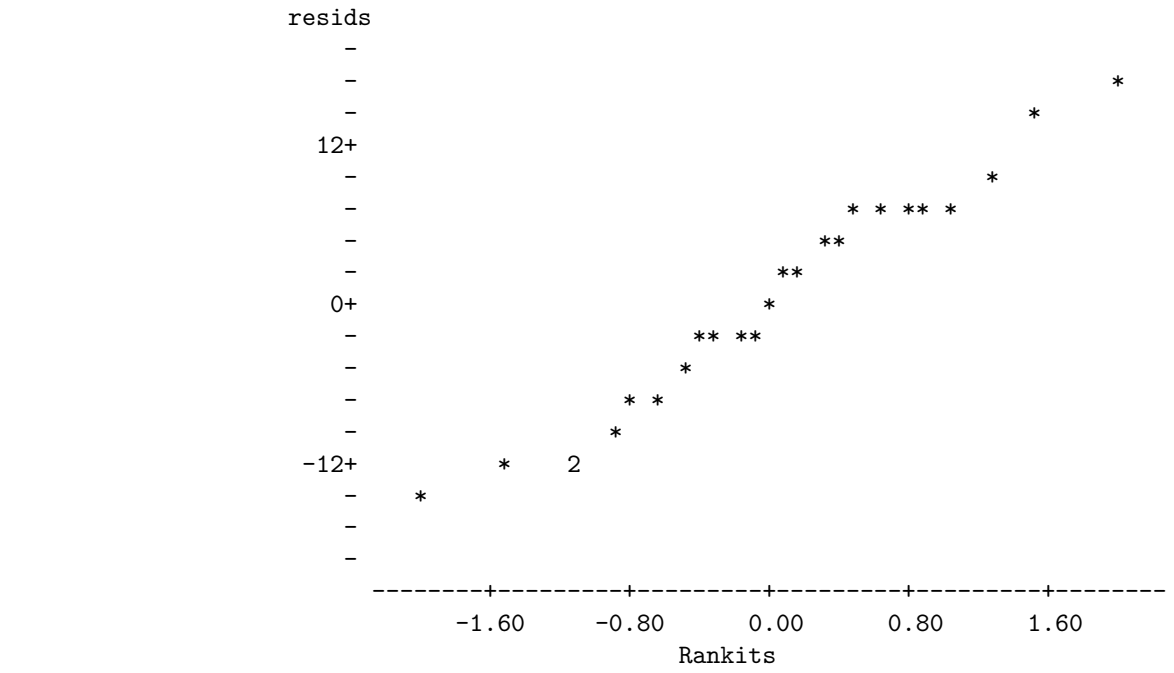



Figure 9.10: *Normal plot of residuals for mangold root data,* $W' = 0.978$.

The 'glm' command can also be used in place of the 'ancova' command but gives more complicated output.

Computer programs for doing balanced analysis of variance are frequently incapable of dealing with Latin squares. For example, Minitab's 'anova' command will not give the analysis. In such cases, a simple trick can obtain the necessary results but the analysis must be constructed out of pieces of the output. The commands are given below. There are many simple ways to get the correct ANOVA table but a key aspect of these commands is that they give the correct residuals.

```
MTB > names c1 'y' c2 'Rows' c3 'Cols' c4 'Trts'
MTB > anova c1 = c2 c3;
SUBC> means c2 c3;
SUBC> resid c10.
MTB > anova c10 = c4;
SUBC> means c4;
SUBC> resid c11.
```

The degrees of freedom, sums of squares, and mean squares for rows, cols, and trts in the two ANOVA tables will be correct. The degrees of freedom total and the sum of squares total from the first ANOVA table (the one computed on c1) will be correct. The first ANOVA is a two-way with $y$ as the dependent variable and rows and columns as the effects. The $SSE$ from the second ANOVA table (on c10) will be correct but $dfE$ and thus $MSE$ will be incorrect. The second ANOVA is a one-way using the treatments as groups and the residuals from the first ANOVA as the dependent variable. From these pieces, a correct ANOVA table can be constructed. In particular, none of the $F$ tests reported by these commands are appropriate. The residuals from the second ANOVA are the appropriate residuals for the Latin square analysis. These are in column c11. The means reported for rows and cols in the first ANOVA will be correct. The means for Trts in the second ANOVA are adjusted for the rows and columns, so they are not the actual treatment means. However, the means for treatments reported in the second ANOVA can be used for treatment comparisons (contrasts)

Table 9.8: *Analysis of variance*

| Source | $df$ | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Trts($\alpha$) | $r-1$ | $r\sum_{i=1}^{r}(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdots})^2$ | $SS(\alpha)/(r-1)$ | $\frac{MS(\alpha)}{MSE}$ |
| Columns($\kappa$) | $r-1$ | $r\sum_{j=1}^{r}(\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdots})^2$ | $SS(\kappa)/(r-1)$ | $\frac{MS(\kappa)}{MSE}$ |
| Rows($\rho$) | $r-1$ | $r\sum_{k=1}^{r}(\bar{y}_{\cdot\cdot k}-\bar{y}_{\cdots})^2$ | $SS(\rho)/(r-1)$ | $\frac{MS(\rho)}{MSE}$ |
| Error | $(r-1)(r-2)$ | $\sum_{i,j}\hat{\varepsilon}_{ijk}^2$ | $SSE/(dfE)$ | |
| Total | $r^2-1$ | $\sum_i\sum_j(y_{ijk}-\bar{y}_{\cdots})^2$ | | |

just as if they were the original treatment means because the row and column adjustments cancel out when performing contrasts. Two additional points should be made. First, the residuals used as the dependent variable in the second ANOVA must be raw residuals, they cannot be standardized. Second, the roles played by rows, columns, and treatments can be interchanged.

*Latin square models*

The model for an $r\times r$ Latin square design is a three-way analysis of variance,

$$y_{ijk}=\mu+\alpha_i+\kappa_j+\rho_k+\varepsilon_{ijk},\quad \varepsilon_{ijk}\text{s independent } N(0,\sigma^2). \tag{9.3.1}$$

The parameter $\mu$ is viewed as a grand mean, $\alpha_i$ is an effect for the $i$th treatment, $\kappa_j$ is an effect for the $j$th column, and $\rho_k$ is an effect for the $k$th row. The subscripting for this model is peculiar. All of the subscripts run from 1 to $r$ but not freely. If you specify a row and a column, the design tells you the treatment. Thus, if you know $k$ and $j$, the design tells you $i$. If you specify a row and a treatment, the design tells you the column, so $k$ and $i$ dictate $j$. In fact, if you know any two of the subscripts, the design tells you the third. The summary statistics necessary for the analysis are the sample variance of all $r^2$ observations and the means for each treatment, column, and row. The predicted values are

$$\hat{y}_{ijk}\equiv\bar{y}_{i\cdot\cdot}+\bar{y}_{\cdot j\cdot}+\bar{y}_{\cdot\cdot k}-2\bar{y}_{\cdots}$$

and the residuals are

$$\hat{\varepsilon}_{ijk}\equiv y_{ijk}-\hat{y}_{ijk}=y_{ijk}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdot\cdot k}+2\bar{y}_{\cdots}.$$

The computations for the $MSE$ can be summarized in an analysis of variance table. The commonly used form for the analysis of variance table is given in Table 9.8. Notice that because of the peculiarity in the subscripting, the sums for error and total are taken over only $i$ and $j$. The choice of these two subscripts is arbitrary; summing over any two subscripts sums over all of the observations in the Latin square. The degrees of freedom and sums of squares for treatments, columns, rows, and error add up to the degrees of freedom and sums of squares total (corrected for the grand mean).

