Chapter 9

Inference

In Chapters 3–8 we developed methods for studying the behavior of random variables. Given a specific probability distribution, we can calculate the probabilities of various events. For example, knowing that $Y \sim \text{Binomial}(n = 100; p = 0.5)$, we can calculate $P(40 \leq Y \leq 60)$. Roughly speaking, statistics is concerned with the opposite sort of problem. For example, knowing that $Y \sim \text{Binomial}(n = 100; p)$, *where the value of* p *is unknown*, and having observed $Y = y$ (say $y = 32$), what can we say about p? The phrase *statistical inference* describes any procedure for extracting information about a probability distribution from an observed sample.

The present chapter introduces the fundamental principles of statistical inference. We will discuss three types of statistical inference—point estimation, hypothesis testing, and set estimation—in the context of drawing inferences about a single population mean. More precisely, we will consider the following situation:

1. $X_1, \ldots, X_n$ are independent and identically distributed random variables. We observe a sample, $\vec{x} = \{x_1, \ldots, x_n\}$.

2. Both $EX_i = \mu$ and $\text{Var}\, X_i = \sigma^2$ exist and are finite. We are interested in drawing inferences about the population mean μ, a quantity that is fixed but unknown.

3. The sample size, n, is sufficiently large that we can use the normal approximation provided by the Central Limit Theorem.

We begin, in Section 9.1, by examining a narrative that is sufficiently nuanced to motivate each type of inferential technique. We then proceed to

discuss point estimation (Section 9.2), hypothesis testing (Sections 9.3 and 9.4), and set estimation (Section 9.5). Although we are concerned exclusively with large-sample inferences about a single population mean, it should be appreciated that this concern often arises in practice. More importantly, the fundamental concepts that we introduce in this context are common to virtually all problems that involve statistical inference.

9.1 A Motivating Example

We consider an artificial example that permits us to scrutinize the precise nature of statistical reasoning. Two siblings, a magician (Arlen) and an attorney (Robin) agree to resolve their disputed ownership of an Erté painting by tossing a penny. Arlen produces a penny and, just as Robin is about to toss it in the air, Arlen smoothly suggests that spinning the penny on a table might ensure better randomization. Robin assents and spins the penny. As it spins, Arlen calls "Tails!" The penny comes to rest with `Tails` facing up and Arlen takes possession of the Erté. Robin is left with the penny.

That evening, Robin wonders if she has been had. She decides to perform an experiment. She spins the same penny on the same table 100 times and observes 68 `Tails`. It occurs to Robin that perhaps spinning this penny was not entirely fair, but she is reluctant to accuse her brother of impropriety until she is convinced that the results of her experiment cannot be dismissed as coincidence. How should she proceed?

It is easy to devise a mathematical model of Robin's experiment: each spin of the penny is a Bernoulli trial and the experiment is a sequence of $n = 100$ trials. Let X_i denote the outcome of spin i, where $X_i = 1$ if `Heads` is observed and $X_i = 0$ if `Tails` is observed. Then $X_1, \ldots, X_{100} \sim \text{Bernoulli}(p)$, where p is the fixed but unknown (to Robin!) probability that a single spin will result in `Heads`. The probability distribution Bernoulli(p) is our mathematical abstraction of a population and the population parameter of interest is $\mu = EX_i = p$, the population mean.

Let

$$Y = \sum_{i=1}^{100} X_i,$$

the total number of `Heads` obtained in $n = 100$ spins. Under the mathematical model that we have proposed, $Y \sim \text{Binomial}(p)$. In performing her

experiment, Robin observes a sample $\vec{x} = \{x_1, \ldots, x_{100}\}$ and computes

$$y = \sum_{i=1}^{100} x_i,$$

the total number of `Heads` in her sample. In our narrative, $y = 32$.

We emphasize that $p \in [0, 1]$ is fixed but unknown. Robin's goal is to draw inferences about this fixed but unknown quantity. We consider three questions that she might ask:

1. What is the true value of p? More precisely, what is a reasonable guess as to the true value of p?
2. Is $p = 0.5$? Specifically, is the evidence that $p \neq 0.5$ so compelling that Robin can comfortably accuse Arlen of impropriety?
3. What are plausible values of p? In particular, is there a subset of $[0, 1]$ that Robin can confidently claim contains the true value of p?

The first set of questions introduces a type of inference that statisticians call *point estimation*. We have already encountered (in Chapter 7) a natural approach to point estimation, the plug-in principle. In the present case, the plug-in principle suggests estimating the theoretical probability of success, p, by computing the observed proportion of successes,

$$\hat{p} = \frac{y}{n} = \frac{32}{100} = 0.32.$$

The second set of questions introduces a type of inference that statisticians call *hypothesis testing*. Having calculated $\hat{p} = 0.32 \neq 0.5$, Robin is inclined to guess that $p \neq 0.5$. But how compelling is the evidence that $p \neq 0.5$? Let us play devil's advocate: perhaps $p = 0.5$, but chance produced "only" $y = 32$ instead of a value nearer $EY = np = 100 \times 0.5 = 50$. This is a possibility that we can quantify. If $Y \sim \text{Binomial}(n = 100; p = 0.5)$, then the probability that Y will deviate from its expected value by at least $|50 - 32| = 18$ is

$$\begin{aligned}
\mathbf{p} &= P\left(|Y - 50| \geq 18\right) \\
&= P(Y \leq 32 \text{ or } Y \geq 68) \\
&= P(Y \leq 32) + P(Y \geq 68) \\
&= P(Y \leq 32) + 1 - P(Y \leq 67) \\
&= \texttt{pbinom(32,100,.5)+1-pbinom(67,100,.5)} \\
&= 0.0004087772.
\end{aligned}$$

This *significance probability* seems fairly small—perhaps small enough to convince Robin that in fact $p \neq 0.5$.

The third set of questions introduces a type of inference that statisticians call *set estimation*. We have just tested the possibility that $p = p_0$ in the special case $p_0 = 0.5$. Now, imagine testing the possibility that $p = p_0$ for each $p_0 \in [0, 1]$. Those p_0 that are not rejected as inconsistent with the observed data, $y = 32$, will constitute a set of plausible values of p.

To implement this procedure, Robin will have to adopt a standard of implausibility. Perhaps she decides to reject p_0 as implausible when the corresponding significance probability,

$$\begin{aligned} \mathbf{p} &= P\left(|Y - 100p_0| \geq |32 - 100p_0|\right) \\ &= P\left(Y - 100p_0 \geq |32 - 100p_0|\right) + P\left(Y - 100p_0 \leq -|32 - 100p_0|\right) \\ &= P\left(Y \geq 100p_0 + |32 - 100p_0|\right) + P\left(Y \leq 100p_0 - |32 - 100p_0|\right), \end{aligned}$$

satisfies $\mathbf{p} \leq 0.1$. Recalling that $Y \sim \text{Binomial}(100; p_0)$ and using the `R` function `pbinom`, some trial and error reveals that $\mathbf{p} > 0.1$ if p_0 lies in the interval $[0.245, 0.404]$. (The endpoints of this interval are included.) Notice that this interval does *not* contain $p_0 = 0.5$, which we had already rejected as implausible.

9.2 Point Estimation

The goal of point estimation is to make a reasonable guess of the unknown value of a designated population quantity, e.g., the population mean. The quantity that we hope to guess is called the *estimand*.

9.2.1 Estimating a Population Mean

Suppose that the estimand is μ, the population mean. The plug-in principle suggests estimating μ by computing the mean of the empirical distribution. This leads to the plug-in estimate of μ, $\hat{\mu} = \bar{x}_n$. Thus, we estimate the mean of the population by computing the mean of the sample, which is certainly a natural thing to do.

We will distinguish between

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

a real number that is calculated from the sample $\vec{x} = \{x_1, \ldots, x_n\}$, and

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

a random variable that is a function of the random variables $X_1, \ldots, X_n$. (Such a random variable is called a *statistic*.) The latter is our rule for guessing, an *estimation procedure* or *estimator*. The former is the guess itself, the result of applying our rule for guessing to the sample that we observed, an *estimate*.

The quality of an individual estimate depends on the individual sample from which it was computed and is therefore affected by chance variation. Furthermore, it is rarely possible to assess how close to correct an individual estimate may be. For these reasons, we study estimation procedures and identify the statistical properties that these random variables possess. In the present case, two properties are worth noting:

1. We know that $E\bar{X}_n = \mu$. Thus, on the average, our procedure for guessing the population mean produces the correct value. We express this property by saying that $\bar{X}_n$ is an *unbiased* estimator of μ.

   The property of unbiasedness is intuitively appealing and sometimes is quite useful. However, many excellent estimation procedures are biased and some unbiased estimators are unattractive. For example, $EX_1 = \mu$ by definition, so X_1 is also an unbiased estimator of μ; but most researchers would find the prospect of estimating a population mean with a single observation to be rather unappetizing. Indeed,

   $$\operatorname{Var} \bar{X}_n = \frac{\sigma^2}{n} < \sigma^2 = \operatorname{Var} X_1,$$

   so the unbiased estimator $\bar{X}_n$ has smaller variance than the unbiased estimator X_1.

2. The Weak Law of Large Numbers states that $\bar{X}_n \xrightarrow{P} \mu$. Thus, as the sample size increases, the estimator $\bar{X}_n$ converges in probability to the estimand μ. We express this property by saying that $\bar{X}_n$ is a *consistent* estimator of μ.

   The property of consistency is essential—it is difficult to conceive a circumstance in which one would be willing to use an estimation procedure that might fail regardless of how much data one collected. Notice that the unbiased estimator X_1 is not consistent.

9.2.2 Estimating a Population Variance

Now suppose that the estimand is σ^2, the population variance. Although we are concerned with drawing inferences about the population mean, we will discover that hypothesis testing and set estimation may require knowing the population variance. If the population variance is not known, then it must be estimated from the sample.

The plug-in principle suggests estimating σ^2 by computing the variance of the empirical distribution. This leads to the plug-in estimate of σ^2,

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 .$$

The plug-in estimator of σ^2 is *biased*; in fact,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{n-1}{n}\sigma^2 < \sigma^2.$$

This does not present any particular difficulties; however, if we desire an unbiased estimator, then we simply multiply the plug-in estimator by the factor $n/(n-1)$, obtaining

$$S_n^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 . \qquad (9.1)$$

The statistic S_n^2 is the most popular estimator of σ^2 and many books refer to the estimate

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

as *the* sample variance. (For example, the `R` command `var` computes s_n^2.) In fact, both estimators are perfectly reasonable, consistent estimators of σ^2. We will prefer S_n^2 for the rather mundane reason that using it will simplify some of the formulas that we will encounter.

9.3 Heuristics of Hypothesis Testing

Hypothesis testing is appropriate for situations in which one wants to guess which of two possible statements about a population is correct. For example, in Section 9.1 we considered the possibility that spinning a penny is fair ($p = 0.5$) versus the possibility that spinning a penny is not fair ($p \neq 0.5$). The logic of hypothesis testing is of a familar sort:

> *If an alleged coincidence seems too implausible, then we tend to believe that it wasn't really a coincidence.*

Man has engaged in this kind of reasoning for millenia. In Cicero's *De Divinatione*, Quintus exclaims:

> "They are entirely fortuitous you say? Come! Come! Do you really mean that? ... When the four dice [astragali] produce the venus-throw you may talk of accident: but suppose you made a hundred casts and the venus-throw appeared a hundred times; could you call that accidental?"[1]

The essence of hypothesis testing is captured by the familiar saying, "Where there's smoke, there's fire." In this section we formalize such reasoning, appealing to three prototypical examples:

1. Assessing circumstantial evidence in a criminal trial.

   For simplicity, suppose that the defendant has been charged with a single count of pre-meditated murder and that the jury has been instructed to either convict of murder in the first degree or acquit. The defendant had motive, means, and opportunity. Furthermore, two types of blood were found at the crime scene. One type was evidently the victim's. Laboratory tests demonstrated that the other type was not the victim's, but failed to demonstrate that it was not the defendant's. What should the jury do?

   The evidence used by the prosecution to try to establish a connection between the blood of the defendant and blood found at the crime scene is probabilistic, i.e., circumstantial. It will likely be presented to the jury in the language of mathematics, e.g., "Both blood samples have characteristics x, y and z; yet only 0.5% of the population has such blood." The defense will argue that this is merely an unfortunate coincidence. The jury must evaluate the evidence and decide whether or not such a coincidence is too extraordinary to be believed, i.e., they must decide if their assent to the proposition that the defendant committed the murder rises to a level of certainty sufficient to convict.

[1] Cicero rejected the conclusion that a run of one hundred venus-throws is so improbable that it must have been caused by divine intervention; however, Cicero was castigating the practice of divination. Quintus was entirely correct in suggesting that a run of one hundred venus-throws should not be rationalized as "entirely fortuitous." A modern scientist might conclude that an unusual set of astragali had been used to produce this remarkable result.

If the combined weight of the evidence against the defendant is a chance of one in ten, then the jury is likely to acquit; if it is a chance of one in a million, then the jury is likely to convict.

2. Assessing data from a scientific experiment.

   A study[2] of termite foraging behavior reached the controversial conclusion that two species of termites compete for scarce food resources. In this study, a site in the Sonoran desert was cleared of dead wood and toilet paper rolls were set out as food sources. The rolls were examined regularly over a period of many weeks and it was observed that only very rarely was a roll infested with both species of termites. Was this just a coincidence or were the two species competing for food?

   The scientists constructed a mathematical model of termite foraging behavior under the assumption that the two species forage independently of each other. This model was then used to quantify the probability that infestation patterns such as the one observed arise due to chance. This probability turned out to be just one in many billions—a coincidence far too extraordinary to be dismissed as such—and the researchers concluded that the two species were competing.

3. Assessing the results of Robin's penny-spinning experiment.

   In Section 9.1, we noted that Robin observed only $y = 32$ `Heads` when she would expect $EY = 50$ `Heads` if indeed $p = 0.5$. This is a discrepancy of $|32 - 50| = 18$, and we considered that possibility that such a large discrepancy might have been produced by chance. More precisely, we calculated $\mathbf{p} = P(|Y - EY| \geq 18)$ under the assumption that $p = 0.5$, obtaining $\mathbf{p} \doteq 0.0004$. On this basis, we speculated that Robin might be persuaded to accuse her brother of cheating.

In each of the preceding examples, a binary decision was based on a level of assent to probabilistic evidence. At least conceptually, this level can be quantified as a *significance probability*, which we loosely interpret to mean the probability that chance would produce a coincidence at least as extraordinary as the phenomenon observed. This begs an obvious question, which we pose now for subsequent consideration: how small should a significance probability be for one to conclude that a phenomenon is not a coincidence?

[2] S.C. Jones and M.W. Trosset (1991). Interference competition in desert subterranean termites. *Entomologia Experimentalis et Applicata*, 61:83–90.

We now proceed to explicate a formal model for statistical hypothesis testing that was proposed by J. Neyman and E. S. Pearson in the late 1920s and 1930s. Our presentation relies heavily on drawing simple analogies to criminal law, which we suppose is a more familiar topic than statistics to most students.

The States of Nature

The states of nature are the possible mechanisms that might have produced the observed phenomenon. Mathematically, they are the possible probability distributions under consideration. Thus, in the penny-spinning example, the states of nature are the Bernoulli trials indexed by $p \in [0, 1]$. In hypothesis testing, the states of nature are partitioned into two sets or *hypotheses*. In the penny-spinning example, the hypotheses that we formulated were $p = 0.5$ (penny-spinning is fair) and $p \neq 0.5$ (penny-spinning is not fair); in the legal example, the hypotheses are that the defendant did commit the murder (the defendant is factually guilty) and that the defendant did not commit the murder (the defendant is factually innocent).

The goal of hypothesis testing is to decide which hypothesis is correct, i.e., which hypothesis contains the true state of nature. In the penny-spinning example, Robin wants to determine whether or not penny-spinning is fair. In the termite example, Jones and Trosset wanted to determine whether or not termites were foraging independently. More generally, scientists usually partition the states of nature into a hypothesis that corresponds to a theory that the experiment is designed to investigate and a hypothesis that corresponds to a chance explanation; the goal of hypothesis testing is to decide which explanation is correct. In a criminal trial, the jury would like to determine whether the defendant is factually innocent of factually guilty—in the words of the United States Supreme Court in *Bullington v. Missouri* (1981):

> Underlying the question of guilt or innocence is an objective truth: the defendant did or did not commit the crime. From the time an accused is first suspected to the time the decision on guilt or innocence is made, our system is designed to enable the trier of fact to discover that truth.

Formulating appropriate hypotheses can be a delicate business. In the penny-spinning example, we formulated hypotheses $p = 0.5$ and $p \neq 0.5$. These hypotheses are appropriate if Robin wants to determine whether or

not penny-spinning is fair. However, one can easily imagine that Robin is not interested in whether or not penny-spinning is fair, but rather in whether or not her brother gained an advantage by using the procedure. If so, then appropriate hypotheses would be $p < 0.5$ (penny-spinning favored Arlen) and $p \geq 0.5$ (penny-spinning did not favor Arlen).

The Actor

The states of nature having been partitioned into two hypotheses, it is necessary for a decisionmaker (the actor) to choose between them. In the penny-spinning example, the actor is Robin; in the termite example, the actor is the team of researchers; in the legal example, the actor is the jury.

Statisticians often describe hypothesis testing as a game that they play against Nature. To study this game in greater detail, it becomes necessary to distinguish between the two hypotheses under consideration. In each example, we declare one hypothesis to be the *null hypothesis* (H_0) and the other to be the *alternative hypothesis* (H_1). Roughly speaking, the logic for determining which hypothesis is H_0 and which is H_1 is the following: H_0 should be the hypothesis to which one defaults if the evidence is equivocal and H_1 should be the hypothesis that one requires compelling evidence to embrace.

We shall have a great deal more to say about distinguishing null and alternative hypotheses, but for now suppose that we have declared the following: (1) H_0: the defendant did not commit the murder, (2) H_0: the termites are foraging independently, and (3) H_0: spinning the penny is fair. Having done so, the game takes the following form:

		State of Nature	
		H_0	H_1
Actor's Choice	H_0		Type II error
	H_1	Type I error	

There are four possible outcomes to this game, two of which are favorable and two of which are unfavorable. If the actor chooses H_1 when in fact H_0 is true, then we say that a Type I error has been committed. If the actor chooses H_0 when in fact H_1 is true, then we say that a Type II error has been committed. In a criminal trial, a Type I error occurs when a jury convicts a factually innocent defendant and a Type II error occurs when a jury acquits a factually guilty defendant.

Innocent Until Proven Guilty

Because we are concerned with probabilistic evidence, any decision procedure that we devise will occasionally result in error. Obviously, we would like to devise procedures that minimize the probabilities of committing errors. Unfortunately, there is an inevitable tradeoff between Type I and Type II error that precludes simultaneously minimizing the probabilities of both types. To appreciate this, consider two juries. The first jury always acquits and the second jury always convicts. Then the first jury *never* commits a Type I error and the second jury *never* commits a Type II error. The only way to simultaneously better both juries is to never commit an error of either type, which is impossible with probabilistic evidence.

The distinguishing feature of hypothesis testing (and Anglo-American criminal law) is the manner in which it addresses the tradeoff between Type I and Type II error. The Neyman-Pearson formulation of hypothesis testing accords the null hypothesis a privileged status: H_0 will be maintained unless there is compelling evidence against it. It is instructive to contrast the asymmetry of this formulation with situations in which neither hypothesis is privileged. In statistics, this is the problem of determining which hypothesis better explains the data. This is *discrimination*, not hypothesis testing. In law, this is the problem of determining whether the defendant or the plaintiff has the stronger case. This is the criterion in civil suits, not in criminal trials.

In the penny-spinning example, Robin required compelling evidence against the privileged null hypothesis that penny-spinning is fair to overcome her scruples about accusing her brother of impropriety. In the termite example, Jones and Trosset required compelling evidence against the privileged null hypothesis that two termite species forage independently in order to write a credible article claiming that two species were competing with each other. In a criminal trial, the principle of according the null hypothesis a privileged status has a familiar characterization: the defendant is "innocent until proven guilty."

According the null hypothesis a privileged status is equivalent to declaring Type I errors to be more egregious than Type II errors. This connection was eloquently articulated by Justice John Harlan in a 1970 Supreme Court decision: "If, for example, the standard of proof for a criminal trial were a preponderance of the evidence rather than proof beyond a reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but a far greater risk of factual errors that result in convicting the innocent."

A preference for Type II errors instead of Type I errors can often be glimpsed in scientific applications. For example, because science is conservative, it is generally considered better to wrongly accept than to wrongly reject the prevailing wisdom that termite species forage independently. Moreover, just as this preference is the foundation of statistical hypothesis testing, so is it a fundamental principle of criminal law. In his famous *Commentaries*, William Blackstone opined that "it is better that ten guilty persons escape, than that one innocent man suffer;" and in his influential *Practical Treatise on the Law of Evidence* (1824), Thomas Starkie suggested that "The maxim of the law... is that it is better that ninety-nine... offenders shall escape than that one innocent man be condemned." In *Reasonable Doubts* (1996), Alan Dershowitz quotes both maxims and notes anecdotal evidence that jurors actually do prefer committing Type II to Type I errors: on *Prime Time Live* (October 4, 1995), O.J. Simpson juror Anise Aschenbach stated, "If we made a mistake, I would rather it be a mistake on the side of a person's innocence than the other way."

Beyond a Reasonable Doubt

To actualize its antipathy to Type I errors, the Neyman-Pearson formulation imposes an upper bound on the maximal probability of Type I error that will be tolerated. This bound is the *significance level*, conventionally denoted α. The significance level is specified (prior to examining the data) and only decision rules for which the probability of Type I error is no greater than α are considered. Such tests are called *level α tests*.

To fix ideas, we consider the penny-spinning example and specify a significance level of α. Let $\mathbf{p}$ denote the significance probability that results from performing the analysis in Section 9.1 and consider a rule that rejects the null hypothesis $H_0 : p = 0.5$ if and only if $\mathbf{p} \leq \alpha$. Then a Type I error occurs if and only if $p = 0.5$ and we observe y such that $\mathbf{p} = P(|Y-50| \geq |y-50|) \leq \alpha$. We claim that the probability of observing such a y is just α, in which case we have constructed a level α test.

To see why this is the case, let $W = |Y-50|$ denote the *test statistic*. The decision to accept or reject the null hypothesis H_0 depends on the observed value, w, of this random variable. Let

$$\mathbf{p}(w) = P_{H_0}(W \geq w)$$

denote the significance probability associated with w. Notice that w is the $1 - \mathbf{p}(w)$ quantile of the random variable W under H_0. Let q denote the

$1 - \alpha$ quantile of W under H_0, i.e.,

$$\alpha = P_{H_0}(W \geq q).$$

We reject H_0 if and only if we observe

$$P_{H_0}(W \geq w) = \mathbf{p}(w) \leq \alpha = P_{H_0}(W \geq q),$$

i.e., if and only $w \geq q$. If H_0 is true, then the probability of committing a Type I error is precisely

$$P_{H_0}(W \geq q) = \alpha,$$

as claimed above. We conclude that α quantifies the level of assent that we require to risk rejecting H_0, i.e., the significance level specifies how small a significance probability is required in order to conclude that a phenomenon is not a coincidence.

In statistics, the significance level α is a number in the interval $[0, 1]$. It is not possible to quantitatively specify the level of assent required for a jury to risk convicting an innocent defendant, but the legal principle is identical: in a criminal trial, the operative significance level is *beyond a reasonable doubt*. Starkie (1824) described the possible interpretations of this phrase in language derived from British empirical philosopher John Locke:

> Evidence which satisfied the minds of the jury of the truth of the fact in dispute, to the entire exclusion of every reasonable doubt, constitute full proof of the fact. ... Even the most direct evidence can produce nothing more than such a high degree of probability as amounts to moral certainty. From the highest it may decline, by an infinite number of gradations, until it produces in the mind nothing more than a preponderance of assent in favour of the particular fact.

The gradations that Starkie described are not intrinsically numeric, but it is evident that the problem of defining reasonable doubt in criminal law is the problem of specifying a significance level in statistical hypothesis testing.

In both criminal law and statistical hypothesis testing, actions typically are described in language that acknowledges the privileged status of the null hypothesis and emphasizes that the decision criterion is based on the probability of committing a Type I error. In describing the action of choosing H_0, many statisticians prefer the phrase "fail to reject the null hypothesis" to the less awkward "accept the null hypothesis" because choosing H_0 does

not imply an affirmation that H_0 is correct, only that the level of evidence against H_0 is not sufficiently compelling to warrant its rejection at significance level α. In precise analogy, juries render verdicts of "not guilty" rather than "innocent" because acquital does not imply an affirmation that the defendant did not commit the crime, only that the level of evidence against the defendant's innocence was not beyond a reasonable doubt.[3]

And To a Moral Certainty

The Neyman-Pearson formulation of statistical hypothesis testing is a mathematical abstraction. Part of its generality derives from its ability to accommodate *any* specified significance level. As a practical matter, however, α must be specified and we now ask how to do so.

In the penny-spinning example, Robin is making a personal decision and is free to choose α as she pleases. In the termite example, the researchers were guided by decades of scientific convention. In 1925, in his extremely influential *Statistical Methods for Research Workers*, Ronald Fisher[4] suggested that $\alpha = 0.05$ and $\alpha = 0.01$ are often appropriate significance levels. These suggestions were intended as practical guidelines, but they have become enshrined (especially $\alpha = 0.05$) in the minds of many scientists as a sort of Delphic determination of whether or not a hypothesized theory is true. While some degree of conformity is desirable (it inhibits a researcher from choosing—after the fact—a significance level that will permit rejecting the null hypothesis in favor of the alternative in which s/he may be invested), many statisticians are disturbed by the scientific community's slavish devotion to a single standard and by its often uncritical interpretation of the resulting conclusions.[5]

The imposition of an arbitrary standard like $\alpha = 0.05$ is possible because of the precision with which mathematics allows hypothesis testing to be formulated. Applying this precision to legal paradigms reveals the issues

[3] In contrast, Scottish law permits a jury to return a verdict of "not proven," thereby reserving a verdict of "not guilty" to affirm a defendant's innocence.

[4] Sir Ronald Fisher is properly regarded as the single most important figure in the history of statistics. It should be noted that he did not subscribe to all of the particulars of the Neyman-Pearson formulation of hypothesis testing. His fundamental objection to it, that it may not be possible to fully specify the alternative hypothesis, does not impact our development, since we are concerned with situations in which both hypotheses are fully specified.

[5] See, for example, J. Cohen (1994). The world is round ($p < .05$). *American Psychologist*, 49:997–1003.

with great clarity, but is of little practical value when specifying a significance level, i.e., when trying to define the meaning of "beyond a reasonable doubt." Nevertheless, legal scholars have endeavored for centuries to position "beyond a reasonable doubt" along the infinite gradations of assent that correspond to the continuum $[0, 1]$ from which α is selected. The phrase "beyond a reasonable doubt" is still often connected to the archaic phrase "to a moral certainty." This connection survived because moral certainty was actually a significance level, intended to invoke an enormous body of scholarly writings and specify a level of assent:

> Throughout this development two ideas to be conveyed to the jury have been central. The first idea is that there are two realms of human knowledge. In one it is possible to obtain the absolute certainty of mathematical demonstration, as when we say that the square of the hypotenuse is equal to the sum of the squares of the other two sides of a right triangle. In the other, which is the empirical realm of events, absolute certainty of this kind is not possible. The second idea is that, in this realm of events, just because absolute certainty is not possible, we ought not to treat everything as merely a guess or a matter of opinion. Instead, in this realm there are levels of certainty, and we reach higher levels of certainty as the quantity and quality of the evidence available to us increase. The highest level of certainty in this empirical realm in which no absolute certainty is possible is what traditionally was called "moral certainty," a certainty which there was no reason to doubt.[6]

Although it is rarely (if ever) possible to quantify a juror's level of assent, those comfortable with statistical hypothesis testing may be inclined to wonder what values of α correspond to conventional interpretations of reasonable doubt. If a juror believes that there is a 5 percent probability that chance alone could have produced the circumstantial evidence presented against a defendant accused of pre-meditated murder, is the juror's level of assent beyond a reasonable doubt and to a moral certainty? We hope not. We may be willing to tolerate a 5 percent probability of a Type I error when studying termite foraging behavior, but the analogous prospect of a 5

[6]Barbara J. Shapiro (1991). *"Beyond Reasonable Doubt" and "Probable Cause": Historical Perspectives on the Anglo-American Law of Evidence*, University of California Press, Berkeley, p. 41.

percent probability of wrongly convicting a factually innocent defendant is abhorrent.[7]

In fact, little is known about how anyone in the legal system quantifies reasonable doubt. Mary Gray cites a 1962 Swedish case in which a judge trying an overtime parking case explicitly ruled that a significance probability of 1/20736 was beyond reasonable doubt but that a significance probability of 1/144 was not.[8] In contrast, Alan Dershowitz relates a provocative classroom exercise in which his students preferred to acquit in one scenario with a significance probability of 10 percent and to convict in an analogous scenario with a significance probability of 15 percent.[9]

9.4 Testing Hypotheses About a Population Mean

We now apply the heuristic reasoning described in Section 9.3 to the problem of testing hypotheses about a population mean. Initially, we consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

The intuition that we are seeking to formalize is fairly straightfoward. By virtue of the Weak Law of Large Numbers, the observed sample mean ought to be fairly close to the true population mean. Hence, if the null hypothesis is true, then $\bar{x}_n$ ought to be fairly close to the hypothesized mean, μ_0. If we observe $\bar{X}_n = \bar{x}_n$ far from μ_0, then we guess that $\mu \neq \mu_0$, i.e., we reject H_0.

Given a significance level α, we want to calculate a significance probability $\mathbf{p}$. The significance level is a real number that is fixed by and known to the researcher, e.g., $\alpha = 0.05$. The significance probability is a real number that is determined by the sample, e.g., $\mathbf{p} \doteq 0.0004$ in Section 9.1. We will reject H_0 if and only if $\mathbf{p} \leq \alpha$.

In Section 9.3, we interpreted the significance probability as the probability that chance would produce a coincidence at least as extraordinary as the phenomenon observed. Our first challenge is to make this notion mathematically precise; how we do so depends on the hypotheses that we

[7] This discrepancy illustrates that the consequences of committing a Type I error influence the choice of a significance level. The consequences of Jones and Trosset wrongly concluding that termite species compete are not commensurate with the consequences of wrongly imprisoning a factually innocent citizen.

[8] M.W. Gray (1983). Statistics and the law. *Mathematics Magazine*, 56:67–81. As a graduate of Rice University, I cannot resist quoting another of Gray's examples of statistics-as-evidence: "In another case, that of millionaire W. M. Rice, the signature on his will was disputed, and the will was declared a forgery on the basis of probability evidence. As a result, the fortune of Rice went to found Rice Institute."

[9] A.M. Dershowitz (1996). *Reasonable Doubts*, Simon & Schuster, New York, p. 40.

want to test. In the present situation, we submit that a natural significance probability is

$$\mathbf{p} = P_{\mu_0}\left(\left|\bar{X}_n - \mu_0\right| \geq \left|\bar{x}_n - \mu_0\right|\right). \tag{9.2}$$

To understand why this is the case, it is essential to appreciate the following details:

1. The hypothesized mean, μ_0, is a real number that is fixed by and known to the researcher.

2. The estimated mean, $\bar{x}_n$, is a real number that is calculated from the observed sample and known to the researcher; hence, the quantity $|\bar{x}_n - \mu_0|$ is a fixed real number.

3. The estimator, $\bar{X}_n$, is a random variable. Hence, the inequality

$$\left|\bar{X}_n - \mu_0\right| \geq \left|\bar{x}_n - \mu_0\right| \tag{9.3}$$

   defines an event that may or may not occur each time the experiment is performed. Specifically, (9.3) is the event that the sample mean assumes a value at least as far from the hypothesized mean as the researcher observed.

4. The significance probability, $\mathbf{p}$, is the probability that (9.3) occurs. The notation P_{μ_0} reminds us that we are interested in the probability that this event occurs *under the assumption that the null hypothesis is true*, i.e., under the assumption that $\mu = \mu_0$.

Having formulated an appropriate significance probability for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, our second challenge is to find a way to compute $\mathbf{p}$. We remind the reader that we have assumed that n is large.

Case 1: The population variance is known or specified by the null hypothesis.

We define two new quantities, the random variable

$$Z_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

and the real number

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}.$$

Under the null hypothesis $H_0 : \mu = \mu_0$, $Z_n \dot{\sim} \text{Normal}(0, 1)$ by the Central Limit Theorem; hence,

$$\begin{aligned}
\mathbf{p} &= P_{\mu_0}\left(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|\right) \\
&= 1 - P_{\mu_0}\left(-|\bar{x}_n - \mu_0| < \bar{X}_n - \mu_0 < |\bar{x}_n - \mu_0|\right) \\
&= 1 - P_{\mu_0}\left(-\frac{|\bar{x}_n - \mu_0|}{\sigma/\sqrt{n}} < \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} < \frac{|\bar{x}_n - \mu_0|}{\sigma/\sqrt{n}}\right) \\
&= 1 - P_{\mu_0}\left(-|z| < Z_n < |z|\right) \\
&\doteq 1 - [\Phi(|z|) - \Phi(-|z|)] \\
&= 2\Phi(-|z|),
\end{aligned}$$

which can be computed by the `R` command

```
> 2*pnorm(-abs(z))
```

or by consulting a table. An illustration of the normal probability of interest is sketched in Figure 9.1.

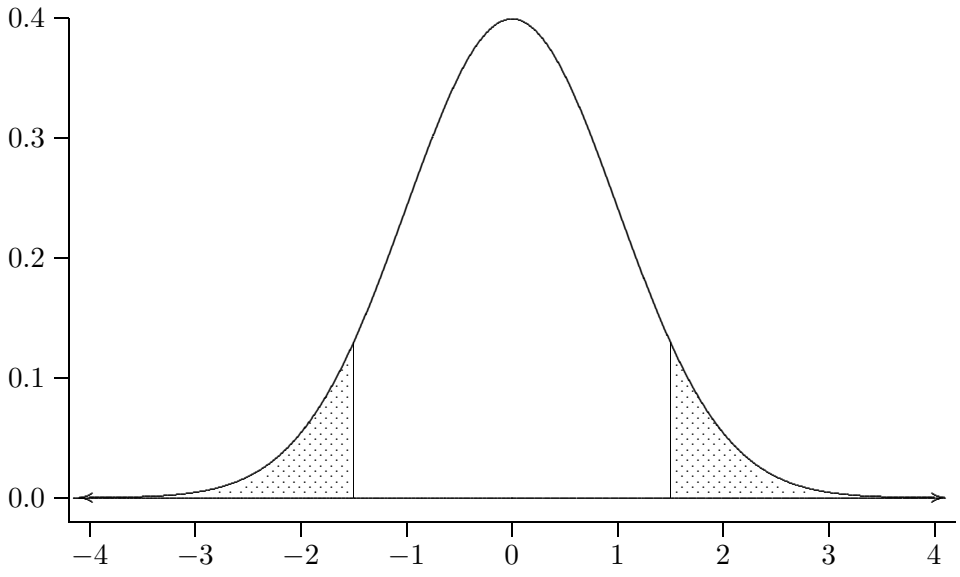

Figure 9.1: $P(|Z| \geq |z| = 1.5)$

An important example of Case 1 occurs when $X_i \sim \text{Bernoulli}(\mu)$. In this case, $\sigma^2 = \text{Var}\, X_i = \mu(1 - \mu)$; hence, under the null hypothesis that $\mu = \mu_0$,

$\sigma^2 = \mu_0(1-\mu_0)$ and

$$z = \frac{\bar{x}_n - \mu_0}{\sqrt{\mu_0(1-\mu_0)/n}}.$$

Example 9.1 *To test $H_0 : \mu = 0.5$ versus $H_1 : \mu \neq 0.5$ at significance level $\alpha = 0.05$, we perform $n = 2500$ trials and observe 1200 successes. Should H_0 be rejected?*

The observed proportion of successes is $\bar{x}_n = 1200/2500 = 0.48$, so the value of the test statistic is

$$z = \frac{0.48 - 0.50}{\sqrt{0.5(1-0.5)/2500}} = \frac{-0.02}{0.5/50} = -2$$

and the significance probability is

$$\mathbf{p} \doteq 2\Phi(-2) \doteq 0.0456 < 0.05 = \alpha.$$

Because $\mathbf{p} \leq \alpha$, we reject H_0.

