2

Theory of Tests, *p*-Values, and Confidence Intervals

2.1 Overview

In this chapter we outline the theory of statistical hypothesis tests. This is important for defining terminology. As we introduce terms, we first show in Section 2.2.1 how they apply to a simple single sample binomial example. In Section 2.2.2 we explore some of the generality of the theory by outlining three different ways of formulating the null and alternative hypotheses for a second data example: a two-sample study with numeric responses. This shows that there is not one "correct" way of formulating a problem. These three formulations can get abstract, for example, the hypotheses can be represented by sets of infinite dimensional parameters. Yet even in this representation, the statistical hypothesis problem can still be solved. We present this level of abstraction this early in the book to show the breadth of application of the statistical hypothesis test theory.

In Section 2.3 we show how hypothesis tests are intimately related to confidence intervals. This relationship is important, because the scientific information in finding a significant effect (i.e., rejecting the null hypothesis) is enhanced when supplemented with important information such as an effect estimate and its confidence interval. In Section 2.4 we define some important properties of hypothesis tests.

Readers who are less theoretically inclined can skip to the end-of-chapter summary (Section 2.5). This summary will provide the essential terminology and information needed for the rest of the book.

2.2 Components of a Hypothesis Test

2.2.1 Main Definitions with a Simple Example

A hypothesis test has three basic parts: (1) the set of possible values of data if the study is repeated, (2) a set of probability models that is partitioned into two sets of models, called the *null hypothesis set* and *alternative hypotheses set*, and (3) the decision rule which uses the observed data to decide whether or not you reject the null hypothesis (i.e., whether or not you reject the statement that the true data generating probability model is in the null hypothesis set). The decision rule is a function of the data and the α-level. We describe each part separately.

Suppose we observe a vector of data, say $\mathbf{x}$. For a statistical hypothesis test, we envision that the data comes from a data generating process, or probability model. If we reran the data generating process, then we would get a different vector of data. Let $\mathcal{X}$ be the set of all

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

possible data vectors associated with this process. The set $\mathcal{X}$ is called the *sample space* or *support*.

Example 2.1 *Suppose we observe $x = 6$ successes out of $n = 10$ tries. Suppose the data generating process is binomial with $n = 10$ and parameter θ with $0 < \theta < 1$. Then here $\mathbf{x} = 6$ and $\mathcal{X} = \{0, 1, \ldots, 10\}$.*

Let the probability model associated with the data generating process be denoted P_θ, where θ represents the parameter vector, which may be infinite dimensional (see Example 2.4). We do not know the value of θ but assume that it is one member in a set of possible values, Θ, and therefore, P_θ is one member of a set of probability models, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ (see Note N1 for set notation explanation). We partition the probability models into two disjoint sets, the null hypothesis models, $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\}$ and the alternative hypothesis models, $\mathcal{P}_1 = \{P_\theta : \theta \in \Theta_1\}$. The null ($H_0$) and alternative ($H_1$) hypothesis statements are

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1,$$

where θ is the true parameter.

Example 2.1 *(continued) For the binomial example, P_θ is a binomial distribution with parameters n and θ. Let $\Theta = \{\theta : \theta \in [0, 1]\}$. We might want to know if θ is bigger than 0.5. In this case we partition Θ as $\Theta_0 = \{\theta : \theta \in [0, 0.5]\}$ and $\Theta_1 = \{\theta : \theta \in (0.5, 1]\}$. We can write the hypotheses in this case as $H_0 : \theta \leq 0.5$ and $H_1 : \theta > 0.5$.*

The final piece of the hypothesis test is the decision rule, which we write as δ. The *decision rule* is a function of $\mathbf{X}$ (a possible element of $\mathcal{X}$) and the *significance level*. The significance level is set by the researcher, and it is usually denoted α, and hence is sometimes called the α-level.[1] The result of the decision rule is the probability of rejecting the null hypothesis. For example, if we observe $\mathbf{x}$, the probability of rejecting the null hypothesis using a significance level of α is $\delta(\mathbf{x}, \alpha)$. Typically, we only use decision rules that only give two values, 0 (do not reject the null) or 1 (reject the null). Other decision rules are called *randomized decision rules* and are not discussed in this chapter and are rarely used in practice (see Note N2). A *valid decision rule* is defined such that, the probability of rejecting the null hypothesis for any one of the null hypothesis models is less than or equal to α. Mathematically, a valid (nonrandomized) decision rule has (for any $\alpha \in (0, 1)$)

$$\sup_{\theta \in \Theta_0} \Pr[\delta(\mathbf{X}, \alpha) = 1 | \theta] \leq \alpha. \tag{2.1}$$

The left-hand side of the above equation is called the *size* of the hypothesis test.