The $F$ statistic $MS(\alpha)/MSE$ is the ratio of the mean square treatments to the mean square error. It is used to test whether there are treatment effects, i.e., it is used to test

$$H_0:\alpha_1=\cdots=\alpha_r.$$

Again if all the $\alpha_i$s are equal, we cannot distinguish between the effects of different treatments, so the treatment $F$ statistic provides a test of whether we can isolate comparative differences in the treatment effects.

The $F$ statistic $MS(\kappa)/MSE$ is the ratio of the mean square columns to the mean square error. It is used to test whether there are column effects, i.e., it is used to test

$$H_0:\kappa_1=\cdots=\kappa_r.$$

The $F$ statistic provides a test of whether we can isolate comparative differences in the column effects. The ratio of the mean square rows to the mean square error gives the $F$ statistic $MS(\rho)/MSE$. This is used to test

$$H_0 : \rho_1 = \cdots = \rho_r.$$

The $F$ statistic provides a test of whether we can isolate comparative differences in the row effects. Some models for Latin square designs do not allow testing for row and column effects but in any case the ratios $MS(\rho)/MSE$ and $MS(\kappa)/MSE$ are of interest in that large values indicate, respectively, that blocking on rows and columns was worthwhile.

The treatments are dealt with exactly as in a one-way ANOVA. A contrast in the treatment effects is, for known $\lambda_i$s that sum to zero,

$$Par = \sum_{i=1}^{r} \lambda_i \alpha_i$$

with

$$Est = \sum_{i=1}^{r} \lambda_i \bar{y}_{i\cdot\cdot}.$$

The treatment means are the average of $r$ observations, so

$$\mathrm{SE}\left(\sum_{i=1}^{r} \lambda_i \bar{y}_{i\cdot\cdot}\right) = \sqrt{MSE \sum_{i=1}^{r} \lambda_i^2/r}.$$

The reference distribution is $t(dfE)$. The sum of squares for the contrast is

$$SS\left(\sum_{i=1}^{r} \lambda_i \alpha_i\right) = \left[\sum_{i=1}^{r} \lambda_i \bar{y}_{i\cdot\cdot}\right]^2 \Bigg/ \left[\sum_{i=1}^{r} \lambda_i^2/r\right].$$

If of interest, similar results hold for the row and column effects. For known $\xi_j$s that add to zero, inferences for, say, the column effects can be based on

$$Par = \sum_{j=1}^{r} \xi_j \kappa_j$$

with

$$Est = \sum_{j=1}^{r} \xi_j \bar{y}_{\cdot j\cdot}.$$

The column means are averages of $r$ observations so

$$\mathrm{SE}\left(\sum_{j=1}^{r} \xi_j \bar{y}_{\cdot j\cdot}\right) = \sqrt{MSE \sum_{j=1}^{r} \xi_j^2/r}.$$

The reference distribution is $t(dfE)$. The sum of squares for the contrast is

$$SS\left(\sum_{j=1}^{r} \xi_j \kappa_j\right) = \left[\sum_{j=1}^{r} \xi_j \bar{y}_{\cdot j\cdot}\right]^2 \Bigg/ \left[\sum_{j=1}^{r} \xi_j^2/r\right].$$

The theoretical justification for the analysis is similar to that for a balanced one-way and a balanced two-way.

*Discussion of Latin squares*

The idea of simultaneously having two distinct sets of complete blocks is quite useful. For example, suppose you wish to compare the performance of four machines in producing something. Productivity is notorious for depending on the day of the week, with Mondays and Fridays often having low productivity; thus we may wish to block on days. The productivity of the machine is also likely to depend on who is operating the machine, so we may wish to block on operators. Thus we may decide to run the experiment on Monday through Thursday with four machine operators and using each operator on a different machine each day. One possible design is

| | Operator | | | |
|---|---|---|---|---|
| Day | 1 | 2 | 3 | 4 |
| Mon | A | B | C | D |
| Tue | B | C | D | A |
| Wed | C | D | A | B |
| Thu | D | A | B | C |

where the numbers 1 through 4 are randomly assigned to the four people who will operate the machines and the letters A through D are randomly assigned to the machines to be examined. Moreover, the days of the week should actually be randomly assigned to the rows of the Latin square. In general, the rows, columns, and treatments should all be randomized in a Latin square.

Another distinct Latin square design for this situation is

| | Operator | | | |
|---|---|---|---|---|
| Day | 1 | 2 | 3 | 4 |
| Mon | A | B | C | D |
| Tue | B | A | D | C |
| Wed | C | D | B | A |
| Thu | D | C | A | B |

This square cannot be obtained from the first one by any interchange of rows, columns, and treatments. Typically, one would randomly choose a possible Latin square design from a list of such squares (see, for example, Cochran and Cox, 1957) in addition to randomly assigning the numbers, letters, and rows to the operators, machines, and days.

The use of Latin square designs can be extended in numerous ways. One modification is the incorporation of a third kind of block; such designs are called *Graeco-Latin squares*. The use of Graeco-Latin squares is explored in the exercises for this chapter. A problem with Latin squares is that small squares give poor variance estimates because they provide few degrees of freedom for error, cf. Table 9.8. For example, a $3 \times 3$ Latin square gives only 2 degrees of freedom for error. In such cases, the Latin square experiment is often performed several times, giving additional replications that provide improved variance estimation. Section 11.4 presents an example in which several Latin squares are used.

## 9.4 Discussion of experimental design

Data are frequently collected with the intention of evaluating a change in the current system of doing things. If you really want to know the effect of a change in the system, you have to execute the change. It is not enough to look at conditions in the past that were similar to the proposed change because, along with the past similarities, there were dissimilarities. For example, suppose you think that instituting a good sex education program in schools will decrease teenage pregnancies. To evaluate this, it is not enough to compare schools that currently have such programs with schools that do not, because along with the differences in sex education programs there are other differences

Table 9.9: *Tensile strength of uniform twill*

| Fabric strips | Machines $m_1$ | $m_2$ | $m_3$ | $m_4$ |
|---|---|---|---|---|
| $s_1$ | 18 | 7 | 5 | 9 |
| $s_2$ | 9 | 11 | 12 | 3 |
| $s_3$ | 7 | 11 | 11 | 1 |
| $s_4$ | 6 | 4 | 10 | 8 |
| $s_5$ | 10 | 8 | 6 | 10 |
| $s_6$ | 7 | 12 | 3 | 15 |
| $s_7$ | 13 | 5 | 15 | 16 |
| $s_8$ | 1 | 11 | 8 | 12 |

in the schools that affect teen pregnancy rates. Such differences may include parents' average socio-economic status and education. While adjustments can be made for any such differences that can be identified, there is no assurance that all important differences can be found. Moreover, initiating the proposed program involves making a change and the very act of change can affect the results. For example, current programs may exist and be effective because of the enthusiasm of the school staff that initiated them. Such enthusiasm is not likely to be duplicated when the new program is mandated from above.