Case 2: The population variance is unknown.

Because σ^2 is unknown, we must estimate it from the sample. We will use the estimator introduced in Section 9.2,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2,$$

and define

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

Because S_n^2 is a consistent estimator of σ^2, i.e., $S_n^2 \xrightarrow{P} \sigma^2$, it follows from Theorem 8.3 that

$$\lim_{n\to\infty} P\left(T_n \leq z\right) = \Phi(z).$$

Just as we could use a normal approximation to compute probabilities involving Z_n, so can we use a normal approximation to compute probabilities involving T_n. The fact that we must estimate σ^2 slightly degrades the quality of the approximation; however, because n is large, we should observe an accurate estimate of σ^2 and the approximation should not suffer much. Accordingly, we proceed as in Case 1, using

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$$

instead of z.

Example 9.2 *To test $H_0 : \mu = 20$ versus $H_1 : \mu \neq 20$ at significance level $\alpha = 0.05$, we collect $n = 400$ observations, observing $\bar{x}_n = 21.82935$ and $s_n = 24.70037$. Should H_0 be rejected?*

The value of the test statistic is

$$t = \frac{21.82935 - 20}{24.70037/20} = 1.481234$$

and the significance probability is

$$\mathbf{p} \doteq 2\Phi(-1.481234) = 0.1385441 > 0.05 = \alpha.$$

Because $\mathbf{p} > \alpha$, we decline to reject H_0.

9.4.1 One-Sided Hypotheses

In Section 9.3 we suggested that, if Robin is not interested in whether or not penny-spinning is fair but rather in whether or not it favors her brother, then appropriate hypotheses would be $p < 0.5$ (penny-spinning favors Arlen) and $p \geq 0.5$ (penny-spinning does not favor Arlen). These are examples of one-sided (as opposed to two-sided) hypotheses.

More generally, we will consider two canonical cases:

$$\begin{array}{lll} H_0 : \mu \leq \mu_0 & \text{versus} & H_1 : \mu > \mu_0 \\ H_0 : \mu \geq \mu_0 & \text{versus} & H_1 : \mu < \mu_0 \end{array}$$

Notice that the possibility of equality, $\mu = \mu_0$, belongs to the null hypothesis in both cases. This is a technical necessity that arises because we compute significance probabilities using the μ in H_0 that is nearest H_1. For such a μ to exist, the boundary between H_0 and H_1 must belong to H_0. We will return to this necessity later in this section.

Instead of memorizing different formulas for different situations, we will endeavor to understand which values of our test statistic tend to undermine the null hypothesis in question. Such reasoning can be used on a case-by-case basis to determine the relevant significance probability. In so doing, sketching crude pictures can be quite helpful!

Consider testing each of the following:

$$\begin{array}{lllll} \text{(a)} & H_0 : \mu = \mu_0 & \text{versus} & H_1 : \mu \neq \mu_0 \\ \text{(b)} & H_0 : \mu \leq \mu_0 & \text{versus} & H_1 : \mu > \mu_0 \\ \text{(c)} & H_0 : \mu \geq \mu_0 & \text{versus} & H_1 : \mu < \mu_0 \end{array}$$

Qualitatively, we will be inclined to reject the null hypothesis if

(a) We observe $\bar{x}_n \ll \mu_0$ or $\bar{x}_n \gg \mu_0$, i.e., if we observe $|\bar{x}_n - \mu_0| \gg 0$.

This is equivalent to observing $|t| \gg 0$, so the significance probability is

$$\mathbf{p}_a = P_{\mu_0}\left(|T_n| \geq |t|\right).$$

(b) We observe $\bar{x}_n \gg \mu_0$, i.e., if we observe $\bar{x}_n - \mu_0 \gg 0$.

This is equivalent to observing $t \gg 0$, so the significance probability is

$$\mathbf{p}_b = P_{\mu_0}\left(T_n \geq t\right).$$

(c) We observe $\bar{x}_n \ll \mu_0$, i.e., if we observe $\bar{x}_n - \mu_0 \ll 0$.

This is equivalent to observing $t \ll 0$, so the significance probability is

$$\mathbf{p}_c = P_{\mu_0}\left(T_n \leq t\right).$$

Example 9.2 (continued) Applying the above reasoning, we obtain the significance probabilities sketched in Figure 9.2. Notice that $\mathbf{p}_b = \mathbf{p}_a/2$ and that $\mathbf{p}_b + \mathbf{p}_c = 1$. The probability $\mathbf{p}_b$ is fairly small, about 7%. This makes sense: we observed $\bar{x}_n \doteq 21.8 > 20 = \mu_0$, so the sample does contain *some* evidence that $\mu > 20$. However, the statistical test reveals that the strength of this evidence is not sufficiently compelling to reject $H_0 : \mu \leq 20$.

In contrast, the probability of $\mathbf{p}_c$ is quite large, about 93%. This also makes sense, because the sample contains *no* evidence that $\mu < 20$. In such instances, performing a statistical test only confirms that which is transparent from comparing the sample and hypothesized means.

9.4.2 Formulating Suitable Hypotheses

Examples 9.1 and 9.2 illustrated the mechanics of hypothesis testing. Once understood, the above techniques for calculating significance probabilities are fairly straightforward and can be applied routinely to a wide variety of problems. In contrast, determining suitable hypotheses to be tested requires one to carefully consider each situation presented. These determinations cannot be reduced to formulas. To make them requires good judgment, which can only be acquired through practice.

We now consider some examples that illustrate some important issues that arise when formulating hypotheses. In each case, there are certain key questions that must be answered: *Why was the experiment performed? Who needs to be convinced of what? Is one type of error perceived as more important than the other?*

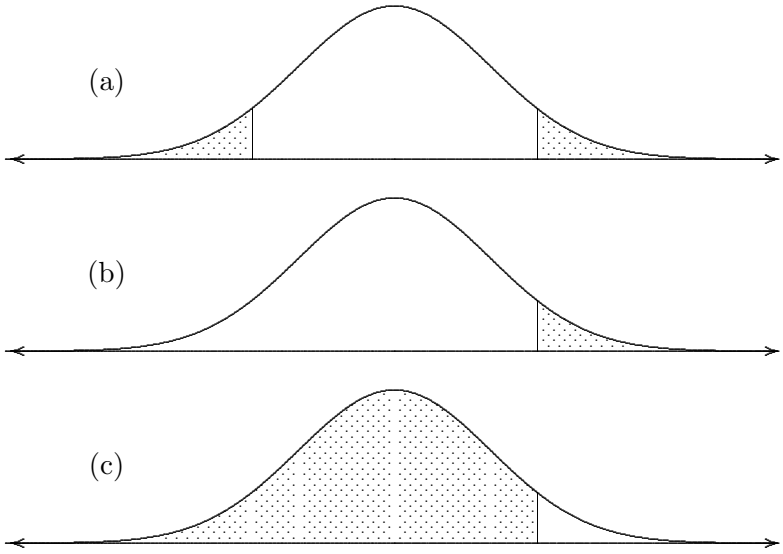

Figure 9.2: Significance probabilities for Example 9.2. Each significance probability is the area of the corresponding shaded region.

Example 9.3 *A group of concerned parents wants speed humps installed in front of a local elementary school, but the city traffic office is reluctant to allocate funds for this purpose. Both parties agree that humps should be installed if the average speed of all motorists who pass the school while it is in session exceeds the posted speed limit of* 15 *miles per hour (mph). Let* μ *denote the average speed of the motorists in question. A random sample of* $n = 150$ *of these motorists was observed to have a sample mean of* $\bar{x} = 15.3$ *mph with a sample standard deviation of* $s = 2.5$ *mph.*

(a) State null and alternative hypotheses that are appropriate from the parents' perspective.

(b) State null and alternative hypotheses that are appropriate from the city traffic office's perspective.

(c) Compute the value of an appropriate test statistic.

(d) Adopting the parents' perspective and assuming that they are willing to risk a 1% *chance of committing a Type I error, what action should be taken? Why?*

(e) Adopting the city traffic office's perspective and assuming that they are willing to risk a 10% chance of committing a Type I error, what action should be taken? Why?

Solution

(a) The parents would prefer to err on the side of protecting their children, so they would rather build unnecessary speed humps than forego necessary speed humps. Hence, they would like to see the hypotheses formulated so that foregoing necessary speed humps is a Type I error. Since speed humps will be built if it is concluded that $\mu > 15$ and will not be built if it is concluded that $\mu < 15$, the parents would prefer a null hypothesis of $H_0 : \mu \geq 15$ and an alternative hypothesis of $H_1 : \mu < 15$.

Equivalently, if we suppose that the purpose of the experiment is to provide evidence to the parents, then it is clear that the parents need to be persuaded that speed humps are unnecessary. The null hypothesis to which they will default in the absence of compelling evidence is $H_0 : \mu \geq 15$. They will require compelling evidence to the contrary, $H_1 : \mu < 15$.

(b) The city traffic office would prefer to err on the side of conserving their budget for important public works, so they would rather forego necessary speed humps than build unnecessary speed humps. Hence, they would like to see the hypotheses formulated so that building unnecessary speed humps is a Type I error. Since speed humps will be built if it is concluded that $\mu > 15$ and will not be built if it is concluded that $\mu < 15$, the city traffic office would prefer a null hypothesis of $H_0 : \mu \leq 15$ and an alternative hypothesis of $H_1 : \mu > 15$.

Equivalently, if we suppose that the purpose of the experiment is to provide evidence to the city traffic, then it is clear that the office needs to be persuaded that speed humps are necessary. The null hypothesis to which it will default in the absence of compelling evidence is $H_0 : \mu \leq 15$. It will require compelling evidence to the contrary, $H_1 : \mu > 15$.

(c) Because the population variance is unknown, the appropriate test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{15.3 - 15}{2.5/\sqrt{150}} \doteq 1.47.$$

(d) We would reject the null hypothesis in (a) if $\bar{x}$ is sufficiently smaller than $\mu_0 = 15$. Since $\bar{x} = 15.3 > 15$, there is no evidence against $H_0 : \mu \geq 15$. The null hypothesis is retained and speed humps are installed.

(e) We would reject the null hypothesis in (b) if $\bar{x}$ is sufficiently larger than $\mu_0 = 15$, i.e., for sufficiently large positive values of t. Hence, the significance probability is

$$\mathbf{p} = P\left(T_n \geq t\right) \doteq P(Z \geq 1.47) = 1 - \Phi(1.47) \doteq 0.071 < 0.10 = \alpha.$$

Because $\mathbf{p} \leq \alpha$, the traffic office should reject $H_0 : \mu \leq 15$ and install speed humps.

Example 9.4 *Imagine a variant of the Lanarkshire milk experiment described in Section 1.2. Suppose that it is known that 10-year-old Scottish schoolchildren gain an average of* 0.5 *pounds per month. To study the effect of daily milk supplements, a random sample of* $n = 1000$ *such children is drawn. Each child receives a daily supplement of 3/4 cups pasteurized milk. The study continues for four months and the weight gained by each student during the study period is recorded. Formulate suitable null and alternative hypotheses for testing the effect of daily milk supplements.*

Solution Let $X_1, \ldots, X_n$ denote the weight gains and let $\mu = EX_i$. Then milk supplements are effective if $\mu > 2$ and ineffective if $\mu < 2$. One of these possibilities will be declared the null hypothesis, the other will be declared the alternative hypothesis. The possibility $\mu = 2$ will be incorporated into the null hypothesis.

The alternative hypothesis should be the one for which compelling evidence is desired. Who needs to be convinced of what? The parents and teachers already believe that daily milk supplements are beneficial and would have to be convinced otherwise. But this is not the purpose of the study! The study is performed for the purpose of obtaining objective scientific evidence that supports prevailing popular wisdom. It is performed to convince government bureaucrats that spending money on daily milk supplements for schoolchildren will actually have a beneficial effect. The parents and teachers hope that the study will provide compelling evidence of this effect. Thus, the appropriate alternative hypothesis is $H_1 : \mu > 2$ and the appropriate null hypothesis is $H_0 : \mu \leq 2$.

9.4.3 Statistical Significance and Material Significance

The significance probability is the probability that a coincidence at least as extraordinary as the phenomenon observed can be produced by chance. The smaller the significance probability, the more confidently we reject the null hypothesis. However, it is one thing to be convinced that the null hypothesis is incorrect—it is something else to assert that the true state of nature is very different from the state(s) specified by the null hypothesis.

Example 9.5 A government agency requires prospective advertisers to provide statistical evidence that documents their claims. In order to claim that a gasoline additive increases mileage, an advertiser must fund an independent study in which n vehicles are tested to see how far they can drive, first without and then with the additive. Let X_i denote the increase in miles per gallon (mpg with the additive minus mpg without the additive) observed for vehicle i and let $\mu = EX_i$. The null hypothesis $H_0 : \mu \leq 1$ is tested against the alternative hypothesis $H_1 : \mu > 1$ and advertising is authorized if H_0 is rejected at a significance level of $\alpha = 0.05$.

Consider the experiences of two prospective advertisers:

1. A large corporation manufactures an additive that increases mileage by an average of $\mu = 1.01$ miles per gallon. The corporation funds a large study of $n = 900$ vehicles in which $\bar{x} = 1.01$ and $s = 0.1$ are observed. This results in a test statistic of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.01 - 1.00}{0.1/\sqrt{900}} = 3$$

and a significance probability of

$$\mathbf{p} = P\left(T_n \geq t\right) \doteq P(Z \geq 3) = 1 - \Phi(3) \doteq 0.00135 < 0.05 = \alpha.$$

The null hypothesis is decisively rejected and advertising is authorized.

2. An amateur automotive mechanic invents an additive that increases mileage by an average of $\mu = 1.21$ miles per gallon. The mechanic funds a small study of $n = 9$ vehicles in which $\bar{x} = 1.21$ and $s = 0.4$ are observed. This results in a test statistic of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.21 - 1.00}{0.4/\sqrt{9}} = 1.575$$

and (assuming that the normal approximation remains valid) a significance probability of

$$\mathbf{p} = P\left(T_n \geq t\right) \doteq P(Z \geq 1.575) = 1 - \Phi(1.575) \doteq 0.05763 > 0.05 = \alpha.$$

The null hypothesis is not rejected and advertising is not authorized.

These experiences are highly illuminating. Although the corporation's mean increase of $\mu = 1.01$ mpg is much closer to the null hypothesis than the mechanic's mean increase of $\mu = 1.21$ mpg, the corporation's study resulted in a much smaller significance probability. This occurred because of the smaller standard deviation and larger sample size in the corporation's study. As a result, the government could be more confident that the corporation's product had a mean increase of more than 1.0 mpg than they could be that the mechanic's product had a mean increase of more than 1.0 mpg.

The preceding example illustrates that a small significance probability does not imply a large physical effect and that a large physical effect does not imply a small significance probability. To avoid confusing these two concepts, statisticians distinguish between statistical significance and *material significance* (importance). To properly interpret the results of hypothesis testing, it is essential that one remember:

Statistical significance is not the same as material significance.

9.5 Set Estimation

Hypothesis testing is concerned with situations that demand a binary decision, e.g., whether or not to install speed humps in front of an elementary school. The relevance of hypothesis testing in situations that do not demand a binary decision is somewhat less clear. For example, many statisticians feel that the scientific community overuses hypothesis testing and that other types of statistical inference are often more appropriate. As we have discussed, a typical application of hypothesis testing in science partitions the states of nature into two sets, one that corresponds to a theory and one that corresponds to chance. Usually the theory encompasses a great many possible states of nature and the mere conclusion that the theory is true only begs the question of which states of nature are actually plausible. Furthermore, it is a rather fanciful conceit to imagine that a single scientific article should attempt to decide whether a theory is or is not true. A more

sensible enterprise for the authors to undertake is simply to set forth the evidence that they have discovered and allow evidence to accumulate until the scientific community reaches a consensus. One way to accomplish this is for each article to identify what its authors consider a set of plausible values for the population quantity in question.

To construct a set of plausible values of μ, we imagine testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ for *every* $\mu_0 \in (-\infty, \infty)$ and eliminating those μ_0 for which $H_0 : \mu = \mu_0$ is rejected. To see where this leads, let us examine our decision criterion in the case that σ is known: we reject $H_0 : \mu = \mu_0$ if and only if

$$\mathbf{p} = P_{\mu_0}\left(\left|\bar{X}_n - \mu_0\right| \geq \left|\bar{x}_n - \mu_0\right|\right) \doteq 2\Phi\left(-\left|z\right|\right) \leq \alpha, \tag{9.4}$$

where $z = (\bar{x}_n - \mu_0)/(\sigma/\sqrt{n})$. Using the symmetry of the normal distribution, we can rewrite condition (9.4) as

$$\alpha/2 \geq \Phi\left(-\left|z\right|\right) = P\left(Z < -\left|z\right|\right) = P\left(Z > \left|z\right|\right),$$

which in turn is equivalent to the condition

$$\Phi\left(\left|z\right|\right) = P\left(Z < \left|z\right|\right) = 1 - P\left(Z > \left|z\right|\right) \geq 1 - \alpha/2, \tag{9.5}$$

where $Z \sim \text{Normal}(0, 1)$.

Now let q denote the $1 - \alpha/2$ quantile of Normal$(0, 1)$, so that

$$\Phi(q) = 1 - \alpha/2.$$

Then condition (9.5) obtains if and only if $|z| \geq q$. We express this by saying that q is the *critical value* of the test statistic $|Z_n|$, where $Z_n = (\bar{X}_n - \mu_0)/(\sigma/\sqrt{n})$. For example, suppose that $\alpha = 0.05$, so that $1 - \alpha/2 = 0.975$. Then the critical value is computed in `R` as follows:

```
> qnorm(.975)
[1] 1.959964
```

Given a significance level α and the corresponding q, we have determined that q is the critical value of $|Z_n|$ for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ at significance level α. Thus, we reject $H_0 : \mu = \mu_0$ if and only if (iff)

$$\begin{aligned}
& \left|\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right| = |z| \geq q \\
\text{iff} \quad & |\bar{x}_n - \mu_0| \geq q\sigma/\sqrt{n} \\
\text{iff} \quad & \mu_0 \notin \left(\bar{x}_n - q\sigma/\sqrt{n},\ \bar{x}_n + q\sigma/\sqrt{n}\right).
\end{aligned}$$

Thus, the desired set of plausible values is the interval

$$\left(\bar{x}_n - q\frac{\sigma}{\sqrt{n}},\ \bar{x}_n + q\frac{\sigma}{\sqrt{n}}\right). \tag{9.6}$$

If σ is unknown, then the argument is identical except that we estimate σ^2 as

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2,$$

obtaining as the set of plausible values the interval

$$\left(\bar{x}_n - q\frac{s_n}{\sqrt{n}},\ \bar{x}_n + q\frac{s_n}{\sqrt{n}}\right). \tag{9.7}$$

Example 9.2 (continued) *A random sample of $n = 400$ observations is drawn from a population with unknown mean μ and unknown variance σ^2, resulting in $\bar{x}_n = 21.82935$ and $s_n = 24.70037$. Using a significance level of $\alpha = 0.05$, determine a set of plausible values of μ.*

First, because $\alpha = 0.05$ is the significance level, $q = 1.959964$ is the critical value. From (9.7), an interval of plausible values is

$$21.82935 \pm 1.959964 \cdot 24.70037/\sqrt{400} = (19.40876, 24.24994).$$

Notice that $20 \in (19.40876, 24.24994)$, meaning that (as we discovered in Section 9.4) we would accept $H_0 : \mu = 20$ at significance level $\alpha = 0.05$.

Now consider the random interval I, defined in Case 1 (population variance known) by

$$I = \left(\bar{X}_n - q\frac{\sigma}{\sqrt{n}},\ \bar{X}_n - q\frac{\sigma}{\sqrt{n}}\right)$$

and in Case 2 (population variance unknown) by

$$I = \left(\bar{X}_n - q\frac{S_n}{\sqrt{n}},\ \bar{X}_n - q\frac{S_n}{\sqrt{n}}\right).$$

The probability that this random interval covers the real number μ_0 is

$$P_\mu\left(I \supset \mu_0\right) = 1 - P_\mu\left(\mu_0 \notin I\right) = 1 - P_\mu\left(\text{reject } H_0 : \mu = \mu_0\right).$$

If $\mu = \mu_0$, then the probability of coverage is

$$1 - P_{\mu_0}\left(\text{reject } H_0 : \mu = \mu_0\right) = 1 - P_{\mu_0}(\text{Type I error}) \geq 1 - \alpha.$$

Thus, the probability that I covers the true value of the population mean is at least $1-\alpha$, which we express by saying that I is a $(1-\alpha)$-*level confidence interval* for μ. The level of confidence, $1-\alpha$, is also called the *confidence coefficient*.

We emphasize that the confidence interval I is random and the population mean μ is fixed, albeit unknown. Each time that the experiment in question is performed, a random sample is observed and an interval is constructed from it. As the sample varies, so does the interval. Any one such interval, constructed from a single sample, either does or does not contain the population mean. However, if this procedure is repeated a great many times, then the proportion of such intervals that contain μ will be at least $1-\alpha$. Actually observing one sample and constructing one interval from it amounts to randomly selecting one of the many intervals that might or might not contain μ. Because most (at least $1-\alpha$) of the intervals do, we can be "confident" that the interval that was actually constructed does contain the unknown population mean.

9.5.1 Sample Size

Confidence intervals are often used to determine sample sizes for future experiments. Typically, the researcher specifies a desired confidence level, $1-\alpha$, and a desired interval length, L. After determining the appropriate critical value, q, one equates L with $2q\sigma/\sqrt{n}$ and solves for n, obtaining

$$n = (2q\sigma/L)^2 . \tag{9.8}$$

Of course, this formula presupposes knowledge of the population variance. In practice, it is usually necessary to replace σ with an estimate—which may be easier said than done if the experiment has not yet been performed. This is one reason to perform a pilot study: to obtain a preliminary estimate of the population variance and use it to design a better study.

Several useful relations can be deduced from equation (9.8):

1. Higher levels of confidence $(1-\alpha)$ correspond to larger critical values (q), which result in larger sample sizes (n).
2. Smaller interval lengths (L) result in larger sample sizes (n).
3. Larger variances (σ^2) result in larger sample sizes (n).

In summary, if a researcher desires high confidence that the true mean of a highly variable population is covered by a small interval, then s/he should plan on collecting a great deal of data!

Example 9.5 (continued) *A rival corporation purchases the rights to the amateur mechanic's additive. How large a study is required to determine this additive's mean increase in mileage to within* 0.05 *mpg with a confidence coefficient of* $1 - \alpha = 0.99$*?*

The desired interval length is $L = 2 \cdot 0.05 = 0.1$ and the critical value that corresponds to $\alpha = 0.01$ is computed in `R` as follows:

```
> qnorm(1-.01/2)
[1] 2.575829
```

From the mechanic's small pilot study, we estimate σ to be $s = 0.4$. Then

$$n = (2 \cdot 2.575829 \cdot 0.4/0.1)^2 \doteq 424.6,$$

so the desired study will require $n = 425$ vehicles.

9.5.2 One-Sided Confidence Intervals

The set of μ_0 for which we would accept the null hypothesis $H_0 : \mu = \mu_0$ when tested against the two-sided alternative hypothesis $H_1 : \mu \neq \mu_0$ is a traditional, 2-sided confidence interval. In situations where 1-sided alternatives are appropriate, we can construct corresponding 1-sided confidence intervals by determining the set of μ_0 for which the appropriate null hypothesis would be accepted.

Example 9.5 (continued) The government test has a significance level of $\alpha = 0.05$. It rejects the null hypothesis $H_0 : \mu \leq \mu_0$ if and only if (iff)

$$\begin{aligned} & \mathbf{p} = P(Z \geq t) \leq 0.05 \\ \text{iff} \quad & P(Z < t) \geq 0.95 \\ \text{iff} \quad & t \geq \texttt{qnorm(0.95)} \doteq 1.645. \end{aligned}$$

Equivalently, the null hypothesis $H_0 : \mu \leq \mu_0$ is accepted if and only if

$$\begin{aligned} & t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < 1.645 \\ \text{iff} \quad & \bar{x} < \mu_0 + 1.645 \cdot \frac{s}{\sqrt{n}} \\ \text{iff} \quad & \mu_0 > \bar{x} - 1.645 \cdot \frac{s}{\sqrt{n}}. \end{aligned}$$

1. In the case of the large corporation, the null hypothesis $H_0 : \mu \leq \mu_0$ is accepted if and only if

$$\mu_0 > 1.01 - 1.645 \cdot \frac{0.1}{\sqrt{900}} \doteq 1.0045,$$

so the 1-sided confidence interval with confidence coefficient $1 - \alpha = 0.95$ is $(1.0045, \infty)$.

2. In the case of the amateur mechanic, the null hypothesis $H_0 : \mu \leq \mu_0$ is accepted if and only if

$$\mu_0 > 1.21 - 1.645 \cdot \frac{0.4}{\sqrt{9}} \doteq 0.9967,$$

so the 1-sided confidence interval with confidence coefficient $1 - \alpha = 0.95$ is $(0.9967, \infty)$.

9.6 Exercises

1. According to *The Justice Project*, "John Spirko was sentenced to death on the testimony of a witness who was '70 percent certain' of his identification." Formulate this case as a problem in hypothesis testing. What can be deduced about the significance level used to convict Spirko? Does this choice of significance level strike you as suitable for a capital murder trial?

2. Blaise Pascal, the French theologian and mathematician, argued that we cannot know whether or not God exists, but that we must behave as though we do. He submitted that the consequences of wrongly behaving as though God does not exist are greater than the consequences of wrongly behaving as though God does exist, concluding that it is better to err on the side of caution and act as though God exists. This argument is known as Pascal's Wager. Formulate Pascal's Wager as a hypothesis testing problem. What are the Type I and Type II errors? On whom did Pascal place the burden of proof, believers or nonbelievers?

3. Dorothy owns a lovely glass dreidl. Curious as to whether or not it is fairly balanced, she spins her dreidl ten times, observing five gimels and five hehs. Surprised by these results, Dorothy decides to compute

the probability that a fair dreidl would produce such aberrant results. Which of the probabilities specified in Exercise 3.7.5 is the most appropriate choice of a significance probability for this investigation? Why?

4. It is thought that human influenza viruses originate in birds. It is quite possible that, several years ago, a human influenza pandemic was averted by slaughtering 1.5 million chickens brought to market in Hong Kong. Because it is impossible to test each chicken individually, such decisions are based on samples. Suppose that a boy has already died of a bird flu virus apparently contracted from a chicken. Several diseased chickens have already been identified. The health officials would prefer to err on the side of caution and destroy all chickens that might be infected; the farmers do not want this to happen unless it is absolutely necessary. Suppose that both the farmers and the health officals agree that all chickens should be destroyed if more than 2 percent of them are diseased. A random sample of $n = 1000$ chickens reveals 40 diseased chickens.

   (a) Let $X_i = 1$ if chicken i is diseased and $X_i = 0$ if it is not. Assume that $X_1, \ldots, X_n \sim P$. To what family of probability distributions does P belong? What population parameter indexes this family? Use this parameter to state formulas for $\mu = EX_i$ and $\sigma^2 = \operatorname{Var} X_i$.

   (b) State appropriate null and alternative hypotheses from the perspective of the health officials.

   (c) State appropriate null and alternative hypotheses from the perspective of the farmers.

   (d) Use the value of μ_0 in the above hypotheses to compute the value of σ^2 under H_0. Then compute the value of the test statistic z.

   (e) Adopting the health officials' perspective, and assuming that they are willing to risk a 0.1% chance of committing a Type I error, what action should be taken? Why?

   (f) Adopting the farmers' perspective, and assuming that they are willing to risk a 10% chance of committing a Type I error, what action should be taken? Why?

5. A company that manufactures light bulbs has advertised that its 75-watt bulbs burn an average of 800 hours before failing. In reaction

to the company's advertising campaign, several dissatisfied customers have complained to a consumer watchdog organization that they believe the company's claim to be exaggerated. The consumer organization must decide whether or not to allocate some of its financial resources to countering the company's advertising campaign. So that it can make an informed decision, it begins by purchasing and testing 100 of the disputed light bulbs. In this experiment, the 100 light bulbs burned an average of $\bar{x} = 745.1$ hours before failing, with a sample standard deviation of $s = 238.0$ hours. Formulate null and alternative hypotheses that are appropriate for this situation. Calculate a significance probability. Do these results warrant rejecting the null hypothesis at a significance level of $\alpha = 0.05$?

6. To study the effects of Alzheimer's disease (AD) on cognition, a scientist administers two batteries of neuropsychological tasks to 60 mildly demented AD patients. One battery is administered in the morning, the other in the afternoon. Each battery includes a task in which discourse is elicited by showing the patient a picture and asking the patient to describe it. The quality of the discourse is measured by counting the number of "information units" conveyed by the patient. The scientist wonders if asking a patient to describe Picture A in the morning is equivalent to asking the same patient to describe Picture B in the afternoon, after having described Picture A several hours earlier. To investigate, she computes the number of information units for Picture A minus the number of information units for Picture B for each patient. She finds an average difference of $\bar{x} = -0.1833$, with a sample standard deviation of $s = 5.18633$. Formulate null and alternative hypotheses that are appropriate for this situation. Calculate a significance probability. Do these results warrant rejecting the null hypothesis at a significance level of $\alpha = 0.05$?

7. Each student in a large statistics class of 600 students is asked to toss a fair coin 100 times, count the resulting number of `Heads`, and construct a 0.95-level confidence interval for the probability of `Heads`. Assume that each student uses a fair coin and constructs the confidence interval correctly. True or False: *We would expect approximately* 570 *of the confidence intervals to contain the number* 0.5.

8. The USGS decides to use a laser altimeter to measure the height μ of Mt. Wrightson, the highest point in Pima County, Arizona. It is

known that measurements made by the laser altimeter have an expected value equal to μ and a standard deviation of 1 meter. How many measurements should be made if the USGS wants to construct a 0.90-level confidence interval for μ that has a length of 20 centimeters?

9. Professor Johnson is interested in the probability that a certain type of randomly generated matrix has a positive determinant. His student attempts to calculate the probability exactly, but runs into difficulty because the problem requires her to evaluate an integral in 9 dimensions. Professor Johnson therefore decides to obtain an approximate probability by simulation, i.e., by randomly generating some matrices and observing the proportion that have positive determinants. His preliminary investigation reveals that the probability is roughly 0.05. At this point, Professor Park decides to undertake a more comprehensive simulation experiment that will, with 0.95-level confidence, correctly determine the probability of interest to within ± 0.00001. How many random matrices should he generate to achieve the desired accuracy?

10. Consider a box that contains 10 tickets, labelled

$$\{1, 1, 1, 1, 2, 5, 5, 10, 10, 10\}.$$

From this box, I propose to draw (with replacement) $n = 40$ tickets. Let Y denote the sum of the values on the tickets that are drawn.

To approximate $p = P(170.5 < Y < 199.5)$, a Math 351 student writes an `R` function `box.model` that simulates the proposed experiment. Evaluating `box.model` is like observing a value, y, of the random variable Y. Then she writes a loop that repeatedly evaluates `box.model` and computes $\hat{p}$, the proportion of times that `box.model` produces $y \in (170.5, 199.5)$. The student intends to construct a 0.95-level confidence interval for p. If she desires an interval of length L, then how many times should she plan to evaluate `box.model`?

Hint: How else might the student estimate p?

11. In September 2003, Lena spun a penny 89 times and observed 2 `Heads`. Let p denote the true probability that one spin of her penny will result in `Heads`.

    (a) The significance probability for testing $H_0 : p \geq 0.3$ versus $H_1 : p < 0.3$ is $\mathbf{p} = P(Y \leq 2)$, where $Y \sim \text{Binomial}(89; 0.3)$.

i. Compute **p** as in Section 9.1, using the binomial distribution and `pbinom`.

ii. Approximate **p** as in Section 9.4, using the normal distribution and `pnorm`. How good is this approximation?

(b) Construct a 1-sided confidence interval for p by determining for which values of p_0 the null hypothesis $H_0 : p \geq p_0$ would be accepted at a significance level of (approximately) $\alpha = 0.05$.

Chapter 10

1-Sample Location Problems

The basic ideas associated with statistical inference were introduced in Chapter 9. We developed these ideas in the context of drawing inferences about a single population mean, and we assumed that the sample was large enough to justify appeals to the Central Limit Theorem for normal approximations. The population mean is a natural measure of centrality, but it is not the only one. Furthermore, even if we are interested in the population mean, our sample may be too small to justify the use of a large-sample normal approximation. The purpose of the next several chapters is to explore more thoroughly how statisticians draw inferences about measures of centrality.

Measures of centrality are sometimes called location parameters. The title of this chapter indicates an interest in a location parameter of a *single* population. More specifically, we assume that $X_1, \ldots, X_n \sim P$ are independently and identically distributed, we observe a random sample $\vec{x} = \{x_1, \ldots, x_n\}$, and we attempt to draw an inference about a location parameter of P. Because it is not always easy to identify the relevant population in a particular experiment, we begin with some examples. Our analysis of these examples is clarified by posing the following four questions:

1. What are the experimental units, i.e., what are the objects that are being measured?

2. From what population (or populations) were the experimental units drawn?

3. What measurements were taken on each experimental unit?

4. What random variables are relevant to the specified inference?

For the sake of specificity, we assume that the location parameter of interest in the following examples is the population median, $q_2(P)$.

Example 10.1 A machine is supposed to produce ball bearings that are 1 millimeter in diameter. To determine if the machine was correctly calibrated, a sample of ball bearings is drawn and the diameter of each ball bearing is measured. For this experiment:

1. An experimental unit is a ball bearing. Notice that we are distinguishing between experimental units, the objects being measured (ball bearings), and units of measurement (e.g., millimeters).

2. There is one population, viz., all ball bearings that might be produced by the designated machine.

3. One measurement (diameter) is taken on each experimental unit.

4. Let X_i denote the diameter of ball bearing i. Then $X_1, \ldots, X_n \sim P$ and we are interested in drawing inferences about $q_2(P)$, the population median diameter. For example, we might test $H_0 : q_2(P) = 1$ against $H_1 : q_2(P) \neq 1$.

Example 10.2 A drug is supposed to lower blood pressure. To determine if it does, a sample of hypertensive patients are administered the drug for two months. Each person's blood pressure is measured before and after the two month period. For this experiment:

1. An experimental unit is a patient.

2. There is one population of hypertensive patients. (It may be difficult to discern the precise population that was actually sampled. All hypertensive patients? All Hispanic male hypertensive patients who live in Houston, TX? All Hispanic male hypertensive patients who live in Houston, TX, and who are sufficiently well-disposed to the medical establishment to participate in the study? In published journal articles, scientists are often rather vague about just what population was actually sampled.)

3. Two measurements (blood pressure before and after treatment) are taken on each experimental unit. Let B_i and A_i denote the blood pressures of patient i before and after treatment.

4. Let $X_i = B_i - A_i$, the decrease in blood pressure for patient i. Then $X_1, \ldots, X_n \sim P$ and we are interested in drawing inferences about $q_2(P)$, the population median decrease. For example, we might test $H_0 : q_2(P) \leq 0$ against $H_1 : q_2(P) > 0$.

Example 10.3 A graduate student investigated the effect of Parkinson's disease (PD) on speech breathing. She recruited 16 PD patients to participate in her study. She also recruited 16 normal control (NC) subjects. Each NC subject was carefully matched to one PD patient with respect to sex, age, height, and weight. The lung volume of each study participant was measured. For this experiment:

1. An experimental unit was a matched PD-NC pair.

2. The population comprises all possible PD-NC pairs that satisfy the study criteria.

3. Two measurements (PD and NC lung volume) were taken on each experimental unit. Let D_i and C_i denote the PD and NC lung volumes of pair i.

4. Let $X_i = \log(D_i/C_i) = \log D_i - \log C_i$, the logarithm of the PD proportion of NC lung volume. (This is not the only way of comparing D_i and C_i, but it worked well in this investigation. Ratios can be difficult to analyze and logarithms convert ratios to differences. Furthermore, lung volume data tend to be skewed to the right. As in Exercise 2 of Section 7.6, logarithmic transformations of such data often have a symmetrizing effect.) Then $X_1, \ldots, X_n \sim P$ and we are interested in drawing inferences about $q_2(P)$. For example, to test the theory that PD restricts lung volume, we might test $H_0 : q_2(P) \geq 0$ against $H_1 : q_2(P) < 0$.

