Chapter 9

Inference

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 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 (Robynne) agree to resolve their disputed ownership of an Erté painting by tossing a penny. Arlen produces a penny and, just as Robynne is about to toss it in the air, Arlen smoothly suggests that spinning the penny on a table might ensure better randomization. Robynne 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. Robynne is left with the penny.

That evening, Robynne 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 Robynne 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 Robynne'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 Robynne!) 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, Robynne 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. Robynne's goal is to draw inferences about this fixed but unknown quantity. We consider three sets of 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 Robynne can comfortably accuse Arlen of impropriety?

3. What are plausible values of p? In particular, is there a subset of $[0, 1]$ that Robynne 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 $\hat{p} = 0.32 \neq 0.5$, Robynne 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 Robynne 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, Robynne 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*. An estimate is an observed value of an estimator.

The quality of an individual estimate depends on the individual sample from which it was computed and therefore is 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 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 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 premeditated murder and that the jury has been instructed that it should 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. 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.

[1] Quoted by F. N. David (1962). *Games, Gods and Gambling: A History of Probability and Statistical Ideas*. Dover Publications, New York, p. 24. 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.

2. Assessing data from a scientific experiment.

   A study of termite foraging behavior reached the controversial conclusion that two species of termites compete for scarce food resources.[2] 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 Robynne's penny-spinning experiment.

   In Section 9.1 we noted that Robynne 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 the 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 Robynne 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?

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.

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

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, Robynne wants to determine whether or not spinning a penny 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 or factually guilty. In the words of the United States Supreme Court:

> Underlying the question of guilt or innocence is an objective truth: the defendant, in fact, did or did not commit the acts constituting the crime charged. From the time an accused is first suspected to the time the decision on guilt or innocence is made, our criminal justice system is designed to enable the trier of fact to discover that truth according to law.[3]

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 Robynne wants to determine whether or not penny-spinning is fair. However, one can easily imagine that Robynne 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).

[3] *Bullington v. Missouri*, 451 U. S. 430 (1981).

The Actor

The states of nature having been partitioned into two hypotheses, it is necessary for a decision maker (the actor) to choose between them. In the penny-spinning example, the actor is Robynne; 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, Robynne 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 Supreme Court Justice John Harlan:

> The standard of proof influences the relative frequency of these two types of erroneous outcomes. 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. Because the standard of proof affects the comparative frequency of these two types of erroneous outcomes, the choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect an assessment of the comparative social disutility of each.[4]

[4] *In re Winship*, 397 U. S. 358 (1970).

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*, 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."[5]

Beyond a Reasonable Doubt

To operationalize an aversion 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

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

[5] A. M. Dershowitz (1996). *Reasonable Doubts: The O. J. Simpson Case and the Criminal Justice System.* Simon & Schuster, New York, pp. 38, 212, 85.

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.[6]

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

[6]T. Starkie (1824). *Practical Treatise on the Law of Evidence*, 2 volumes. London, p. 478 of the 1833 edition. Quoted by B. J. Shapiro (1991). *"Beyond Reasonable Doubt" and "Probable Cause": Historical Perspectives on the Anglo-American Law of Evidence*. University of California Press, Berkeley, p. 35.

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.[7]

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, Robynne 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[8] suggested that $\alpha = 0.10$, $\alpha = 0.05$, and $\alpha = 0.02$ 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.[9] While some degree of conformity is desirable (it inhibits a researcher from choosing—after the fact—a significance level that will permit rejecting

[7] 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.

[8] R. A. Fisher (1925). *Statistical Methods for Research Workers.* Re-issued in *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford University Press, Oxford, 1995. 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, as we are concerned with situations in which both hypotheses are fully specified. Fisher, Neyman, and Pearson all accepted the fundamental principle that the null hypothesis should be accorded a privileged status and maintained unless there is compelling evidence against it.

[9] Perhaps this development was inevitable. For decades, one could perform a test only by comparing the value of one's test statistic to a table of critical values. Fisher (1925) was obliged to choose a small number of significance levels in constructing his tables of critical values; researchers who used those tables were obliged to adopt one of Fisher's

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.[10]

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 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.[11]

Although it is rarely (if ever) possible to quantify a juror's level of assent, those comfortable with statistical hypothesis testing may be inclined

signficance levels. Our use of R instead of tables affords us considerably greater freedom.

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

[11] B. 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. I am greatly indebted to Shapiro's fascinating study.

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 premeditated 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 percent probability of wrongly convicting a factually innocent defendant is abhorrent.[12]

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.[13] 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.[14]

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 **p**. The significance level is a real number that is fixed by and known to

[12]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.

[13]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."

[14]A. M. Dershowitz (1996). *Reasonable Doubts: The O. J. Simpson Case and the Criminal Justice System.* Simon & Schuster, New York, p. 40.

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

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

which can be computed by the following `R` command:

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

An illustration of the normal probability of interest is sketched in Figure 9.1.