To establish the effect of instituting a sex education program in a population of schools, you really need to (randomly) choose schools and actually institute the program. The schools at which the program is instituted should be chosen randomly, so no (unconscious) bias creeps in due to the selection of schools. For example, the people conducting the investigation are likely to favor or oppose the project. They could (perhaps unconsciously) choose the schools in such a way that makes the evaluation likely to reflect their prior attitudes. Unconscious bias occurs frequently and should *always* be assumed. Other schools without the program should be monitored to establish a base of comparison. These other schools should be treated as similarly as possible to the schools with the new program. For example, if the district school administration or the news media pay a lot of attention to the schools with the new program but ignore the other schools, we will be unable to distinguish the effect of the program from the effect of the attention. In addition, blocking similar schools together can improve the precision of the experimental results.

One of the great difficulties in learning about human populations is that obtaining the best data often requires morally unacceptable behavior. We object to having our lives randomly changed for the benefit of experimental science and typically the more important the issue under study, the more we object to such changes. Thus we find that in studying humans, the best data available are often historical. In our example we might have to accept that the best data available will be an historical record of schools with and without sex education programs. We must then try to identify and adjust for *all* differences in the schools that could potentially affect our conclusions. It is the extreme difficulty of doing this that leads to the relative unreliability of many studies in the social sciences. On the other hand, it would be foolish to give up the study of interesting and important phenomena just because they are difficult to study.

## 9.5 Exercises

EXERCISE 9.5.1. Garner (1956) presented data on the tensile strength of fabrics. Here we consider a subset of the data. The complete data and a more extensive discussion of the experimental procedure are given in Exercise 11.5.2. The experiment involved testing fabric strengths on four different machines. Eight homogeneous strips of cloth were divided into four samples. Each sample was tested on one of four machines. The data are given in Table 9.9.

(a) Identify the design for this experiment and give an appropriate model. List all of the assumptions made in the model.

Table 9.10: *Dead adult flies*

| | Units of active ingredient | | | |
|---|---|---|---|---|
| Medium | 0 | 4 | 8 | 16 |
| A | 423 | 445 | 414 | 247 |
| B | 326 | 113 | 127 | 147 |
| C | 246 | 122 | 206 | 138 |
| D | 141 | 227 | 78 | 148 |
| E | 208 | 132 | 172 | 356 |
| F | 303 | 31 | 45 | 29 |
| G | 256 | 177 | 103 | 63 |

(b) Analyze the data. Give an appropriate analysis of variance table. Examine appropriate contrasts using Tukey's method with $\alpha = .05$

(c) Check the assumptions of the model and adjust the analysis appropriately.

EXERCISE 9.5.2. Snedecor (1945b) presented data on a spray for killing adult flies as they emerged from a breeding medium. The data were numbers of adults found in cages that were set over the medium containers. The treatments were different levels of the spray's active ingredient, namely 0, 4, 8, and 16 units. (Actually, it is not clear whether a spray with 0 units was actually applied or whether no spray was applied. The former might be preferable.) Seven different sources for the breeding mediums were used and each spray was applied on each distinct breeding medium. The data are presented in Table 9.10.

(a) Identify the design for this experiment and give an appropriate model. List all the assumptions made in the model.

(b) Analyze the data. Give an appropriate analysis of variance table. Examine a contrast that compares the treatment with no active ingredient to the average of the three treatments that contain the active ingredient. Ignoring the treatment with no active ingredient, the other three treatments are quantitative levels of the active ingredient. On the log scale, these levels are equally spaced, so the tabled polynomial contrasts can be used to examine the polynomial regression of numbers killed on the log of the amount of active ingredient. The contrasts are given below.

| | Contrasts | | |
|---|---|---|---|
| Treatment | Active vs. inactive | log(active) linear | log(active) quadratic |
| 0 | 3 | 0 | 0 |
| 4 | −1 | −1 | 1 |
| 8 | −1 | 0 | −2 |
| 16 | −1 | 1 | 1 |

Examine these contrasts. Compare the results given by the LSD, Bonferroni, and Scheffé methods. Use $\alpha = .10$ for LSD and Scheffé and something close to .05 for Bonferroni. Are the polynomial contrasts orthogonal to the first contrast?

(c) Check the assumptions of the model and adjust the analysis appropriately.

EXERCISE 9.5.3. Cornell (1988) considered data on scaled thickness values for five formulations of vinyl designed for use in automobile seat covers. Eight groups of material were prepared. The production process was then set up and the five formulations run with the first group. The production process was then reset and another group of five was run. In all, the production process was set eight times and a group of five formulations was run with each setting. The data are displayed in Table 9.11.

Table 9.11: *Cornell's scaled vinyl thickness values*

| | Production setting | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Formulation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 8 | 7 | 12 | 10 | 7 | 8 | 12 | 11 |
| 2 | 6 | 5 | 9 | 8 | 7 | 6 | 10 | 9 |
| 3 | 10 | 11 | 13 | 12 | 9 | 10 | 14 | 12 |
| 4 | 4 | 5 | 6 | 3 | 5 | 4 | 6 | 5 |
| 5 | 11 | 10 | 15 | 11 | 9 | 7 | 13 | 9 |

(a) From the information given, identify the design for this experiment and give an appropriate model. List all the assumptions made in the model.

(b) Analyze the data. Give an appropriate analysis of variance table. Examine appropriate contrasts using the Bonferroni method with an $\alpha$ of about .05.

(c) Check the assumptions of the model and adjust the analysis appropriately.

EXERCISE 9.5.4. In data related to that of the previous problem, Cornell (1988) has scaled thickness values for vinyl under four different process conditions. The process conditions were A, high rate of extrusion, low drying temperature; B, low rate of extrusion, high drying temperature; C, low rate of extrusion, low drying temperature; D, high rate of extrusion, high drying temperature. An initial set of data with these conditions was collected and later a second set was obtained. The data are given below.

| | Treatments | | | |
|---|---|---|---|---|
| | A | B | C | D |
| Rep 1 | 7.8 | 11.0 | 7.4 | 11.0 |
| Rep 2 | 7.6 | 8.8 | 7.0 | 9.2 |

Identify the design, give the model, check the assumptions, give the analysis of variance table and interpret the $F$ test for treatments.

The structure of the treatments suggest some interesting contrasts. These are given below.

| | Treatments | | | |
|---|---|---|---|---|
| Contrast | A | B | C | D |
| Rate | 1 | −1 | −1 | 1 |
| Temp | −1 | 1 | −1 | 1 |
| RT | −1 | −1 | 1 | 1 |

The rate contrast examines the difference between the two treatments with a high rate of extrusion and those with a low rate. The temp contrast examines the difference between the two treatments with a high drying temperature and those with a low temperature. The RT contrast is an interaction contrast that examines whether the effect of extrusion rate is the same for high drying temperatures as for low temperatures. Show that the contrasts are orthogonal and use the contrasts to analyze the data.

EXERCISE 9.5.5. Johnson (1978) and Mandel and Lashof (1987) present data on measurements of $P_2O_5$ (phosphorous pentoxide) in fertilizers. Table 9.12 presents data for five fertilizers, each analyzed in five labs. Our interest is in differences among the labs. Analyze the data.