This chapter is divided into sections according to distributional assumptions about the X_i:

10.1 If the data are assumed to be normally distributed, then we will be interested in inferences about the population's center of symmetry, which we will identify as the population mean.

10.3 If the data are only assumed to be symmetrically distributed, then we will also be interested in inferences about the population's center of symmetry, but we will identify it as the population median.

10.2 If the data are only assumed to be continuously distributed, then we will be interested in inferences about the population median.

Each section is subdivided into subsections, according to the type of inference (point estimation, hypothesis testing, set estimation) at issue.

10.1 The Normal 1-Sample Location Problem

In this section we assume that $P = \text{Normal}(\mu, \sigma^2)$. As necessary, we will distinguish between cases in which σ is known and cases in which σ is unknown.

10.1.1 Point Estimation

Because normal distributions are symmetric, the location parameter μ is the center of symmetry and therefore both the population mean and the population median. Hence, there are (at least) two natural estimators of μ, the sample mean $\bar{X}_n$ and the sample median $q_2(\hat{P}_n)$. Both are consistent, unbiased estimators of μ. We will compare them by considering their *asymptotic relative efficiency* (ARE). A rigorous definition of ARE is beyond the scope of this book, but the concept is easily interpreted.

If the true distribution is $P = N(\mu, \sigma^2)$, then the ARE of the sample median to the sample mean for estimating μ is

$$e(P) = \frac{2}{\pi} \doteq 0.64.$$

This statement has the following interpretation: for large samples, using the sample median to estimate a normal population mean is equivalent to randomly discarding approximately 36% of the observations and calculating the sample mean of the remaining 64%. Thus, the sample mean is substantially more efficient than is the sample median at extracting location information from a normal sample.

In fact, if $P = \text{Normal}(\mu, \sigma^2)$, then the ARE of *any* estimator of μ to the sample mean is ≤ 1. This is sometimes expressed by saying that the sample mean is *asymptotically efficient* for estimating a normal mean. The sample mean also enjoys a number of other optimal properties in this case. The sample mean is unquestionably the preferred estimator for the normal 1-sample location problem.

10.1.2 Hypothesis Testing

If σ is known, then the possible distributions of X_i are

$$\left\{\text{Normal}(\mu, \sigma^2) : -\infty < \mu < \infty\right\}.$$

If σ is unknown, then the possible distributions of X_i are

$$\left\{\text{Normal}(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma > 0\right\}.$$

We partition the possible distributions into two subsets, the null and alternative hypotheses. For example, if σ is known then we might specify

$$H_0 = \left\{\text{Normal}(0, \sigma^2)\right\} \quad \text{and} \quad H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq 0\right\},$$

which we would typically abbreviate as $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$. Analogously, if σ is unknown then we might specify

$$H_0 = \left\{\text{Normal}(0, \sigma^2) : \sigma > 0\right\}$$

and

$$H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq 0,\ \sigma > 0\right\},$$

which we would also abbreviate as $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$.

More generally, for any real number μ_0 we might specify

$$H_0 = \left\{\text{Normal}(\mu_0, \sigma^2)\right\} \quad \text{and} \quad H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq \mu_0\right\}$$

if σ is known, or

$$H_0 = \left\{\text{Normal}(\mu_0, \sigma^2) : \sigma > 0\right\}$$

and

$$H_1 = \left\{\text{Normal}(\mu, \sigma^2) : \mu \neq \mu_0,\ \sigma > 0\right\}$$

if σ in unknown. In both cases, we would typically abbreviate these hypotheses as $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

The preceding examples involve two-sided alternative hypotheses. Of course, as in Section 9.4, we might also specify one-sided hypotheses. However, the material in the present section is so similar to the material in Section 9.4 that we will only discuss two-sided hypotheses.

The intuition that underlies testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ was discussed in Section 9.4:

- If H_0 is true, then we would expect the sample mean to be close to the population mean μ_0.
- Hence, if $\bar{X}_n = \bar{x}_n$ is observed far from μ_n, then we are inclined to reject H_0.

To make this reasoning precise, we reject H_0 if and only if the significance probability

$$\mathbf{p} = P_{\mu_0}\left(\left|\bar{X}_n - \mu_0\right| \geq |\bar{x}_n - \mu_0|\right) \leq \alpha. \tag{10.1}$$

The first equation in (10.1) is a formula for a significance probability. Notice that this formula is identical to equation (9.2). The one difference between the material in Section 9.4 and the present material lies in how one computes **p**. For emphasis, we recall the following:

1. The hypothesized mean μ_0 is a fixed number specified by the null hypothesis.
2. The estimated mean, $\bar{x}_n$, is a fixed number computed from the sample. Therefore, so is $|\bar{x}_n - \mu_0|$, the difference between the estimated mean and the hypothesized mean.
3. The estimator, $\bar{X}_n$, is a random variable.
4. The subscript in P_{μ_0} reminds us to compute the probability under $H_0 : \mu = \mu_0$.
5. The significance level α is a fixed number specified by the researcher, preferably before the experiment was performed.

To apply (10.1), we must compute **p**. In Section 9.4, we overcame that technical difficulty by appealing to the Central Limit Theorem. This allowed us to approximate **p** even when we did not know the distribution of the X_i, but only for reasonably large sample sizes. However, if we know that $X_1, \ldots, X_n$ are normally distributed, then it turns out that we can calculate **p** exactly, even when n is small.

Case 1: The Population Variance is Known

Under the null hypothesis that $\mu = \mu_0$, $X_1, \ldots, X_n \sim \text{Normal}(\mu_0, \sigma^2)$ and

$$\bar{X}_n \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

This is the exact distribution of $\bar{X}_n$, not an asymptotic approximation. We convert $\bar{X}_n$ to standard units, obtaining

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0,1)\,. \tag{10.2}$$

The observed value of Z is

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}.$$

The significance probability is

$$\begin{aligned} \mathbf{p} &= P_{\mu_0}\left(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|\right) \\ &= P_{\mu_0}\left(\left|\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}\right| \geq \left|\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right|\right) \\ &= P\left(|Z| \geq |z|\right) \\ &= 2P\left(Z \geq |z|\right). \end{aligned}$$

In this case, the test that rejects H_0 if and only if $\mathbf{p} \leq \alpha$ is sometimes called the 1-*sample z-test*. The random variable Z is the *test statistic*.

Before considering the case of an unknown population variance, we remark that it is possible to derive point estimators from hypothesis tests. For testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, the test statistics are

$$Z(\mu_0) = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

If we observe $\bar{X}_n = \bar{x}_n$, then what value of μ_0 minimizes $|z(\mu_0)|$? Clearly, the answer is $\mu_0 = \bar{x}_n$. Thus, our preferred point estimate of μ is the μ_0 for which it is most difficult to reject $H_0 : \mu = \mu_0$. This type of reasoning will be extremely useful for analyzing situations in which we know how to test but don't know how to estimate.

Case 2: The Population Variance is Unknown

Statement (10.2) remains true if σ is unknown, but it is no longer possible to compute z. Therefore, we require a different test statistic for this case. A natural approach is to modify Z by replacing the unknown σ with an estimator of it. Toward that end, we introduce the test statistic

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}},$$

where S_n^2 is the unbiased estimator of the population variance defined by equation (9.1). Because T_n and Z are different random variables, they have different probability distributions and our first order of business is to determine the distribution of T_n.

We begin by stating a useful fact:

Theorem 10.1 *If $X_1, \ldots, X_n \sim$ Normal(μ, σ^2), then*

$$\frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 / \sigma^2 \sim \chi^2(n-1).$$

The χ^2 (chi-squared) distribution was described in Section 5.5 and Theorem 10.1 is closely related to Theorem 5.3.

Next we write

$$\begin{aligned} T_n &= \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \cdot \frac{\sigma/\sqrt{n}}{S_n/\sqrt{n}} \\ &= Z \cdot \frac{\sigma}{S_n} = Z/\sqrt{S_n^2/\sigma^2} \\ &= Z/\sqrt{\left[(n-1)S_n^2/\sigma^2\right]/(n-1)}. \end{aligned}$$

Using Theorem 10.1, we see that T_n can be written in the form

$$T_n = \frac{Z}{\sqrt{Y/\nu}},$$

where $Z \sim \text{Normal}(0,1)$ and $Y \sim \chi^2(\nu)$. If Z and Y are independent random variables, then it follows from Definition 5.7 that $T_n \sim t(n-1)$.

Both Z and $Y = (n-1)S_n^2/\sigma^2$ depend on $X_1, \ldots, X_n$, so one would be inclined to think that Z and Y are dependent. This is usually the case, but it turns out that they are independent if $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. This is another remarkable property of normal distributions, usually stated as follows:

Theorem 10.2 *If $X_1, \ldots, X_n \sim$ Normal(μ, σ^2), then $\bar{X}_n$ and S_n^2 are independent random variables.*

The result that interests us can then be summarized as follows:

Corollary 10.1 *If $X_1, \ldots, X_n \sim$ Normal(μ_0, σ^2), then*

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1).$$

Now let

$$t_n = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}},$$

the observed value of the test statistic T_n. The significance probability is

$$\mathbf{p} = P_{\mu_0}\left(|T_n| \geq |t_n|\right) = 2P_{\mu_0}\left(T_n \geq |t_n|\right).$$

In this case, the test that rejects H_0 if and only if $\mathbf{p} \leq \alpha$ is called *Student's 1-sample t-test*. Because it is rarely the case that the population variance is known when the population mean is not, Student's 1-sample t-test is used much more frequently than the 1-sample z-test. We will use the `R` function `pt` to compute significance probabilities for Student's 1-sample t-test, as illustrated in the following examples.

Example 10.4 Suppose that, to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ (a 2-sided alternative), we draw a sample of size $n = 25$ and observe $\bar{x} = 1$ and $s = 3$. Then $t = (1-0)/(3/\sqrt{25}) = 5/3$ and the 2-tailed significance probability is computed using both tails of the $t(24)$ distribution, i.e., $\mathbf{p} = 2 * \texttt{pt}(-5/3, \texttt{df} = 24) \doteq 0.1086$.

Example 10.5 Suppose that, to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ (a 1-sided alternative), we draw a sample of size $n = 25$ and observe $\bar{x} = 2$ and $s = 5$. Then $t = (2-0)/(5/\sqrt{25}) = 2$ and the 1-tailed significance probability is computed using one tail of the $t(24)$ distribution, i.e., $\mathbf{p} = 1 - \texttt{pt}(2, \texttt{df} = 24) \doteq 0.0285$.

10.1.3 Interval Estimation

As in Section 9.5, we will derive confidence intervals from tests. We imagine testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ for every $\mu_0 \in (-\infty, \infty)$. The μ_0 for which $H_0 : \mu = \mu_0$ is rejected are implausible values of μ; the μ_0 for which $H_0 : \mu = \mu_0$ is accepted constitute the confidence interval. To accomplish this, we will have to derive the critical values of our tests. A significance level of α will result in a confidence coefficient of $1 - \alpha$.

Case 1: The Population Variance is Known

If σ is known, then we reject $H_0 : \mu = \mu_0$ if and only if

$$\mathbf{p} = P_{\mu_0}\left(\left|\bar{X}_n - \mu_0\right| \geq |\bar{x}_n - \mu_0|\right) = 2\Phi\left(-|z_n|\right) \leq \alpha,$$

where $z_n = (\bar{x}_n - \mu_0)/(\sigma/\sqrt{n})$. By the symmetry of the normal distribution, this condition is equivalent to the condition

$$1 - \Phi(-|z_n|) = P(Z > -|z_n|) = P(Z < |z_n|) = \Phi(|z_n|) \geq 1 - \alpha/2,$$

where $Z \sim \text{Normal}(0, 1)$, and therefore to the condition $|z_n| \geq q_z$, where q_z denotes the $1 - \alpha/2$ quantile of Normal$(0, 1)$. The quantile q_z is the critical value of the two-sided 1-sample z-test. Thus, given a significance level α and a corresponding critical value q_z, we reject $H_0 : \mu = \mu_0$ if and only if (iff)

$$\begin{aligned} & \left| \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \right| = |z_n| \geq q_z \\ \text{iff} \quad & |\bar{x}_n - \mu_0| \geq q_z \sigma/\sqrt{n} \\ \text{iff} \quad & \mu_0 \notin (\bar{x}_n - q_z\sigma/\sqrt{n},\ \bar{x}_n + q_z\sigma/\sqrt{n}) \end{aligned}$$

and we conclude that the desired set of plausible values is the interval

$$\left(\bar{x}_n - q_z \frac{\sigma}{\sqrt{n}},\ \bar{x}_n + q_z \frac{\sigma}{\sqrt{n}} \right).$$

Notice that both the preceding derivation and the resulting confidence interval are identical to the derivation and confidence interval in Section 9.5. The only difference is that, because we are now assuming that $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$ instead of relying on the Central Limit Theorem, no approximation is required.

Example 10.6 Suppose that we desire 90% confidence about μ and $\sigma = 3$ is known. Then $\alpha = 0.10$ and $q_z \doteq 1.645$. Suppose that we draw $n = 25$ observations and observe $\bar{x}_n = 1$. Then

$$1 \pm 1.645 \frac{3}{\sqrt{25}} = 1 \pm 0.987 = (0.013, 1.987)$$

is a 0.90-level confidence interval for μ.

Case 2: The Population Variance is Unknown

If σ is unknown, then it must be estimated from the sample. The reasoning in this case is the same, except that we rely on Student's 1-sample t-test.

As before, we use S_n^2 to estimate σ^2. The critical value of the 2-sided 1-sample t-test is q_t, the $1-\alpha/2$ quantile of a t distribution with $n-1$ degrees of freedom, and the confidence interval is

$$\left(\bar{x}_n - q_t \frac{s_n}{\sqrt{n}},\ \bar{x}_n + q_t \frac{s_n}{\sqrt{n}} \right).$$

Example 10.7 Suppose that we desire 90% confidence about μ and σ is unknown. Suppose that we draw $n = 25$ observations and observe $\bar{x}_n = 1$ and $s = 3$. Then $q_t = \texttt{qt}(.95, \texttt{df} = 24) \doteq 1.711$ and

$$1 \pm 1.711 \times 3/\sqrt{25} = 1 \pm 1.027 = (-0.027, 2.027)$$

is a 90% confidence interval for μ. Notice that the confidence interval is larger when we use $s = 3$ instead of $\sigma = 3$.

10.2 The General 1-Sample Location Problem

In Section 10.1 we assumed that $X_1, \ldots, X_n \sim P$ and $P = \text{Normal}(\mu, \sigma^2)$. In this section, we again assume that $X_1, \ldots, X_n \sim P$, but now we assume only that the X_i are continuous random variables.

Because P is not assumed to be symmetric, we must decide which location parameter to study. The population median, $q_2(P)$, enjoys several advantages. Unlike the population mean, the population median always exists and is not sensitive to the influence of outliers. Furthermore, it turns out that one can develop fairly elementary ways to study medians, even when little is known about the probability distribution P. For simplicity, we will denote the population median by θ.

10.2.1 Hypothesis Testing

It is convenient to begin our study of the general 1-sample location problem with a discussion of hypothesis testing. As is Section 10.1, we initially consider testing a 2-sided alternative, $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We will explicate a procedure known as the *sign test*.

The intuition that underlies the sign test is elementary. If the population median is $\theta = \theta_0$, then when we sample P we should observe roughly half the x_i above θ_0 and half the x_i below θ_0. Hence, if we observe proportions of x_i above/below θ_0 that are very different from one half, then we are inclined to reject the possibility that $\theta = \theta_0$.

More formally, let $p_+ = P_{H_0}(X_i > \theta_0)$ and $p_- = P_{H_0}(X_i < \theta_0)$. Because the X_i are continuous, $P_{H_0}(X_i = \theta_0) = 0$ and therefore $p_+ = p_- = 0.5$. Hence, under H_0, observing whether $X_i > \theta_0$ or $X_i < \theta_0$ is equivalent to tossing a fair coin, i.e., to observing a Bernoulli trial with success probability $p = 0.5$. The sign test is the following procedure:

1. Let $\vec{x} = \{x_1, \ldots, x_n\}$ denote the observed sample. If the X_i are continuous random variables, then $P(X_i = \theta_0) = 0$ and it should be that each $x_i \neq \theta_0$. In practice, of course, it may happen that we do observe one or more $x_i = \theta_0$. For the moment, we assume that $\vec{x}$ contains no such values.

2. Let
$$Y = \#\{X_i > \theta_0\} = \#\{X_i - \theta_0 > 0\}$$
be the test statistic. Under $H_0 : \theta = \theta_0$, $Y \sim \text{Binomial}(n; p = 0.5)$. The observed value of the test statistic is
$$y = \#\{x_i > \theta_0\} = \#\{x_i - \theta_0 > 0\}.$$

3. Notice that $EY = n/2$. The significance probability is
$$\mathbf{p} = P_{\theta_0}\left(\left|Y - \frac{n}{2}\right| \geq \left|y - \frac{n}{2}\right|\right).$$
The sign test rejects $H_0 : \theta = \theta_0$ if and only if $\mathbf{p} \leq \alpha$.

4. To compute $\mathbf{p}$, we first note that
$$\left|Y - \frac{n}{2}\right| \geq \left|y - \frac{n}{2}\right|$$
is equivalent to the event
   (a) $\{Y \leq y \text{ or } Y \geq n - y\}$ if $y \leq n/2$;
   (b) $\{Y \geq y \text{ or } Y \leq n - y\}$ if $y \geq n/2$.

   To accomodate both cases, let $c = \min(y, n - y)$. Then
$$\mathbf{p} = P_{\theta_0}(Y \leq c) + P_{\theta_0}(Y \geq c) = 2P_{\theta_0}(Y \leq c) = \texttt{2*pbinom(c,n,.5)}.$$

Example 10.8(a) Suppose that we want to test $H_0 : \theta = 100$ versus $H_1 : \theta \neq 100$ at significance level $\alpha = 0.05$, having observed the sample
$$\vec{x} = \{98.73, 97.17, 100.17, 101.26, 94.47, 96.39, 99.67, 97.77, 97.46, 97.41\}.$$
Here $n = 10$, $y = \#\{x_i > 100\} = 2$, and $c = \min(2, 10 - 2) = 2$, so
$$\mathbf{p} = \texttt{2*pbinom(2,10,.5)} = 0.109375 > 0.05$$
and we decline to reject H_0.

Example 10.8(b) Now suppose that we want to test $H_0 : \theta \leq 97$ versus $H_1 : \theta > 97$ at significance level $\alpha = 0.05$, using the same data. Here $n = 10$, $y = \#\{x_i > 97\} = 8$, and $c = \min(8, 10 - 8) = 2$. Because large values of Y are evidence against $H_0 : \theta \leq 97$,

$$\begin{aligned} \mathbf{p} &= P_{\theta_0}(Y \geq y) = P_{\theta_0}(Y \geq 8) = 1 - P_{\theta_0}(Y \leq 7) \\ &= \texttt{1-pbinom(7,10,.5)} = 0.0546875 > 0.05 \end{aligned}$$

and we decline to reject H_0.

Thus far we have assumed that the sample contains no values for which $x_i = \theta_0$. In practice, we may well observe such values. For example, if the measurements in Example 10.8(a) were made less precisely, then we might have observed the following sample:

$$\vec{x} = \{99, 97, 100, 101, 94, 96, 100, 98, 97, 97\}. \tag{10.3}$$

If we want to test $H_0 : \theta = 100$ versus $H_1 : \theta \neq 100$, then we have two values that equal θ_0 and the sign test requires modification.

We assume that $\#\{x_i = \theta_0\}$ is fairly small; otherwise, the assumption that the X_i are continuous is questionable. We consider two possible ways to proceed:

1. Perhaps the most satisfying solution is to compute all of the significance probabilities that correspond to different ways of counting the $x_i = \theta_0$ as larger or smaller than θ_0. If there are k observations $x_i = \theta_0$, then this will produce 2^k significance probabilities, which we might average to obtain a single $\mathbf{p}$.

2. Alternatively, let $\mathbf{p}_0$ denote the significance probability obtained by counting in the way that is most favorable to H_0 (least favorable to H_1). This is the largest of the possible significance probabilities, so if $\mathbf{p}_0 \leq \alpha$ then we reject H_0. Similarly, let $\mathbf{p}_1$ denote the significance probability obtained by counting in the way that is least favorable to H_0 (most favorable to H_1). This is the smallest of the possible significance probabilities, so if $\mathbf{p}_1 > \alpha$ then we decline to reject H_0. If $\mathbf{p}_0 > \alpha \geq \mathbf{p}_1$, then we simply declare the results to be equivocal.

Example 10.8(c) Suppose that we want to test $H_0 : \theta = 100$ versus $H_1 : \theta \neq 100$ at significance level $\alpha = 0.05$, having observed the sample

(10.3). Here $n = 10$ and $y = \#\{x_i > 100\}$ depends on how we count the observations $x_3 = x_7 = 100$. There are $2^2 = 4$ possibilities:

possibility	$y = \#\{x_i > 100\}$	$c = \min(y, 10 - y)$	$\mathbf{p}$
$y_3 < 100, y_7 < 100$	1	1	0.021484
$y_3 < 100, y_7 > 100$	2	2	0.109375
$y_3 > 100, y_7 < 100$	2	2	0.109375
$y_3 > 100, y_7 > 100$	3	3	0.343750

Noting that $\mathbf{p}_0 \doteq 0.344 > 0.05 > 0.021 \doteq \mathbf{p}_1$, we might declare the results to be equivocal. However, noting that 3 of the 4 possibilities lead us to accept H_0 (and that the average $\mathbf{p} \doteq 0.146$), we might conclude—somewhat more decisively—that there is insufficient evidence to reject H_0. The distinction between these two interpretations is largely rhetorical, as the fundamental logic of hypothesis testing requires that we decline to reject H_0 unless there is compelling evidence against it.

10.2.2 Point Estimation

Next we consider the problem of estimating the population median. A natural estimate is the plug-in estimate, the sample median. Another approach begins by posing the following question: For what value of θ_0 is the sign test least inclined to reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$? The answer to this question is also a natural estimate of the population median.

In fact, the plug-in and sign-test approaches lead to the same estimation procedure. To understand why, we focus on the case that n is even, in which case $n/2$ is a possible value of $Y = \#\{X_i > \theta_0\}$. If $|y - n/2| = 0$, then

$$\mathbf{p} = P\left(\left|Y - \frac{n}{2}\right| \geq 0\right) = 1.$$

We see that the sign test produces the maximal significance probability of $\mathbf{p} = 1$ when $y = n/2$, i.e., when θ_0 is chosen so that precisely half the observations exceed θ_0. This means that the sign test is least likely to reject $H_0 : \theta = \theta_0$ when θ_0 is the sample median. (A similar argument leads to the same conclusion when n is odd.)

Thus, using the sign test to test hypotheses about population medians corresponds to using the sample median to estimate population medians, just as using Student's t-test to test hypotheses about population means corresponds to using the sample mean to estimate population means. One

consequence of this remark is that, when the population mean and median are identical, the "Pitman efficiency" of the sign test to Student's t-test equals the asymptotic relative efficiency of the sample median to the sample median. For example, using the sign test on normal data is asymptotically equivalent to randomly discarding 36% of the observations, then using Student's t-test on the remaining 64%.

10.2.3 Interval Estimation

Finally, we consider the problem of constructing a $(1-\alpha)$-level confidence interval for the population median. Again we rely on the sign test, determining for which θ_0 the level-α sign test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ will accept H_0.

The sign test will reject $H_0 : \theta = \theta_0$ if and only if

$$y(\theta_0) = \#\{x_i > \theta_0\}$$

is either too large or too small. Equivalently, H_0 will be accepted if θ_0 is such that the numbers of observations above and below θ_0 are roughly equal.

To determine the critical value for the desired sign test, we suppose that $Y \sim \text{Binomial}(n; 0.5)$. We would like to find k such that $\alpha = 2P(Y \leq k)$, or $\alpha/2 = \texttt{pbinom}(k, n, 0.5)$. In practice, we won't be able to solve this equation exactly. We will use the `qbinom` function plus trial-and-error to solve it approximately, then modify our choice of α accordingly.

Having determined an acceptable (α, k), the sign test rejects $H_0 : \theta = \theta_0$ at level α if and only if either $y(\theta_0) \leq k$ or $y(\theta_0) \geq n - k$. We need to translate these inequalities into an interval of plausible values of θ_0. To do so, it is helpful to sort the values observed in the sample.

Definition 10.1 *The* order statistics *of* $\vec{x} = \{x_1, \ldots, x_n\}$ *are any permutation of the* x_i *such that*

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}.$$

If $\vec{x}$ *contains* n *distinct values, then there is a unique set of order statistics and the above inequalities are strict; otherwise, we say that* $\vec{x}$ *contains* ties.

Thus, $x_{(1)}$ is the smallest value in $\vec{x}$ and $x_{(n)}$ is the largest. If $n = 2m+1$ (n is odd), then the sample median is $x_{(m+1)}$; if $n = 2m$ (n is even), then the sample median is $[x_{(m)} + x_{(m+1)}]/2$.

For simplicity we assume that $\vec{x}$ contains no ties. If $\theta_0 < x_{(k+1)}$, then at least $n - k$ observations exceed θ_0 and the sign test rejects $H_0 : \theta = \theta_0$. Similarly, if $\theta_0 > x_{(n-k)}$, then no more than k observations exceed θ_0 and the sign test rejects $H_0 : \theta = \theta_0$. We conclude that the sign test accepts $H_0 : \theta = \theta_0$ if and only if θ_0 lies in the $(1 - \alpha)$-level confidence interval

$$\left(x_{(k+1)}, x_{(n-k)}\right).$$

Example 10.8(d) Using the $n = 10$ observations from Example 10.8(a), we endeavor to construct a 0.90-level confidence interval for the population median. We begin by determining a suitable choice of (α, k). If $1-\alpha = 0.90$, then $\alpha/2 = 0.05$. `R` returns `qbinom(.05,10,.5)` $= 2$. Next we experiment:

k	`pbinom`$(k, 10, 0.5)$
2	0.0546875
1	0.01074219

We choose $k = 2$, resulting in a confidence level of

$$1 - \alpha = 1 - 2 \cdot 0.0546875 = 0.890625 \doteq 0.89,$$

nearly equal to the requested level of 0.90. Now, upon sorting the data (the `sort` function in `R` may be useful), we quickly discern that the desired confidence interval is

$$\left(x_{(3)}, x_{(8)}\right) = (97.17, 99.67).$$

10.3 The Symmetric 1-Sample Location Problem

10.4 A Case Study from Neuropsychology

10.5 Exercises

Problem Set A

1. Assume that a large number, $n = 400$, of observations are independently drawn from a normal distribution with unknown population mean μ and unknown population variance σ^2. The resulting sample, $\vec{x}$, is used to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ at significance level $\alpha = 0.05$.

   (a) What test should be used in this situation? If we observe $\vec{x}$ that results in $\bar{x} = 3.194887$ and $s^2 = 104.0118$, then what is the value of the test statistic?

   (b) If we observe $\vec{x}$ that results in a test statistic value of 1.253067, then which of the following `R` expressions best approximates the significance probability?

      i. `2*pnorm(1.253067)`
      ii. `2*pnorm(-1.253067)`
      iii. `1-pnorm(1.253067)`
      iv. `1-pt(1.253067,df=399)`
      v. `pt(1.253067,df=399)`

   (c) True of False: if we observe $\vec{x}$ that results in a significance probability of $\mathbf{p} = 0.03044555$, then we should reject the null hypothesis.

2. A device counts the number of ions that arrive in a given time interval, unless too many arrive. An experiment that relies on this device produces the following counts, where `Big` means that the count exceeded 255.

251	238	249	`Big`	243	248	229	`Big`	235	244
254	251	252	244	230	222	224	246	`Big`	239

   Use these data to construct a confidence interval for the population median number of ions with a confidence coefficient of approximately 0.95.

Problem Set B The following data are from Darwin (1876), *The Effect of Cross- and Self-Fertilization in the Vegetable Kingdom, Second Edition*, London: John Murray. They appear as Data Set 3 in *A Handbook of Small Data Sets*, accompanied by the following description:

"Pairs of seedlings of the same age, one produced by cross-fertilization and the other by self-fertilization, were grown together so that the members of each pair were reared under nearly identical conditions. The aim was to demonstrate the greater vigour of the cross-fertilized plants. The data are the final heights [in inches] of each plant after a fixed period of time. Darwin consulted [Francis] Galton about the analysis of these data, and they were discussed further in [Ronald] Fisher's *Design of Experiments*."

Pair	Fertilized	
	Cross	Self
1	23.5	17.4
2	12.0	20.4
3	21.0	20.0
4	22.0	20.0
5	19.1	18.4
6	21.5	18.6
7	22.1	18.6
8	20.4	15.3
9	18.3	16.5
10	21.6	18.0
11	23.3	16.3
12	21.0	18.0
13	22.1	12.8
14	23.0	15.5
15	12.0	18.0

1. Show that this problem can be formulated as a 1-sample location problem. To do so, you should:

   (a) Identify the experimental units and the measurement(s) taken on each unit.

   (b) Define appropriate random variables $X_1, \ldots, X_n \sim P$. Remember that the statistical procedures that we will employ assume that these random variables are independent and identically distributed.

   (c) Let θ denote the location parameter (measure of centrality) of interest. Depending on which statistical procedure we decide to use, either $\theta = EX_i = \mu$ or $\theta = q_2(X_i)$. State appropriate null and alternative hypotheses about θ.

2. Does it seem reasonable to assume that the sample $\vec{x} = (x_1, \ldots, x_n)$, the observed values of $X_1, \ldots, X_n$, were drawn from:

   (a) a normal distribution? Why or why not?

   (b) a symmetric distribution? Why or why not?

3. Assume that $X_1, \ldots, X_n$ are normally distributed and let $\theta = EX_i = \mu$.

   (a) Test the null hypothesis derived above using Student's 1-sample t-test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Construct a (2-sided) confidence interval for θ with a confidence coefficient of approximately 0.90.

4. Now we drop the assumption of normality. Assume that $X_1, \ldots, X_n$ are symmetric (but not necessarily normal), continuous random variables and let $\theta = q_2(X_i)$.

   (a) Test the null hypothesis derived above using the Wilcoxon signed rank test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Estimate θ by computing the median of the Walsh averages.

   (c) Construct a (2-sided) confidence interval for θ with a confidence coefficient of approximately 0.90.

5. Finally we drop the assumption of symmetry, assuming only that $X_1, \ldots, X_n$ are continuous random variables, and let $\theta = q_2(X_i)$.

   (a) Test the null hypothesis derived above using the sign test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Estimate θ by computing the sample median.

   (c) Construct a (2-sided) confidence interval for θ with a confidence coefficient of approximately 0.90.

Problem Set C The ancient Greeks greatly admired rectangles with a height-to-width ratio of

$$1 : \frac{1 + \sqrt{5}}{2} = 0.618034.$$

They called this number the "golden ratio" and used it repeatedly in their art and architecture, e.g., in building the Parthenon. Furthermore, golden rectangles are often found in the art of later western cultures.

A cultural anthropologist wondered if the Shoshoni, a native American civilization, also used golden rectangles. The following measurements, which appear as Data Set 150 in *A Handbook of Small Data Sets*, are height-to-width ratios of beaded rectangles used by the Shoshoni in decorating various leather goods:

0.693	0.662	0.690	0.606	0.570
0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576
0.670	0.606	0.611	0.553	0.933

We will analyze the Shoshoni rectangles as a 1-sample location problem.

1. There are two natural scales that we might use in analyzing these data. One possibility is to analyze the ratios themselves; the other is to analyze the (natural) logarithms of the ratios. For which of these possibilities would an assumption of normality seem more plausible? Please justify your answer.

2. Choose the possibility (ratios or logarithms of ratios) for which an assumption of normality seems more plausible. Formulate suitable null and alternative hypotheses for testing the possibility that the Shoshoni were using golden rectangles. Using Student's 1-sample t-test, compute a significance probability for testing these hypotheses. Would you reject or accept the null hypothesis using a significance level of 0.05?

3. Suppose that we are unwilling to assume that either the ratios or the log-ratios were drawn from a normal distribution. Use the sign test to construct a 0.90-level confidence interval for the population median of the ratios.

Problem Set D Researchers studied the effect of the drug caprotil on essential hypertension, reporting their findings in the *British Medical Journal*. They measured the supine systolic and diastolic blood pressures of 15 patients with moderate essential hypertension, immediately before and two hours after administering caprotil. The following measurements are Data

Set 72 in *A Handbook of Small Data Sets*:

Patient	Systolic		Diastolic	
	before	after	before	after
1	210	201	130	125
2	169	165	122	121
3	187	166	124	121
4	160	157	104	106
5	167	147	112	101
6	176	145	101	85
7	185	168	121	98
8	206	180	124	105
9	173	147	115	103
10	146	136	102	98
11	174	151	98	90
12	201	168	119	98
13	198	179	106	110
14	148	129	107	103
15	154	131	100	82

We will consider the question of whether or not caprotil affects systolic and diastolic blood pressure differently.

1. Let SB and SA denote before and after systolic blood pressure; let DB and DB denote before and after diastolic blood pressure. There are several random variables that might be of interest:

$$X_i = (SB_i - SA_i) - (DB_i - DA_i) \tag{10.4}$$

$$X_i = \frac{SB_i - SA_i}{SB_i} - \frac{DB_i - DA_i}{DB_i} \tag{10.5}$$

$$X_i = \frac{SB_i - SA_i}{SB_i} \div \frac{DB_i - DA_i}{DB_i} \tag{10.6}$$

$$X_i = \log\left(\frac{SB_i - SA_i}{SB_i} \div \frac{DB_i - DA_i}{DB_i}\right) \tag{10.7}$$

   Suggest rationales for considering each of these possibilities.

2. Which (if any) of the above random variables appear to be normally distributed? Which appear to be symmetrically distributed?

3. Does caprotil affect systolic and diastolic blood pressure differently? Write a brief report that summarizes your investigation and presents your conclusion(s).

Chapter 11

2-Sample Location Problems

Thus far, in Chapters 9 and 10, we have studied inferences about a single population. In contrast, the present chapter is concerned with comparing *two* populations with respect to some measure of centrality, typically the population mean or the population median. Specifically, we assume the following:

1. $X_1, \ldots, X_{n_1} \sim P_1$ and $Y_1, \ldots, Y_{n_2} \sim P_2$ are continuous random variables. The X_i and the Y_j are mutually independent. In particular, there is no natural pairing of X_1 with Y_1, X_2 with Y_2, etc.

2. P_1 has location parameter θ_1 and P_2 has location parameter θ_2. We assume that comparisons of θ_1 and θ_2 are meaningful. For example, we might compare population means, $\theta_1 = \mu_1 = EX_i$ and $\theta_2 = \mu_2 = EY_j$, or population medians, $\theta_1 = q_2(X_i)$ and $\theta_2 = q_2(Y_j)$, but we would not compare the mean of one population and the median of another population. The *shift parameter*, $\Delta = \theta_1 - \theta_2$, measures the difference in population location.

3. We observe random samples $\vec{x} = \{x_1, \ldots, x_{n_1}\}$ and $\vec{y} = \{y_1, \ldots, y_{n_2}\}$, from which we attempt to draw inferences about Δ. Notice that we do *not* assume that $n_1 = n_2$.

The same four questions that we posed at the beginning of Chapter 10 can be asked here. What distinguishes 2-sample problems from 1-sample problems is the number of populations from which the experimental units were drawn. The prototypical case of a 2-sample problem is the case of a treatment population and a control population. We begin by considering some examples.

Example 11.1 A researcher investigated the effect of Alzheimer's disease (AD) on ability to perform a confrontation naming task. She recruited 60 mildly demented AD patients and 60 normal elderly control subjects. The control subjects resembled the AD patients in that the two groups had comparable mean ages, years of education, and (estimated) IQ scores; however, the control subjects were not individually matched to the AD patients. Each person was administered the Boston Naming Test (BNT), on which higher scores represent better performance. For this experiment:

1. An experimental unit is a person.

2. The experimental units belong to one of two populations: AD patients or normal elderly persons.

3. One measurement (score on BNT) is taken on each experimental unit.

4. Let X_i denote the BNT score for AD patient i. Let Y_j denote the BNT score for control subject j. Then $X_1, \ldots, X_{n_1} \sim P_1$, $Y_1, \ldots, Y_{n_2} \sim P_2$, and we are interested in drawing inferences about $\Delta = \theta_1 - \theta_2$. Notice that $\Delta < 0$ if and only if $\theta_1 < \theta_2$. Thus, to document that AD compromises confrontation naming ability, we might test $H_0 : \Delta \geq 0$ against $H_1 : \Delta < 0$.