When the alternative represents parameters only in one direction, we call this a one-sided hypothesis test. Example 2.1 is a one-sided hypothesis test, because the alternatives are all $\theta > 0.5$. We can formulate one-sided hypotheses in the other direction: $H_0 : \theta \geq 0.5$ and

[1] Lehmann and Romano (2005) use the term significance level, but Greenland (2019) prefers "α-level" because Fisher and other British authors sometimes use "significance level" to mean p-value. We never use "significance level" to mean p-value, and we use "significance level" and "α-level" interchangeably. We do not always use α-level because it can sometimes lead to awkward phrases like "for the one-sided test we use $\alpha/2$ for the α-level."

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

$H_1 : \theta < 0.5$. For the binomial problem, a two-sided formulation is $H_0 : \theta = 0.5$ and $H_1 : \theta \neq 0.5$. For setting significance levels, in medical publications the tradition has been to use $\alpha = 0.05$ for two-sided tests. Often the two-sided test rejects at the $\alpha = 0.05$ level if either of the one-sided tests rejects at the $\alpha = 0.025$ level. Thus, we would argue that a significance level of 0.025 for a one-sided test is appropriate when an $\alpha = 0.05$ for two-sided tests is appropriate. Of course, not all situations demand using two-sided $\alpha = 0.05$ (see Note N3).

Aside from tradition, the α-level is somewhat arbitrary. Because of this, a useful approach is to give the smallest α-level for which we would reject the null hypothesis for the observed data at that level and all larger levels. This is known as the *p-value*. In mathematical notation, for nonrandomized decision rules the p-value is

$$p(\mathbf{x}) = \inf\{\alpha : \delta(\mathbf{x}, \alpha^*) = 1 \text{ for all } \alpha^* \geq \alpha\}. \tag{2.2}$$

So that for a valid hypothesis test, we have

$$\sup_{\theta \in \Theta_0} \Pr[p(\mathbf{X}) \leq \alpha | \theta] \leq \alpha. \tag{2.3}$$

We call a p-value function, $p(\cdot)$, that satisfies Equation 2.3 a valid p-value.[2]

Except with rare exceptions, applied decision rules are monotonic in α, rejecting more often as α increases (see Note N4). For monotonic decision rules, we can think of the p-value function as providing an ordering of the possible data from the most extreme (least likely under the null) to the least extreme (most likely under the null), where least and most likely are defined using a $\theta \in \Theta_0$ that maximizes the left-hand side of expression 2.3. Then the p-value can be interpreted as the probability of observing equal or more extreme data under the null hypothesis (see Note N5).

We always want valid decision rules but sometimes we settle for approximately valid decision rules. For example, we might base a decision rule on asymptotics, so that as the sample size gets larger and larger the decision rule gets closer and closer to valid. These are asymptotically valid decision rules. These are discussed in Section 2.4.

Sometimes when using nonasymptotic decision rules or tests we use the term *exact* to mean *valid*. In some statistical literature, *exact* means that the inequality in the definition of validity (expression 2.1) is an equality; however, in this book, we use the term *exact* as synonymous with *valid*. This latter usage is more common in the biostatistics literature (for example, Fisher's exact test is valid, but the size is generally not equal to the α-level).

Now consider decision rules related to our example.

Example 2.1 *(continued) For the binomial testing* $H_0 : \theta \leq 0.5$ *versus* $H_1 : \theta > 0.5$*, we use the binomial distribution to define* δ*. First, note that larger values of* x *suggest larger values of* θ*, so we want to reject when* $x \geq c$*, where* c *is a constant that depends on* n *and* α*. We want to choose* c *to give a valid test and have as much power as possible, so we want the smallest* c *that gives a size less than or equal to* α*. To calculate the size, in this case we only need to calculate the probabilities under the model on the boundary between the null and alternative hypotheses, the binomial with* $\theta = 0.5$*. When* $n = 10$ *and* $c = 8$ *or* $c = 9$ *we get sizes (i.e., probabilities of rejection at the boundary) of* $\Pr[X \geq 8 | \theta = 0.5] = 0.0547$

[2] The term *p-value* can refer to either $p(\mathbf{x})$ (the value of the p-value function evaluated at the data $\mathbf{x}$), or $p(\cdot)$ (the p-value function itself); the meaning is inferred from the context.