EXERCISE 9.5.6. Table 9.13 presents data on yields of cowpea hay. Four treatments are of interest, variety I of hay was planted 4 inches apart (I4), variety I of hay was planted 8 inches apart (I8), variety II of hay was planted 4 inches apart (II4), and variety II of hay was planted 8 inches apart (II8). Three blocks of land were each divided into four plots and one of the four treatments

Table 9.12: *Phosphorous fertilizer data*

| | Laboratory | | | | |
|---|---|---|---|---|---|
| Fertilizer | 1 | 2 | 3 | 4 | 5 |
| F | 20.20 | 19.92 | 20.91 | 20.65 | 19.94 |
| G | 30.20 | 30.09 | 29.10 | 29.85 | 30.29 |
| H | 31.40 | 30.42 | 30.18 | 31.34 | 31.11 |
| I | 45.88 | 45.48 | 45.51 | 44.82 | 44.63 |
| J | 46.75 | 47.14 | 48.00 | 46.37 | 46.63 |

Table 9.13: *Cowpea hay yields*

| | Block | | | Trt. |
|---|---|---|---|---|
| Treatment | 1 | 2 | 3 | means |
| I4 | 45 | 43 | 46 | $44.66\bar{6}$ |
| I8 | 50 | 45 | 48 | $47.66\bar{6}$ |
| II4 | 61 | 60 | 63 | $61.33\bar{3}$ |
| II8 | 58 | 56 | 60 | 58.000 |
| Block means | 53.50 | 51.00 | 54.25 | $52.91\bar{6}$ |

was randomly applied to each plot. These data are actually a subset of a larger data set given by Snedecor and Cochran (1980, p. 309) that involves three varieties and three spacings in four blocks. Analyze the data. Check your assumptions. Examine appropriate contrasts.

EXERCISE 9.5.7. In the study of the optical emission spectrometer discussed in Example 9.2.1 and Table 9.1, the target value for readings was .89. Subtract .89 from each observation and repeat the analysis. What new questions are of interest? Which aspects of the analysis have changed and which have not?

EXERCISE 9.5.8. An experiment was conducted to examine differences among operators of Suter hydrostatic testing machines. These machines are used to test the water repellency of squares of fabric. One large square of fabric was available but its water repellency was thought to vary along the length (warp) and width (fill) of the fabric. To adjust for this, the square was divided into four equal parts along the length of the fabric and four equal parts along the width of the fabric, yielding 16 smaller pieces. These pieces were used in a Latin square design to investigate differences among four operators: A, B, C, D. The data are given in Table 9.14. Construct an analysis of variance table. What, if any, differences can be established among the operators? Compare the results of using the Tukey, Newman–Keuls, and Bonferroni methods for comparing the operators.

EXERCISE 9.5.9. Table 9.15 contains data similar to that in the previous exercise except that in

Table 9.14: *Hydrostatic pressure tests: operator, yield*

| | | | |
|---|---|---|---|
| A | B | C | D |
| 40.0 | 43.5 | 39.0 | 44.0 |
| B | A | D | C |
| 40.0 | 42.0 | 40.5 | 38.0 |
| C | D | A | B |
| 42.0 | 40.5 | 38.0 | 40.0 |
| D | C | B | A |
| 40.0 | 36.5 | 39.0 | 38.5 |

Table 9.15: *Hydrostatic pressure tests: machine, yield*

| | | | |
|---|---|---|---|
| 2 | 4 | 3 | 1 |
| 39.0 | 39.0 | 41.0 | 41.0 |
| 1 | 3 | 4 | 2 |
| 36.5 | 42.5 | 40.5 | 38.5 |
| 4 | 2 | 1 | 3 |
| 40.0 | 39.0 | 41.5 | 41.5 |
| 3 | 1 | 2 | 4 |
| 41.5 | 39.5 | 39.0 | 44.0 |

Table 9.16: *Hydrostatic pressure tests: operator, machine*

| | | | |
|---|---|---|---|
| B,2 | A,4 | D,3 | C,1 |
| A,1 | B,3 | C,4 | D,2 |
| D,4 | C,2 | B,1 | A,3 |
| C,3 | D,1 | A,2 | B,4 |

Operators are A, B, C, D.
Machines are 1, 2, 3, 4.

this Latin square differences among four machines: 1, 2, 3, 4, were investigated rather than differences among operators. Machines 1 and 2 were operated with a hand lever, while machines 3 and 4 were operated with a foot lever. Construct an analysis of variance table. What, if any, differences can be established among the machines? To this end, construct appropriate orthogonal contrasts.

EXERCISE 9.5.10. Table 9.15 is incomplete. The data were actually obtained from a Graeco-Latin square that incorporates four different operators as well as the four different machines. The correct design is given in Table 9.16. Note that this is a Latin square for machines when we ignore the operators and a Latin square for operators when we ignore the machines. Moreover, every operator works once with every machine. Using the four operator means, compute a sum of squares for operators and subtract this from the error computed in Exercise 9.5.9. Give the new analysis of variance table. How do the results on machines change? What evidence is there for differences among operators. Was the analysis for machines given earlier incorrect or merely inefficient?

EXERCISE 9.5.11. Table 9.17 presents data given by Nelson (1993) on disk drives from a Graeco-Latin square design (see Exercise 9.5.10). The experiment was planned to investigate the effect of four different substrates on the drives. The dependent variable is the amplitude of a signal read from the disk where the signal written onto the disk had a fixed amplitude. Blocks were constructed from machines, operators, and day of production. (In Table 9.17, Days are indicated by lower case Latin letters.) The substrata consist of A, aluminum; B, nickel plated aluminum; and two types of glass, C and D. Analyze the data. In particular, check for differences between aluminum and glass, between the two types of glass, and between the two types of aluminum. Check your assumptions.

Table 9.17: *Amplitudes of disk drives*

| | Machine | | | |
|---|---|---|---|---|
| Operator | 1 | 2 | 3 | 4 |
| I | Aa 8 | Cd 7 | Db 3 | Bc 4 |
| II | Cc 11 | Ab 5 | Bd 9 | Da 5 |
| III | Dd 2 | Ba 2 | Ac 7 | Cb 9 |
| IV | Bb 8 | Dc 4 | Ca 9 | Ad 3 |

Chapter 10

# Analysis of covariance

Analysis of covariance incorporates one or more regression variables into an analysis of variance. The regression variables are referred to as covariates (relative to the dependent variable), hence the name analysis of covariance. Covariates are also known as supplementary or concomitant observations. Cox (1958, chapter 4) gives a particularly nice discussion of the ideas behind analysis of covariance and illustrates various useful plotting techniques. In 1957 and 1982, *Biometrics* devoted entire issues to the analysis of covariance. In this chapter, we only examine the use of a single covariate. We begin our discussion with an example that involves one-way analysis of variance and a covariate.

In Sections 1 and 4 of this chapter, we make extensive use of model comparisons. To simplify the discussions within these sections, we will often refer to a model such as (10.1.1) as simply model (1).

## 10.1 An example

Fisher (1947) gives data on the body weights (in kilograms) and heart weights (in grams) for domestic cats of both sexes that were given digitalis. A subset of the data is presented in Table 10.1. Our primary interest is to determine whether females' heart weights differ from males' heart weights when both have received digitalis.