Example 11.2 A drug is supposed to lower blood pressure. To determine if it does, $n_1 + n_2$ hypertensive patients are recruited to participate in a *double-blind* study. The patients are randomly assigned to a treatment group of n_1 patients and a control group of n_2 patients. Each patient in the treatment group receives the drug for two months; each patient in the control group receives a *placebo* for the same period. Each patient's blood pressure is measured before and after the two month period, and neither the patient nor the technician know to which group the patient was assigned. For this experiment:

1. An experimental unit is a patient.

2. The experimental units belong to one of two populations: hypertensive patients who receive the drug and hypertensive patients who receive the placebo. Notice that there are two populations despite the fact that all $n_1 + n_2$ patients were initially recruited from a single population. *Different treatment protocols create different populations.*

3. Two measurements (blood pressure before and after treatment) are taken on each experimental unit.

4. Let B_{1i} and A_{1i} denote the before and after blood pressures of patient i in the treatment group. Similarly, let B_{2j} and A_{2j} denote the before and after blood pressures of patient j in the control group. Let $X_i = B_{1i} - A_{1i}$, the decrease in blood pressure for patient i in the treatment group, and let $Y_j = B_{2j} - A_{2j}$, the decrease in blood pressure for patient j in the control group. Then $X_1, \ldots, X_{n_1} \sim P_1$, $Y_1, \ldots, Y_{n_2} \sim P_2$, and we are interested in drawing inferences about $\Delta = \theta_1 - \theta_2$. Notice that $\Delta > 0$ if and only if $\theta_1 > \theta_2$, i.e., if the decrease in blood pressure is greater for the treatment group than for the control group. Thus, a drug company required to produce compelling evidence of the drug's efficacy might test $H_0 : \Delta \leq 0$ against $H_1 : \Delta > 0$.

This chapter is divided into three sections:

11.1 If the data are assumed to be normally distributed, then we will be interested in inferences about the difference in population means. We will distinguish three cases, corresponding to what is known about the population variances.

11.2 If the data are only assumed to be continuously distributed, then we will be interested in inferences about the difference in population medians. We will assume a *shift model*, i.e., we will assume that P_1 and P_2 only differ with respect to location.

11.3 If the data are also assumed to be symmetrically distributed, then we will be interested in inferences about the difference in population centers of symmetry. If we assume symmetry, then we need not assume a shift model.

11.1 The Normal 2-Sample Location Problem

In this section we assume that

$$P_1 = \text{Normal}\left(\mu_1, \sigma_1^2\right) \quad \text{and} \quad P_2 = \text{Normal}\left(\mu_2, \sigma_2^2\right).$$

In describing inferential methods for $\Delta = \mu_1 - \mu_2$, we emphasize connections with material in Chapter 9 and Section 10.1. For example, the natural

estimator of a single normal population mean μ is the plug-in estimator $\hat{\mu}$, the sample mean, an unbiased, consistent, asymptotically efficient estimator of μ. In precise analogy, the natural estimator of $\Delta = \mu_1 - \mu_2$, the difference in populations means, is $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X} - \bar{Y}$, the difference in sample means. Because

$$E\hat{\Delta} = E\bar{X} - E\bar{Y} = \mu_1 - \mu_2 = \Delta,$$

$\hat{\Delta}$ is an unbiased estimator of Δ. It is also consistent and asymptotically efficient.

In Chapter 9 and Section 10.1, hypothesis testing and set estimation for a single population mean were based on knowing the distribution of the standardized natural estimator, a random variable of the form

$$\frac{\begin{array}{c}\text{sample} \\ \text{mean}\end{array} \; - \; \begin{array}{c}\text{hypothesized} \\ \text{mean}\end{array}}{\begin{array}{c}\text{standard deviation} \\ \text{of sample mean}\end{array}}.$$

The denominator of this random variable, often called the *standard error*, was either known or estimated, depending on our knowledge of the population variance σ^2. For σ^2 known, we learned that

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \qquad \left\{ \begin{array}{lll} \sim & \text{Normal}(0,1) & \text{if} \quad X_1, \ldots, X_n \sim \text{Normal}\left(\mu_0, \sigma^2\right) \\ \dot{\sim} & \text{Normal}(0,1) & \text{if} \quad n \text{ large} \end{array} \right\}.$$

For σ^2 unknown and estimated by S^2, we learned that

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \qquad \left\{ \begin{array}{lll} \sim & t(n-1) & \text{if} \quad X_1, \ldots, X_n \sim \text{Normal}\left(\mu_0, \sigma^2\right) \\ \dot{\sim} & \text{Normal}(0,1) & \text{if} \quad n \text{ large} \end{array} \right\}.$$

These facts allowed us to construct confidence intervals for and test hypotheses about the population mean. The confidence intervals were of the form

$$\left(\begin{array}{c}\text{sample} \\ \text{mean}\end{array} \right) \; \pm \; q \cdot \left(\begin{array}{c}\text{standard} \\ \text{error}\end{array} \right),$$

where the critical value q is the appropriate quantile of the distribution of Z or T. The tests also were based on Z or T, and the significance probabilities were computed using the corresponding distribution.

The logic for drawing inferences about two populations means is identical to the logic for drawing inferences about one population mean—we simply

replace "mean" with "difference in means" and base inferences about Δ on the distribution of

$$\frac{\begin{array}{c}\text{sample} \\ \text{difference}\end{array} - \begin{array}{c}\text{hypothesized} \\ \text{difference}\end{array}}{\begin{array}{c}\text{standard deviation} \\ \text{of sample difference}\end{array}} = \frac{\hat{\Delta} - \Delta_0}{\text{standard error}}.$$

Because $X_i \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2)$,

$$\bar{X} \sim \text{Normal}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \bar{Y} \sim \text{Normal}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Because $\bar{X}$ and $\bar{Y}$ are independent, it follows from Theorem 5.2 that

$$\hat{\Delta} = \bar{X} - \bar{Y} \sim \text{Normal}\left(\Delta = \mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

We now distinguish three cases:

1. Both σ_i are known (and possibly unequal). The inferential theory for this case is easy; unfortunately, population variances are rarely known.

2. The σ_i are unknown, but necessarily equal ($\sigma_1 = \sigma_2 = \sigma$). This case should strike the student as somewhat implausible. If the population variances are not known, then under what circumstances might we reasonably assume that they are equal? Although such circumstances do exist, the primary importance of this case is that the corresponding theory is elementary. Nevertheless, it is important to study this case because the methods derived from the assumption of an unknown common variance are widely used—and abused.

3. The σ_i are unknown and possibly unequal. This is clearly the case of greatest practical importance, but the corresponding theory is somewhat unsatisfying. The problem of drawing inferences when the population variances are unknown and possibly unequal is sufficiently notorious that it has a name: the *Behrens-Fisher problem*.

11.1.1 Known Variances

If $\Delta = \Delta_0$, then

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \text{Normal}(0, 1).$$

Given $\alpha \in (0,1)$, let q_z denote the $1-\alpha/2$ quantile of Normal$(0,1)$. We construct a $(1-\alpha)$-level confidence interval for Δ by writing

$$\begin{aligned} 1-\alpha &= P\left(|Z| < q_z\right) \\ &= P\left(|\hat{\Delta} - \Delta| < q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \\ &= P\left(\hat{\Delta} - q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \Delta < \hat{\Delta} - q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \end{aligned}$$

The desired confidence interval is

$$\hat{\Delta} \pm q_z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Example 11.3 For the first population, suppose that we know that the population standard deviation is $\sigma_1 = 5$ and that we observe a sample of size $n_1 = 60$ with sample mean $\bar{x} = 7.6$. For the second population, suppose that we know that the population standard deviation is $\sigma_2 = 2.5$ and that we observe a sample of size $n_2 = 15$ with sample mean $\bar{y} = 5.2$. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_z = \texttt{qnorm}(.975) = 1.959964 \doteq 1.96,$$

then

$$(7.6 - 5.2) \pm 1.96\sqrt{\frac{5^2}{60} + \frac{2.5^2}{15}} \doteq 2.4 \pm 1.79 = (0.61, 4.21).$$

Example 11.4 For the first population, suppose that we know that the population variance is $\sigma_1^2 = 8$ and that we observe a sample of size $n_1 = 10$ with sample mean $\bar{x} = 9.7$. For the second population, suppose that we know that the population variance is $\sigma_2^2 = 96$ and that we observe a sample of size $n_2 = 5$ with sample mean $\bar{y} = 2.6$. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_z = \texttt{qnorm}(.975) = 1.959964 \doteq 1.96,$$

then

$$(9.7 - 2.6) \pm 1.96\sqrt{\frac{8}{10} + \frac{96}{5}} \doteq 7.1 \pm 8.765 = (-1.665, 15.865).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the fact that $Z \sim \text{Normal}(0, 1)$ under H_0. Let z denote the observed value of Z. Then a natural level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |z|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|z| \geq q_z$. This test is sometimes called the 2-sample z-test.

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$z = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} \doteq 2.629.$$

Because $|2.629| > 1.96$, we reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |2.629|) = 2 * \texttt{pnorm}(-2.629) \doteq 0.008562.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$z = \frac{(9.7 - 2.6) - 0}{\sqrt{8/10 + 96/5}} \doteq 1.5876.$$

Because $|1.5876| < 1.96$, we decline to reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |1.5876|) = 2 * \texttt{pnorm}(-1.5876) \doteq 0.1124.$$

11.1.2 Unknown Common Variance

Now we assume that $\sigma_1 = \sigma_2 = \sigma$, but that the common variance σ^2 is unknown. Because σ^2 is unknown, we must estimate it. Let

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

denote the sample variance for the X_i and let

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$$

denote the sample variance for the Y_j. If we only sampled the first population, then we would use S_1^2 to estimate the first population variance, σ_1^2. Likewise, if we only sampled the second population, then we would use S_2^2 to estimate the second population variance, σ_2^2. Neither is appropriate in the present situation, as S_1^2 does not use the second sample and S_2^2 does not use the first sample. Therefore, we create a weighted average of the separate sample variances,

$$\begin{aligned} S_P^2 &= \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)} \\ &= \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2\right], \end{aligned}$$

the *pooled sample variance*. Then

$$ES_P^2 = \frac{(n_1-1)ES_1^2+(n_2-1)ES_2^2}{(n_1-1)+(n_2-1)} = \frac{(n_1-1)\sigma^2+(n_2-1)\sigma^2}{(n_1-1)+(n_2-1)} = \sigma^2,$$

so the pooled sample variance is an unbiased estimator of a common population variance. It is also consistent and asymptotically efficient for estimating a common normal variance.

Instead of

$$Z = \frac{\hat{\Delta}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{\hat{\Delta}-\Delta_0}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\sigma^2}},$$

we now rely on

$$T = \frac{\hat{\Delta}-\Delta_0}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_P^2}}.$$

The following result allows us to construct confidence intervals and test hypotheses about the shift parameter $\Delta = \mu_1 - \mu_2$.

Theorem 11.1 *If $\Delta = \Delta_0$, then $T \sim t(n_1+n_2-2)$.*

Given $\alpha \in (0,1)$, let q_t denote the $1-\alpha/2$ quantile of $t(n_1+n_2-2)$. Exploiting Theorem 11.1, a $(1-\alpha)$-level confidence interval for Δ is

$$\hat{\Delta} \pm q_t\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_P^2}.$$

Example 11.3 (continued) Now suppose that, instead of knowing population standard deviations $\sigma_1 = 5$ and $\sigma_2 = 2.5$, we observe sample standard deviations $s_1 = 5$ and $s_2 = 2.5$. The ratio of sample variances, $s_1^2/s_2^2 = 4 \neq 1$, strongly suggests that the population variances are unequal. We proceed under the assumption that $\sigma_1 = \sigma_2$ for the purpose of illustration. The pooled sample variance is

$$S_P^2 = \sqrt{\frac{59 \cdot 5^2 + 14 \cdot 2.5^2}{59 + 14}} = 21.40411.$$

To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 73) = 1.992997 \doteq 1.993,$$

then

$$(7.6 - 5.2) \pm 1.993\sqrt{\left(\frac{1}{60} + \frac{1}{15}\right) \cdot 21.40411} \doteq 2.4 \pm 2.66 = (-0.26, 5.06).$$

Example 11.4 (continued) Now suppose that, instead of knowing population variances $\sigma_1^2 = 8$ and $\sigma_2^2 = 96$, we observe sample variances $s_1^2 = 8$ and $s_2^2 = 96$. Again, the ratio of sample variances, $s_2^2/s_1^2 = 12 \neq 1$, strongly suggests that the population variances are unequal. We proceed under the assumption that $\sigma_1 = \sigma_2$ for the purpose of illustration. The pooled sample variance is

$$S_P^2 = \sqrt{\frac{9 \cdot 8 + 4 \cdot 96}{9 + 4}} = 35.07692.$$

To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 13) = 2.160369 \doteq 2.16,$$

then

$$(9.7 - 2.6) \pm 2.16\sqrt{\left(\frac{1}{10} + \frac{1}{5}\right) \cdot 35.07692} \doteq 7.1 \pm 7.01 = (0.09, 14.11).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the fact that $T \sim t(n_1 + n_2 - 2)$ under H_0. Let t denote the observed value of T. Then a natural level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |t|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|t| \geq q_t$. This test is called *Student's 2-sample t-test.*

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t = \frac{(7.6 - 5.2) - 0}{\sqrt{(1/60 + 1/15) \cdot 21.40411}} \doteq 1.797.$$

Because $|1.797| < 1.993$, we decline to reject H_0 at significance level $\alpha = .05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |1.797|) = 2 * \texttt{pt}(-1.797, 73) \doteq 0.0764684.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t = \frac{(9.7 - 2.6) - 0}{\sqrt{(1/10 + 1/5) \cdot 35.07692}} \doteq 2.19.$$

Because $|2.19| > 2.16$, we reject H_0 at significance level $\alpha = .05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |2.19|) = 2 * \texttt{pt}(-2.19, 13) \doteq 0.04747.$$

11.1.3 Unknown Variances

Now we drop the assumption that $\sigma_1 = \sigma_2$. We must then estimate each population variance separately, σ_1^2 with S_1^2 and σ_2^2 with S_2^2. Instead of

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

we now rely on

$$T_W = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Unfortunately, there is no analogue of Theorem 11.1—the exact distribution of T_W is not known.

The exact distribution of T_W appears to be intractable, but Welch (1937, 1947) argued that $T_W \dot{\sim} t(\nu)$, with

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}}.$$

Because σ_1^2 and σ_2^2 are unknown, we estimate ν by

$$\hat{\nu} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

Simulation studies have revealed that the approximation $T_W \dot{\sim} t(\hat{\nu})$ works well in practice.

Given $\alpha \in (0,1)$, let q_t denote the $1-\alpha/2$ quantile of $t(\hat{\nu})$. Using Welch's approximation, an approximate $(1-\alpha)$-level confidence interval for Δ is

$$\hat{\Delta} \pm q_t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Example 11.3 (continued) Now we estimate the unknown population variances separately, σ_1^2 by $s_1^2 = 5^2$ and σ_2^2 by $s_2^2 = 2.5^2$. Welch's approximation involves

$$\hat{\nu} = \frac{\left(\frac{5^2}{60} + \frac{2.5^2}{15}\right)^2}{\frac{(5^2/60)^2}{60-1} + \frac{(2.5^2/15)^2}{15-1}} = 45.26027 \doteq 45.26$$

degrees of freedom. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 45.26) \doteq 2.014,$$

then

$$(7.6 - 5.2) \pm 2.014\sqrt{5^2/60 + 2.5^2/15} \doteq 2.4 \pm 1.84 = (0.56, 4.24).$$

Example 11.4 (continued) Now we estimate the unknown population variances separately, σ_1^2 by $s_1^2 = 8$ and σ_2^2 by $s_2^2 = 96$. Welch's approximation involves

$$\hat{\nu} = \frac{\left(\frac{8}{10} + \frac{96}{5}\right)^2}{\frac{(8/10)^2}{10-1} + \frac{(96/5)^2}{5-1}} = 4.336931 \doteq 4.337$$

degrees of freedom. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 4.337) \doteq 2.6934,$$

then

$$(9.7 - 2.6) \pm 2.6934\sqrt{8/10 + 96/5} \doteq 7.1 \pm 13.413 = (-6.313, 20.513).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the approximation $T_W \dot{\sim} t(\hat{\nu})$ under H_0. Let t_W denote the observed value of T_W. Then a natural approximate level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|T_W| \geq |t_W|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|t_W| \geq q_t$. This test is sometimes called Welch's approximate t-test.

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t_W = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} \doteq 2.629.$$

Because $|2.629| > 2.014$, we reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T_W| \geq |2.629|) = 2 * \texttt{pt}(-2.629, 45.26) \doteq 0.011655.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t_W = \frac{(9.7 - 2.6) - 0}{\sqrt{8/10 + 96/5}} \doteq 1.4257.$$

Because $|1.4257| < 2.6934$, we decline to reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T_W| \geq |1.4257|) = 2 * \texttt{pt}(-1.4257, 4.337) \doteq 0.2218.$$

Examples 11.3 and 11.4 were carefully constructed to reveal the sensitivity of Student's 2-sample t-test to the assumption of equal population variances. Welch's approximation is good enough that we can use it to benchmark Student's test when variances are unequal. In Example 11.3, Welch's approximate t-test produced a significance probability of $\mathbf{p} \doteq 0.012$, leading us to reject the null hypothesis at $\alpha = 0.05$. Student's 2-sample t-test produced a misleading significance probability of $\mathbf{p} \doteq 0.076$, leading

us to commit a Type II error. In Example 11.4, Welch's approximate t-test produced a significance probability of $\mathbf{p} \doteq 0.222$, leading us to accept the null hypothesis at $\alpha = 0.05$. Student's 2-sample t-test produced a misleading significance probability of $\mathbf{p} \doteq 0.047$, leading us to commit a Type I error.

Evidently, Student's 2-sample t-test (and the corresponding procedure for constructing confidence intervals) should not be used unless one is convinced that the population variances are identical. The consequences of using Student's test when the population variances are unequal may be exacerbated when the sample sizes are unequal. In general:

- If $n_1 = n_2$, then $t = t_W$.
- If the population variances are (approximately) equal, then t and t_W tend to be (approximately) equal.
- If the larger sample is drawn from the population with the larger variance, then t will tend to be less than t_W. All else equal, this means that Student's test will tend to produce significance probabilities that are too large.
- If the larger sample is drawn from the population with the smaller variance, then t will tend to be greater than t_W. All else equal, this means that Student's test will tend to produce significance probabilities that are too small.
- If the population variances are (approximately) equal, then $\hat{\nu}$ will be (approximately) $n_1 + n_2 - 2$.
- It will *always* be the case that $\hat{\nu} \leq n_1 + n_2 - 2$. All else equal, this means that Student's test will tend to produce significance probabilities that are too large.

From these observations we draw the following conclusions:

1. If the population variances are unequal, then Student's 2-sample t-test may produce misleading significance probabilities.
2. If the population variances are equal, then Welch's approximate t-test is approximately equivalent to Student's 2-sample t-test. Thus, if one uses Welch's test in the situation for which Student's test is appropriate, one is not likely to be led astray.

141	148	132	138	154	142	150	146	155	158	150	140
147	148	144	150	149	145	149	158	143	141	144	144
126	140	144	142	141	140	145	135	147	146	141	136
140	146	142	137	148	154	137	139	143	140	131	143
141	149	148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138	142	149
142	137	134	144	146	147	140	142	140	137	152	145
133	138	130	138	134	127	128	138	136	131	126	120
124	132	132	125	139	127	133	136	121	131	125	130
129	125	136	131	132	127	129	132	116	134	125	128
139	132	130	132	128	139	135	133	128	130	130	143
144	137	140	136	135	126	139	131	133	138	133	137
140	130	137	134	130	148	135	138	135	138		

Table 11.1: Maximum breadth (in millimeters) of 84 skulls of Etruscan males (top) and 70 skulls of modern Italian males.

3. *Don't use Student's 2-sample t-test!* I remember how shocked I was when I first heard this advice as a first-year graduate student in a course devoted to the theory of hypothesis testing. The instructor, Erich Lehmann, one of the great statisticians of the 20th century and the author of a famous book on hypothesis testing, told us: "If you get just one thing out of this course, I'd like it to be that you should *never* use Student's 2-sample t-test."

11.2 The Case of a General Shift Family

11.3 The Symmetric Behrens-Fisher Problem

11.4 Case Study: Etruscan versus Italian Head Breadth

In a collection of essays on the origin of the Etruscan empire, N.A. Barnicott and D.R. Brothwell compared measurements on ancient and modern bones.[1]

[1]N.A. Barnicott and D.R. Brothwell (1959). The evaluation of metrical data in the comparison of ancient and modern bones. In *Medical Biology and Etruscan Origins*, edited

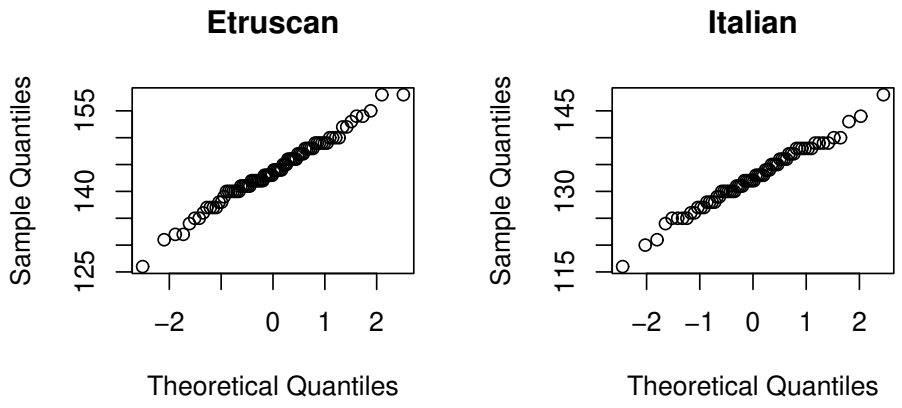

Figure 11.1: Normal probability plots of two samples of maximum skull breadth.

Measurements of the maximum breadth of 84 Etruscan skulls and 70 modern Italian skulls were subsequently reproduced as Data Set 155 in *A Handbook of Small Data Sets* and are displayed in Table 11.1. We use these data to explore the difference (if any) between Etruscan and modern Italian males with respect to head breadth. In the discussion that follows, x will denote Etruscans and y will denote modern Italians.

We begin by asking if it is reasonable to assume that maximum skull breadth is normally distributed. Normal probability plots of our two samples are displayed in Figure 11.1. The linearity of these plots conveys the distinct impression of normality. Kernel density estimates constructed from the two samples are superimposed in Figure 11.2, created by the following `R` commands:

```
> plot(density(x),type="l",xlim=c(100,180),
+ xlab="Maximum Skull Breadth",
+ main="Kernel Density Estimates")
> lines(density(y),type="l")
```

Not only do the kernel density estimates reinforce our impression of nor-

by G.E.W. Wolstenholme and C.M. O'Connor, Little, Brown & Company, p. 136.

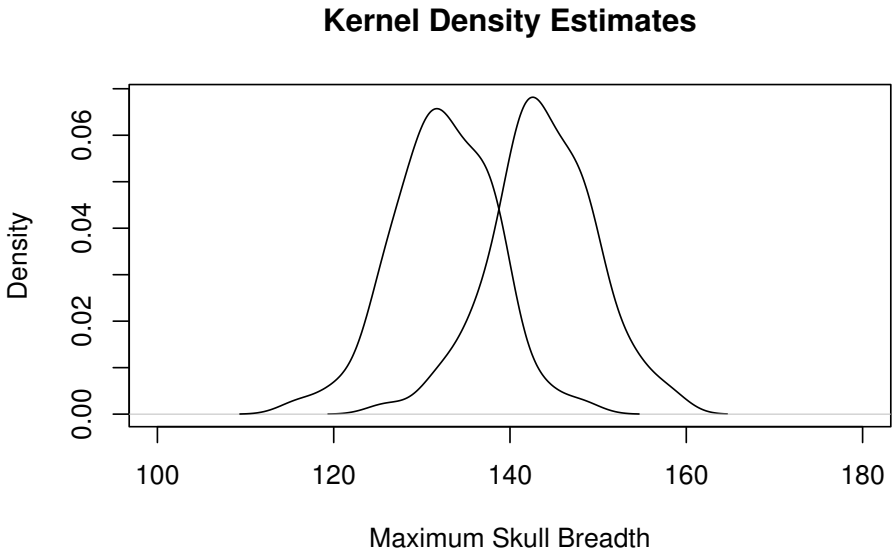

Figure 11.2: Kernel density estimates constructed from two samples of maximum skull breadth. The sample mean for the Etruscan skulls is $\bar{x} \doteq 143.8$; the sample mean for the modern Italian skulls is $\bar{y} \doteq 132.4$.

mality, they also suggest that the two populations have comparable variances. (The ratio of sample variances is $s_1^2/s_2^2 = 1.07819$.) The difference is maximum breadth between Etruscan and modern Italian skulls is nicely summarized by a shift parameter.

Now we construct a probability model. This is a 2-sample location problem in which an experimental unit is a skull. The skulls were drawn from two populations, Etruscan males and modern Italian males, and one measurement (maximum breadth) was made on each experimental unit. Let X_i denote the maximum breadth of Etruscan skull i and let Y_j denote the maximum breadth of Italian skull j. We assume that the X_i and Y_j are independent, with $X_i \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2)$. Notice that, although the sample variances are nearly equal, we do not assume that the population variances are identical. Instead, we will use Welch's approximation to construct an approximate 0.95-level confidence interval for

$\Delta = \mu_1 - \mu_2$.

Because the confidence coefficient $1 - \alpha = 0.95$, $\alpha = 0.05$. The desired confidence interval is of the form

$$\hat{\Delta} \pm q\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where q is the $1 - \alpha/2 = 0.975$ quantile of a t distribution with $\hat{\nu}$ degrees of freedom. We can easily compute these quantities in `R`. To compute $\hat{\Delta}$, the estimated shift parameter:

```
> Delta <- mean(x)-mean(y)
```

To compute the standard error:

```
> n1 <- length(x)
> n2 <- length(y)
> v1 <- var(x)/n1
> v2 <- var(y)/n2
> se <- sqrt(v1+v2)
```

To compute $\hat{\nu}$, the estimated degrees of freedom:

```
> nu <- (v1+v2)^2/(v1^2/(n1-1)+v2^2/(n2-1))
```

To compute q, the desired quantile:

```
> q <- qt(.975,df=nu)
```

Finally, to compute the lower and upper endpoints of the desired confidence interval:

```
> lower <- Delta-q*se
> upper <- Delta+q*se
```

These calculations result in a 0.95-level confidence interval for $\Delta = \mu_1 - \mu_2$ of $(9.459782, 13.20212)$, so that we can be fairly confident that the maximum breadth of Etruscan male skulls is, on average, roughly a centimeter greater than the maximum breadth of modern Italian male skulls.

11.5 Exercises

Problem Set A

1. We have been using various mathematical symbols in our study of 1- and 2-sample location problems. Each of the symbols listed below is used to represent a real number. State which of the following statements applies to each symbol:

   i. The real number represented by this symbol is an unknown population parameter.

   ii. The real number represented by this symbol is calculated from the observed data.

   iii. The real number represented by this symbol is specified by the experimenter.

   Here are the symbols:

   $$\mu \quad \mu_0 \quad \bar{x} \quad s^2 \quad t \quad \alpha \quad \Delta \quad \Delta_0 \quad \mathbf{p} \quad \hat{\nu}$$

2. Assume that $X_1, \ldots, X_{10} \sim \text{Normal}(\mu_1, \sigma_1^2)$ and that $Y_1, \ldots, Y_{20} \sim \text{Normal}(\mu_2, \sigma_2^2)$. None of the population parameters are known. Let $\Delta = \mu_1 - \mu_2$. To test $H_0 : \Delta \geq 0$ versus $H_1 : \Delta < 0$ at significance level $\alpha = 0.05$, we observe samples $\vec{x}$ and $\vec{y}$.

   (a) What test should be used in this situation? If we observe $\vec{x}$ and $\vec{y}$ that result in $\bar{x} = -0.82$, $s_1 = 4.09$, $\bar{y} = 1.39$, and $s_2 = 1.22$, then what is the value of the test statistic?

   (b) If we observe $\vec{x}$ and $\vec{y}$ that result in $s_1 = 4.09$, $s_2 = 1.22$,, and a test statistic value of 1.76, then which of the following `R` expressions best approximates the significance probability?

   i. `2*pnorm(-1.76)`

   ii. `pt(-1.76,df=28)`

   iii. `pt(1.76,df=10)`

   iv. `pt(-1.76,df=10)`

   v. `2*pt(1.76,df=28)`

   (c) True of False: if we observe $\vec{x}$ and $\vec{y}$ that result in a significance probability of $\mathbf{p} = 0.96$, then we should reject the null hypothesis.

Problem Set B Each of the following scenarios can be modelled as a 1- or 2-sample location problem. For 1-sample problems, let X_i denote the random variables of interest and let $\mu = EX_i$. For 2-sample problems, let X_i and Y_j denote the random variables of interest; let $\mu_1 = EX_i$, $\mu_2 = EY_j$, and $\Delta = \mu_1 - \mu_2$. For each scenario, you should answer/do the following:

(a) What is the experimental unit?

(b) From how many populations were the experimental units drawn? Identify the population(s). How many units were drawn from each population? Is this a 1- or a 2-sample problem?

(c) How many measurements were taken on each experimental unit? Identify them.

(d) Define the parameter(s) of interest for this problem. For 1-sample problems, this should be μ; for 2-sample problems, this should be Δ.

(e) State appropriate null and alternative hypotheses.

Here are the scenarios:

1. A mathematics/education concentrator theorizes that learning mathematics and statistics is sometimes impeded by the widespread use of odd symbols like α, χ, and ω. She reasons that, if her theory is correct, then students who belong to sororities and fraternities—who she presumes are more familiar with Greek letters—should have an easier time learning the mathematical subjects that use such symbols. To investigate, she obtains a list of all William & Mary students who are enrolled in Math 111 (calculus) and a list of all William & Mary students who belong to a sorority or fraternity. She uses this information to choose (at random) 20 calculus students who do belong to a sorority or fraternity and 20 calculus students who do not. She persuades each of these students to take a calculus quiz, specially designed to use lots of Greek letters. How might she use the resulting data to test her theory? (Respond to (a)–(e) above.)

2. Umberto theorizes that living with a dog diminishes depression in the elderly, here defined as more than 70 years of age. To investigate his theory, he recruits 15 single elderly men who own dogs and 15 single elderly men who do not own any pets. The Hamilton instrument for

measuring depressive tendency is administered to each subject. High scores indicate depression. How might Umberto use the resulting data to test his theory? (Respond to (a)–(e) above.)

3. The William & Mary women's tennis team uses championship balls in their matches and less expensive practice balls in their team practices. The players have formed a strong impression that the practice balls do not wear as well as the championship balls, i.e., that the practice balls lose their bounce more quickly than the championship balls. To investigate this perception, Nina and Delphine conceive the following experiment. Before one practice, the team opens new cans of championship balls and practice balls, which they then use for that day's practice. After practice, Nina and Delphine randomly select 10 of the used championship balls and 10 of the used practice balls. They drop each ball from a height of 1 meter and measure the height of its first bounce. How might Nina and Delphine test the team's impression that practice balls do not wear as well as championship balls? (Respond to (a)–(e) above.)

4. A political scientist theorizes that women tend to be more opposed to military intervention than do men. To investigate this theory, he devises an instrument on which a subject responds to several recent U.S. military interventions on a 5-point Likert scale (1="strongly support,"...,5="strongly oppose"). A subject's score on this instrument is the sum of his/her individual responses. The scientist randomly selects 50 married couples in which neither spouse has a registered party affiliation and administers the instrument to each of the 100 individuals so selected. How might he use his results to determine if his theory is correct? (Respond to (a)–(e) above.)

5. A shoe company claims that wearing its racing flats will typically improve one's time in a 10K road race by more than 30 seconds. A running magazine sponsors an event to test this claim. It arranges for 120 runners to enter two road races, held two weeks apart on the same course. For the second race, each of these runners is supplied with the new racing flat. How might the race results be used to determine the validity of the shoe company's claim? (Respond to (a)–(e) above.)

6. Susan theorizes that impregnating wood with an IGR (insect growth regulator) will reduce wood consumption by termites. To investigate

this theory, she impregnates 60 wood blocks with a solvent containing the IGR and 60 wood blocks with just the solvent. Each block is weighed, then placed in a separate container with 100 ravenous termites. After two weeks, she removes the blocks and weighs them again to determine how much wood has been consumed. How might Susan use her results to determine if her theory is correct? (Respond to (a)–(e) above.)

7. To investigate the effect of swing dancing on cardiovascular fitness, an exercise physiologist recruits 20 couples enrolled in introductory swing dance classes. Each class meets once a week for ten weeks. Participants are encouraged to go out dancing on at least two additional occasions each week. In general, lower resting pulses are associated with greater cardiovascular fitness. Accordingly, each participant's resting pulse is measured at the beginning and at the end of the ten-week class. How might the resulting data be used to determine if swing dancing improves cardiovascular fitness? (Respond to (a)–(e) above.)

8. It is thought that Alzheimer's disease (AD) impairs short-term memory more than it impairs long-term memory. To test this theory, a psychologist studied 60 mildly demented AD patients and 60 normal elderly control subjects. Each subject was administered a short-term and a long-term memory task. On each task, high scores are better than low scores. How might the psychologist use the resulting task scores to determine if the theory is correct? (Respond to (a)–(e) above.)

9. According to an article in *Newsweek* (May 10, 2004, page 89), recent "studies have shown consistently that women are better than men at reading and responding to subtle cues about mood and temperament." Some psychologists believe that such differences can be explained in part by biological differences between male and female brains. One such psychologist conducts a study in which day-old babies are shown three human faces and three mechanical objects. The time that the baby stares at each face/object is recorded. Of interest is how much time the baby spends staring at faces versus how much time the baby spends staring at objects. The psychologist's theory predicts that this comparison will differ by sex, with female babies preferring faces to objects to a greater extent than do male babies. How might the psychologist use his results to determine if his theory is correct? (Respond to (a)–(e) above.)

Problem Set C In the early 1960s, the Western Collaborative Group Study investigated the relation between behavior and risk of coronary heart disease in middle-aged men. Type A behavior is characterized by urgency, aggression and ambition; Type B behavior is noncompetitive, more relaxed and less hurried. The following data, which appear in Table 2.1 of Selvin (1991) and Data Set 47 in *A Handbook of Small Data Sets*, are the cholesterol measurements of 20 heavy men of each behavior type. (In fact, these 40 men were the heaviest in the study. Each weighed at least 225 pounds.) We consider whether or not they provide evidence that heavy Type A men have higher cholesterol levels than heavy Type B men.

Cholesterol Levels for Heavy Type A Men									
233	291	312	250	246	197	268	224	239	239
254	276	234	181	248	252	202	218	212	325

Cholesterol Levels for Heavy Type B Men									
344	185	263	246	224	212	188	250	148	169
226	175	242	252	153	183	137	202	194	213

1. Respond to (a)–(e) in Problem Set B.
2. Does it seem reasonable to assume that the samples $\vec{x}$ and $\vec{y}$, the observed values of $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, were drawn from normal distributions? Why or why not?
3. Assume that the X_i and the Y_j are normally distributed.
   (a) Test the null hypothesis derived above using Welch's approximate t-test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?
   (b) Construct a (2-sided) confidence interval for Δ with a confidence coefficient of approximately 0.90.