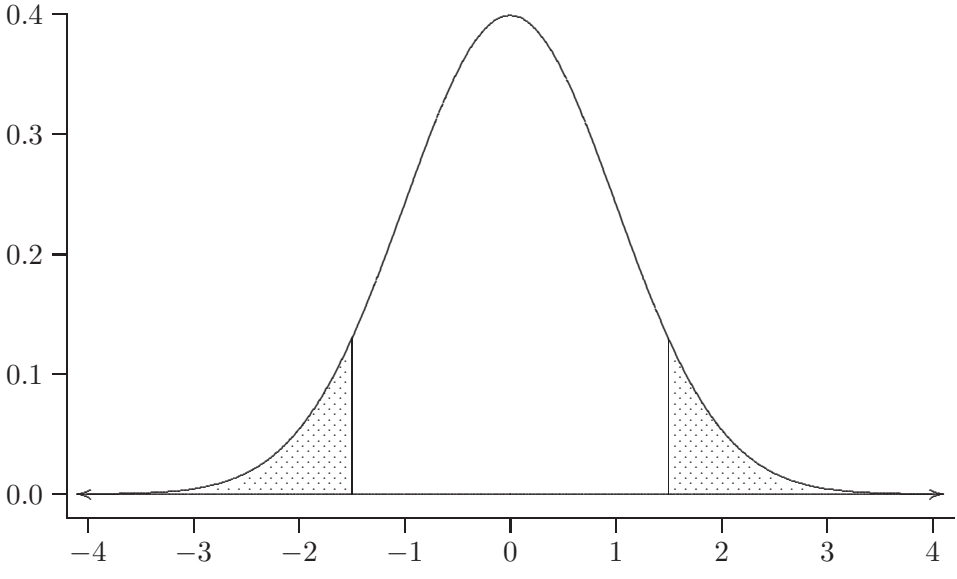

Figure 9.1: $P(|Z| \geq |z| = 1.5)$

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

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

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

$$\mathbf{p} \approx 2\Phi(-1.481234) \doteq 0.1385 > 0.05 = \alpha.$$

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

9.4.1 One-Sided Hypotheses

In Section 9.3 we suggested that, if Robynne 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 and only 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 confirms only 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

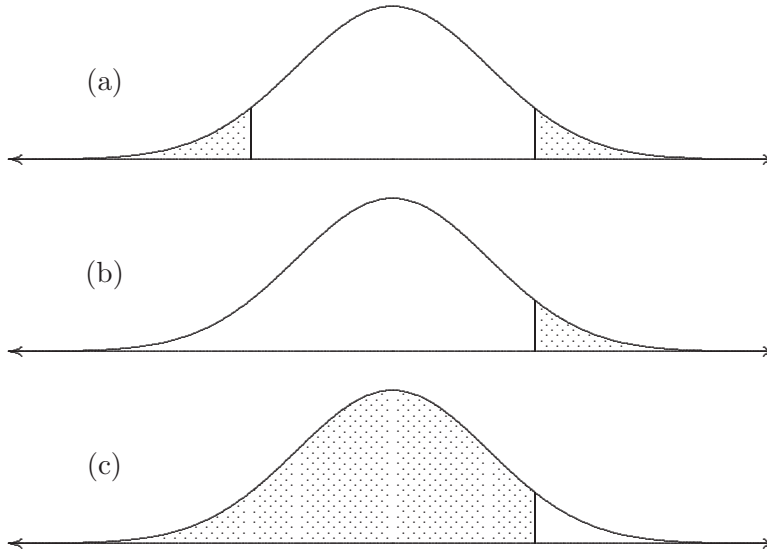

Figure 9.2: Significance probabilities for Example 9.2. Each significance probability is the area of the corresponding shaded region.

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?*

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 forgo necessary speed humps. Hence, they would like to see the hypotheses formulated so that forgoing necessary speed humps is a Type I error. Because 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 forgo 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. Because 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 $\mu_0 = 15$. Because $\bar{x} = 15.3 > 15$, 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

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

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

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

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

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

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

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

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 not reject $H_0 : \mu = 20$ 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

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

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

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

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 $1-\alpha$. 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 retain 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 retained.

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 retained 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 retained 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 retained 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 3.7.6 is the most appropriate choice of a significance probability for this investigation? Why?

4. The U.S. Food and Drug Administration requires evaporated milk to contain "not less than 23 percent by weight of total milk solids." A company that sells evaporated milk is sued by a group of consumers who are concerned that the company's product does not meet FDA standards. The two parties agree to binding arbitration. If the consumers win, the company will pay damages and enhance its product; if the company wins, then the consumers will issue a public apology.

   To resolve the dispute, the arbiter commissions a neutral study in which the percent by weight of total milk solids will be measured in a random sample of $n = 225$ packages produced by the company. Both parties agree to a standard of proof ($\alpha = 0.05$), but they disagree on which party should bear the burden of proof.

   (a) State appropriate null and alternative hypotheses from the perspective of the consumers.

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

   (c) Suppose that the random sample reveals a sample mean of $\bar{x} = 22.8$ percent with a sample standard deviation of $s = 3$ percent. Compute t, the value of the test statistic.

   (d) From the consumers' perspective, what action should be taken? Why?

   (e) From the company's perspective, what action should be taken? Why?

5. 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 the population is 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 $\sigma^2 = \text{Var}\, X_i$.

(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 z, the value of the test statistic.

(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?

6. 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 $\alpha = 0.05$?

7. 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 $\alpha = 0.05$?

8. 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. Explain.

9. Mt. Wrightson, the fifth highest summit in Arizona and the highest in Pima County, has a reputed elevation of 9453 feet. To amuse its members, the Southern Arizona Hiking Club (SAHC) decides to construct its own confidence interval for μ, the true elevation of Mt. Wrightson's summit. SAHC acquires an altimeter whose measurements will have an expected value of μ with a standard deviation of 6 feet. How many measurements should SAHC plan to take if it wants to construct a 0.99-level confidence interval for μ that has a length of 2 feet?

10. 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?

11. Refer to Exercise 8.4.5. A third Math 351 student wants to use the function `urn.model` to construct a 0.95-level confidence interval for $p = P(170.5 < Y < 199.5)$. If he desires an interval of length L, then how many times should he plan to evaluate `urn.model`?

    Hint: How else might the student estimate p?

12. 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 $\mathbf{p}$ as in Section 9.1, using the binomial distribution and `pbinom`.

        ii. Approximate $\mathbf{p}$ 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 retained at a significance level of (approximately) $\alpha = 0.05$.