As a first step, we might fit a one-way ANOVA model,

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij}, \end{aligned} \tag{10.1.1}$$

where the $y_{ij}$s are the heart weights, $i = 1,2$, and $j = 1,\ldots,24$. This model yields the analysis of variance given in Table 10.2. Note the overwhelming effect due to sexes.

Table 10.1: *Fisher's data on body weights (kg) and heart weights (g) of domestic cats given digitalis*

| Females | | | | Males | | | |
|---|---|---|---|---|---|---|---|
| Body | Heart | Body | Heart | Body | Heart | Body | Heart |
| 2.3 | 9.6 | 2.0 | 7.4 | 2.8 | 10.0 | 2.9 | 9.4 |
| 3.0 | 10.6 | 2.3 | 7.3 | 3.1 | 12.1 | 2.4 | 9.3 |
| 2.9 | 9.9 | 2.2 | 7.1 | 3.0 | 13.8 | 2.2 | 7.2 |
| 2.4 | 8.7 | 2.3 | 9.0 | 2.7 | 12.0 | 2.9 | 11.3 |
| 2.3 | 10.1 | 2.1 | 7.6 | 2.8 | 12.0 | 2.5 | 8.8 |
| 2.0 | 7.0 | 2.0 | 9.5 | 2.1 | 10.1 | 3.1 | 9.9 |
| 2.2 | 11.0 | 2.9 | 10.1 | 3.3 | 11.5 | 3.0 | 13.3 |
| 2.1 | 8.2 | 2.7 | 10.2 | 3.4 | 12.2 | 2.5 | 12.7 |
| 2.3 | 9.0 | 2.6 | 10.1 | 2.8 | 13.5 | 3.4 | 14.4 |
| 2.1 | 7.3 | 2.3 | 9.5 | 2.7 | 10.4 | 3.0 | 10.0 |
| 2.1 | 8.5 | 2.6 | 8.7 | 3.2 | 11.6 | 2.6 | 10.5 |
| 2.2 | 9.7 | 2.1 | 7.2 | 3.0 | 10.6 | 2.5 | 8.6 |

Table 10.2: *One-way analysis of variance on heart weights*

| Source | $df$ | $SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Sex | 1 | 56.117 | 56.117 | 23.44 | .0000 |
| Error | 46 | 110.11 | 2.3936 | | |
| Total | 47 | 166.223 | | | |

Table 10.3: *Analysis of variance for heart weights based on model (2)*

| Source | $df$ | $Adj.\ SS$ | $MS$ | $F$ | $P$ |
|---|---|---|---|---|---|
| Body weights | 1 | 37.828 | 37.828 | 23.55 | 0.000 |
| Sex | 1 | 4.499 | 4.499 | 2.80 | 0.101 |
| Error | 45 | 72.279 | 1.606 | | |
| Total | 47 | 166.223 | | | |

Fisher provided both heart weights and body weights, so we can ask a more complex question, 'Is there a sex difference in the heart weights over and above the fact that male cats are naturally larger than female cats?' To examine this we add a regression term to model (1) and fit the traditional *analysis of covariance model*,

$$\begin{aligned} y_{ij} &= \mu_i + \gamma z_{ij} + \varepsilon_{ij} \\ &= \mu + \alpha_i + \gamma z_{ij} + \varepsilon_{ij}. \end{aligned} \tag{10.1.2}$$

Here the $z_{ij}$s are the body weights and $\gamma$ is a slope parameter associated with body weights. Note that model (2) is an extension of the simple linear regression between the $y$s and the $z$s in which we allow a different intercept $\mu_i$ for each sex. An analysis of variance table for model (2) is given as Table 10.3. The interpretation of this table is different from the ANOVA tables examined earlier. For example, the sums of squares for body weights, sex, and error *do not* add up to the sum of squares total. The sums of squares in Table 10.3 are referred to as *adjusted sums of squares* ($Adj.\ SS$) because the body weight sum of squares is adjusted for sexes and the sex sum of squares is adjusted for body weights. In this section, we focus on the interpretation of Table 10.3; in Section 10.3 we discuss its computation.

The error line in Table 10.3 is simply the error from fitting model (2). The body weights line comes from comparing model (2) with the reduced model (1). Note that the only difference between models (1) and (2) is that (1) does not involve the regression on body weights, so by testing the two models we are testing whether there is a significant effect due to the regression on body weights. The standard way of comparing a full and a reduced model is by comparing their error terms. Model (2) has one more parameter, $\gamma$, than model (1), so there is one more degree of freedom for error in model (1) than in model (2), hence one degree of freedom for body weights. The adjusted sum of squares for body weights is the difference between the sum of squares error in model (1) and the sum of squares error in model (2). Given the sum of squares and the mean square, the $F$ statistic for body weights is constructed in the usual way, cf. Section 5.5. Examining Table 10.3, we see a major effect due to the regression on body weights.

The sex line in Table 10.3 provides a test of whether there are differences in sexes *after adjusting for the regression on body weights*. This comes from comparing model (2) to a similar model in which sex differences have been eliminated. In model (2), the sex differences are incorporated as $\mu_1$ and $\mu_2$ in the first version and as $\alpha_1$ and $\alpha_2$ in the second version. To eliminate sex differences in model (2), we simply eliminate the distinctions between the $\mu$s (the $\alpha$s). Such a model can be written as

$$y_{ij} = \mu + \gamma z_{ij} + \varepsilon_{ij}.$$

In this example, the analysis of covariance model without treatment effects is just a simple linear

regression of heart weight on body weight. We have reduced the two sex parameters to one overall parameter, so the difference in degrees of freedom between this model and model (2) is 1. The difference in the sums of squares error between this model and model (2) is the adjusted sum of squares for sex. Examining Table 10.3 we see that the evidence for a sex effect over and above the effect due to the regression on body weights is not great.

Our current data come from an observational study rather than a designed experiment. It is difficult to take a group of cats and randomly assign them to sex groups. As discussed in the next section, principles of experimental design focus attention on models such as (2). However, these data are from an observational study, so yet another model is of interest. There is little reason to assume that when regressing heart weight on body weight the relationships are the same for females and males. Model (2) allows different intercepts for these regressions but uses the same slope $\gamma$. We should test the assumption of a common slope by fitting the more general model that allows different slopes for females and males, i.e.,

$$\begin{aligned} y_{ij} &= \mu_i + \gamma_i z_{ij} + \varepsilon_{ij} \\ &= \mu + \alpha_i + \gamma_i z_{ij} + \varepsilon_{ij}. \end{aligned} \tag{10.1.3}$$

In model (3) the $\gamma$s depend on $i$ and thus the slopes are allowed to differ between the sexes. While model (3) may look complicated, it consists of nothing more than fitting a simple linear regression to each group: one to the female data and a separate simple linear regression to the male data. The sum of squares error for model (3) comes from adding the error sums of squares for the two simple linear regressions. It is easily seen that for females the simple linear regression has an error sum of squares of 22.459 on 22 degrees of freedom and the males have an error sum of squares of 49.614 also on 22 degrees of freedom. Thus model (3) has an error sum of squares of $22.459 + 49.614 = 72.073$ on $22 + 22 = 44$ degrees of freedom. The mean squared error for model (3) is

$$MSE(3) = \frac{72.073}{44} = 1.638$$

and using results from Table 10.3, the test of model (3) against the reduced model (2) has

$$F = \frac{[72.279 - 72.073]/[45 - 44]}{1.638} = \frac{.206}{1.638} = .126.$$

The $F$ statistic is very small; there is no evidence that we need to fit different slopes for the two sexes. We now return to the analysis of model (2).