Problem Set D Researchers measured urinary β-thromboglobulin excretion in 12 diabetic patients and 12 normal control subjects, reporting their findings in *Thrombosis and Haemostasis*. The following measurements are Data Set 313 in *A Handbook of Small Data Sets*:

Normal	4.1	6.3	7.8	8.5	8.9	10.4
	11.5	12.0	13.8	17.6	24.3	37.2
Diabetic	11.5	12.1	16.1	17.8	24.0	28.8
	33.9	40.7	51.3	56.2	61.7	69.2

1. Do these measurements appear to be samples from symmetric distributions? Why or why not?

2. Both samples of positive real numbers appear to be drawn from distributions that are skewed to the right, i.e., the upper tail of the distribution is longer than the lower tail of the distribution. Often, such distributions can be symmetrized by applying a suitable data transformation. Two popular candidates are:

   (a) The natural logarithm: $u_i = \log(x_i)$ and $v_j = \log(y_j)$.

   (b) The square root: $u_i = \sqrt{x_i}$ and $v_j = \sqrt{y_j}$.

   Investigate the effect of each of these transformations on the above measurements. Do the transformed measurements appear to be samples from symmetric distributions? Which transformation do you prefer?

3. Do the transformed measurements appear to be samples from normal distributions? Why or why not?

4. The researchers claimed that diabetic patients have increased urinary β-thromboglobulin excretion. Assuming that the transformed measurements are samples from normal distributions, how convincing do you find the evidence for their claim?

Problem Set E

1. Chemistry lab partners Arlen and Stuart collaborated on an experiment in which they measured the melting points of 20 specimens of two types of sealing wax. Twelve of the specimens were of one type (A); eight were of the other type (B). Each student then used Welch's approximate t-test to test the null hypothesis of no difference in mean melting point between the two methods:

   - Arlen applied Welch's approximate t-test to the original melting points, which were measured in degrees Fahrenheit.
   - Stuart first converted each melting point to degrees Celsius (by subtracting 32, then multiplying by 5/9), then applied Welch's approximate t-test to the converted melting points.

Comment on the potential differences between these two analyses. In particular, is it *True* or *False* that (ignoring round-off error) Arlen and Stuart will obtain identical significance probabilities? Please justify your comments.

2. A graduate student in ornithology would like to determine if created marshes differ from natural marches in their appeal to avian communities. He plans to observe $n_1 = 9$ natural marshes and $n_2 = 9$ created marshes, counting the number of red-winged blackbirds per acre that inhabit each marsh. His thesis committee wants to know how much he thinks he will be able to learn from this experiment.

   Let X_i denote the number of blackbirds per acre in natural marsh i and let Y_j denote the number of blackbirds per acre in created marsh j. In order to respond to his committee, the student makes the simplifying assumptions that $X_i \sim \text{Normal}(\mu_1, \sigma^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma^2)$. He estimates that $\text{iqr}(X_i) = \text{iqr}(Y_j) = 10$. Calculate L, the length of the 0.90-level confidence interval for $\Delta = \mu_1 - \mu_2$ that he can expect to construct.

3. A film buff has formed the vague impression that movies tend to be longer than they used to be. Are they really longer? Or do they just *seem* longer? To investigate, he randomly samples U.S. feature films made in 1956 and U.S. feature films made in 1996, obtaining the following results:

Year	Title	Minutes
1956	*Accused of Murder*	74
	Away All Boats	114
	Baby Doll	114
	The Bold and the Brave	87
	Come Next Spring	92
	The Flaming Teen-Age	55
	Gun Girls	67
	Helen of Troy	118
	The Houston Story	79
	Patterns	83
	The Price of Fear	79
	The Revolt of Mamie Stover	92
	Written on the Wind	99
	The Young Guns	87
1996	*$40,000*	70
	Barb Wire	98
	Breathing Room	90
	Daddy's Girl	95
	Ed's Next Move	88
	From Dusk to Dawn	108
	Galgameth	110
	The Glass Cage	96
	Kissing a Dream	91
	Love & Sex etc.	88
	Love is All There Is	120
	Making the Rules	96
	Spirit Lost	90
	Work	90

Do these data provide convincing evidence that 1996 movies are longer than 1956 movies? Compute a significance probability that may be used to encourage or discourage the film buff's impression. Explain how this number should be interpreted. Identify and defend any assumptions that you made in your calculations.

Chapter 12

k-Sample Location Problems

Now we generalize our study of location problems from two to $k \geq 3$ populations. Again we are concerned with comparing the populations with respect to some measure of centrality, typically the population mean or the population median. We designate the populations by $P_1, \ldots, P_k$ and the corresponding sample sizes by $n_1, \ldots, n_k$. Our bookkeeping will be facilitated by the use of double subscripts, e.g.,

$$\begin{aligned} X_{11}, \ldots, X_{1n_1} &\sim P_1, \\ X_{21}, \ldots, X_{2n_2} &\sim P_2, \\ &\vdots \\ X_{k1}, \ldots, X_{kn_k} &\sim P_k. \end{aligned}$$

These expressions can be summarized succinctly by writing

$$X_{ij} \sim P_i.$$

We assume the following:

1. The X_{ij} are mutually independent continuous random variables.
2. P_i has location parameter θ_i, e.g., $\theta_i = \mu_i = EX_{ij}$ or $\theta_i = q_2(X_{ij})$.
3. We observe random samples $\vec{x}_i = \{x_{i1}, \ldots, x_{in_i}\}$, from which we attempt to draw inferences about $(\theta_1, \ldots, \theta_k)$. In general, we do *not* assume that $n_1 = \cdots = n_k$. However, certain procedures do require equal sample sizes. Furthermore, certain procedures that can be used with unequal sample sizes are greatly simplified when the sample sizes are equal.

The same four questions that we posed at the beginning of Chapter 10 and asked in Chapters 10–11 can be asked here. What distinguishes k-sample problems from 1-sample and 2-sample problems is the number of populations from which the experimental units were drawn. The prototypical case of a k-sample problem is the case of several treatment populations.

One may wonder why we distinguish between $k = 2$ and $k \geq 3$ populations. In fact, many methods for k-sample problems can be applied to 2-sample problems, in which case they often simplify to methods studied in Chapter 11. However, many issues arise with $k \geq 3$ populations that do not arise with two populations, so the problem of comparing more than two location parameters is considerably more complicated than the problem of comparing only two. For this reason, our study of k-sample location problems will be less comprehensive than our previous studies of 1-sample and 2-sample location problems.

12.1 The Case of a Normal Shift Family

In this section we assume that $P = \text{Normal}(\mu_i, \sigma^2)$. This is sometimes called the fixed effects model for the oneway analysis of variance (ANOVA). Notice that we are assuming that each normal population has the same variance. Recall that we criticized the assumption of equal variances for the normal 2-sample problem. In that setting, however, Welch's approximate t-test provides a viable alternative that is available in many popular statistical software packages. In the more complicated setting of k normal populations, the assumption of equal variances (sometimes called the assumption of *homoscedasticity*) is fairly standard, if only because it is less clear how to proceed when the variances are unequal. The problem of unequal variances is discussed in Section 12.3.

12.1.1 The Fundamental Null Hypothesis

The fundamental problem of the analysis of variance is the problem of testing the null hypothesis that all of the population means are the same, i.e.,

$$H_0 : \mu_1 = \cdots = \mu_k, \tag{12.1}$$

against the alternative hypothesis that they are not all the same. Notice that the statement that the population means are not identical does *not* imply that each population mean is distinct. For example, if $\mu_1 = \mu_2 = 1.5$

and $\mu_3 = 2.2$, then H_0 is false. We stress that the analysis of variance is concerned with inferences about means, not variances.

To motivate our test of H_0, we formulate another null hypothesis that is equivalent to H_0. First, let

$$N = \sum_{i=1}^{k} n_i$$

denote the sum of the sample sizes and let

$$\bar{\mu}_\cdot = \sum_{i=1}^{k} \frac{n_i}{N} \mu_i$$

denote the *population grand mean*. The population grand mean is a weighted average of the individual population means, each population weighted in proportion to how many of the observations were drawn from it. If H_0 is true, then $\mu_1 = \cdots = \mu_k$ have a common value, say μ, and the population grand mean equals that common value:

$$\bar{\mu}_\cdot = \sum_{i=1}^{k} \frac{n_i}{N} \mu = \frac{\mu}{N} \sum_{i=1}^{k} n_i = \mu.$$

Next we introduce a quantity that measures how nearly the individual population means equal the population grand mean. Let

$$\gamma = \sum_{i=1}^{k} n_i \left(\mu_i - \bar{\mu}_\cdot\right)^2. \tag{12.2}$$

Notice that $\gamma \geq 0$ and that $\gamma = 0$ if and only if each $\mu_i = \bar{\mu}_\cdot$. But each $\mu_i = \bar{\mu}_\cdot$ if and only if each individual mean assumes a common value, which occurs if and only if the individual means are identical. Thus, H_0 is equivalent to the null hypothesis

$$H_0' : \gamma = 0,$$

which is to be tested against the alternative hypothesis

$$H_1' : \gamma > 0.$$

12.1.2 Testing the Fundamental Null Hypothesis

The idea that underlies our test is to estimate γ and reject H_0' when the estimate is sufficiently larger than zero. To estimate γ, we need only estimate

the population means that appear in (12.2). The individual sample means,

$$\bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij},$$

are unbiased estimators of the individual population means, and the sample grand mean,

$$\bar{X}_{\cdot\cdot} = \sum_{i=1}^{k} \frac{n_i}{N} \bar{X}_{i\cdot} = \sum_{i=1}^{k} \frac{n_i}{N} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \right) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

is an unbiased estimator of the population grand mean. Hence, a natural estimator of γ is the *between-groups* or *treatment* sum of squares,

$$SS_B = \sum_{i=1}^{k} n_i \left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} \right)^2,$$

the variation of the individual sample means about the sample grand mean. A useful formula for computing the observed value of SS_B from the observed values of the individual sample means is

$$ss_B = \sum_{i=1}^{k} n_i \bar{x}_{i\cdot}^2 - \frac{1}{N} \left(\sum_{i=1}^{k} n_i \bar{x}_{i\cdot} \right)^2.$$

What remains is to determine when SS_B is "sufficiently larger than zero." We consider two cases, depending on whether or not the common population variance σ^2 is known.

Known Population Variance

Situations in which σ^2 is known are rarely encountered, but it is useful to consider how to proceed in this case. Here is the key fact that we require:

Theorem 12.1 *Under the fundamental null hypothesis (12.1), the random variable*

$$SS_B/\sigma^2 \sim \chi^2(k-1),$$

where $\chi^2(\nu)$ denotes the chi-squared distribution with ν degrees of freedom, introduced in Section 5.5. The quantity $k-1$ is the between-groups degrees of freedom.

Theorem 12.1 suggests a way to determine whether or not SS_B is "sufficiently larger than zero." Under H_0,

$$P\left(SS_B \geq q\right) = P\left(SS_B/\sigma^2 \geq q/\sigma^2\right) = P\left(Y \geq q/\sigma^2\right),$$

where $Y \sim \chi^2(k-1)$; hence, we can use the chi-squared distribution to compute significance probabilities and/or critical values.

Example 12.1 Suppose that we draw samples of $n_1 = 20$, $n_2 = 25$, and $n_3 = 30$ observations from normal populations with unknown means and common variance $\sigma^2 = 9$, obtaining sample means of $\bar{x}_1 = 1.489$, $\bar{x}_2 = 1.712$, and $\bar{x}_3 = 3.082$. To test the fundamental null hypothesis that the individual population means are identical, we first compute $N = 20 + 25 + 30 = 75$ and evaluate SS_B, obtaining

$$\begin{aligned} ss_B &= \left(20 \cdot 1.489^2 + 25 \cdot 1.712^2 + 30 \cdot 3.082^2\right) - \\ &\quad \left(20 \cdot 1.489 + 25 \cdot 1.712 + 30 \cdot 3.082\right)^2 / 75 \\ &\doteq 39.402. \end{aligned}$$

Now we use the `R` function `pchisq` to compute a significance probability **p**:

```
> 1-pchisq(39.402/9,df=2)
[1] 0.1120287
```

For conventional levels of significance, $\mathbf{p} > 0.10$ is too large to warrant rejecting the null hypothesis.

Unknown Population Variance

Now we consider the more realistic case of an unknown population variance. Our development will mimic the case of a known population variance, but it is complicated by the need to estimate σ^2. Recall that, in Section 11.1.2, we estimated the unknown common population variance of $k = 2$ normal populations with the pooled sample variance,

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)},$$

where S_i^2 is the sample variance for sample i. This procedure is easily extended to the present case of $k \geq 3$ by defining the pooled sample variance

as

$$\begin{aligned} S_P^2 &= \frac{(n_1-1)S_1^2+\cdots+(n_k-1)S_k^2}{(n_1-1)+\cdots+(n_2-1)} \\ &= \frac{1}{n_1+\cdots+n_k-k}\sum_{i=1}^{k}(n_i-1)\,S_i^2 \\ &= \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2. \end{aligned}$$

As in the case of $k=2$,

$$\begin{aligned} ES_P^2 &= \frac{(n_1-1)ES_1^2+\cdots+(n_k-1)ES_k^2}{(n_1-1)+\cdots+(n_k-1)} \\ &= \frac{(n_1-1)\sigma^2+\cdots(n_k-1)\sigma^2}{(n_1-1)+\cdots+(n_k-1)}=\sigma^2, \end{aligned}$$

so the pooled sample variance is an unbiased estimator of a common population variance. It is also consistent and asymptotically efficient for estimating a common normal variance.

In the previous case of a known population variance, our statistic for testing the fundamental null hypothesis was SS_B/σ^2. In the present case of an unknown population variance, we estimate σ^2 with S_P^2. Our test statistic will turn out to be SS_B/S_P^2 multiplied by a constant.

In order to simplify the formulas that follow, we multiply S_P^2 by $N-k$, obtaining the *within-groups* or *error* sum of squares

$$SS_W=(N-k)S_P^2=\sum_{i=1}^{k}(n_i-1)\,S_i^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\cdot}\right)^2.$$

In contrast to SS_B, which measures the variation of the individual sample means about the sample grand mean, SS_W measures the variations of the individual observations about the corresponding sample means. For completeness, we also define the *total* sum of squares,

$$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{\cdot\cdot}\right)^2,$$

which measures the variation of the individual observations about the sample grand mean.

There is a beautiful relationship between SS_B, SS_W, and SS_T, viz.,

Theorem 12.2 $SS_B + SS_W = SS_T$

This formula turns out to be a corollary of the Pythagorean Theorem in N-dimensional Euclidean space! (In Section 14.2, we will explore a similar formula in greater detail.) The reason that our method for testing the fundamental null hypothesis is called the analysis of variance is that the method relies on decomposing total squared error into squared error between groups and squared error within groups. This elegant—and extremely useful—decomposition is only possible when we use *squared* error.

The quantities SS_B, SS_W, and SS_T are random variables. The following facts, which subsume Theorem 12.1, summarize the statistical behavior of these random variables.

Theorem 12.3 *The random variable*

$$SS_T/\sigma^2 \sim \chi^2(N-1).$$

The quantity $N-1$ is the total degrees of freedom.

Under the fundamental null hypothesis (12.1), SS_B and SS_W are independent random variables and

$$\begin{aligned} SS_B/\sigma^2 &\sim \chi^2(k-1), \\ SS_W/\sigma^2 &\sim \chi^2(N-k). \end{aligned}$$

The quantity $k-1$ is the between-groups degrees of freedom and the quantity $N-k$ is the within-groups degrees of freedom.

We have already remarked that the random variable

$$\frac{SS_B}{S_P^2} = \frac{SS_B}{SS_W/(N-k)}$$

would seem to be a natural statistic for testing the fundamental null hypothesis. Although sound in theory, this approach fails in practice because the distribution of SS_B/S_P^2 is not tractable. Fortunately, this approach can be salvaged by a trivial modification. Applying the definition of Fisher's F distribution in Section 5.5 to the independent χ^2 random variables SS_B/σ^2 and SS_W/σ^2, we discover

Corollary 12.1 *Under the fundamental null hypothesis (12.1),*

$$F = \frac{\frac{SS_B}{\sigma^2}/(k-1)}{\frac{SS_W}{\sigma^2}/(N-k)} = \frac{SS_B/(k-1)}{SS_W/(N-k)} \sim F(k-1, N-k),$$

where $F(\nu_1, \nu_2)$ denotes Fisher's F distribution with ν_1 and ν_2 degrees of freedom.

The random variable F is the desired test statistic; notice that

$$F = \frac{SS_B/(k-1)}{SS_W/(N-k)} = \frac{1}{k-1}\frac{SS_B}{S_P^2}.$$

Appealing to Corollary 12.1, we see that the ANOVA F-test of the fundamental null hypothesis of equal population means is to reject H_0 at significance level α if and only if the significance probability

$$\mathbf{p} = P(Y \geq f) \leq \alpha,$$

where f denotes the observed value of F and $Y \sim F(k-1, N-k)$. Of course, we can also formulate the test using critical values instead of significance probabilities, in which case we reject H_0 at significance level α if and only if $f \geq q$, where q is the $1-\alpha$ quantile of the $F(k-1, N-k)$ distribution.

Example 12.2 Suppose that we draw samples of $n_1 = 25$, $n_2 = 20$, and $n_3 = 20$ observations from normal populations with unknown means and unknown common variance, obtaining the following sample quantities:

	$i=1$	$i=2$	$i=3$
n_i	25	20	20
$\bar{x}_{i\cdot}$	9.783685	10.908170	15.002820
s_i^2	29.89214	18.75800	51.41654

To test the null hypothesis of equal population means at significance level $\alpha = 0.05$, we begin by computing the observed values of SS_B and SS_W, obtaining $ss_B \doteq 322.4366$ and

$$ss_W = (25-1)\cdot 29.89214 + (20-1)\cdot 18.75800 + (20-1)\cdot 51.41654 \doteq 2050.7280.$$

It follows that the observed value of the test statistic is

$$f = \frac{ss_B/(k-1)}{ss_W/(N-k)} \doteq \frac{322.4366/2}{2050.7280/62} \doteq 4.874141.$$

Now we use the `R` function `pf` to compute a significance probability $\mathbf{p}$:

```
> 1-pf(4.874141,df1=2,df2=62)
[1] 0.01081398
```

Because $\mathbf{p} < \alpha$, we reject the null hypothesis. Equivalently, we might use the `R` function `qf` to compute a critical value q:

```
> qf(1-.05,df1=2,df2=62)
[1] 3.145258
```

Because $f > q$, we reject the null hypothesis.

The information related to an ANOVA F-test is usually collected in an ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test Statistic	Significance Probability
Between	SS_B	$k-1$	MS_B	F	$\mathbf{p}$
Within	SS_W	$N-k$	$MS_W = S_P^2$		
Total	SS_T	$N-1$			

Note that we have introduced new notation for the *mean squares*, $MS_B = SS_B/(k-1)$ and $MS_W = SS_W/(N-k)$, allowing us to write $F = MS_B/MS_W$. It is also helpful to examine $R^2 = SS_B/SS_T$, the proportion of total variation "explained" by differences in the sample means.

Example 12.2 (continued) For the ANOVA performed in Example 12.2, the ANOVA table is

Source	SS	df	MS	F	$\mathbf{p}$
Between	322.4366	2	161.21830	4.874141	0.01081398
Within	2050.7280	62	33.07625		
Total	2373.1640	64			

The proportion of total variation explained by differences in the sample means is $322.4366/2373.1640 \doteq 0.1358678$. Thus, although there is sufficient variation between the sample means for us to infer that the population means are not identical, this variation accounts for a fairly small proportion of the total variation in the data.

12.1.3 Planned Comparisons

Rejecting the fundamental null hypothesis of equal population means leaves numerous alternatives. Typically, a scientist would like to say more than simply "$H_0 : \mu_1 = \cdots = \mu_k$ is false." Concluding that the population

means are not identical naturally invites investigation of how they differ. Sections 12.1.3 and 12.1.4 describe several useful inferential procedures for performing more elaborate comparisons of population means. Section 12.1.3 describes two procedures that are appropriate when the scientist has determined specific comparisons of interest *in advance of the experiment*. For reasons that will become apparent, this is the preferred case. However, it is often the case that a specific comparison occurs to a scientist *after examining the results of the experiment*. Although statistical inference in such cases is rather tricky, a variety of procedures for *a posteriori* inference have been developed. Two such procedures are described in Section 12.1.4.

Inspired by K.A. Brownlee's classic statistics text,[1] we motivate the concept of a *planned comparison* by considering a famous physics experiment.

Example 12.3 Heyl (1930) attempted to determine the gravitational constant using $k = 3$ different materials—gold, platinum, and glass. It seems natural to ask not just if the three materials lead to identical determinations of the gravitational constant, by testing $H_0 : \mu_1 = \mu_2 = \mu_3$, but also to ask:

1. If glass differs from the two heavy metals, by testing

$$H_0 : \frac{\mu_1 + \mu_2}{2} = \mu_3 \quad \text{vs.} \quad H_1 : \frac{\mu_1 + \mu_2}{2} \neq \mu_3,$$

or, equivalently,

$$H_0 : \mu_1 + \mu_2 = 2\mu_3 \quad \text{vs.} \quad H_1 : \mu_1 + \mu_2 \neq 2\mu_3,$$

or, equivalently,

$$H_0 : \mu_1 + \mu_2 - 2\mu_3 = 0 \quad \text{vs.} \quad H_1 : \mu_1 + \mu_2 - 2\mu_3 \neq 0,$$

or, equivalently,

$$H_0 : \theta_1 = 0 \quad \text{vs.} \quad H_1 : \theta_1 \neq 0,$$

where $\theta_1 = \mu_1 + \mu_2 - 2\mu_3$.

2. If the two heavy metals differ from each other, by testing

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2,$$

[1] K.A. Brownlee, *Statistical Theory and Methodology in Science and Engineering, Second Edition*, John Wiley & Sons, 1965.

or, equivalently,

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0,$$

or, equivalently,

$$H_0 : \theta_2 = 0 \quad \text{vs.} \quad H_1 : \theta_2 \neq 0,$$

where $\theta_2 = \mu_1 - \mu_2$.

Notice that both of the planned comparisons proposed in Example 12.3 have been massaged into testing a null hypothesis of the form $\theta = 0$. For this construction to make sense, θ must have a special structure, which statisticians identify as a *contrast*.

Definition 12.1 *A contrast is a linear combination (weighted sum) of the k population means,*

$$\theta = \sum_{i=1}^{k} c_i \mu_i,$$

for which $\sum_{i=1}^{k} c_i = 0$.

Example 12.3 (continued) In the contrasts suggested previously,

1. $\theta_1 = 1 \cdot \mu_1 + 1 \cdot \mu_2 + (-2) \cdot \mu_3$ and $1 + 1 - 2 = 0$; and
2. $\theta_2 = 1 \cdot \mu_1 + (-1) \cdot \mu_2 + 0 \cdot \mu_3$ and $1 - 1 + 0 = 0$.

We usually identify different contrasts by their coefficients, e.g., $c = (1, 1, -2)$ or $c = (1, -1, 0)$.

The methods of Section 12.1.2 are easily extended to the problem of testing a single contrast, $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. In Definition 12.1, each population mean μ_i can be estimated by the unbiased estimator $\bar{X}_{i\cdot}$; hence, an unbiased estimator of θ is

$$\hat{\theta} = \sum_{i=1}^{k} c_i \bar{X}_{i\cdot}.$$

We will reject H_0 if $\hat{\theta}$ is observed sufficiently far from zero.

Once again, we rely on a squared error criterion and ask if the observed quantity $(\hat{\theta})^2$ is sufficiently far from zero. However, the quantity $(\hat{\theta})^2$ is not a satisfactory measure of departure from $H_0 : \theta = 0$ because its magnitude depends on the magnitude of the coefficients in the contrast. To remove this dependency, we form a ratio that does not depend on how the coefficients were scaled. The sum of squares associated with the contrast θ is the random variable

$$SS_\theta = \frac{\left(\sum_{i=1}^k c_i \bar{X}_{i\cdot}\right)^2}{\sum_{i=1}^k c_i^2/n_i}.$$

The following facts about the distribution of SS_θ lead to a test of $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

Theorem 12.4 *Under the fundamental null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$, SS_θ is independent of SS_W, $SS_\theta/\sigma^2 \sim \chi^2(1)$, and*

$$F(\theta) = \frac{\frac{SS_\theta}{\sigma^2}/1}{\frac{SS_W}{\sigma^2}/(N-k)} = \frac{SS_\theta}{SS_W/(N-k)} \sim F(1, N-k).$$

The F-test of $H_0 : \theta = 0$ is to reject H_0 if and only if

$$\mathbf{p} = P_{H_0}\left(F(\theta) \geq f(\theta)\right) \leq \alpha,$$

i.e., if and only if

$$f(\theta) \geq q = \texttt{qf(1-}\alpha\texttt{,df1=1,df2=N-k)},$$

where $f(\theta)$ denotes the observed value of $F(\theta)$.

Example 12.3 (continued) Heyl (1930) collected the following data:

Gold	83	81	76	78	79	72
Platinum	61	61	67	67	64	
Glass	78	71	75	72	74	

Applying the methods of Section 12.1.2, we obtain the following ANOVA table:

Source	SS	df	MS	F	**p**
Between	565.1	2	282.6	26.1	0.000028
Within	140.8	13	10.8		
Total	705.9	15			

To test $H_0 : \theta_1 = 0$ versus $H_1 : \theta_1 \neq 0$, we first compute

$$ss_{\theta_1} = \frac{[1 \cdot \bar{x}_{1\cdot} + 1 \cdot \bar{x}_{2\cdot} + (-2) \cdot \bar{x}_{3\cdot}]^2}{1^2/6 + 1^2/5 + (-2)^2/5} \doteq 29.16667,$$

then

$$f(\theta_1) = \frac{ss_\theta}{ss_W/(N-k)} \doteq \frac{29.16667}{140.8333/(16-3)} \doteq 2.692308.$$

Finally, we use the `R` function `pf` to compute a significance probability **p**:

```
> 1-pf(2.692308,df1=1,df2=13)
[1] 0.1247929
```

Because $\mathbf{p} > 0.05$, we decline to reject the null hypothesis at significance level $\alpha = 0.05$. Equivalently, we might use the `R` function `qf` to compute a critical value q:

```
> qf(1-.05,df1=1,df2=13)
[1] 4.667193
```

Because $f < q$, we decline to reject the null hypothesis.

In practice, one rarely tests a single contrast. However, testing multiple contrasts involves more than testing each contrast as though it was the only contrast. Entire books have been devoted to the problem of *multiple comparisons*; the remainder of Section 12.1 describes four popular procedures for testing multiple contrasts.

Orthogonal Contrasts

When it can be used, the *method of orthogonal contrasts* is generally preferred. It is quite elegant, but has certain limitations. We begin by explaining what it means for contrasts to be orthogonal.

Definition 12.2 *Two contrasts with coefficient vectors* $(c_1, \ldots, c_k)$ *and* $(d_1, \ldots, d_k)$ *are orthogonal if and only if*

$$\sum_{i=1}^{k} \frac{c_i d_i}{n_i} = 0.$$

A collection of contrasts is mutually orthogonal if and only if each pair of contrasts in the collection is orthogonal.

Notice that, if $n_1 = \cdots = n_k$, then the orthogonality condition simplifies to

$$\sum_{i=1}^{k} c_i d_i = 0.$$

Students who know some linear algebra should recognize that this condition states that the dot product between the vectors c and d vanishes, i.e., that the vectors c and d are orthogonal (perpendicular) to each other.

Example 12.3 (continued) Whether or not two contrasts are orthogonal depends not only on their coefficient vectors, but also on the size of the samples drawn from each population.

- Suppose that Heyl (1930) had collected samples of equal size for each of the three materials that he used. If $n_1 = n_2 = n_3$, then θ_1 and θ_2 are orthogonal because

$$1 \cdot 1 + 1 \cdot (-1) + (-2) \cdot 0 = 0.$$

- In fact, Heyl (1930) collected samples with sizes $n_1 = 6$ and $n_2 = n_3 = 5$. In this case, θ_1 and θ_2 are *not* orthogonal because

$$\frac{1 \cdot 1}{6} + \frac{1 \cdot (-1)}{5} + \frac{(-2) \cdot 0}{5} = \frac{1}{6} - \frac{1}{5} \neq 0.$$

  However, θ_1 is orthogonal to $\theta_3 = 18\mu_1 - 17\mu_2 - \mu_3$ because

$$\frac{1 \cdot 18}{6} + \frac{1 \cdot (-17)}{5} + \frac{(-2) \cdot (-1)}{5} = 3 - 3.2 + 0.2 = 0.$$

It turns out that the number of mutually orthogonal contrasts cannot exceed $k - 1$. Obviously, this fact limits the practical utility of the method; however, families of mutually orthogonal contrasts have two wonderful properties that commend their use.

First, any family of $k - 1$ mutually orthogonal contrasts partitions SS_B into $k - 1$ separate components,

$$SS_B = SS_{\theta_1} + \cdots + SS_{\theta_{k-1}},$$

each with one degree of freedom. This information is usually incorporated into an expanded ANOVA table, as in...

Example 12.3 (continued) In the case of Heyl's (1930) data, the orthogonal contrasts θ_1 and θ_3 partition the between-groups sum-of-squares:

Source	SS	df	MS	F	**p**
Between	565.1	2	282.6	26.1	0.000028
θ_1	29.2	1	29.2	2.7	0.124793
θ_3	535.9	1	535.9	49.5	0.000009
Within	140.8	13	10.8		
Total	705.9	15			

Testing the fundamental null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3$, results in a tiny signficance probability, leading us to conclude that the population means are not identical. The decomposition of the variation between groups into contrasts θ_1 and θ_3 provides insight into the differences between the population means. Testing the null hypothesis, $H_0 : \theta_1 = 0$, results in a large signficance probability, leading us to conclude that the heavy metals do not, in tandem, differ from glass. However, testing the null hypothesis, $H_0 : \theta_3 = 0$, results in a tiny signficance probability, leading us to conclude that the heavy metals do differ from each other. This is only possible if the glass mean lies between the gold and platinum means. For this simple example, our conclusions are easily checked by examining the raw data.

The second wonderful property of mutually orthogonal contrasts is that tests of mutually orthogonal contrasts are mutually independent. As we shall demonstrate, this property provides us with a powerful way to address a crucial difficulty that arises whenever we test multiple hypotheses. The difficulty is as follows. When testing a single null hypothesis that is true, there is a small chance (α) that we will falsely reject the null hypothesis and commit a Type I error. When testing multiple null hypotheses, each of which are true, there is a much larger chance that we will falsely reject at least one of them. We desire control of this *family-wide error rate*, often abbreviated FWER.

Definition 12.3 *The family-wide error rate (FWER) of a family of contrasts is the probability under the fundamental null hypothesis* $H_0 : \mu_1 = \cdots = \mu_k$ *of falsely rejecting at least one null hypothesis.*

The fact that tests of mutually orthogonal contrasts are mutually independent allows us to deduce a precise relation between the significance level(s) of the individual tests and the FWER.

1. Let E_r denote the event that $H_0 : \theta_r = 0$ is falsely rejected. Then $P(E_r) = \alpha$ is the rate of Type I error for an individual test.

2. Let E denote the event that at least one Type I error is committed, i.e.,
$$E = \bigcup_{r=1}^{k-1} E_r.$$
The family-wide rate of Type I error is FWER $= P(E)$.

3. The event that no Type I errors are committed is
$$E^c = \bigcap_{r=1}^{k-1} E_r^c,$$
and the probability of this event is $P(E^c) = 1 - \text{FWER}$.

4. By independence,
$$1 - \text{FWER} = P\left(E^c\right) = P\left(E_1^c\right) \times \cdots \times P\left(E_{k-1}^c\right) = (1-\alpha)^{k-1};$$
hence,
$$\text{FWER} = 1 - (1-\alpha)^{k-1}.$$

Notice that FWER $> \alpha$, i.e., the family rate of Type I error is greater than the error rate for an individual test. For example, if $k = 3$ and $\alpha = 0.05$, then
$$\text{FWER} = 1 - (1 - .05)^2 = 0.0975.$$
This phenomenon is sometimes called "alpha slippage." To protect against alpha slippage, we usually prefer to specify the family rate of Type I error that will be tolerated, then compute a significance level that will ensure the specified family rate. For example, if $k = 3$ and we desire FWER $= 0.05$, then we solve
$$0.05 = 1 - (1-\alpha)^2$$
to obtain a significance level of
$$\alpha = 1 - \sqrt{0.95} \doteq 0.0253.$$

Bonferroni t-Tests

It is often the case that one desires to test contrasts that are not mutually orthogonal. This can happen with a small family of contrasts. For example, suppose that we want to compare a control mean μ_1 to each of two treatment means, μ_2 and μ_3, in which case the natural contrasts have coefficient vectors $c = (1, -1, 0)$ and $d = (1, 0, -1)$. In this case, the orthogonality condition simplifies to $1/n_1 = 0$, which is impossible. Furthermore, as we have noted, families of more than $k - 1$ contrasts cannot be mutually orthogonal.

Statisticians have devised a plethora of procedures for testing multiple contrasts that are not mutually orthogonal. Many of these procedures address the case of multiple pairwise contrasts, i.e., contrasts for which each coefficient vector has exactly two nonzero components. We describe one such procedure that relies on *Bonferroni's inequality*.

Suppose that we plan m pairwise comparisons. These comparisons are defined by contrasts $\theta_1, \ldots, \theta_m$, each of the form $\mu_i - \mu_j$, not necessarily mutually orthogonal. Notice that each $H_0 : \theta_r = 0$ versus $H_1 : \theta_r \neq 0$ is a normal 2-sample location problem with equal variances. From this observation, the following facts can be deduced.

Theorem 12.5 *Under the fundamental null hypothesis* $H_0 : \mu_1 = \cdots = \mu_k$,

$$Z = \frac{\bar{X}_{i\cdot} - \bar{X}_{j\cdot}}{\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\sigma^2}} \sim N(0, 1)$$

and

$$T(\theta_r) = \frac{\bar{X}_{i\cdot} - \bar{X}_{j\cdot}}{\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) MS_W}} \sim t(N - k).$$

From Theorem 12.5, the t-test of $H_0 : \theta_r = 0$ is to reject if and only if

$$\mathbf{p} = P(|T(\theta_r)| \geq |t(\theta_r)|) \leq \alpha,$$

i.e., if and only if

$$|t(\theta_r)| \geq q = \texttt{qt(1-}\alpha\texttt{/2,df=N-k)},$$

where $t(\theta_r)$ denotes the observed value of $T(\theta_r)$. This t-test is virtually identical to Student's 2-sample t-test, described in Section 11.1.2, except that it pools all k samples to estimate the common variance instead of only pooling the two samples that are being compared.

At this point, you may recall that Section 11.1 strongly discouraged the use of Student's 2-sample t-test, which assumes a common population variance. Instead, we recommended Welch's approximate t-test. In the present case, our test of the fundamental null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ has already imposed the assumption of a common population variance, so our use of the T statistic in Theorem 12.5 is theoretically justified. But this justification is rather too glib, as it merely begs the question of why we assumed a common population variance in the first place. The general answer to this question is that the ANOVA methodology is extremely powerful and that comparable procedures in the case of unequal population variances may not exist. (Fortunately, ANOVA often provides useful insights even when its assumptions are violated. In such cases, however, one should interpret significance probabilities with extreme caution.) In the present case, a good procedure does exist, viz., the pairwise application of Welch's approximate t-test. The following discussion of how to control the family-wide error rate in such cases applies equally to either type of pairwise t-test.

Unless the pairwise contrasts are mutually orthogonal, we cannot use the multiplication rule for independent events to compute the family rate of Type I error. However, Bonferroni's inequality states that

$$\text{FWER} = P(E) = P\left(\bigcup_{r=1}^{m} E_r\right) \leq \sum_{r=1}^{m} P\left(E_r\right) = m\alpha;$$

hence, we can ensure that the family rate of Type I error is no greater than a specified FWER by testing each contrast at significance level $\alpha = \text{FWER}/m$.