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

and $\Pr[X \geq 9|\theta = 0.5] = 0.0107$. *So when* $\alpha = 0.05$, *a valid test uses the decision rule,* $\delta(X, 0.05) = 1$ *if* $X \geq 9$ *and* 0 *otherwise.*

2.2.2 Hypothesis Test Components for a Two-Sample Example

Suppose we have two samples with five independent responses in each sample. Let the data be $\mathbf{x} = [\mathbf{y}, \mathbf{z}]$, where $\mathbf{y}$ is a vector of responses, and $\mathbf{z}$ is the vector of associated group indicators. For example,

$$\begin{aligned} \mathbf{y} &= \{3.1, 5.2, 6.6, 5.2, 1.3, 6.8, 9.2, 11.1, 9.5, 2.2\}, \\ \mathbf{z} &= \{2, 1, 2, 1, 1, 1, 2, 2, 2, 1\}. \end{aligned} \tag{2.4}$$

Suppose we want to know if the responses from Group 1 (i.e., those with $z_i = 1$) are in general larger or smaller than the responses from Group 2. There are many ways to formulate the null and alternative hypotheses; we present three formulations as Examples 2.2–2.4. Here we do not examine all the features that determine the process that created the data (for that see Chapter 3). The data generating and analysis context may favor one formulation over another, (for that see Chapter 9), although there are situations where many different formulations would be acceptable. The purpose of presenting three hypothesis test formulations here is to get a flavor for the different ways the components of the hypothesis test may be formulated, and to emphasize that there is a choice for each data set.

The first example formulation is the t-test. We give a standard set of assumptions associated with this test, although other formulations of the t-test are possible (see Chapter 9).

Example 2.2 *In this t-test formulation, we assume that the group indicators are fixed, and the responses are samples from a normal distribution. Each normal response can be any real number, so we write* $Y_i \in \Re$ *for the* i*th response, or* $\mathbf{Y} \in \Re^{10}$ *for all 10 responses. The sample space is* $\mathcal{X} = \{\mathbf{Y}, \mathbf{z} : \mathbf{Y} \in \Re^{10}\}$.

If $z_i = 1$, *then we assume* Y_i *is normal with mean* μ_1 *and variance* σ_1^2, *or* $Y_i \sim N(\mu_1, \sigma_1^2)$, *while if* $z_i = 2$, *then* $Y_i \sim N(\mu_2, \sigma_2^2)$. *Here* P_θ *represents the probability model for the* 10 *responses of the* $\mathbf{Y}$ *vector. The parameter* θ *is four-dimensional,* $\theta = [\mu_1, \sigma_1, \mu_2, \sigma_2]$. *The possible values of* θ *are*

$$\Theta = \{\mu_i, \sigma_i : \mu_i \in (-\infty, \infty), \sigma_i \in (0, \infty), i = 1, 2\}.$$

One formulation of the null and alternative hypotheses is $H_0 : \mu_1 = \mu_2$ *and* $H_1 : \mu_1 \neq \mu_2$. *This is shorthand for expressing the two sets of possible parameters,*

$$\begin{aligned} \Theta_0 &= \{\mu_i, \sigma_i : \mu_1 = \mu_2, \mu_i \in (-\infty, \infty), \sigma_i \in (0, \infty), i = 1, 2\} \\ \Theta_1 &= \{\mu_i, \sigma_i : \mu_1 \neq \mu_2, \mu_i \in (-\infty, \infty), \sigma_i \in (0, \infty), i = 1, 2\}. \end{aligned}$$

We can calculate p-values by using Welch's t-test, which takes the difference in means of the groups, divides that difference by an estimate of the standard error of that difference, and compares the resulting ratio to a t-distribution. We show in Chapter 9 that Welch's t-test is an approximately valid way to calculate p-values and confidence intervals when we allow $\sigma_1 \neq \sigma_2$ *as in this formulation.*

A second way to formulate statistical hypotheses for the data of expression 2.4 is as a randomization test. This switches the randomness from the responses to the group indicators.