Frequently, computer programs for fitting model (2) give information on the regression parameter $\gamma$. Often, this is presented in the same way the program gives information on parameters in pure regression problems, e.g.,

| Covariate | $\hat{\gamma}$ | SE($\hat{\gamma}$) | $t$ | $P$ |
|---|---|---|---|---|
| Body weight | 2.7948 | 0.5759 | 4.853 | 0.000. |

Note that the $t$ statistic here is the square root of the $F$ statistic for body weights in Table 10.3. The $P$ values are identical. Again, we find clear evidence for the effect of body weights. A 95% confidence interval for $\gamma$ has end points

$$2.7948 \pm 2.014(0.5759)$$

which yields the interval $(1.6, 4.0)$. We are 95% confident that, for data comparable to the data in this study, an increase in body weight of one kilogram corresponds to a mean increase in heart weight of between 1.6 g and 4.0 g.

In model (2), comparing treatments by comparing the treatment means $\bar{y}_{i\cdot}$ is inappropriate because of the complicating effect of the covariate. Adjusted means are often used to compare treatments. The formula and the actual values for the adjusted means are given below along with the raw means for body weights and heart rates.

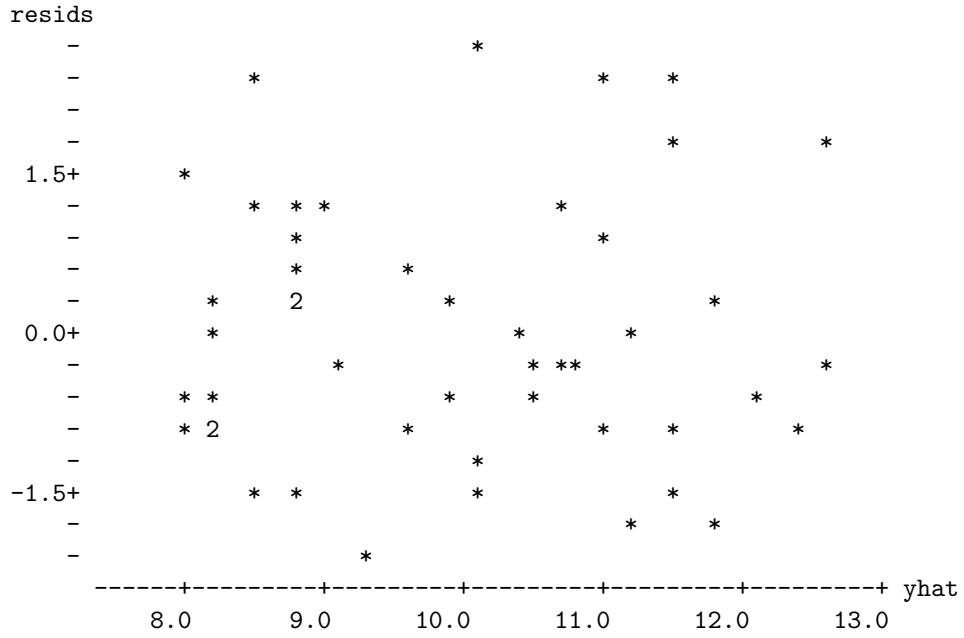



Figure 10.1: *Residuals versus predicted values.*

Adjusted means $\equiv \bar{y}_{i\cdot} - \hat{\gamma}(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})$

| Sex | $N$ | Body | Heart | Adj. heart |
|---|---|---|---|---|
| Female | 24 | 2.333 | 8.887 | 9.580 |
| Male | 24 | 2.829 | 11.050 | 10.357 |
| Combined | 48 | 2.581 | 9.969 | |

We have seen previously that there is little evidence of a differential effect on heart weights due to sexes after adjusting for body weights. Nonetheless, from the adjusted means what evidence exists suggests that, even after adjusting for body weights, a typical heart weight for males, 10.357, is larger than a typical heart weight for females, 9.580.

Figures 10.1 through 10.3 contain residual plots. The plot of residuals versus predicted values looks exceptionally good. The plot of residuals versus sexes shows slightly less variability for females than for males. The difference is probably not enough to worry about. The normal plot of the residuals is alright with $W'$ above the appropriate percentile.

*Minitab commands*

The following Minitab commands were used to generate the analysis of these data. The means given by the 'ancova' subcommand 'means' are the adjusted treatment means.

```
MTB > names c1 'body' c2 'heart' c3 'sex'
MTB > note   Fit model (1).
MTB > oneway c2 c3
MTB > note   Fit model (2).
MTB > ancova c2 = c3;
SUBC> covar c1;
SUBC> resid c10;
SUBC> fits c11;
SUBC> means c3.
```