Example 12.3 (continued) Instead of planning θ_1 and θ_3, suppose that we had planned θ_4 and θ_5, defined by coefficient vectors $c = (-1, 0, 1)$ and $d = (0, -1, 1)$ respectively. To test θ_4 and θ_5 with a family-wide error rate of FWER ≤ 0.10, we first compute

$$t\left(\theta_4\right) = \frac{\bar{x}_{3\cdot} - \bar{x}_{1\cdot}}{\sqrt{\left(\frac{1}{n_3} + \frac{1}{n_1}\right) ms_W}} \doteq -2.090605$$

and

$$t\left(\theta_5\right) = \frac{\bar{x}_{3\cdot} - \bar{x}_{2\cdot}}{\sqrt{\left(\frac{1}{n_3} + \frac{1}{n_2}\right) ms_W}} \doteq 4.803845,$$

resulting in the following significance probabilities:

```
> 2*pt(-2.090605,df=13)
[1] 0.0567719
> 2*pt(-4.803845,df=13)
[1] 0.0003444588
```

There are $m = 2$ pairwise comparisons. To ensure FWER ≤ 0.10, we compare the significance probabilities to $\alpha = 0.10/2 = 0.05$, which leads us to reject $H_0 : \theta_5 = 0$ and to decline to reject $H_0 : \theta_4 = 0$.

What do we lose by using Bonferroni's inequality instead of the multiplication rule? Without the assumption of independence, we must be slightly more conservative in choosing a significance level that will ensure a specified family-wide rate of error. For the same FWER, Bonferroni's inequality leads to a slightly smaller α than does the multiplication rule. The discrepancy grows as m increases.

12.1.4 Post Hoc Comparisons

We now consider situations in which we determine that a comparison is of interest *after* inspecting the data. For example, suppose that we had decided to compare gold to platinum *after* inspecting Heyl's (1930) data. This ought to strike you as a form of cheating. Almost every randomly generated data set will have an appealing pattern in it that may draw the attention of an interested observer. To allow such patterns to determine what the scientist will investigate is to invite abuse. Fortunately, statisticians have devised procedures that protect ethical scientists from the heightened risk of Type I error when the null hypothesis was constructed after the data were examined. The present section describes two such procedures.

Bonferroni t-Tests

To fully appreciate the distinction between planned and *post hoc* comparisons, it is highly instructive to examine the method of Bonferroni t-tests. Suppose that only pairwise comparisons are of interest. Because we are testing *after* we have had the opportunity to inspect the data (and therefore to construct the contrasts that appear to be nonzero), we suppose that *all* pairwise contrasts were of interest *a priori*. Hence, whatever the number of pairwise contrasts actually tested *a posteriori*, we set

$$m = \binom{k}{2} = \frac{k(k-1)}{2}$$

and proceed as before.

The difference between planned and *post hoc* comparisons is especially sobering when k is large. For example, suppose that we desire that the family-wide error rate does not exceed 0.10 when testing two pairwise contrasts among $k = 10$ groups. If the comparisons were planned, then $m = 2$ and we can perform each test at signficance level $\alpha = 0.10/2 = 0.05$. However, if the comparisons were constructed after examining the data, then $m = 45$ and we must perform each test at signficance level $\alpha = 0.10/45 \doteq 0.0022$. Obviously, much stronger evidence is required to reject the same null hypothesis when the comparison is chosen after examining the data.

Scheffé F-Tests

The reasoning that underlies Scheffé F-Tests for *post hoc* comparisons is analogous to the reasoning that underlies Bonferroni t-tests for *post hoc* comparisons. To accommodate the possibility that a general contrast was constructed after examining the data, Scheffé's procedure is predicated on the assumption that *all possible* contrasts were of interest *a priori*. This makes Scheffé's procedure the most conservative of all multiple comparison procedures.

Scheffé's F-test of $H_0 : \theta_r = 0$ versus $H_1 : \theta_r \neq 0$ is to reject H_0 if and only if

$$\mathbf{p} = \texttt{1-pf(}f(\theta)\texttt{/(k-1),df1=k-1,df2=N-k)} \leq \alpha,$$

i.e., if and only if

$$\frac{f\left(\theta_r\right)}{k-1} \geq q = \texttt{qf(1-}\alpha\texttt{,k-1,N-k)},$$

where $f(\theta_r)$ denotes the observed value of the $F(\theta_r)$ defined for the method of planned orthogonal contrasts. It can be shown that, no matter how many $H_0 : \theta_r = 0$ are tested by this procedure, the family-wide rate of Type I error is no greater than α.

Example 12.3 (continued) Let $\theta_6 = \mu_1 - \mu_3$. Scheffé's F-test produces the following results:

Source	$f(\theta_r)/2$	**p**
θ_1	1.3	0.294217
θ_2	25.3	0.000033
θ_3	24.7	0.000037
θ_6	2.2	0.151995

For the first three comparisons, our conclusions are not appreciably affected by whether the contrasts were constructed before or after examining the data. However, if θ_6 had been planned, we would have obtained $f(\theta_6) = 4.4$ and $\mathbf{p} = 0.056772$, which might easily lead to a different conclusion.

12.2 The Case of a General Shift Family

12.2.1 The Kruskal-Wallis Test

12.3 The Behrens-Fisher Problem

12.4 Exercises

1. Jean Kerr devoted an entire chapter of *Please Don't Eat the Daisies* (1959) to the subject of dieting, observing that...

   > "Today, with the science of nutrition advancing so rapidly, there is plenty of food for conversation, if for nothing else. We have the Rockefeller diet, the Mayo diet, high-protein diets, low-protein diets, "blitz" diets which feature cottage cheese and something that tastes like very thin sandpaper, and—finally—a liquid diet that duplicates all the rich, nourishing goodness of mother's milk. I have no way of knowing which of these is the most efficacious for losing weight, but there's no question in my mind that as a conversation-stopper the "mother's milk diet" is quite a ways out ahead."

   For her master's thesis, a nutrition student at the University of Arizona decides to compare several weight loss strategies. She recruits 140 moderately obese adult women and randomly assigns each woman to one of the following diets: Rockefeller, Mayo, Atkins (high-protein), a low-protein diet, a blitz diet, a liquid diet, and—as a control—Aunt Jean's marshmallow fudge diet. Each woman is weighed before dieting, asked to follow the prescribed diet for eight weeks, then weighed again. The resulting data will be analyzed using the analysis of variance and related statistical techniques.

   (a) This is a k-sample problem. What is the value of k?

   (b) What null hypothesis is tested by an analysis of variance? (Your answer should specify relations between certain population parameters. Be sure to define these parameters!)

   (c) How many pairwise comparisons are possible?

   (d) The student is especially interested in three pairwise comparisons: Atkins versus low-protein, low-protein versus fudge, and fudge versus liquid. Specify contrasts that correspond to each of these comparisons.

   (e) Are the preceding contrasts orthogonal? Why or why not?

2. As part of her senior thesis, a William & Mary physics major decides to repeat Heyl's (1930) experiment for determining the gravitational

constant using 4 different materials: silver, copper, topaz, and quartz. She plans to test 10 specimens of each material.

(a) Three comparisons are planned:

i. Metal (silver & copper) versus Gem (topaz & quartz)

ii. Silver versus Copper

iii. Topaz versus Quartz

What contrasts correspond to these comparisons? Are they orthogonal? Why or why not? If the desired family rate of Type I error is 0.05, then what significance level should be used for testing the null hypotheses $H_0 : \theta_r = 0$?

(b) After analyzing the data, an ANOVA table is constructed. Complete the table from the information provided.

Source	SS	df	MS	F	**p**
Between					
θ_1					0.001399
θ_2					0.815450
θ_3					0.188776
Within			9.418349		
Total					

(c) Referring to the above table, explain what conclusion the student should draw about each of her planned comparisons.

(d) Assuming that the ANOVA assumption of homoscedasticity is warranted, use the above table to estimate the common population variance.

3. R. R. Sokal observed 25 females of each of three genetic lines (RS, SS, NS) of the fruitfly *Drosophila melanogaster* and recorded the number of eggs laid per day by each female for the first 14 days of her life. The lines labelled RS and SS were selectively bred for resistance and for susceptibility to the insecticide DDT. A nonselected control line is labelled NS. The purpose of the experiment was to investigate the following research questions:

- Do the two selected lines (RS and SS) differ in fecundity from the nonselected line (NS)?

- Does the line selected for resistance (RS) differ in fecundity from the line selected for susceptibility (SS)?

The data are presented in Table 12.1.

RS	12.8	21.6	14.8	23.1	34.6	19.7	22.6	29.6	16.4	20.3
	29.3	14.9	27.3	22.4	27.5	20.3	38.7	26.4	23.7	26.1
	29.5	38.6	44.4	23.2	23.6					
SS	38.4	32.9	48.5	20.9	11.6	22.3	30.2	33.4	26.7	39.0
	12.8	14.6	12.2	23.1	29.4	16.0	20.1	23.3	22.9	22.5
	15.1	31.0	16.9	16.1	10.8					
NS	35.4	27.4	19.3	41.8	20.3	37.6	36.9	37.3	28.2	23.4
	33.7	29.2	41.7	22.6	40.4	34.4	30.4	14.9	51.8	33.8
	37.9	29.5	42.4	36.6	47.4					

Table 12.1: Fecundity of Female Fruitflies

(a) Use side-by-side boxplots and normal probability plots to investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible? Why or why not?

(b) Construct constrasts that correspond to the research questions framed above. Verify that these constrasts are orthogonal. At what significance level should the contrasts be tested in order to maintain a family rate of Type I error equal to 5%?

(c) Use ANOVA and the method of orthogonal contrasts to construct an ANOVA table. State the null and alternative hypotheses that are tested by these methods. For each null hypothesis, state whether or not it should be rejected. (Use $\alpha = 0.05$ for the ANOVA hypothesis and the significance level calculated above for the contrast hypotheses.)

4. A number of Byzantine coins were discovered in Cyprus. These coins were minted during the reign of King Manuel I, Comnenus (1143–1180). It was determined that $n_1 = 9$ of these coins were minted in an early coinage, $n_2 = 7$ were minted several years later, $n_3 = 4$ were minted in a third coinage, and $n_4 = 7$ were minted in a fourth coinage.

   The silver content (percentage) of each coin was measured, with the results presented in Table 12.2.

1	5.9	6.8	6.4	7.0	6.6	7.7	7.2	6.9	6.2
2	6.9	9.0	6.6	8.1	9.3	9.2	8.6		
3	4.9	5.5	4.6	4.5					
4	5.3	5.6	5.5	5.1	6.2	5.8	5.8		

Table 12.2: Silver Content of Byzantine Coins

(a) Investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible? Why or why not?

(b) Construct an ANOVA table. State the null and alternative hypotheses tested by this method. Should the null hypothesis be rejected at the $\alpha = 0.10$ level?

(c) Examining the data, it appears that coins minted early in King Manuel's reign (the first two coinages) tended to contain more silver than coins minted later in his reign (the last two coinages). Construct a contrast that is suitable for investigating if this is the case. State appropriate null and alternative hypotheses and test them using Scheffé's F-test for multiple comparisons with a significance level of 5%.

5. R. E. Dolkart and colleagues compared antibody responses in normal and alloxan diabetic mice. Three groups of mice were studied: normal, alloxan diabetic, and alloxan diabetic treated with insulin. Several comparisons are of interest:

   - Does the antibody response of alloxan diabetic mice differ from the antibody response of normal mice?
   - Does the antibody response of alloxan diabetic mice treated with insulin differ from the antibody response of normal mice?
   - Does treating alloxan diabetic mice with insulin affect their antibody response?

   Table 12.3 contains the measured amounts of nitrogen-bound bovine serum albumen produced by the mice.

   (a) Using the above data, investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible for these data? Why or why not?

Normal	156	282	197	297	116	127	119	29	253	122
	349	110	143	64	26	86	122	455	655	14
Alloxan	391	46	469	86	174	133	13	499	168	62
	127	276	176	146	108	276	50	73		
Alloxan	82	100	98	150	243	68	228	131	73	18
+insulin	20	100	72	133	465	40	46	34	44	

Table 12.3: Antibody Responses of Diabetic Mice

(b) Now transform the data by taking the square root of each measurement. Using the transformed data, investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible for the transformed data? Why or why not?

(c) Using the transformed data, construct an ANOVA table. State the null and alternative hypotheses tested by this method. Should the null hypothesis be rejected at the $\alpha = 0.05$ level?

(d) Using the transformed data, construct suitable contrasts for investigating the research questions framed above. State appropriate null and alternative hypotheses and test them using the method of Bonferroni t-tests. At what significance level should these hypotheses be tested in order to maintain a family rate of Type I error equal to 5%? Which null hypotheses should be rejected?

Chapter 13

Association

13.1 Categorical Random Variables

13.2 Normal Random Variables

The continuous random variables (X, Y) define a function that assigns a pair of real numbers to each experimental outcome. Let

$$B = [a, b] \times [c, d] \subset \Re^2$$

be a rectangular set of such pairs and suppose that we want to compute

$$P\left((X, Y) \in B\right) = P\left(X \in [a, b], Y \in [c, d]\right).$$

Just as we compute $P(X \in [a, b])$ using the pdf of X, so we compute $P((X, Y) \in B)$ using the *joint probability density function* of (X, Y). To do so, we must extend the concept of area under the graph of a function of one variable to the concept of volume under the graph of a function of two variables.

Theorem 13.1 *Let X be a continuous random variable with pdf f_x and let Y be a continuous random variable with pdf f_y. In this context, f_x and f_y are called the* marginal pdfs *of (X, Y). Then there exists a function $f : \Re^2 \to \Re$, the* joint pdf *of (X, Y), such that*

$$P\left((X, Y) \in B\right) = \text{Volume}_B(f) = \int_a^b \int_c^d f(x, y)\, dy\, dx \tag{13.1}$$

for all rectangular subsets B. If X and Y are independent, then

$$f(x, y) = f_x(x) f_y(y).$$

Remark: If (13.1) is true for all rectangular subsets of $\Re^2$, then it is true for all subsets in the sigma-field generated by the rectangular subsets.

We can think of the joint pdf as a function that assigns an elevation to a point identified by two coordinates, longitude (x) and latitude (y). Noting that topographic maps display elevations via contours of constant elevation, we can describe a joint pdf by identifying certain of its contours, i.e., subsets of $\Re^2$ on which $f(x, y)$ is contant.

Definition 13.1 *Let f denote the joint pdf of (X, Y) and fix $c > 0$. Then*

$$\Big\{(x, y) \in \Re^2 \ : \ f(x, y) = c\Big\}$$

is a contour *of f.*

13.2.1 Bivariate Normal Distributions

Suppose that $X \sim \text{Normal}(0, 1)$ and $Y \sim \text{Normal}(0, 1)$, not necessarily independent. To measure the degree of dependence between X and Y, we consider the quantity $E(XY)$.

- If there is a *positive association* between X and Y, then experimental outcomes that have...
  - positive values of X will tend to have positive values of Y, so XY will tend to be positive;
  - negative values of X will tend to have negative values of Y, so XY will tend to be positive.

  Hence, $E(XY) > 0$ indicates positive association.

- If there is a *negative association* between X and Y, then experimental outcomes that have...
  - positive values of X will tend to have negative values of Y, so XY will tend to be negative;
  - negative values of X will tend to have positive values of Y, so XY will tend to be negative.

  Hence, $E(XY) < 0$ indicates negative association.

If $X \sim \text{Normal}(\mu_x, \sigma_x^2)$ and $Y \sim \text{Normal}(\mu_y, \sigma_y^2)$, then we measure dependence after converting to standard units:

Definition 13.2 *Let $\mu_x = EX$ and $\sigma_x^2 = \text{Var}\, X < \infty$. Let $\mu_y = EY$ and $\sigma_y^2 = \text{Var}\, Y < \infty$. The* population product-moment correlation coefficient *of X and Y is*

$$\rho = \rho(X, Y) = E\left[\left(\frac{X - \mu_x}{\sigma_x}\right)\left(\frac{Y - \mu_y}{\sigma_y}\right)\right].$$

The product-moment correlation coefficient has the following properties:

Theorem 13.2 *If X and Y have finite variances, then*

1. $-1 \leq \rho \leq 1$
2. $\rho = \pm 1$ *if and only if*

$$\frac{Y - \mu_y}{\sigma_y} = \pm \frac{X - \mu_x}{\sigma_x},$$

   in which case Y is completely determined by X.

3. *If X and Y are independent, then $\rho = 0$.*
4. *If X and Y are normal random variables for which $\rho = 0$, then X and Y are independent.*

If $\rho = \pm 1$, then the values of (X, Y) fall on a straight line. If $|\rho| < 1$, then the five population parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ determine a unique bivariate normal pdf. The contours of this joint pdf are concentric ellipses centered at (μ_x, μ_y). We use one of these ellipses to display the basic features of the bivariate normal pdf in question.

Definition 13.3 *Let f denote a nondegenerate ($|\rho| < 1$) bivariate normal pdf. The* population concentration ellipse *is the contour of f that contains the four points*

$$(\mu_x \pm \sigma_x, \mu_y \pm \sigma_y).$$

It is not difficult to create an `R` function that plots concentration ellipses. The function `binorm.ellipse` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 The following `R` commands produce the population concentration ellipse for a bivariate normal distribution with parameters $\mu_x = 10$, $\mu_y = 20$, $\sigma_x^2 = 4$, $\sigma_y^2 = 16$ and $\rho = 0.5$:

```
> pop <- c(10,20,4,16,.5)
> binorm.ellipse(pop)
```

The ellipse plotted by these commands is displayed in Figure 13.1.

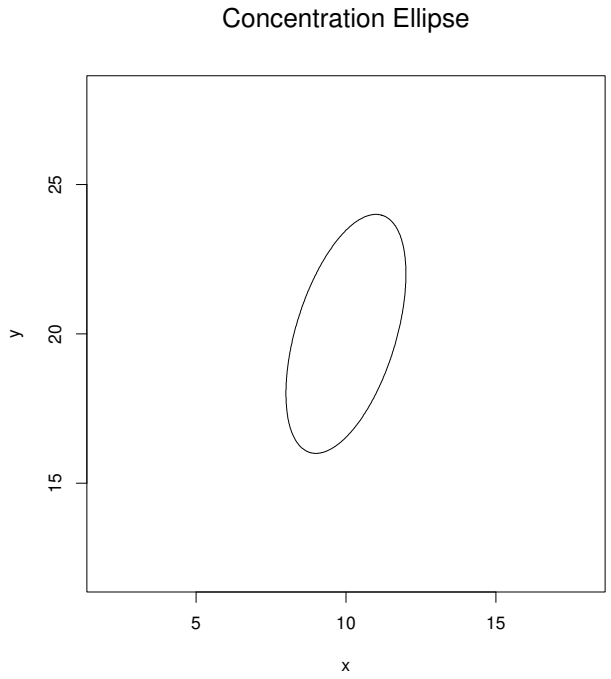

Figure 13.1: The population concentration ellipse for a bivariate normal distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho) = (10, 20, 4, 16, 0.5)$.

Unless the population concentration ellipse is circular, it has a unique major axis. The line that coincides with this axis is the *first principal component* of the population and plays an important role in multivariate statistics. We will encounter this line again in Chapter 14.

13.2.2 Bivariate Normal Samples

A bivariate sample is a set of paired observations:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

We assume that each pair (x_i, y_i) was independently drawn from the same bivariate distribution. Bivariate samples are usually stored in an $n \times 2$ *data matrix*,

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \\ x_n & y_n \end{bmatrix},$$

and are often displayed by plotting each (x_i, y_i) in the Cartesian plane. The resulting figure is called a *scatter diagram*.

Example 13.2 Twenty students enrolled in Math 351 (Applied Statistics) at the College of William & Mary produced the following scores on two midterm tests:

x	y
87	87
25	57
76	91
84	67
91	67
82	66
94	86
89	74
92	92
76	85
84	75
99	92
92	55
74	74
84	74
94	69
99	98
63	81
82	80
91	85

A scatter diagram of these data is displayed in Figure 13.2. Typically, it is easier to discern patterns by inspecting a scatter diagram than by inspecting a table of numbers. In particular, note the presence of an apparent outlier.

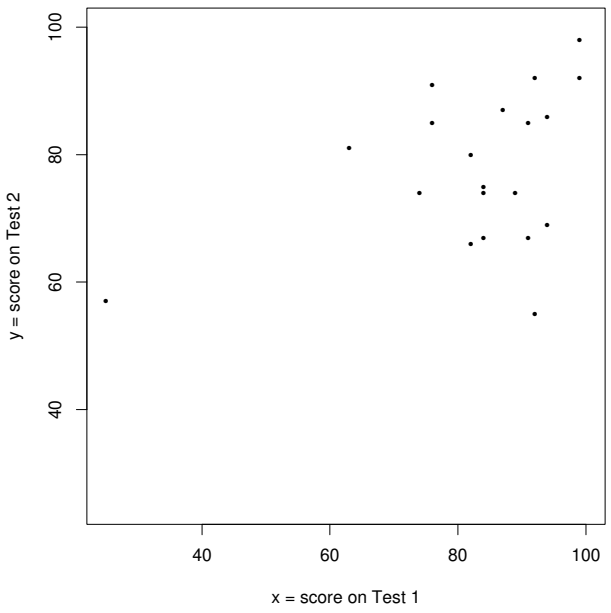

Figure 13.2: A scatter diagram of a bivariate sample. Each point corresponds to a student. The horizontal position of the point represents the student's score on the first midterm test; the vertical position of the point represents the student's score on the second midterm test.

The population from which the bivariate sample in Example 13.2 was drawn is not known, so this sample should not be interpreted as a typical example of a bivariate normal sample. However, it is not difficult to create an `R` function that simulates sampling from a specified bivariate normal population. The function `binorm.sample` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 (continued) The following R command draws $n = 5$ observations from the previously specified bivariate normal distribution:

```
> binorm.sample(pop,5)
          [,1]     [,2]
[1,] 12.293160 24.07643
[2,] 11.819520 24.13076
[3,] 11.529582 17.28637
[4,]  6.912459 23.39430
[5,] 11.043991 18.12538
```

Notice that `binorm.sample` returns the sample in the form of a data matrix.

Having observed a bivariate normal sample, we inquire how to estimate the five population parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. We have already discussed how to estimate the population means (μ_x, μ_y) with the sample means $(\bar{x}, \bar{y})$ and the population variances (σ_x^2, σ_y^2) with the sample variances (s_x^2, s_y^2). The plug-in estimate of ρ is

$$\begin{aligned}
\hat{\rho} &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \hat{\mu}_x}{\hat{\sigma}_x}\right)\left(\frac{y_i - \hat{\mu}_y}{\hat{\sigma}_y}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{\sqrt{(n-1)s_x^2/n}}\right)\left(\frac{y_i - \bar{y}}{\sqrt{(n-1)s_y^2/n}}\right)\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)\right],
\end{aligned}$$

where

$$\hat{\sigma}_x = \sqrt{\widehat{\sigma_x^2}} \quad \text{and} \quad \hat{\sigma}_y = \sqrt{\widehat{\sigma_y^2}}.$$

This quantity is *Pearson's product-moment correlation coefficient*, usually denoted r.

It is not difficult to create an R function that computes the estimates $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$ from a bivariate data matrix. The function `binorm.estimate` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 (continued) The following `R` commands draw $n = 100$ observations from a bivariate normal distribution with parameters $\mu_x = 10$, $\mu_y = 20$, $\sigma_x^2 = 4$, $\sigma_y^2 = 16$ and $\rho = 0.5$, then estimate the parameters from the sample:

```
> Data <- binorm.sample(pop,100)
> binorm.estimate(Data)
[1]  9.8213430 20.3553502  4.2331147 16.7276819  0.5632622
```

Naturally, the estimates do not equal the estimands because of sampling variation.

Finally, it is not difficult to create an `R` function that plots a scatter diagram and overlays the *sample concentration ellipse*, i.e., the concentration ellipse constructed using the computed sample quantities $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$ instead of the unknown population quantities $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. The function `binorm.scatter` is described in Appendix R and/or can be obtained from the web page for this book/course.

Example 13.1 (continued) The following `R` command creates the overlaid scatter diagram displayed in Figure 13.3:

```
> binorm.scatter(Data)
```

When analyzing bivariate data, it is good practice to examine both the scatter diagram and the sample concentration ellipse in order to ascertain how well the latter summarizes the former. A poor summary suggests that the sample may not have been drawn from a bivariate normal distribution, as in Figure 13.4.

13.2.3 Inferences about Correlation

We have already observed that $\hat{\rho} = r$ is the plug-in estimate of ρ. In this section, we consider how to test hypotheses about and construct confidence intervals for ρ.

Given normal random variables X and Y, an obvious question is whether or not they are uncorrelated. To answer this question, we test the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis $H_1 : \rho \neq 0$. (One might also be interested in one-sided hypotheses and ask, for example, whether or not there is convincing evidence of positive correlation.) We can derive a test from the following fact about the plug-in estimator of ρ.

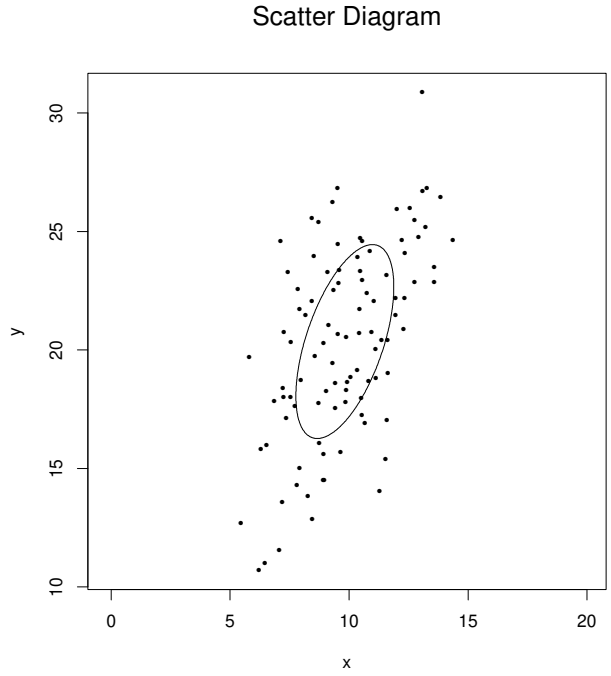

Figure 13.3: A scatter diagram of a bivariate normal sample, with the sample concentration ellipse overlaid.

Theorem 13.3 *Suppose that* (X_i, Y_i), $i = 1, \ldots, n$, *are independent pairs of random variables with a bivariate normal distribution. Let* $\hat{\rho}$ *denote the plug-in estimator of* ρ. *If* X_i *and* Y_i *are uncorrelated, i.e.,* $\rho = 0$, *then*

$$\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t(n-2).$$

Assuming that (X_i, Y_i) have a bivariate normal distribution, Theorem 13.3 allows us to compute a significance probability for testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$. Let $T \sim t(n-2)$. Then the probability of observing $|\hat{\rho}| \geq |r|$ under H_0 is

$$\mathbf{p} = P\left(|T| \geq \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)$$

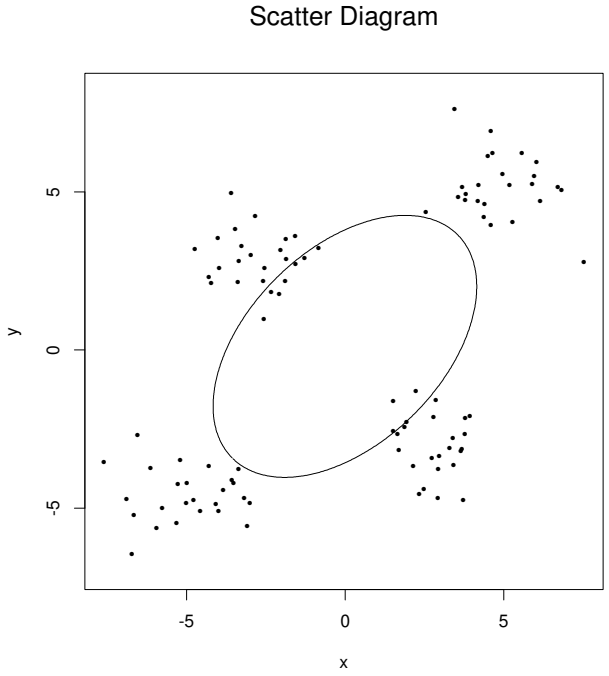

Figure 13.4: A scatter diagram for which the sample concentration ellipse is a poor summary. These data were not drawn from a bivariate normal distribution.

and we reject H_0 if and only if $\mathbf{p} \leq \alpha$. Equivalently, we reject H_0 if and only if (iff)

$$\left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| \geq q_t \quad \text{iff} \quad \frac{r^2(n-2)}{1-r^2} \geq q_t^2 \quad \text{iff} \quad r^2 \geq \frac{q_t^2}{n-2+q_t^2},$$

where $q_t = \texttt{qt}(1-\alpha/2, n-2)$.

When testing hypotheses about correlation, it is important to appreciate the distinction between statistical significance and material significance. *Strong evidence that an association exists is not the same as evidence of a strong association.* The following examples illustrate the distinction.

Example 13.3 I used `binorm.sample` to draw a sample of $n = 300$ observations from a bivariate normal distribution with a population correlation coefficient of $\rho = 0.1$. This is a rather weak association. I then used `binorm.estimate` to compute a sample correlation coefficient of $r = 0.16225689$. The test statistic is

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 2.838604$$

and the significance probability is

$$\underline{\mathrm{p}} = 2 * \mathtt{pt}(-2.838604, 298) = 0.004842441.$$

This is fairly decisive evidence that $\rho \neq 0$, but concluding that X and Y are correlated does not warrant concluding that X and Y are strongly correlated.

Example 13.4 I used `binorm.sample` to draw a sample of $n = 10$ observations from a bivariate normal distribution with a population correlation coefficient of $\rho = 0.8$. This is a fairly strong association. I then used `binorm.estimate` to compute a sample correlation coefficient of $r = 0.3759933$. The test statistic is

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 1.147684$$

and the significance probability is

$$\mathbf{p} = 2 * \mathtt{pt}(-1.147684, 8) = 0.2842594.$$

There is scant evidence that $\rho \neq 0$, despite the fact that X and Y are strongly correlated.

Although testing whether of not $\rho = 0$ is an important decision, it is not the only inference of interest. For example, if we want to construct confidence intervals for ρ, then we need to test $H_0 : \rho = \rho_0$ versus $H_1 : \rho \neq \rho_0$. To do so, we rely on an approximation due to Ronald Fisher. Let

$$\zeta = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

and rewrite the hypotheses as $H_0 : \zeta = \zeta_0$ versus $H_1 : \zeta \neq \zeta_0$. This is sometimes called Fisher's z-transformation. Fisher discovered that

$$\hat{\zeta} = \frac{1}{2}\log\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \stackrel{\sim}{\sim} \text{Normal}\left(\zeta, \frac{1}{n-3}\right),$$

which allows us to compute an approximate significance probability. Let $Z \sim \text{Normal}(0, 1)$ and set

$$z = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right).$$

Then

$$\mathbf{p} \doteq P\left(|Z| \geq |z - \zeta_0| \sqrt{n-3}\right)$$

and we reject $H_0 : \zeta = \zeta_0$ if and only if $\mathbf{p} \leq \alpha$. Equivalently, we reject $H_0 : \zeta = \zeta_0$ if and only if

$$|z - \zeta_0| \sqrt{n-3} \geq q_z,$$

where $q_z = \texttt{qnorm}(1 - \alpha/2)$.

To construct an approximate $(1-\alpha)$-level confidence interval for ρ, we first observe that

$$z \pm \frac{q_z}{\sqrt{n-3}} \tag{13.2}$$

is an approximate $(1-\alpha)$-level confidence interval for ζ. We then use the inverse of Fisher's z-transformation,

$$\rho = \frac{e^{2\zeta} - 1}{e^{2\zeta} + 1},$$

to transform (13.2) to a confidence interval for ρ.

Example 13.5 Suppose that we draw $n = 100$ observations from a bivariate normal distribution and observe $r = 0.5$. To construct a 0.95-level confidence interval, we use $q_z \doteq 1.96$. First we compute

$$z = \frac{1}{2} \log\left(\frac{1+0.5}{1-0.5}\right) = 0.5493061$$

and

$$z \pm \frac{q_z}{\sqrt{n-3}} \doteq 0.5493061 \pm \frac{1.96}{\sqrt{97}} = (0.350302, 0.7483103)$$

to obtain a confidence interval (a, b) for ζ. The corresponding confidence interval for ρ is

$$\left(\frac{e^{2a} - 1}{e^{2a} + 1}, \frac{e^{2b} - 1}{e^{2b} + 1}\right) = (0.3366433, 0.6341398).$$

Notice that the plug-in estimate $\hat{\rho} = r = 0.5$ is *not* the midpoint of this interval.

13.3 Monotonic Association

13.4 Spurious Association

13.5 Exercises

1. Consider the following data matrix:

4.81310497776088	5.50546805210632
3.20790912734096	3.23537831017746
2.03360531141548	1.57466192734915
3.80353555823225	4.0777212868518
3.44874039566775	3.57596515608872
4.02513467455476	4.39110976256498
4.18921274133904	4.62315118989928
1.57765999081644	0.929857871257454
2.55801286069007	2.31628619574412
3.30197349607145	3.36840541617217
3.49344457748324	3.63918641630698
3.84773963203205	4.14023528753161
1.6571339655711	1.04225104421118
2.01676932918443	1.55085225294214
3.26802020797819	3.32038821566353
3.21119453633111	3.24002458012926
3.98834405943784	4.33907997569859
3.39396169865743	3.49849637984759
3.98470335590536	4.33393124338638
2.92484672005844	2.83506761480053
3.24990948234283	2.98840952533401
4.48210022495756	1.24582866569767
2.49246311350902	4.05960045290903
2.5490793094774	3.97953306072058
3.56806772786439	2.53846581953658
2.58341332552653	3.93097742957316
3.00614070448958	3.33315063705718
3.59845899773574	2.49548607350678
3.24798603840268	2.99112968584062
3.27071210738312	2.95899017086906
3.61265049129421	2.47541627084607
3.98487089689919	1.94901712504748
2.92139406397179	3.453000485443
2.10733672639563	4.60425141279254
3.20304499253985	3.05468592240708
1.84295811639769	4.97813922865297
3.11571443259585	3.17818998468951
3.5505950180758	2.56317596269101
3.41454250084746	2.75558327775034
2.6505463184044	3.83603704056258

(a) Do the x values appear to have been drawn from a normal distribution? Why or why not?

(b) Do the y values appear to have been drawn from a normal distribution? Why or why not?

(c) Do the (x, y) values appear to have been drawn from a bivariate normal distribution? Why or why not?

(d) Suggest an explanation for the phenomena observed in (a)–(c). Is this a paradox? How do you think that these (x, y) pairs were obtained?

Hint: Do *not* try to type these data into `R`! They are available electronically. Assuming that the data matrix is stored in a text file named `ex131.dat`, located in the root directory of a diskette, the following command reads the data into the Windows version of `R`:

```
> Data <- matrix(scan("a:\\ex131.dat"),byrow=T,ncol=2)
```

The following `R` commands then create vectors of x and y values:

```
> x <- Data[,1]
> y <- Data[,2]
```

2. Consider the test score data reported in Example 13.2.

(a) Quantify the association between midterm test scores by computing Pearson's product-moment correlation coefficient. Is the association positive or negative?

(b) Examining the scatter diagram displayed in Figure 13.2, one student appears to be an outlier. Omitting the corresponding row of the data matrix, re-compute Pearson's product-moment correlation coefficient. How does the outlier affect the value of r?

Hint: If `Data` is a complete data matrix, then `Data[-17,]` is the same data matrix without row 17.

3. Pearson and Lee reported the following heights (in inches) of eleven pairs of siblings:

sister	brother
69	71
64	68
65	66
63	67
65	70
62	71
65	70
64	73
66	72
59	65
62	66

Assuming that these pairs were drawn from a bivariate normal population, construct a confidence interval for ρ, the population product-moment correlation coefficient, that has a confidence level of approximately 0.90.

Hint: If `x` is the vector of sister heights and `y` is the vector of brother heights (in the same order), then the following `R` command creates the above data matrix:

```
> Data <- cbind(x,y)
```

4. Let $\alpha = 0.05$.

   (a) Suppose that we sample from a bivariate normal distribution with $\rho = 0.5$. Assuming that we observe $r = 0.5$, how large a sample will be needed to reject $H_0 : \rho = 0$ in favor of $H_0 : \rho \neq 0$?

   (b) Suppose that we sample from a bivariate normal distribution with $\rho = 0.1$. Assuming that we observe $r = 0.1$, how large a sample will be needed to reject $H_0 : \rho = 0$ in favor of $H_0 : \rho \neq 0$?