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

Example 2.3 *Suppose the data generating process keeps the responses the same every time new data is generated, but the group indicator vector is randomly shuffled. For example, suppose we do an experiment where we randomize 10 individuals to two interventions ($z = 1$ or $z = 2$). We imagine redoing the experiment such that the group indicators were different each time, but the interventions had the same effects, so that the ith response would be the same no matter which of the two interventions was applied. Then the sample space is $\mathcal{X} = \{\mathbf{y}, \mathbf{Z} : \mathbf{Z}$ is a vector with five ones and five twos$\}$. There are 10 choose 5, or 252 elements in $\mathcal{X}$.*

For this example, P_θ represents the probability model for the $\mathbf{Z}$ vectors, vectors of length 10 with 5 zeros and 5 ones. So θ could be a vector of length 252 giving the probability for each of the 252 elements in $\mathcal{X}$. Typically, Θ_0 is a set with one value, a vector of length 252 with each element of the vector equal to $1/252$. The set Θ_1 would be all other possible probability vectors, in other words,

$$\Theta_1 = \left\{\theta = [\theta_1, \ldots, \theta_{252}] : 0 \leq \theta_i \leq 1 \text{ for all } i, \ \sum \theta_i = 1, \text{ and } \theta \neq \left[\frac{1}{252}, \ldots, \frac{1}{252}\right]\right\}.$$

In this case, typically we do not refer to θ in describing the hypothesis test (since it is not a parameter that is useful for describing the data), but only use it in the probability calculations under the null hypothesis.

In this example, we have that under the null hypothesis each of the 252 ways of picking the vector $\mathbf{Z}$ are equally likely. To choose a decision rule, we first choose a test statistic, T. The test statistic is a function of the possible values of the data, $\mathbf{X}$, that returns a scalar value. So T allows us to order the sample space. Let us assume that larger values of $T(\mathbf{X})$ indicate more extreme values under the null hypothesis. For example, we may use T which is the mean of responses from one group ($z_i = 2$) minus the mean of responses from the other group ($z_i = 1$). Note that because θ is such a large dimension, we cannot look at the null and alternative hypotheses and automatically see how to define more extreme values. There are many choices for defining that extremeness, and choosing T implicitly defines a direction for measuring extremeness under the null hypothesis. Given T we want the smallest value c such that under the null hypothesis

$$\Pr[T(\mathbf{X}) \geq c] \leq \alpha.$$

The calculations that find c depend on $\mathbf{x}$ and can be computationally difficult. Tests using this kind of decision rule are called permutation tests (see Chapter 10 for more details).

Finally, we consider a proportional odds formulation. This allows us to estimate a parameter for an effect (the proportional odds parameter) yet does not require nearly as many assumptions as the t-test (Example 2.2). In fact, the number of parameters is infinite, yet we can still use a permutation test to calculate the p-value.

Example 2.4 *Again, suppose we have observed data vector as listed in expression 2.4. Now, let the ith elements of $\mathbf{y}$ and $\mathbf{z}$ be y_i and z_i. Assume that if $z_i = 1$, then Y_i, the random variable associated with y_i, is distributed with distribution F. We write this as $Y_i \sim F$, so $\Pr[Y_i \leq y] = F(y)$ for any y. Also, if $z_i = 2$, then $Y_i \sim G$. We assume F and G are unspecified distributions with support $[0, \infty)$. In other words, because somehow we know that the response variable in this setting cannot have a negative value, the possible values of*

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

Y_i are in $[0,\infty)$. Thus, $\mathcal{X} = \{\mathbf{Y}, \mathbf{z} : \mathbf{Y} \in \left(\Re^+\right)^{10}\}$. There are an uncountably infinite number of elements in $\mathcal{X}$.

The probability model, P_θ, is represented by the two distributions F and G with support on $[0,\infty)$. Consider a proportional odds model where the odds of one group are proportional to the odds of the other group. Letting F and G be cumulative distribution functions (cdfs), assume that for each $y \in (0,\infty)$,

$$\frac{G(y)}{1-G(y)} = \beta \frac{F(y)}{1-F(y)},$$

where $\beta \in (0,\infty)$. This implies that

$$G(y) = \frac{\beta F(y)}{1-F(y)+\beta F(y)}.$$

So the parameters, θ, can be represented by the scalar β which is in $(0,\infty)$, and the (infinite dimensional) cdf, F. We can write $\Theta = \{\beta, F\}$. A typical partition is $\Theta_0 = \{\beta, F : \beta = 1\}$ and $\Theta_1 = \{\beta, F : \beta \neq 1\}$, with hypotheses usually written as $H_0 : \beta = 1$ and $H_1 : \beta \neq 1$.