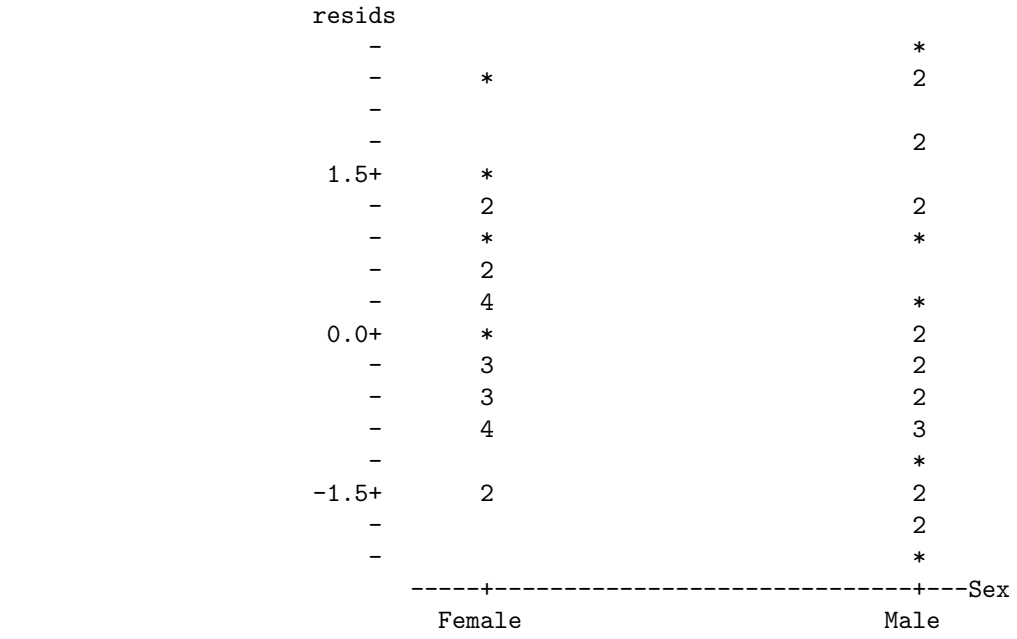



Figure 10.2: *Residuals versus sex.*

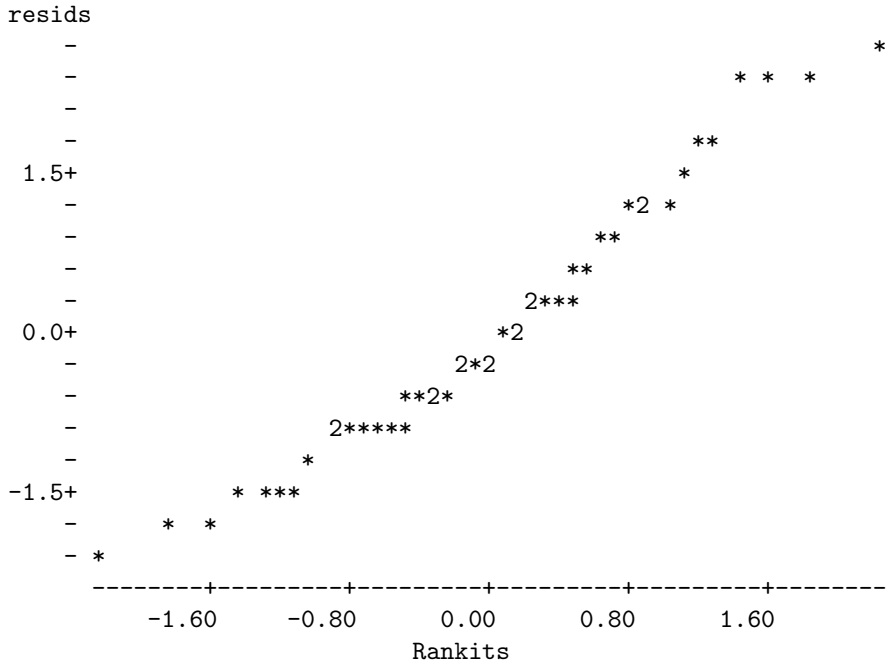



Figure 10.3: *Normal plot of residuals,* $W' = 0.968$.

```
MTB > plot c10 c11
MTB > plot c10 c3
MTB > note   Split the data into females and males and
MTB > note   perform two regressions to fit model (3).
MTB > copy c1 c2 to c11 c12;
SUBC> use c3=1.
MTB > regress c12 on 1 c11
MTB > copy c1 c2 to c21 c22;
SUBC> use c3=2.
MTB > regress c22 on 1 c21
```

### 10.2 Analysis of covariance in designed experiments

In designing an experiment to investigate a group of treatments, covariates are used to reduce the error of treatment comparisons. One way to use the concomitant observations is to define blocks based on them. For example, income, IQ, and heights can all be used to collect people into similar groups for a randomized complete block design. In fact, any construction of blocks must be based on information not otherwise incorporated into the ANOVA model, so any experiment with blocking uses concomitant information. In analysis of covariance we use the concomitant observations more directly, as regression variables in the statistical model.

Obviously, for a covariate to help our analysis it must be related to the dependent variable. Unfortunately, improper use of concomitant observations can invalidate, or at least alter, comparisons among the treatments. In the example of Section 10.1, the original ANOVA demonstrated an effect on heart weights due to sex but after adjusting for body weights, there was little evidence for a sex difference. The very nature of what we were comparing changed when we adjusted for body weights. Originally, we investigated whether heart weights were different for females and males. The analysis of covariance examined whether there were differences between female heart weights and male heart weights *beyond what could be accounted for by the regression on body weights*. These are very different interpretations. In a designed experiment, we want to investigate the effects of the treatments and not the treatments adjusted for some covariates. To this end, in a designed experiment we require that the covariates be logically independent of the treatments. In particular, we require that

the concomitant observations be made before assigning the treatments to the experimental units,

the concomitant observations be made after assigning treatments to experimental units but before the effect of the treatments has developed, or

the concomitant observations be such that they are unaffected by treatment differences.

For example, suppose the treatments are five diets for cows and we wish to investigate milk production. Milk production is related to the size of the cow, so we might pick height of the cow as a covariate. For immature cows over a long period of time, diet may well affect both height and milk production. Thus to use height as a covariate we should measure heights before treatments begin or we could measure heights, say, two days after treatments begin. Two days on any reasonable diet should not affect a cow's height. Alternatively, if we use only mature cows their heights should be unaffected by diet and thus the heights of mature cows could be measured at any time during the experiment. Typically, *one should be very careful when claiming that a covariate measured near the end of an experiment is unaffected by treatments*.

The requirements listed above on the nature of covariates in a designed experiment are imposed so that the treatment effects do not depend on the presence or absence of covariates in the analysis. The treatment effects are logically identical regardless of whether covariates are actually measured or incorporated into the analysis. Recall that in the observational study of Section 10.1, the nature of the treatment (sex) effects changed depending on whether covariates were incorporated in the model. The role of the covariates in the analysis of a designed experiment is solely to reduce the

error. In particular, using good covariates should reduce both the variance of the observations $\sigma^2$ and its estimate, the *MSE*. On the other hand, we will see in the next section that one pays a price for using covariates. Variances of treatment comparisons are $\sigma^2$ times a constant. With covariates in the model, the constant is larger than when they are not present. However, with well chosen covariates the appropriate value of $\sigma^2$ should be sufficiently smaller that the reduction in *MSE* overwhelms the increase in the multiplier. Nonetheless, in designing an experiment we need to play off these aspects against one another. We need covariates whose reduction in *MSE* more than makes up for the increase in the constant.

The requirements imposed on the nature of the covariates in a designed experiment have little affect on the analysis illustrated in the Section 10.1. The analysis focuses on a model such as (10.1.2). In Section 10.1, we also considered model (10.1.3) that has different slope parameters for the different treatments (sexes). The requirements on the covariates in a designed experiment imply that the relationship between the dependent variable $y$ and the covariate $z$ *cannot* depend on the treatments. Thus with covariates chosen for a designed experiment *it is inappropriate to have slope parameters that depend on the treatment*. There is one slope that is valid for the entire analysis and the treatment effects do not depend on the presence or absence of the covariates. If a model such as (10.1.3) fits better than (10.1.2) when the covariate has been chosen appropriately, it suggests that the effects of treatments may differ from experimental unit to experimental unit. In such cases a treatment cannot really be said to have *an* effect, it has a variety of effects depending on which units it is applied to. A suitable transformation of the dependent variable may alleviate the problem.

## 10.3 Computations and contrasts

Analysis of covariance begins with an analysis of variance model and adds a regressor to the model. The original idea in computing an analysis of covariance was to use the simple computations available for one-way ANOVAs and balanced higher-way ANOVAs to expedite the computations for the more complicated analysis of covariance model. With modern computing machines this is less crucial, but the original computational methods deserve consideration. In particular, they are extremely useful in statistical theory. To illustrate, consider a randomized complete block (RCB) experiment with a covariate $z$. The model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma z_{ij} + \varepsilon_{ij} \tag{10.3.1}$$

with $i = 1,\ldots,a$ indicating treatments, $j = 1,\ldots,b$ indicating blocks, and independent $N(0,\sigma^2)$ errors. The computational method involves performing RCB analyses on both the $y_{ij}$s and the $z_{ij}$s. In addition to computing the usual sums of *squares* for the $y$s and the $z$s, we need to compute sums of *cross products*. The formulae are given in Table 10.4. The entire analysis of covariance can be computed from the sums of squares and cross products in Table 10.4 along with the mean values needed to perform the two RCB analyses. In particular, the analysis focuses on the three error lines in Table 10.4. *For analysis of covariance models other than (10.3.1), similar methods applied to the error lines yield the appropriate analysis*. The analogous computations for model (10.1.2) are illustrated at the end of the section.