Chapter 14

Simple Linear Regression

One way to quantify the association between two random variables, X and Y, is to quantify the extent to which knowledge of X allows one to predict values of Y. Notice that this approach to association is asymmetric: one variable (conventionally denoted X) is the *predictor variable* and the other variable (conventionally denoted Y) is the *predicted variable*. The predictor variable is often called the *independent variable* and the predicted variable is often called the *dependent variable*. We will eschew this terminology, as it has nothing to do with the probabilistic (in)dependence of events and random variables.

14.1 The Regression Line

Suppose that $Y \sim \text{Normal}(\mu_y, \sigma_y^2)$ and that we want to predict the outcome of an experiment in which we observe Y. If we know μ_y, then the obvious value of Y to predict is $EY = \mu_y$. The expected value of the squared error of this prediction is $E(Y - \mu_y)^2 = \text{Var}\, Y = \sigma_y^2$.

Now suppose that $X \sim \text{Normal}(\mu_x, \sigma_x^2)$ and that we observe $X = x$. Again we want to predict Y. Does knowing $X = x$ allow us to predict Y more accurately? The answer depends on the association between X and Y. If X and Y are independent, then knowing $X = x$ will not help us predict Y. If X and Y are dependent, then knowing $X = x$ should help us predict Y.

Example 14.1 Suppose that we want to predict the adult height to which a male baby will grow. Knowing only that adult male heights are normally distributed, we would predict the average height of this population.

However, if we knew that the baby's father had attained a height of $6'$-$11''$, then we surely would be inclined to revise our prediction and predict that the baby will grow to a greater-than-average height.

When X and Y are normally distributed, the key to predicting Y from $X = x$ is the following result.

Theorem 14.1 *Suppose that (X, Y) have a bivariate normal distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. Then the conditional distribution of Y given $X = x$ is*

$$Y|X = x \sim \text{Normal}\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}\left(x - \mu_x\right), \left(1 - \rho^2\right)\sigma_y^2\right).$$

Because $Y|X = x$ is normally distributed, the obvious value of Y to predict when $X = x$ is

$$\hat{y}(x) = E(Y|X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}\left(x - \mu_x\right). \tag{14.1}$$

Interpreting (14.1) as a function that assigns a predicted value of Y to each value of x, we see that the prediction function (14.1) corresponds to a line that passes through the point (μ_x, μ_y) with slope $\rho\,\sigma_y/\sigma_x$. The prediction function (14.1) is the *population regression function* and the corresponding line is the *population regression line*.

The expected squared error of the prediction (14.1) is

$$\text{Var}(Y|X = x) = (1 - \rho^2)\sigma_y^2.$$

Notice that this quantity does not depend on the value of x. If X and Y are strongly correlated, then $\rho \approx \pm 1$, $(1-\rho^2)\sigma_y^2 \approx 0$, and prediction is extremely accurate. If X and Y are uncorrelated, then $\rho = 0$, $(1 - \rho^2)\sigma_y^2 = \sigma_y^2$, and the accuracy of prediction is not improved by knowing $X = x$. These remarks suggest a natural way of interpreting what ρ actually measures: the proportion by which the expected squared error of prediction is reduced by virtue of knowing $X = x$ is

$$\frac{\sigma_y^2 - \left(1 - \rho^2\right)\sigma_y^2}{\sigma_y^2} = \rho^2,$$

the *population coefficient of determination*. Statisticians often express this interpretation by saying that ρ^2 is "the proportion of variation explained by linear regression." Of course, as we emphasized in Section 13.4, this is not an explanation in the sense of articulating a causal mechanism.

Example 14.2 Suppose that $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho) = (10, 20, 2^2, 4^2, 0.5)$. Then

$$\hat{y}(x) = 20 + 0.5 \cdot \frac{4}{2}(x - 10) = x + 10$$

and $\rho^2 = 0.25$.

Rewriting (14.1), the equation for the population regression line, as

$$\frac{\hat{y}(x) - \mu_y}{\sigma_y} = \rho \, \frac{x - \mu_x}{\sigma_x},$$

we discern an important fact:

Corollary 14.1 *Suppose that (x, y) lies on the population regression line. If x lies z standard deviations above μ_x, then y lies ρz standard deviations above μ_y.*

Example 14.2 (continued) The value $x = 12$ lies $(12 - 10)/2 = 1$ standard deviations above the X-population mean, $\mu_x = 10$. The predicted y-value that corresponds to $x = 12$, $\hat{y}(12) = 12 + 10 = 22$, lies $(22 - 20)/4 = 0.5$ standard deviations above the Y-population mean, $\mu_y = 20$.

Example 14.2 (continued) The 0.90 quantile of X is

$$x = \texttt{qnorm(.9,mean=10,sd=2)} = 12.5631.$$

The predicted y-value that corresponds to $x = 12.5631$ is $\hat{y}(12.5631) = 22.5631$. At what quantile of Y does the predicted y-value lie? The answer is

$$P\left(Y \leq \hat{y}(x)\right) = \texttt{pnorm(22.5631,mean=20,sd=4)} = 0.7391658.$$

At first, most students find the preceding example counterintuitive. If x lies at the 0.90 quantile of X, then should we not predict $\hat{y}(x)$ to lie at the 0.90 quantile of Y? This is a natural first impression, but one that must be dispelled. We begin by considering two familiar situations:

1. Consider the case of a young boy whose father is extremely tall, at the 0.995 quantile of adult male heights. We surely would predict that the boy will grow to be quite tall. But precisely how tall? A father's height does not completely determine his son's height. Height is also

affected by myriad other factors, considered here as chance variation. Statistically speaking, it's more likely that the boy will grow to an adult height slightly shorter than his extremely tall father than that he will grow to be even taller.

2. Consider the case of two college freshman, William and Mary, who are enrolled in an introductory chemistry class of 250 students. On the first midterm examination, Mary attains the 5th highest score and William obtains the 245th highest (5th lowest) score. How should we predict their respective performances on the second midterm examonation? There is undoubtedly a strong, positive correlation between scores on the two tests. We surely will predict that Mary will do quite well on the second test and that William will do rather badly. But how well and how badly? One test score does not completely determine another—if it did, then computing semester grades would be easy! Mary can't do much better on the second test than she did on the first, but she might easily do worse. Statistically speaking, it's likely that she'll rank slightly below 5th on the second test. Likewise, William can't do much worse on the second test than he did on the first. Statistically speaking, it's likely that he'll rank slightly above 245th on the second test.

The phenomenon that we have just described, that experimental units with extreme X quantiles will tend to have less extreme Y quantiles, is purely statistical. It was first discerned by Sir Francis Galton, who called it "regression to mediocrity." Modern statisticians call it *regression to the mean*, or simply *the regression effect*.

Having refined our intuition, we can now explain the regression effect by examining the population concentration ellipse in Figure 14.1. For simplicity, we assume that X and Y have been converted to standard units. The bivariate normal population represented in Figure 14.1 has population parameters $\mu_x = \mu_y = 0$, $\sigma_x^2 = \sigma_y^2 = 1$, and $\rho = 0.5$. Recall that the line that coincides with the major axis of the ellipse is called the first principal component. In Figure 14.1, the first principal component is the line $y = x$ and the regression line is the line $y = x/2$. Both lines pass through the point $(\mu_x, \mu_y) = (0, 0)$, but their slopes differ by a factor of $|\rho| = 0.5$.

Let us explore the implications of the fact that, if $|\rho| < 1$, then *the regression line does not coincide with the major axis of the concentration ellipse*. Given $X = x$, it might seem tempting to predict $Y = x$. But this

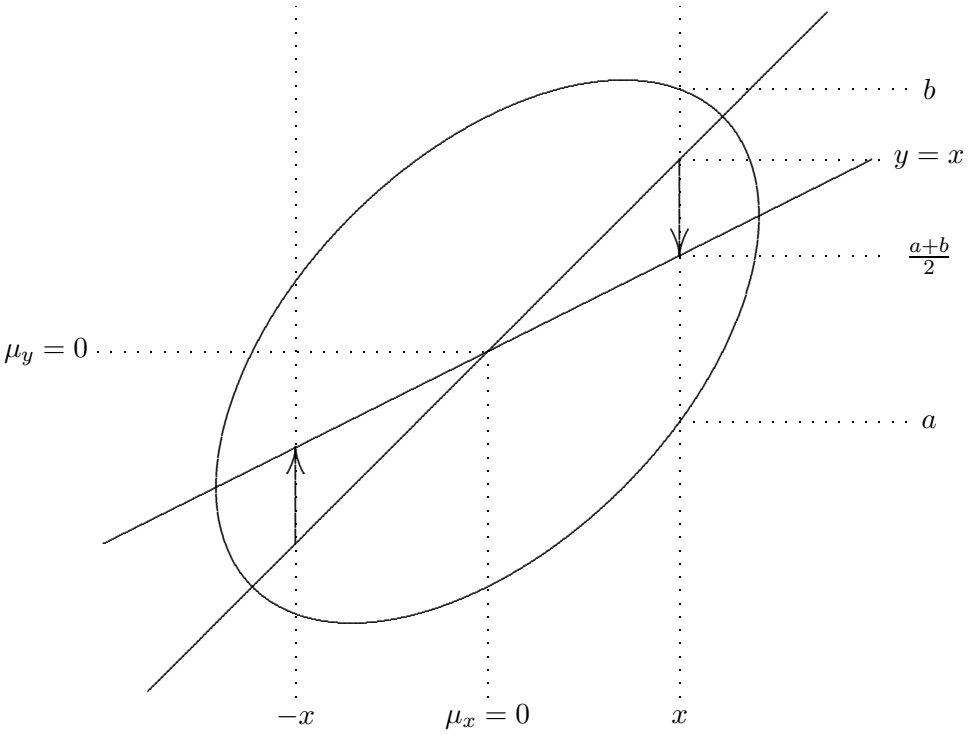

Figure 14.1: The Regression Effect.

would be a mistake! Here, $x > \mu_x$ and clearly

$$P\left(Y > x | X = x\right) < \frac{1}{2},$$

so $\hat{y}(x) = x$ overpredicts $Y|X = x$. Similarly, $\hat{y}(-x) = -x$ underpredicts $Y|X = -x$.

The population regression line is the line of conditional expected values, $y = E(Y|X = x)$. Let (x, a) and (x, b) denote the lower and upper points at which the vertical line $X = x$ intersects the population concentration ellipse. As one might guess, it turns out that

$$\hat{y}(x) = E(Y|X = x) = \frac{a + b}{2}.$$

However, the midpoint of the vertical line segment that connects (x, a) and (x, b) is *not* (x, x). The discrepancy between using the first principal component to predict $\hat{y}(x) = x$ and using the regression line to predict $\hat{y}(x) = (a + b)/2$, indicated by an arrow in Figure 14.1, is the regression effect.

The correlation coefficient ρ mediates the strength of the regression effect. If $\rho = \pm 1$, then

$$\frac{Y - \mu_x}{\sigma_y} = \pm \frac{X - \mu_x}{\sigma_x}$$

and Y is completely determined by X. In this case there is no regression effect: if x lies z standard deviations above μ_x, then we know that y lies z standard deviations above μ_y. At the other extreme, if $\rho = 0$, then knowing $X = x$ does not reduce the expected squared error of prediction at all. In this case, we regress all the way to the mean: regardless of where x lies, we predict $\hat{y} = \mu_y$.

Thus far, we have focussed on predicting Y from $X = x$ in the case that the population concentration ellipse is known. We have done so in order to emphasize that the regression effect is an inherent property of prediction, not a statistical anomaly caused by chance variation. In practice, however, the population concentration ellipse typically is not known and we must rely on the sample concentration ellipse, estimated from bivariate data. This means that we must substitute $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$ for $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. The *sample regression function* is

$$\hat{y}(x) = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x}) \tag{14.2}$$

and the corresponding line is the *sample regression line*. Notice that the slope of the sample regression line does not depend on whether we use plug-in or unbiased estimates of the population variances. The variances affect the regression line through the (square root of) their ratio,

$$\frac{\widehat{\sigma_y^2}}{\widehat{\sigma_x^2}} = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_y^2}{s_x^2},$$

which is not affected by the choice of plug-in or unbiased.

Example 14.2 (continued) I used `binorm.sample` to draw a sample of $n = 100$ observations from a bivariate normal distribution with parameters

$$\texttt{pop} = \left(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\right) = \left(10, 20, 2^2, 4^2, 0.5\right).$$

I then used `binorm.estimate` to compute sample estimates of `pop`, obtaining

$$\begin{aligned} \texttt{est} &= \left(\bar{x}, \bar{y}, s_x^2, s_y^2, r\right) \\ &= (10.0006837, 19.3985929, 4.4512393, 14.1754248, 0.4707309)\,. \end{aligned}$$

The resulting formula for the sample regression line is

$$\hat{y}(x) = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x}) = \bar{y} + 1.784545\,(x - \bar{x}) = 1.55192 + 1.784545x.$$

It is not difficult to create an `R` function that plots a scatter diagram of the sample and overlays both the sample concentration ellipse and the sample regression line. The function `binorm.regress` is described in Appendix R and/or can be obtained from the web page for this book/course. The commands used in this example are as follows:

```
> pop <- c(10,20,4,16,.5)
> Data <- binorm.sample(pop,100)
> est <- binorm.estimate(Data)
> binorm.regress(Data)
```

The scatter diagram created by `binorm.regress` is displayed in Figure 14.2.

14.2 The Method of Least Squares

In Section 14.1 we derived the regression line from properties of bivariate normal distributions. Having derived it, we now note that the sample regression line can be computed from *any* set of $n \geq 2$ points $(x_i, y_i) \in \Re^2$ for which the x_i assume more than one distinct value (and therefore $s_x > 0$). In this section, we derive the regression line in this more general setting.

Given points $(x_i, y_i) \in \Re^2$, $i = 1, \ldots, n$, we ask two conceptually distinct questions:

1. What line best *summarizes* the (x, y) pairs?

2. What line best *predicts* values of y from values of x?

We will answer each of these questions by applying the method of least squares. The possible lines are of the form $y = a + bx$. Given a candidate line, we measure the error between the line and each (x_i, y_i), then sum the

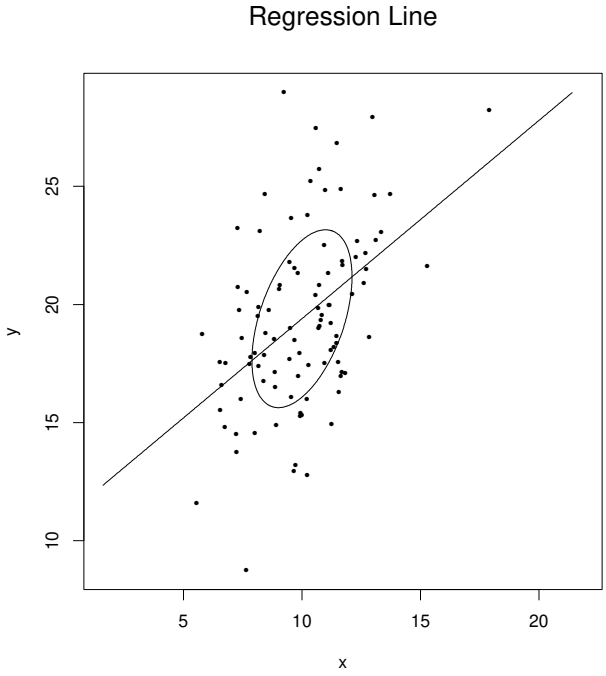

Figure 14.2: Scatter diagram, sample concentration ellipse, and sample regression line of $n = 100$ observations sampled from a bivariate normal distribution. Notice that the sample regression line is *not* the major axis of the sample concentration ellipse.

squared errors from $i = 1, \ldots, n$. The best line is the one that minimizes this sum of squared errors:

$$\min_{a,b} \sum_{i=1}^{n} \left[\text{error} \left(\begin{array}{c} (x_i, y_i) \\ y = a + bx \end{array} \right) \right]^2 \tag{14.3}$$

The distinction between (1) summary and (2) prediction lies in how we define error.

To define the line that best summarizes the (x, y) pairs, it is natural to define the error between a point and a line as the Euclidean distance from the point to the line. This is found by measuring the length of the perpendicular

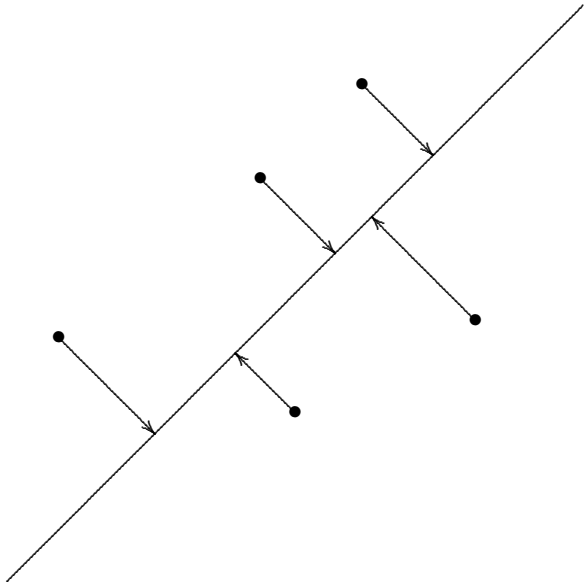

Figure 14.3: Perpendicular Errors for Summary

line segment that connects them, as in Figure 14.3. Thus,

$$\begin{array}{c}\text{summary}\\ \text{error}\end{array}\left(\begin{array}{c}(x_i,y_i)\\ y=a+bx\end{array}\right)=\begin{array}{c}\text{perpendicular}\\ \text{distance}\end{array}\left(\begin{array}{c}(x_i,y_i)\\ y=a+bx\end{array}\right).$$

Using this definition of error, the solution of Problem 14.3 is the major axis of the sample concentration ellipse, the first principal component of the sample. We emphasize: *the first principal component is used for summary, not prediction.*

In contrast, to define the line that best predicts y values from x values, it is natural to define the error between a point (x_i, y_i) and a line $y = a + bx$ as the difference between the observed value $y = y_i$ and the predicted value

$$y = \hat{y}\,(x_i) = a + bx_i.$$

The difference $y_i - \hat{y}(x_i)$ is a *residual error* and the absolute difference $|y_i - \hat{y}(x_i)|$ is the length of the vertical line segment that connects (x_i, y_i) and $y = a + bx$, as in Figure 14.4. Using this definition of error, the solution of

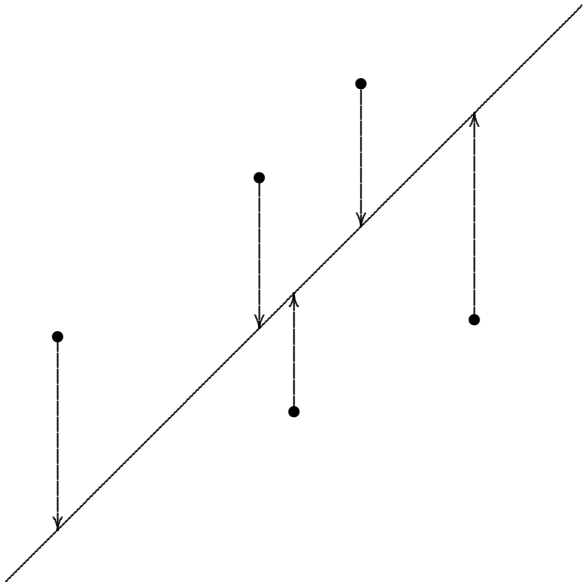

Figure 14.4: Vertical Errors for Prediction

Problem 14.3 is the sample regression line. We emphasize: *the regression line is used for prediction, not summary.*

The remainder of this section provides a more detailed exposition of the squared error approach to prediction. Let

$$\mathrm{SS}(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 ,$$

the sum of the squared residual errors that result from the prediction function $\hat{y}(x) = a + bx$. The method of least squares chooses (a, b) to minimize $\mathrm{SS}(a, b)$. Before analyzing this problem, we first consider an easier problem. If we knew $\{y_1, \ldots, y_n\}$ but not the corresponding $\{x_1, \ldots, x_n\}$, then it would be impossible to measure errors associated with prediction functions that involve x. In this situation we would be forced to restrict attention to prediction functions of the form $\hat{y} = a$, which corresponds to restricting attention to lines with zero slope. The method of least squares then chooses a to minimize

$$\sum_{i=1}^{n} (y_i - a)^2 = \mathrm{SS}(a, 0).$$

Theorem 14.2 *The value of a that minimizes $\mathrm{SS}(a,0)$ is $a=\bar{y}$.*

Proof We can conclude that $\mathrm{SS}(a,0)/n$ is minimal when $a=\bar{y}$ by applying part (2) of Theorem 6.1 to the empirical distribution of $\{y_1,\ldots,y_n\}$; however, it is instructive to verify this conclusion by direct calculation:

$$\begin{aligned}
\mathrm{SS}(a,0) &= \sum_{i=1}^{n}(y_i-a)^2=\sum_{i=1}^{n}(y_i-\bar{y}+\bar{y}-a)^2 \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^2+\sum_{i=1}^{n}2(y_i-\bar{y})(\bar{y}-a)+\sum_{i=1}^{n}(\bar{y}-a)^2 \\
&= (n-1)s_y^2+2(\bar{y}-a)\left[\sum_{i=1}^{n}y_i-n\bar{y}\right]+n(\bar{y}-a)^2 \\
&= (n-1)s_y^2+n(\bar{y}-a)^2
\end{aligned}$$

The second term in this expression is the only term that involves a. It achieves its minimal value of zero when $a=\bar{y}$. □

For future reference, we define the *total sum of squares* to be

$$\mathrm{SS}_T=\mathrm{SS}(\bar{y},0)=\sum_{i=1}^{n}(y_i-\bar{y})^2=(n-1)s_y^2.$$

This is the smallest squared error possible when predicting y without information about x.

Now we consider the problem of finding the line $y=a+bx$ that best predicts values of y from values of x. The method of least squares chooses (a,b) to minimize $\mathrm{SS}(a,b)$. Let (a^*,b^*) denote the minimizing values of (a,b) and define the *error sum of squares* to be

$$\mathrm{SS}_E=\mathrm{SS}(a^*,b^*).$$

Because we have not restricted attention to $b=0$, $\hat{y}(x)=a^*+b^*x$ must predict at least as well as $\hat{y}=\bar{y}$. Thus,

$$\mathrm{SS}_E=\mathrm{SS}(a^*,b^*)\le\mathrm{SS}(\bar{y},0)=\mathrm{SS}_T.$$

We have already stated that $y=a^*+b^*x$ is the sample regression line. We can verify that statement by a calculation that resembles the proof of Theorem 14.2.

Theorem 14.3 *Let $(x_i, y_i) \in \Re^2$, $i = 1, \ldots, n$, be a set of (x, y) pairs with at least two distinct values of x. Let*

$$b^* = r\frac{s_y}{s_x} \quad \textit{and} \quad a^* = \bar{y} - b^*\bar{x}.$$

Then

$$\mathrm{SS}\left(a^*, b^*\right) \leq \mathrm{SS}(a, b)$$

for all choices of (a, b).

Proof First, write

$$\begin{aligned}
\mathrm{SS}(a,b) &= \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y} + \bar{y} - b\bar{x} + b\bar{x} - a - bx_i\right)^2 \\
&= \sum_{i=1}^{n} \left[\left(y_i - \bar{y}\right) + \left(\bar{y} - b\bar{x} - a\right) - b\left(x_i - \bar{x}\right)\right]^2.
\end{aligned}$$

Expanding the square in this expression results in six terms. The three squared terms are:

$$\begin{aligned}
\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 &= (n-1)s_y^2, \\
\sum_{i=1}^{n} \left(\bar{y} - b\bar{x} - a\right)^2 &= n\left(\bar{y} - b\bar{x} - a\right)^2, \\
\sum_{i=1}^{n} (-b)^2 \left(x_i - \bar{x}\right)^2 &= b^2 \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = b^2(n-1)s_x^2.
\end{aligned}$$

The three cross-product terms are:

$$\begin{aligned}
&\sum_{i=1}^{n} 2\left(y_i - \bar{y}\right)\left(\bar{y} - b\bar{x} - a\right) = 2\left(\bar{y} - b\bar{x} - a\right) \sum_{i=1}^{n} \left(y_i - \bar{y}\right) \\
&\quad = 2\left(\bar{y} - b\bar{x} - a\right)\left[\sum_{i=1}^{n} y_i - n\bar{y}\right] = 0, \\
&\sum_{i=1}^{n} 2\left(y_i - \bar{y}\right)(-b)\left(x_i - \bar{x}\right) = -2b \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \\
&\quad = -2b(n-1)s_x s_y \frac{\frac{1}{n-1}\sum_{i=1}^{n} \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{s_x s_y} = -2b(n-1)s_x s_y r, \\
&\sum_{i=1}^{n} 2\left(\bar{y} - b\bar{x} - a\right)(-b)\left(x_i - \bar{x}\right) = -2b\left(\bar{y} - b\bar{x} - a\right) \sum_{i=1}^{n} \left(x_i - \bar{x}\right) = 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{SS}(a,b) &= (n-1)s_y^2 + n\left(\bar{y} - b\bar{x} - a\right)^2 + b^2(n-1)s_x^2 - 2b(n-1)s_x s_y r \\
&= n\left(\bar{y} - b\bar{x} - a\right)^2 + (n-1)\left[b^2 s_x^2 - 2bs_x r s_y + r^2 s_y^2\right] \\
&\quad -(n-1)r^2 s_y^2 + (n-1)s_y^2 \\
&= n\left(\bar{y} - b\bar{x} - a\right)^2 + (n-1)\left[bs_x - rs_y\right]^2 + \left(1 - r^2\right)(n-1)s_y^2.
\end{aligned}$$

The third term in this expression does not involve b or a. The second term achieves its minimal value of zero when $b = rs_y/s_x = b^*$. The first term is the only term that involves a. Whatever the value of b, the first term achieves its minimal value of zero when $a = \bar{y} - b\bar{x}$. Hence, for $b = b^*$, the minimizing value of a is $a = \bar{y} - b^*\bar{x} = a^*$. □

The total sum of squares, SS_T, measures the prediction error from $\hat{y} = \bar{y}$. The error sum of squares,

$$\begin{aligned}
\mathrm{SS}_E &= \mathrm{SS}\left(a^*, b^*\right) = \sum_{i=1}^{n}\left[y_i - \left(\bar{y} - b^*\bar{x}\right) - b^* x_i\right]^2 \\
&= \sum_{i=1}^{n}\left[y_i - \bar{y} - b^*\left(x_i - \bar{x}\right)\right]^2 = \sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right) - r\frac{s_y}{s_x}\left(x_i - \bar{x}\right)\right]^2 \\
&= \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 - 2r\frac{s_y}{s_x}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) + r^2\frac{s_y^2}{s_x^2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \\
&= (n-1)s_y^2 - 2rs_y^2(n-1)\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{s_x s_y} + r^2 s_y^2(n-1) \\
&= (n-1)s_y^2 - 2(n-1)s_y^2 r^2 + r^2(n-1)s_y^2 \\
&= (n-1)s_y^2\left(1 - r^2\right) \\
&= \left(1 - r^2\right)\mathrm{SS}_T,
\end{aligned}$$

measures the prediction error from the sample regression line. Now we define the *regression sum of squares* to be the sum of the squared differences between the two predictions,

$$\begin{aligned}
\mathrm{SS}_R &= \sum_{i=1}^{n}\left[\hat{y} - \hat{y}\left(x_i\right)\right]^2 = \sum_{i=1}^{n}\left[\bar{y} - \bar{y} - r\frac{s_y}{s_x}\left(x_i - \bar{x}\right)\right]^2 \\
&= r^2\frac{s_y^2}{s_x^2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = r^2 s_y^2(n-1) = r^2\mathrm{SS}_T.
\end{aligned}$$

The three sums of squares $(\mathrm{SS}_R, \mathrm{SS}_E, \mathrm{SS}_T)$ are precisely analogous to the three sums of squares $(\mathrm{SS}_B, \mathrm{SS}_W, \mathrm{SS}_T)$ that arise in the analysis of variance and they enjoy an identical property:

$$\mathrm{SS}_R + \mathrm{SS}_E = r^2 \mathrm{SS}_T + \left(1 - r^2\right) \mathrm{SS}_T = \mathrm{SS}_T$$

This is the Pythagorean Theorem in n-dimensional Euclidean space! The points

$$A = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}, \quad B = \begin{bmatrix} \bar{y} - r\frac{s_y}{s_x}(x_1 - \bar{x}) \\ \vdots \\ \bar{y} - r\frac{s_y}{s_x}(x_n - \bar{x}) \end{bmatrix}, \quad C = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

are the vertices of a right triangle in $\Re^n$. The right angle occurs at vertex B. The squared Euclidean distances of the sides that meet at B are

$$d^2(A, B) = \mathrm{SS}_R \quad \text{and} \quad d^2(B, C) = \mathrm{SS}_E$$

and the squared Euclidean distance of the hypotenuse is

$$d^2(A, C) = \mathrm{SS}_T,$$

so

$$d^2(A, B) + d^2(B, C) = \mathrm{SS}_R + \mathrm{SS}_E = \mathrm{SS}_T = d^2(A, C).$$

To quantify the extent to which knowledge of x improves our ability to predict y, we measure the proportion by which the squared error of prediction is reduced when we use the sample regression line instead of the constant prediction $\hat{y} = \bar{y}$. This proportion is just

$$\frac{\mathrm{SS}(\bar{y}, 0) - \mathrm{SS}(a, b)}{\mathrm{SS}(\bar{y}, 0)} = \frac{\mathrm{SS}_T - \mathrm{SS}_E}{\mathrm{SS}_T} = \frac{\mathrm{SS}_R}{\mathrm{SS}_T} = \frac{r^2 \mathrm{SS}_T}{\mathrm{SS}_T} = r^2,$$

the sample coefficient of determination. Again, we conclude that the square of Pearson's product-moment correlation coefficient measures the proportion of variation "explained" by simple linear regression.

Example 14.2 (continued) For the bivariate sample displayed in Figure 14.2, the total sum of squares is

$$\mathrm{SS}_T = (n-1)s_y^2 = 99 \cdot 14.1754248 = 1403.3671$$

and the coefficient of determination is

$$r^2 = 0.4707309^2 = 0.2215876.$$

Hence, the regression sum of squares is

$$\mathrm{SS}_R = r^2 \mathrm{SS}_T = 0.2215876 \cdot 1403.367 = 310.9688$$

and the error sum of squares is

$$\mathrm{SS}_E = \mathrm{SS}_T - \mathrm{SS}_R = 1403.3671 - 310.9688 = 1092.3983.$$

14.3 Computation

A bivariate sample consists of $2n$ numbers. However, all of the quantities used in the preceding sections can be computed from just six fundamental quantities:

$$n \quad \sum_{i=1}^n x_i \quad \sum_{i=1}^n y_i \quad \sum_{i=1}^n x_i^2 \quad \sum_{i=1}^n y_i^2 \quad \sum_{i=1}^n x_i y_i$$

These quantities are used by many calculators. One reason that they are so convenient is that they are easily incremented as new (x, y) pairs are observed.

Example 14.2 (continued) For the bivariate sample displayed in Figure 14.2, the six fundamental quantities are as follows:

$$n = 100 \quad \sum_{i=1}^n x_i = 1000.068 \quad \sum_{i=1}^n y_i = 1939.859$$
$$\sum_{i=1}^n x_i^2 = 10442.04 \quad \sum_{i=1}^n y_i^2 = 39033.91 \quad \sum_{i=1}^n x_i y_i = 19770.1$$

Now suppose that we draw another (x, y) pair from the same population, say $(8.9, 13.5)$. Then the new sample has the following fundamental quantities:

$$n = 100 + 1 \qquad \sum_{i=1}^n x_i^2 = 10442.04 + 8.9^2$$
$$\sum_{i=1}^n x_i = 1000.068 + 8.9 \qquad \sum_{i=1}^n y_i^2 = 39033.91 + 13.5^2$$
$$\sum_{i=1}^n y_i = 1939.859 + 13.5 \qquad \sum_{i=1}^n x_i y_i = 19770.1 + 8.9 \cdot 13.5$$

Three useful quantities are easily computed from the six fundamental quantities:

$$\begin{aligned}
t_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n} \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2 \right) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \\
t_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2 \\
t_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} \left(y_i^2 - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y} \right) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)
\end{aligned}$$

These quantities are useful because all of the important quantities derived in the preceding sections are easily computed from them. Here are the formulas:

1. Sample variances:

$$s_x^2 = \frac{t_{xx}}{n-1} \qquad s_y^2 = \frac{t_{yy}}{n-1}$$

2. Pearson's correlation coefficient:

$$\begin{aligned}
r &= \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{t_{xy}}{\sqrt{t_{xx}} \sqrt{t_{yy}}} \\
r^2 &= \frac{t_{xy}^2}{t_{xx} t_{yy}}
\end{aligned}$$

3. Sample regression coefficients:

$$\begin{aligned}
b^* &= r \frac{s_y}{s_x} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x^2} = \frac{t_{xy}}{t_{xx}} \\
a^* &= \bar{y} - b^* \bar{x} = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{t_{xy}}{t_{xx}} \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

4. Sums of squares:

$$\begin{aligned} \mathrm{SS}_T &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = t_{yy} \\ \mathrm{SS}_R &= r^2 \mathrm{SS}_T = \frac{t_{xy}^2}{t_{xx} t_{yy}} t_{yy} = \frac{t_{xy}^2}{t_{xx}} \\ \mathrm{SS}_E &= \mathrm{SS}_T - \mathrm{SS}_R = t_{yy} - \frac{t_{xy}^2}{t_{xx}} \end{aligned}$$

14.4 The Simple Linear Regression Model

Let $x_1, \ldots, x_n$ be a list of real numbers for which $s_x > 0$. Suppose that:

1. Associated with each x_i is a random variable

$$Y_i \sim \text{Normal}\left(\mu_i, \sigma^2\right).$$

   Notice that the Y_i have a common population variance $\sigma^2 > 0$. This is analogous to the homoscedasticity assumption of the analysis of variance.

2. The population means μ_i satisfy the linear relation

$$\mu_i = \beta_0 + \beta_1 x_i$$

   for some $\beta_0, \beta_1 \in \Re$. The population parameters (β_0, β_1) are called the population regression coeffcients.

These assumptions define the *simple linear regression model*. Suppose that we sample from a bivariate normal distribution, then condition on the observed values $x_1, \ldots, x_n$. It follows from Theorem 14.1 that this is a special case of the simple linear regression model in which

$$\begin{aligned} \beta_1 &= \rho \frac{\sigma_y}{\sigma_x}, \\ \beta_0 &= \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x = \mu_y - \beta_1 \mu_x, \\ \sigma^2 &= \left(1 - \rho^2\right) \sigma_y^2. \end{aligned}$$

The simple linear regression model has three unknown parameters. The method of least squares estimates (β_0, β_1) by

$$\begin{aligned} \hat{\beta}_1 &= b^* = r\frac{s_y}{s_x} = \frac{t_{xy}}{t_{xx}}, \\ \hat{\beta}_0 &= a^* = \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

These are also the plug-in estimates of (β_0, β_1), and the plug-in estimate of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \frac{1}{n}\mathrm{SS}_E.$$

We proceed to explore some properties of the corresponding estimators. These properties are consequences of the following key facts:

Theorem 14.4 *Under the assumptions of the simple linear regression model, the random variables $\hat{\beta}_1$ and SS_E are independent and satisfy*

$$\hat{\beta}_1 \sim \text{Normal}\left(\beta_1, \frac{\sigma^2}{t_{xx}}\right) \tag{14.4}$$

$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-2). \tag{14.5}$$

It follows from (14.4) that $E\hat{\beta}_1 = \beta_1$, and consequently that

$$\begin{aligned} E\hat{\beta}_0 &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i - \hat{\beta}_1 x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - \beta_1 x_i) = \frac{1}{n}\sum_{i=1}^{n}\beta_0 = \beta_0. \end{aligned}$$

Thus, $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased estimators of (β_0, β_1). Furthermore, it follows from (14.5) and Corollary 5.1 that $E(\mathrm{SS}_E/\sigma^2) = n-2$. Hence, $E[\mathrm{SS}_E/(n-2)] = \sigma^2$ and

$$\mathrm{MS}_E = \frac{1}{n-2}\mathrm{SS}_E$$

is an unbiased estimator of σ^2.