Under the null hypothesis that $\beta = 1$, then $F = G$, and we can use a test statistic, T, similarly to what was done in Example 2.3. Under the set of null hypothesis probability models, it is not possible to calculate $\Pr[T(\mathbf{X}) \geq c]$ for a constant c, without making some assumptions about F. But we can calculate the conditional probability given the ordered vector of responses. When we use the conditional probability to calculate the p-value, this is a permutation test (just as was done in Example 2.3). For the proportional odds model, we usually rank all of the 10 responses and use the difference in means of the ranks between the groups as the test statistic, T. This gives us the Wilcoxon-Mann-Whitney test (see Chapter 9).

So, although it is a bit complicated to show, we can calculate p-values using the permutation test methods, and this creates valid p-values for all sample sizes (see Note N6).

So the same permutation method for deriving decision rules can be applied to the assumptions of Example 2.3 as well as Example 2.4.

When we are being formal, we write the hypothesis test as the set $\{\delta, A\}$, where $A = \{\mathcal{X}, \mathcal{P}_0, \mathcal{P}_1\}$ is the set of assumptions defining the sample space, $\mathcal{X}$, the null hypothesis model or models, $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\}$, and the alternative hypothesis models, $\mathcal{P}_1 = \{P_\theta : \theta \in \Theta_1\}$. This formality is often not necessary, and the hypothesis test is sometimes defined by the δ with the assumptions A implied. The formality is useful when talking about how the same decision rule may be used for different sets of assumptions. For example, a permutation test with the same test statistic, $T(\mathbf{x})$, is valid for the assumptions of either Example 2.3 or Example 2.4 (see also Chapter 8).

2.3 Confidence Sets: Inverting a Series of Hypothesis Tests

It is possible to create a confidence set by inverting a series of hypothesis tests. For one-dimensional parameters, in many cases, the set is an interval and is known as a *confidence interval*. We start with confidence sets for the general case, where θ may be a vector of parameters. Then most of the rest of this section focuses on the case when θ is a scalar.

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

2.3.1 Inverting Point Null Hypotheses

First, consider the general case where θ may be a vector. Let

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

be a series of point null hypotheses and their alternatives indexed by θ_0. Suppose you have a series of valid (nonrandomized) decision rules associated with each set of hypotheses, $\delta(X, \alpha; \theta_0)$. Let the p-value associated with $H_0 : \theta = \theta_0$ be $p(X, \theta_0)$ then $\delta(X, \alpha; \theta_0) = I\,(p(X, \theta_0) \leq \alpha)$. For each X let the set of θ_0 values for which we fail to reject the null hypotheses be

$$\begin{aligned} C_S(X, 1-\alpha) &= \{\theta_0 : \delta(X, \alpha; \theta_0) = 0\} \\ &= \{\theta_0 : p(X, \theta_0) > \alpha\}. \end{aligned} \tag{2.5}$$

Then $C_S(X, 1-\alpha)$ is a $100(1-\alpha)\%$ *confidence set* (also known as a confidence region) and

$$\Pr[\theta \in C_S(X, 1-\alpha)|\theta] \geq 1-\alpha \text{ for all } \theta, \tag{2.6}$$

so that $C_S(X, 1-\alpha)$ is valid.

When θ is a scalar, the alternative hypothesis ($H_1 : \theta \neq \theta_0$) associated with the point null hypothesis ($H_0 : \theta = \theta_0$) is called the *two-sided hypothesis*, and the confidence set given by Equation 2.5 will often be an interval and is called the *confidence interval*. But the confidence set is not guaranteed to be an interval for scalar θ, since there may be gaps. In that case, a confidence interval can be defined by filling in the gaps. In other words, the $100(1-\alpha)\%$ confidence interval is

$$C(X, 1-\alpha) = (\theta_L, \theta_U), \tag{2.7}$$

where $\theta_L = \min[C_S(X, 1-\alpha)]$ if it exists and $-\infty$ (or the minimum possible value of θ) otherwise, and similarly $\theta_U = \max[C_S(X, 1-\alpha)]$ if it exists or ∞ (or the maximum possible value of θ) otherwise. Since $C_S(X, 1-\alpha)$ is valid by Equation 2.6, the associated confidence interval, $C(X, 1-\alpha)$, will be valid as well. For a given p-value function for a series of two-sided hypotheses, $p(X, \theta_0)$, define the *matching confidence set* as $C_S(X, 1-\alpha)$ of Equation 2.5, and the *matching confidence interval* as $C(X, 1-\alpha)$ of Equation 2.7. If the matching confidence set is an interval, i.e., $C_S(X, 1-\alpha) = C(X, 1-\alpha)$, then we say that $p(X, \theta_0)$ and $C(X, 1-\alpha)$ are *compatible*, meaning that for any $\mathbf{x}$, the rejection of $H_0 : \theta = \theta_0$ when $p(\mathbf{x}, \theta_0) \leq \alpha$ occurs if and only if $C(\mathbf{x}, 1-\alpha)$ excludes θ_0.