From the error sums of squares and cross products in Table 10.4, compute the *SSE* for model (10.3.1) as

$$SSE = SSE_{yy} - \frac{(SSE_{yz})^2}{SSE_{zz}}.$$

Model (10.3.1) has one more parameter ($\gamma$) than the corresponding RCB model, so

$$dfE = (a-1)(b-1) - 1$$

and

$$MSE = \frac{SSE}{(a-1)(b-1)-1}.$$

Table 10.4: *RCB analysis of covariance, one covariate*

| Source | $df$ | $SS_{yy}$ |
|---|---|---|
| Trt | $a-1$ | $b\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$ |
| Blocks | $b-1$ | $a\sum_{j=1}^{b}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$ |
| Error | $(a-1)(b-1)$ | subtraction |
| Total | $ab-1$ | $\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij}-\bar{y}_{\cdot\cdot})^2$ |

| Source | $df$ | $SS_{yz}$ |
|---|---|---|
| Trt | $a-1$ | $b\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})(\bar{z}_{i\cdot}-\bar{z}_{\cdot\cdot})$ |
| Blocks | $b-1$ | $a\sum_{j=1}^{b}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})(\bar{z}_{\cdot j}-\bar{z}_{\cdot\cdot})$ |
| Error | $(a-1)(b-1)$ | subtraction |
| Total | $ab-1$ | $\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij}-\bar{y}_{\cdot\cdot})(\bar{z}_{ij}-\bar{z}_{\cdot\cdot})$ |

| Source | $df$ | $SS_{zz}$ |
|---|---|---|
| Trt | $a-1$ | $b\sum_{i=1}^{a}(\bar{z}_{i\cdot}-\bar{z}_{\cdot\cdot})^2$ |
| Blocks | $b-1$ | $a\sum_{j=1}^{b}(\bar{z}_{\cdot j}-\bar{z}_{\cdot\cdot})^2$ |
| Error | $(a-1)(b-1)$ | subtraction |
| Total | $ab-1$ | $\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{z}_{ij}-\bar{z}_{\cdot\cdot})^2$ |

The estimate of $\gamma$ is computed as

$$\hat{\gamma}=\frac{SSE_{yz}}{SSE_{zz}}$$

and the standard error is

$$\mathrm{SE}(\hat{\gamma})=\sqrt{\frac{MSE}{SSE_{zz}}}.$$

The sum of squares for the covariate can be computed as

$$SS(\hat{\gamma})=\hat{\gamma}^2 SSE_{zz}.$$

All of these formulae are very similar to formulae used in simple linear regression. In fact, if model (10.3.1) had no treatment or block effects, it would be a simple linear regression model and these formula would give the usual analysis for a simple linear regression.

The estimate of a contrast in the treatment effects, say $\sum_{i=1}^{a}\lambda_i\alpha_i$, is

$$\sum_{i=1}^{a}\lambda_i\left(\bar{y}_{i\cdot}-\hat{\gamma}\bar{z}_{i\cdot}\right).$$

The variance of the estimated contrast is

$$\mathrm{Var}\left(\sum_{i=1}^{a}\lambda_i\left(\bar{y}_{i\cdot}-\bar{z}_{i\cdot}\hat{\gamma}\right)\right)=\sigma^2\left[\frac{\sum_{i=1}^{a}\lambda_i^2}{b}+\frac{\left(\sum_{i=1}^{a}\lambda_i\bar{z}_{i\cdot}\right)^2}{SSE_{zz}}\right].$$

Recall that in an RCB model without covariates, the variance of the estimate of $\sum_{i=1}^{a}\lambda_i\alpha_i$ is

$\sigma^2\left[\sum_{i=1}^a \lambda_i^2/b\right]$ which is the variance given above without the term involving the $z_{ij}$s. The RCB variance appears to be strictly smaller than the variance from model (10.3.1). This illusion occurs because the variance parameters $\sigma^2$ are not the same in the covariate model (10.3.1) and the RCB model without covariates. With good covariates, the variance $\sigma^2$ in the covariate model should be much smaller than the corresponding variance in the model without covariates. In fact, the $\sigma^2$ in the covariate model should be sufficiently small to *more than make up for* the increase in the term that is multiplying $\sigma^2$.

The standard error of the estimated contrast is obtained immediately from the variance formula. It is

$$\mathrm{SE}\left(\sum_{i=1}^a \lambda_i\left(\bar{y}_{i\cdot} - \bar{z}_{i\cdot}\hat{\gamma}\right)\right) = \sqrt{MSE\left[\frac{\sum_{i=1}^a \lambda_i^2}{b} + \frac{\left(\sum_{i=1}^a \lambda_i \bar{z}_{i\cdot}\right)^2}{SSE_{zz}}\right]}.$$

Because of the complications caused by having the regressor $z_{ij}$ in the model, orthogonal contrasts are difficult to specify and of little interest.

For what they are worth, adjusted treatment means are often defined as

$$\bar{y}_{i\cdot} - \hat{\gamma}(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot}).$$

These adjusted treatment means can be used in place of the values $\bar{y}_{i\cdot} - \hat{\gamma}\bar{z}_{i\cdot}$ when estimating contrasts. Using adjusted treatment means has no affect on the standard error of the estimated contrast.

To test for the existence of treatment effects, we test model (10.3.1) against the reduced model

$$y_{ij} = \mu + \beta_j + \gamma z_{ij} + e_{ij}. \tag{10.3.2}$$

In model (10.3.2) the treatment effects $\alpha_i$ have been eliminated, so the sum of squares for treatments is incorporated into the error term of model (10.3.2). To find the *SSE* for model (10.3.2) we combine the treatment and error lines in Table 10.4 and use the standard formula.

$$SSE(2) = \left[(SSTrt_{yy} + SSE_{yy}) - \frac{(SSTrt_{yz} + SSE_{yz})^2}{SSTrt_{zz} + SSE_{zz}}\right].$$

The sum of squares used in the numerator of the $F$ statistic for testing treatments is

$$\begin{aligned} SSTrt &= SSE(2) - SSE(1) \\ &= \left[(SSTrt_{yy} + SSE_{yy}) - \frac{(SSTrt_{yz} + SSE_{yz})^2}{SSTrt_{zz} + SSE_{zz}}\right] - \left[SSE_{yy} - \frac{SSE_{yz}^2}{SSE_{zz}}\right]. \end{aligned}$$

This has the standard number of degrees of freedom, $a-1$. The $F$ statistic for treatments is

$$F = \frac{SSTrt/(a-1)}{MSE}$$

with *MSE* coming from model (10.3.1).

If it is of interest to test for block effects or investigate block contrasts, the methods given above apply with appropriate substitutions.

*While the discussion in this section has been in terms of analyzing randomized complete block designs, analogous procedures work for any analysis of covariance model in which the corresponding analysis of variance computations are tractable*. We illustrate the computations with the balanced one-way ANOVA data of Section 10.1.

EXAMPLE 10.3.1. Table 10.1 gave Fisher's heart and body weight data, Table 10.2 gave the one-way analysis of variance for heart weights, and Table 10.3 gave an analysis of variance table for the covariate model. Table 10.5 is analogous to Table 10.4; it gives the standard analysis of covariance