Converting (14.4) to standard units results in

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/t_{xx}}} \sim \text{Normal}(0,1). \tag{14.6}$$

Dividing (14.6) by (14.5), it follows from Definition 5.7 that

$$\frac{\left(\hat{\beta}_1 - \beta_1\right)/\sqrt{\sigma^2/t_{xx}}}{\sqrt{\frac{\mathrm{SS}_E}{\sigma^2}/(n-2)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{MS}_E/t_{xx}}} \sim t(n-2).$$

This fact allows us to construct confidence intervals for β_1. Given α, we first compute the critical value

$$q_t = \texttt{qt}(1 - \alpha/2, n-2).$$

Then

$$\hat{\beta}_1 \pm q_t \sqrt{\frac{\mathrm{MS}_E}{t_{xx}}}$$

is a $(1-\alpha)$-level confidence interval for β_1.

Remark: It may be helpful to write

$$\begin{aligned} \frac{\mathrm{MS}_E}{t_{xx}} &= \frac{(1-r^2)\,\mathrm{SS}_T/(n-2)}{(n-1)s_x^2} = \frac{(1-r^2)\,(n-1)s_y^2/(n-2)}{(n-1)s_x^2} \\ &= \left(1-r^2\right)\frac{s_y^2}{s_x^2}/(n-2). \end{aligned}$$

Example 14.3 Suppose that $n = 100$ bivariate observations produce the following estimates:

$$\begin{aligned} \bar{x} &= 97.255564 \\ \bar{y} &= 103.872210 \\ s_x^2 &= 425.062476 \\ s_y^2 &= 872.229230 \\ r &= -0.485857 \end{aligned}$$

To construct a 0.95-level confidence interval for β_1, we first compute

$$\hat{\beta}_1 = r\frac{s_y}{s_x} = -0.5070697 \cdot \sqrt{\frac{414.7388683}{434.9825540}} = -0.695981,$$

$q_t = \texttt{qt}(.975, \texttt{df} = 98) = 1.984467$, and

$$\frac{\mathrm{MS}_E}{t_{xx}} = \frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2} = \frac{1 - 0.485857^2}{98} \cdot \frac{872.229230^2}{425.062476^2} = 0.01599605.$$

The desired confidence interval is then

$$\begin{aligned}\hat{\beta}_1 \pm q_t\sqrt{\frac{\mathrm{MS}_E}{t_{xx}}} &= -0.695981 \pm 1.984467 \cdot \sqrt{0.01599605} \\ &= (-0.9469675, -0.4449945).\end{aligned}$$

Next we consider how to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. This is an important decision because rejecting $H_0 : \beta_1 = 0$ means that we are convinced that values of x help us to predict values of y. Furthermore, if we sampled from a bivariate normal population, then

$$\beta_1 = \rho\frac{\sigma_y}{\sigma_x} = 0$$

if and only if $\rho = 0$. Because normal random variables X and Y are independent if and only if they are uncorrelated, the null hypothesis $H_0 : \beta_1 = 0$ is equivalent to the null hypothesis that X and Y are independent.

If $\beta_1 = 0$, then

$$\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E/t_{xx}}} \sim t(n-2).$$

Hence, the significance probability for testing $H_0 : \beta_1 = 0$ is

$$\mathbf{p} = P\left(|T| \geq \left|\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E/t_{xx}}}\right|\right),$$

where the random variable $T \sim t(n-2)$, and we reject $H_0 : \beta_1 = 0$ if and only if $\mathbf{p} \leq \alpha$. Equivalently, we reject $H_0 : \beta_1 = 0$ if and only if we observe

$$\left|\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E/t_{xx}}}\right| \geq q_t,$$

where q_t is the critical value defined above. Notice that

$$\begin{aligned}\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E/t_{xx}}} &= \frac{t_{xy}/t_{xx}}{\sqrt{\mathrm{MS}_E/t_{xx}}} = \frac{t_{xy}}{\sqrt{t_{xx}}}\frac{1}{\sqrt{\mathrm{SS}_E/(n-2)}} \\ &= \frac{t_{xy}}{\sqrt{t_{xx}}\sqrt{t_{yy}}}\frac{\sqrt{t_{yy}}\sqrt{n-2}}{\sqrt{t_{yy} - t_{xy}^2/t_{xx}}} \\ &= r\frac{\sqrt{n-2}}{\sqrt{1 - t_{xy}^2/(t_{xx}t_{yy})}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},\end{aligned}$$

so this is the same t-test that we described in Section 13.2.3 for testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

It follows from Theorem 5.5 that

$$\left(\frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E / t_{xx}}} \right)^2 \sim F(1, n-2).$$

Hence, an F-test that is equivalent to the t-test derived in the preceding paragraph rejects $H_0 : \beta_1 = 0$ if and only if we observe

$$(n-2)\frac{r^2}{1-r^2} \geq q_F,$$

where the critical value q_F is defined by

$$q_F = \texttt{qf}(1-\alpha, 1, n-2).$$

Equivalently, we reject $H_0 : \beta_1 = 0$ if and only if the significance probability

$$\mathbf{p} = P\left(F \geq (n-2)\frac{r^2}{1-r^2} \right) \leq \alpha,$$

where the random variable $F \sim F(1, n-2)$. The results of the F-test of $H_0 : \beta_1 = 0$ are traditionally presented in the form of an ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Test Statistic	**p**-Value
Regression	$r^2\mathrm{SS}_T$	1	$r^2\mathrm{SS}_T$	$(n-2)\frac{r^2}{1-r^2}$	**p**
Error	$(1-r^2)\mathrm{SS}_T$	$n-2$	$\frac{1-r^2}{n-2}\mathrm{SS}_T$		
Total	SS_T				

Example 14.3 (continued) Let us now test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ at a significance level of $\alpha = 0.05$. Of course, we know that we will reject H_0 because the 0.95-level confidence interval constructed from these data did not contain the hypothesized slope $\beta_1 = 0$.

The t-test statistic is

$$t = \frac{\hat{\beta}_1}{\sqrt{\mathrm{MS}_E / t_{xx}}} = \frac{-0.695981}{\sqrt{0.01599605}} = -5.502893,$$

which results in a significance probability of

$$\mathbf{p} = 2 * \texttt{pt}(-5.502893, \texttt{df} = 98) = 2.989589 \times 10^{-7}.$$

Because $\mathbf{p} < \alpha$, we reject $H_0 : \beta_1 = 0$.

Equivalently, we can compute $SS_T = (n-1)s_x^2 = 99 \cdot 425.062476 \doteq 42081.19$ and $r^2 = 0.236057$, then construct the following ANOVA table:

Source of Variation	Sum of Squares	DF	Mean Square	F-Test Statistic	**p**-Value
Regression	20383.689	1	20383.6885	30.28183	2.989589×10^{-7}
Error	65967.005	98	673.1327		
Total	86350.694				

Again, we reject $H_0 : \beta_1 = 0$ because $\mathbf{p} < \alpha$. Notice that we obtain the same significance probability with either test.

Although equivalent, the t-test and F-test of $H_0 : \beta_1 = 0$ each enjoy certain advantages. The former is more flexible, as it is easily adapted to test 1-sided hypotheses. The F-test is more readily generalized to testing a variety of hypotheses that naturally arise when studying more complicated regression models.

14.5 Regression Diagnostics

14.6 Exercises

1. According to Stanford University Professor Claude M. Steele (Not just a test, *The Nation*, May 3, 2004, page 40),

   > "The SAT, for example, correlates .42 with freshman grades... This means that it measures about 18 percent of the characteristics, whatever they are, that determine freshman grades."

   Comment on this passage. Do you agree with Professor Steele's interpretation of what $r = 0.42$ means?

2. Suppose that (X, Y) have a bivariate normal distribution with parameters $(5, 3, 1, 4, 0.5)$. Compute the following quantities:

   (a) $P(Y > 6)$
   (b) $E(Y|X = 6.5)$
   (c) $P(Y > 6|X = 6.5)$

3. Assume that the population of all sister-brother heights has a bivariate normal distribution and that the data in Exercise 13.5.3 were sampled from this population. Use these data in the following:

   (a) Consider the population of all sister-brother heights. Estimate the proportion of all brothers who are at least $5'\ 10''$.
   (b) Suppose that Carol is $5'\ 1''$. Predict her brother's height.
   (c) Consider the population of all sister-brother heights for which the sister is $5'\ 1''$. Estimate the proportion of these brothers who are at least $5'\ 10''$.

4. Assume that the population of all sister-brother heights has a bivariate normal distribution and that the data in Exercise 13.5.2 were sampled from this population. Use these data in the following:

   (a) Compute the sample coefficient of determination, the proportion of variation "explained" by simple linear regression.
   (b) Let $\alpha = 0.05$. Do these data provide convincing evidence that knowing a sister's height (x) helps one predict her brother's height (y)?
   (c) Construct a 0.90-level confidence interval for the slope of the population regression line for predicting y from x.

(d) Suppose that you are planning to conduct a more comprehensive study of sibling heights. Your goal is to better estimate the slope of the population regression line for predicting y from x. If you want to construct a 0.95-level confidence interval of length 0.1, then how many sister-brother pairs should you plan to observe?

Hint:

$$\frac{\mathrm{MS}_E}{t_{xx}} = \left(1 - r^2\right) \frac{s_y^2}{s_x^2} / (n-2).$$

5. A class of 35 students took two midterm tests. Jack missed the first test and Jill missed the second test. The 33 students who took both tests scored an average of 75 points on the first test, with a standard deviation of 10 points, and an average of 64 points on the second test, with a standard deviation of 12 points. The scatter diagram of their scores is roughly ellipsoidal, with a correlation coefficient of $r = 0.5$.

   Because Jack and Jill each missed one of the tests, their professor needs to guess how each would have performed on the missing test in order to compute their semester grades.

   (a) Jill scored 80 points on Test 1. She suggests that her missing score on Test 2 be replaced with her score on Test 1, 80 points. What do you think of this suggestion? What score would you advise the professor to assign?

   (b) Jack scored 76 points on Test 2, precisely one standard deviation above the Test 2 mean. He suggests that his missing score on Test 1 be replaced with a score of 85 points, precisely one standard deviation above the Test 1 mean. What do you think of this suggestion? What score would you advise the professor to assign?

6. In a study of "Heredity of head form in man" (*Genetica*, 3:193–384, 1921), G.P. Frets reported two head measurements (in millimeters) for each of the first two adult sons of 25 families. These data are reproduced as Data Set 111 in *A Handbook of Small Data Sets*, and

can also be downloaded from the web page for this couse.

First Son		Second Son	
Length	Breadth	Length	Breadth
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151
179	158	186	148
183	147	174	147
174	150	185	152
190	159	195	157
188	151	187	158
163	137	161	130
195	155	183	158
186	153	173	148
181	145	182	146
175	140	165	137
192	154	185	152
174	143	178	147
176	139	176	143
197	167	200	158
190	163	187	150

For each head, we will compute two variables:

```
size <- length+breadth
shape <- length-breadth
```

(a) Consider head size. Investigate the relation between first son head size and second son head size. Can we reject the null hypothesis that these variables are uncorrelated? Of the variation in second son head size, what proportion is explained by variation in first son head size?

(b) Consider head shape. Investigate the relation between first son head shape and second son head shape. Can we reject the null hypothesis that these variables are uncorrelated? Of the variation in second son head shape, what proportion is explained by variation in first son head shape?

(c) In another family from the same era, the first adult son's head had a length of 195 millimeters and a breadth of 160 millimeters. Use this information to guess the size of the second adult son's head.

7. In the athletics event known as the shot put, male competitors "put" the "shot," a 16-pound metal ball. (Female competitors use a smaller shot.) In the United States, high school male competitors put a 12-pound shot, then graduate to the 16-pound shot used in NCAA, USATF, and IAAF competition. In its August 2002 "Stat Corner," the respected athletics periodical *Track & Field News* proclaimed an "Inverse Relationship Between 12 & 16lb Shots:"

> "A look at the accompanying all-time Top 11 lists for high schoolers with the 12lb shot—11 because there have been 11 of them over 70 [feet]—and for U.S. men with the 16 sends two messages to aspiring prep putters:
>
> - If you're not very good in high school, don't worry about it; few of the big guys were either.
> - If you're great in high school, that may be about as good as you'll ever get.
>
> "The numbers are astounding. We'll leave it to a technical expert to figure out why..."

The numbers follow.[1] Do you agree with T&FN's two messages?

[1]Perhaps the most astounding number is Michael Carter's prodigious heave of 81-3.50, arguably the most formidable record in all of track and field. Carter broke an 11-year-old record by *nine feet*! He went on to a sensational college career at SMU, winning the NCAA championship and a silver medal at the 1984 Olympic Games. He then opted for a career in professional football, becoming an All-Pro defensive lineman for the NFL Champion San Francisco 49er's.

ALL-TIME HIGH SCHOOL 70-FOOTERS			
	12	*16*	*16–12*
1. Michael Carter '79	81-3.5	71-4.75	–9-10.75
2. Brent Noon '90	76-2	70-5.75	–5-8.25
3. Arnold Campbell '84	74-10.5	64-3	–10-7.5
4. Charles Moye '87	72-8	57-1	–15-7
5. Sam Walker '68	72-3.25	66-9.5	–5-5.75
6. Jesse Stuart '70	71-11i	68-11.5i	–2-11.5
7. Roger Roesler '96	71-2	61-6.25	–11-7.75
8. Kevin Bookout '02	71-1.5	(too early still)	
9. Doug Lane '68	70-11	66-11.25	–3-11.75
10. Dennis Black '91	70-7	68-10	–1-9
11. Ron Semkiw '72	70-1.75	70-0.5	–0-1.25

ALL-TIME U.S. TOP 11			
	16	*12*	*16–12*
1. Randy Barnes '90	75-10.25	66-9.5	+9-0.75
2. Brian Oldfield '75	75-0	58-10	+16-2
3. John Brenner '87	73-10.75	64-5.5	+9-5.25
4. Adam Nelson '02	73-10.25	63-2.25	+10-8
5. Kevin Toth '02	72-9.75	58-11	+13-10.75
6. George Woods '74	72-3i	60-11	+11-4
6. Dave Laut '82	72-3	65-9	+6-6
6. John Godina '99	72-3	64-1.25	+8-1.75
9. Gregg Trafalis '92	72-1.5	57-0	+15-1.5
10. Terry Albritton '76	71-8.5	67-9	+3-11.5
11. Andy Bloom '00	71-7.25	64-2.5	+7-4.74

Chapter 15

Simulation-Based Inference

15.1 Termite Foraging Revisited

Appendix R

A Statistical Programming Language

R.1 Introduction

R.1.1 What is R?

In the 1970s, researchers at AT&T Bell Laboratories developed `S`, a high-level statistical programming language that became popular with academic statisticians. Bell Labs subsequently licensed `S` to a company that added a variety of capabilities, creating the commercial product `S-Plus`. `R` is yet another implementation of `S`. The R Project for Statistical Computing is an ongoing effort by a group of statisticians to extend and improve `R`.

`R` is free, Open Source software, that can be downloaded in compiled or source code form. It runs on a variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and MacOS. The primary web site for information about `R` is:

`http://www.r-project.org/`

R.1.2 Why Use R?

This question encompasses several issues. First, there is the question of what role statistical software is to play in the course. Introductory statistics courses may use software in different ways. Once upon a time, many instructors (myself included) avoided using software in the first semester. The rationale for this approach is that one should begin one's study of statistics by focussing on basic concepts and learn what the computer is doing

before one uses the computer to do it. Unfortunately, this approach condemns one to analyzing fairly trivial data sets, and even then calculating by hand and/or calculator quickly becomes extremely tedious. As a result, this approach has fallen from favor.

At the other end of the spectrum, many introductory statistics courses use statistics packages like Minitab, SPSS, or SAS to analyze data. Such packages are extremely useful and every statistician should have some familiarity with at least one such package. However, if one begins to rely on such packages too quickly, the package may be viewed as a black box and the student may never really learn what that black box is doing.

There are many different ways to introduce the subject of statistics, and no one way is best for all students. This book is intended for students who want to *understand* what is going on inside the black box procedures available in so many statistics packages. This intention determines the use that we shall make of the computer. We will strive for an intermediate approach, in which the computer is used to relieve the tedium of calculation, but in which the student is obliged to tell the computer what intermediate steps need to be performed in order to obtain the desired output. Such an approach requires a high-level, interactive programming language. Several such languages are available, but `S-Plus` and `R` have achieved the greatest popularity within the statistics community. Acquiring some familiarity with `S-Plus` and/or `R` will benefit students who continue to study statistics and/or analyze data in the future.

Why `R` instead of `S-Plus`? For most of the examples in this book, `R` and `S-Plus` are interchangeable—the same commands work for both. But `R` has two compelling advantages. First, `R` is available for certain operating systems for which `S-Plus` is not, e.g., MacOS. Second, `R` is free! As a result, students who begin using `R` in this course can be confident that they will always have access to `R`.

R.1.3 Installing R

To efficiently download software, documentation, etc., you should use a nearby CRAN (Comprehensive R Archive Network) mirror site, e.g., Statlib at Carnegie Mellon University:

`http://lib.stat.cmu.edu/R/CRAN/`

Most students will want to install `R` in compiled form by downloading executable binary files. On-line documentation and several manuals are in-

cluded, although you may find it easier to get started using the examples provided in this book.

R.1.4 Learning About R

R is far too complicated to learn in one (or even several) lessons. I doubt that any one person—including the R developers—knows everything about R! But don't be intimidated: *the best way to learn* R *is to just start using* R. And, the best time to use R is when you're trying to accomplish a specific task. Try to learn bits and pieces of R as they're introduced in the text and/or you develop an interest in a specific capability.

Of course, it's hard to learn anything without documentation. The material in this book, both the examples scattered throughout various chapters to illustrate various statistical methods and the tutorial material in this appendix, is a good way to get started. Once you know the name of one R function, you can learn more about it and discover related functions using various utilities included in your R installation. If you're using the Windows version of R then you can start by exploring the `Help` menu in `RGui`, which will lead you to manuals, search utilities, and web pages. I tend to use `R functions (text)` for help on specific functions.

R.2 Using R

R is an interpreted language, designed to be used interactively. The user is prompted to issue a command as follows:

```
>
```

The cursor-up key allows the user to recall previous commands.

Except for a few standard arithmetic operations, R accomplishes things by executing various functions. For example, to exit R one executes the `quit` function:

```
> q()
```

When you quit, R will inquire if you want to "Save workspace image?" If you answer `yes` (`y`), then all of the objects in your current workspace, e.g., any data sets and functions that you created, will be saved and restored the next time that you start R.

R.2.1 Vectors

R can store and manipulate a variety of data objects, the most basic of which is a vector. In R a vector is an ordered list of numbers, i.e., a list of numbers with a designated first element, second element, etc. Vectors can be created in various ways. In each of the following examples, the created vector is assigned the name x.

Note that R has a large number of built-in functions. Assigning their names to user-created objects will mask the built-in functions. For this reason certain simple names, e.g., c and t, should be avoided.

Example R.1 To enter a list of numbers from the keyboard, use the concatenate function:

```
> x <- c(20,5,15,18,5,13,1)
```

Notice that this can be done recursively, e.g.,

```
> x <- c(20,5,15)
> x <- c(x,18,5)
> x <- c(x,13,1)
```

To display the vector, type its name:

```
> x
```

Just typing

```
> c(20,5,15,18,5,13,1)
```

causes R to display the vector without saving it for future use.

Example R.2 To read a list of numbers from an ascii text file, say data.txt, use the scan function. In most situations, you will need to specify the complete path of data.txt. How one does this depends on which operating system your computer uses.

For example, suppose that you are using the Windows version of R and data.txt resides in the directory c:\Courses\Math351. Then the following command will read the contents of data.txt into the vector x:

```
> x <- scan("c:\\Courses\\Math351\\data.txt")
```

Notice that the single slashes in the path name must be entered as double slashes in R.

Example R.3 Several functions are useful for creating sequences of numbers, e.g.,

```
> x <- seq(from=1,to=15,by=2)
> x <- rep(1,times=10)
```

Consecutive integers are especially easy, e.g.,

```
x <- 11:20
```

Example R.4 `R` has a variety of functions for generating pseudorandom samples.[1]

To draw 10 numbers from a uniform distribution on $(0, \pi)$:

```
> x <- runif(10,min=0,max=pi)
```

To draw 20 numbers from a normal distribution with mean 5 and standard deviation 1.5:

```
> x <- rnorm(20,mean=5,sd=1.5)
```

To simulate rolling a fair die 30 times:

```
> die <- 1:6
> x <- sample(x=die,size=30,replace=T)
```

A subset of a vector can be identified by a vector of index values. For example, to extract the 2nd, 3rd, and 5th elements of the vector x, one might type:

```
> k <- c(2,3,5)
> x[k]
```

To extract the other elements, just type:

```
> x[-k]
```

One may wish to rearrange the elements, e.g.,

```
> y <- sort(x)
```

The preceding command is equivalent to

```
> y <- x[order(x)]
```

[1] The precise meanings of the phrases that follow are explained in Chapters 3–5.

R.2.2 R is a Calculator!

R provides a variety of arithmetical operations and mathematical functions. These operations/functions have been vectorized, i.e., they work on entire vectors, not just individual numbers. Several examples follow.

First, let's create two vectors:

```
> x <- 10:20
> y <- seq(from=1.8,to=2.2,length=length(x))
```

Now, each of the following is a valid R command:

```
> x+100
> x-20
> x*10
> x/10
> x^2
> sqrt(x)
> exp(x)
> log(x)
> x+y
> x-y
> x*y
> x/y
> x^y
```

R.2.3 Some Statistics Functions

R provides hundreds of functions that perform or facilitate a variety of statistical analyses. Most R functions are not used in this book. (You may enjoy discovering and using some of them on your own initiative.) Tables R.1 and R.2 list some of the R functions that are used.

R.2.4 Creating New Functions

The full power of R emerges when one writes one's own functions. To illustrate, I've written a short function named Edist that computes the Euclidean distance between two vectors. When I type Edist, R displays the function:

```
> Edist
```

Function	Distribution	Section
`pgeom`	Geometric	4.2
`phyper`	Hypergeometric	4.2
`pbinom`	Binomial	4.4
`punif`	Uniform	5.3
`pnorm`	Normal	5.4
`pchisq`	Chi-Squared	5.5
`pt`	Student's t	5.5
`pf`	Fisher's F	5.5

Table R.1: Some `R` functions that evaluate the cumulative distribution function (cdf) for various families of probability distributions. The prefix `p` designates a cdf function; the remainder of the function name specifies the distribution. For the analogous quantile functions, use the prefix `q`, e.g., `qnorm`. To evaluate the analogous probability mass function (pmf) or probability density function (pdf), use the prefix `d`, e.g., `dnorm`. To generate a pseudorandom sample, use the prefix `r`, e.g., `rnorm`.

```
function(u,v){
return(sqrt(sum((u-v)^2)))
}
>
```

`Edist` has two arguments, `u` and `v`, which it interprets as vectors of equal length. `Edist` computes the vector of differences, squares each difference, sums the squares, then takes the square root of the sum to obtain the distance. Finally, it returns the computed distance. I could have written `Edist` as a sequence of intermediate steps, but there's no need to do so.

I might have created `Edist` in any of the following ways:

Example R.5

```
> Edist <- function(u,v){ return(sqrt(sum((u-v)^2))) }
>
```

Example R.6

```
> Edist <- function(u,v){
```

Function	Used to Compute/Display
`sum`	sample sum
`mean`	sample mean
`median`	sample median
`var`	sample variance
`quantile`	sample quantile(s)
`summary`	several useful quantities
`plot.ecdf`	empirical cdf
`boxplot`	box plot(s)
`qqnorm`	normal probability plot
`plot`, `density`	kernel estimate of pdf

Table R.2: Some `R` functions that compute or display useful information about one or more univariate samples. See Chapter 7.

```
+ return(sqrt(sum((u-v)^2)))
+ }
>
```

Notice that `R` recognizes that the command creating `Edist` is not complete and provides continuation prompts (`+`) until it is.

Examples R.5 and R.6 are useful for very short functions, but not for anything complicated. Be warned: ff you mistype and `R` cannot interpret what you did type, then `R` ignores the command and you have to retype it. Using the cursor-up key to recall what you typed may help, but for anything complicated it is best to create a permanent file that you can edit. This can be done within `R` or outside of `R`.

Example R.7 To create moderately complicated functions in `R`, use the `edit` function. For example, I might start by typing

```
> Edist <- function(u,v){u-v}
```

This creates an `R` object called `Edist`, but not the `Edist` that we want—this `Edist` returns the vector of differences.[2] So, I use `edit` to modify `Edist`.[3] This process is initiated with the command

```
> Edist <- edit(Edist)
```

After making and saving the desired changes to `Edist`, I close the editor, thereby returning control to `R`. `R` checks the edited version of `Edist`: if `R` can interpret the edited version, then `R` replaces the previous version with the edited version; if `R` cannot interpret the edited version, e.g., because of typographical errors, then `R` issues an error message and retains the previous version. Fortunately, `R` also retains a temporary version of whatever modifications I attempted to make, so I have another chance at getting it right. To access the temporary version, I type

```
> Edist <- edit()
```

Note that I should *not* retype

```
> Edist <- edit(Edist)
```

as this command returns to the original unedited version and discards whatever changes I attempted to make.

Example R.8 Objects created in `R` can be lost, e.g., if one forgets to save one's workspace image when one quits `R`. For this reason, I prefer to create my `R` functions outside of `R`. To accomplish this, I first use a text editor to create an ascii text file that contains whatever `R` commands I want to execute, e.g., the command that creates `Edist`. For example, I might use the Windows notepad editor to create an ascii text file that contains the following:

```
Edist <- function(u,v)
{
return(sqrt(sum((u-v)^2)))
}
```

[2]Using the `return` function is good practice, but often unnecessary. An `R` function will automatically return the last quantity that it computes.

[3]Each installation has a default editor. For the Windows operating system, the default editor is the Windows notepad editor.

Let's suppose that I call this file `myRfcns.txt` and save it in the directory `c:\Courses\Math351`. Then, I can start `R` and use the `source` function to execute the commands in `myRfcns.txt`:

```
> source("c:\\Courses\\Math351\\myRfcns.txt")
```

To check that I succeeded in creating `Edist`, I can produce a list of all the objects in my workspace by typing

```
> objects()
```

R.2.5 Exploring Bivariate Normal Data

In Sections 13.2 and 14.1, we explored the structure of bivariate normal data using five `R` functions:

```
binorm.ellipse
binorm.sample
binorm.estimate
binorm.scatter
binorm.regress
```

These functions are not part of your `R` installation—I created them for this book/course. To obtain them, download the ascii text file `binorm.R` from the web page for this book/course, then `source` its contents into your `R` workspace. For example, suppose that you have a Windows operating system and that you save `binorm.R` in the directory `c:\Courses\Math351`. Then the following command instructs `R` to execute the commands in `binorm.R` that create the five `binorm` functions:

```
> source("c:\\Courses\\Math351\\binorm.R")
```

Tables R.3–R.7 reproduce the commands in `binorm.R`. Notice that the `#` symbol is used to insert comments, as `R` ignores lines that begin with `#`.

R.2.6 Simulating Termite Foraging

Sections 1.1.3 and 15.1 describe a study of termite foraging behavior. A test statistic, T, assumes a small value when each subsequently attacked roll is near a previously attacked roll. Thus, small values of T are evidence against a null hypothesis of random foraging, under which each unattacked roll is equally likely to be attacked next. To compute a significance probability

for a particular plot, e.g., Plot 20 depicted in Figure 1.1, we require the probability distribution of T. This discrete distribution cannot be calculated by the methods of Chapter 4; instead, we resort to computer simulation.

Dana Ranschaert (a former student) and I created an `R` function, `forage`, that approximates the pmf of T by simulation. To obtain `forage`, download the ascii text file `termites.R` from the web page for this book/course, then `source` its contents into your `R` workspace. Table R.8 reproduces the commands in `termites.R`.

To use `forage`, you must specify four arguments:

1. `initial`, a vector that contains the numbers of the initially attacked rolls. The 5×5 rolls are numbered as follows:

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

   In Figure 1.1, the vector of initially attacked rolls is `c(3,5)`.

2. `nsubsequent`, the number of subsequently attacked rolls. In Figure 1.1, there are 13 such rolls.

3. `nsim`, the number of simulated foraging histories. In the original study, each plot was simulated 1 million times.

4. `maxT`, the largest value of T to be tabulated.

For example, the command

```
> pmf20 <- forage(c(3,5),13,10000,30)
```

computes a matrix with $30-13+1$ rows and 2 columns. The first column of `pmf20` contains values of T, from 13 to 30. Corresponding to each value of T, the corresponding number in the second column of `pmf20` tabulates how many of the 10000 simulated foraging histories produced that value of T.

```
binorm.ellipse <- function(pop) {
#
#  This function plots the concentration ellipse of a bivariate
#  normal distribution.  The 5 bivariate normal parameters are
#  specified in the vector pop in the following order:
#    mean of X, mean of Y, variance of X, variance of Y,
#    correlation of (X,Y).
#  For example:  pop <- c(0,0,1,4,.5)
#
n <- 628
m <- matrix(pop[1:2],nrow=2)
off <- pop[5] * sqrt(pop[3]*pop[4])
C <- matrix(c(pop[3],off,off,pop[4]),nrow=2)
E <- eigen(C,symmetric=T)
a <- 0:n/100
X <- cbind(cos(a),sin(a))
X <- X %*% diag(sqrt(E$values)) %*% t(E$vectors)
X <- X + matrix(rep(1,n+1),ncol=1) %*% t(m)
xmin <- min(X[,1])
xmax <- max(X[,1])
ymin <- min(X[,2])
ymax <- max(X[,2])
dif <- max(xmax-xmin,ymax-ymin)
xlim <- c(m[1]-dif,m[1]+dif)
ylim <- c(m[2]-dif,m[2]+dif)
par(pty="s")
plot(X,type="l",xlab="x",ylab="y",xlim=xlim,ylim=ylim)
title("Concentration Ellipse")
}
```

Table R.3: The command that creates the `R` function `binorm.ellipse`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.sample <- function(pop,n) {
#
#  This function returns a sample of n observations drawn from a
#  bivariate normal distribution.  The 5 bivariate normal
#  parameters are specified in the vector pop in the following
#  order: mean of X, mean of Y, variance of X, variance of Y,
#  correlation of (X,Y).  For example:  pop <- c(0,0,1,4,.5)
#  The sample is returned in the form of an n-by-2 data matrix,
#  each row of which is an observed value of (X,Y).
#
m <- matrix(pop[1:2],nrow=2)
off <- pop[5] * sqrt(pop[3]*pop[4])
C <- matrix(c(pop[3],off,off,pop[4]),nrow=2)
E <- eigen(C,symmetric=T)
Data <- matrix(rnorm(2*n),nrow=n)
Data <- Data %*% diag(sqrt(E$values)) %*% t(E$vectors)
Data + matrix(rep(1,n),nrow=n) %*% t(m)
}
```

Table R.4: The command that creates the `R` function `binorm.sample`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.estimate <- function(Data) {
#
#  This function estimates bivariate normal parameters from a
#  bivariate data matrix.  Each row of the n-by-2 matrix Data
#  contains a single observation of (X,Y).  The function returns
#  a vector of 5 estimated parameters: mean of X, mean of Y,
#  variance of X, variance of Y, correlation of (X,Y).
#
n <- nrow(Data)
m <- c(sum(Data[,1]),sum(Data[,2]))/n
v <- c(var(Data[,1]),var(Data[,2]))
z1 <- (Data[,1]-m[1])/sqrt(v[1])
z2 <- (Data[,2]-m[2])/sqrt(v[2])
r <- sum(z1*z2)/(n-1)
c(m,v,r)
}
```

Table R.5: The command that creates the `R` function `binorm.estimate`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.scatter <- function(Data) {
#
#  This function produces a scatter diagram of the bivariate data
#  contained in the n-by-2 data matrix Data.  It also superimposes
#  the sample concentration ellipse.
#
n <- 628
xmin <- min(Data[,1])
xmax <- max(Data[,1])
xmid <- (xmin+xmax)/2
ymin <- min(Data[,2])
ymax <- max(Data[,2])
ymid <- (ymin+ymax)/2
dif <- max(xmax-xmin,ymax-ymin)/2
xlim <- c(xmid-dif,xmid+dif)
ylim <- c(ymid-dif,ymid+dif)
par(pty="s")
plot(Data,xlab="x",ylab="y",xlim=xlim,ylim=ylim)
title("Scatter Diagram")
v <- binorm.estimate(Data)
m <- matrix(v[1:2],nrow=2)
off <- v[5] * sqrt(v[3]*v[4])
C <- matrix(c(v[3],off,off,v[4]),nrow=2)
E <- eigen(C,symmetric=T)
a <- 1:n/100
Y <- cbind(cos(a),sin(a))
Y <- Y %*% diag(sqrt(E$values)) %*% t(E$vectors)
Y <- Y + matrix(rep(1,n),nrow=n) %*% t(m)
lines(Y)
}
```

Table R.6: The command that creates the `R` function `binorm.scatter`, described in Section 13.2. This command is included in the file `binorm.R`.

```
binorm.regress <- function(Data) {
#
#  This function produces a scatter diagram of the bivariate data
#  contained in the n-by-2 data matrix Data.  It also superimposes
#  the sample concentration ellipse and the regression line.
#
n <- 628
xmin <- min(Data[,1])
xmax <- max(Data[,1])
xmid <- (xmin+xmax)/2
ymin <- min(Data[,2])
ymax <- max(Data[,2])
ymid <- (ymin+ymax)/2
dif <- max(xmax-xmin,ymax-ymin)/2
xlim <- c(xmid-dif,xmid+dif)
ylim <- c(ymid-dif,ymid+dif)
par(pty="s")
plot(Data,xlab="x",ylab="y",xlim=xlim,ylim=ylim)
title("Regression Line")
v <- binorm.estimate(Data)
m <- matrix(v[1:2],nrow=2)
off <- v[5] * sqrt(v[3]*v[4])
C <- matrix(c(v[3],off,off,v[4]),nrow=2)
E <- eigen(C,symmetric=T)
a <- 0:n/100
Y <- cbind(cos(a),sin(a))
Y <- Y %*% diag(sqrt(E$values)) %*% t(E$vectors)
Y <- Y + matrix(rep(1,n+1),ncol=1) %*% t(m)
lines(Y)
x <- xlim[1] + (2*dif*(0:n))/n
slope <- v[5] * sqrt(v[4]/v[3])
y <- v[2] + slope*(x-v[1])
Y <- cbind(x,y)
Y <- Y[Y[,2] < ymax,]
Y <- Y[Y[,2] > ymin,]
lines(Y)
}
```

Table R.7: The command that creates the `R` function `binorm.regress`, described in Section 14.1. This command is included in the file `binorm.R`.

```
forage <- function(initial,nsubsequent,nsim,maxT) {
#
# This function simulates nsim termite foraging histories.
# initial is the vector of initially attacked rolls;
# nsim is the number of subsequently attacked rolls.
# The function returns a matrix in which the first column
# contains values of the test statistic T (from nsubsequent
# to maxT) and the second column contains the corresponding
# number of histories that produced that value of T.
#
v <- rep(1:5,5)
w <- rep(1:5,rep(5,5))
D <- cbind(v,w)
D <- (diag(25) - matrix(1/25,25,25)) %*% D
D <- D %*% t(D)
v <- diag(D)
H <- diag(v) %*% matrix(1,25,25)
D <- H+t(H)-2*D
D[D<0] <- 0
H <- matrix(100,25,25)
for (rowi in 2:25)
  for (colj in 1:(rowi-1)) {
    H[rowi,colj] <- 0
    }
v <- 1:length(initial)
w <- 1:(length(initial)+nsubsequent)
pmf <- rep(0,maxT)
for (isim in 1:nsim){
  rolls <- c(initial, sample(x=(1:25)[-initial],
           size=nsubsequent, replace=F))
  D0 <- D[rolls,rolls] + H[w,w]
  distance <- apply(D0,1,min)
  total <- round(sum(distance[-v]))
  if (total < maxT+0.5) {
    pmf[total] <- pmf[total]+1
    }
return(cbind(nsubsequent:maxT,pmf[-(1:(nsubsequent-1))]))
}
```

Table R.8: The command that creates the `R` function `forage`, described in Section 15.1. This command is included in the file `termites.R`.