For some p-value functions the matching confidence intervals are not compatible. In other words, sometimes there is a gap in the matching confidence sets. To show that this noncompatibility problem can occur in practice, we consider an example with Fisher's exact test, which can be used to test for differences between two groups when the responses are binary.

Example 2.5 *For the two-sided test there are two versions of Fisher's exact test: the Fisher–Irwin test and the central Fisher's exact test. The noninterval confidence set can occur with the Fisher–Irwin version (the default version in* fisher.test *in R, see Note N7). Suppose we observe proportions of* $7/262$ *from one group and* $30/494$ *from another. We are interested in comparing the population proportions, say* θ_1 *and* θ_2. *Let* $\beta = \frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}$ *represent*

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

an odds ratio. The value $\beta = 1$ *if* $\theta_1 = \theta_2$ *and* $\beta > 1$ *if* $\theta_2 > \theta_1$. *If we test* $H_0 : \beta = 0.99$ *using the Fisher–Irwin test we fail to reject at the 0.05 level (*$p = 0.05005$*), when we test* $H_0 : \beta = 1.00$ *we do reject (*$p = 0.04996$*), but when we test* $H_0 : \beta = 1.01$ *we fail to reject again (*$p = 0.05006$*). So the 95% confidence set for the odds ratio has a hole in it:* $C_S(x, 0.95) = \{\beta : \beta \in (0.177, 0.993)$ *or* $\beta \in (1.006, 1.014)\}$.

The details of the example are not important at this juncture (see Chapter 7). The critical point is that for a commonly used test, inverting a series of two-sided hypothesis test can give a confidence set that is not an interval. However, noninterval confidence sets rarely occur, even when using a procedure (e.g., the Fisher–Irwin test) with this possibility. If the confidence set is not an interval, we create one by filling in the gap. For Example 2.5 this gives a 95% confidence interval of $(0.177, 1.014)$.

2.3.2 Central Intervals: Inverting Two One-Sided Hypotheses

An alternative strategy to create a two-sided confidence interval than inverting a series of point null hypotheses (such as $H_0 : \theta = \theta_0$) is to combine the confidence intervals from two series of one-sided hypothesis tests. Typically, we combine two $100(1 - \alpha/2)\%$ one-sided confidence intervals, to get one $100(1 - \alpha)\%$ two-sided confidence interval, which we call a *central* confidence interval. This strategy has the advantage that we can automatically infer either one-sided hypothesis test by using either the lower or the upper limits of the interval. Start with the alternative-is-greater series of one-sided hypotheses,

$$H_0 : \theta \leq \theta_B$$
$$H_1 : \theta > \theta_B,$$

indexed by θ_B, the parameter value on the boundary between the null and alternative hypotheses. Suppose the associated (nonrandomized) decision rule associated with the valid hypothesis test is $\delta_{AG}(\mathbf{X}, \alpha/2; \theta_B)$, where AG represents the alternative-is-greater hypotheses. Using the validity property of the hypothesis test, we can show that for any $\theta \in \Theta_0$, the probability of failing to reject the null hypothesis is at least $1 - \alpha/2$ (see Problem P1). We can use this property to define a confidence set

$$C_L(\mathbf{x}, 1 - \alpha/2) = \{\theta_B : \delta_{AG}(\mathbf{x}, \alpha; \theta_B) = 0\}. \tag{2.8}$$

The set $C_L(X, 1-\alpha/2)$ will almost always be an interval (see Problem P2). If it is not an interval we can write an interval that contains it as $[\theta_L(X, 1 - \alpha/2), \infty)$, where $\theta_L(X, 1-\alpha/2) = \min C_L(\mathbf{x}, 1-\alpha/2)$ (or $-\infty$ if no minimum exists) and if appropriate, we may replace ∞ with θ_{max}, the maximum possible value of θ. Similarly, we can create an analogous confidence set starting with the alternative-is-less series of one-sided hypothesis tests (i.e., with $H_0 : \theta \geq \theta_B$ and $H_1 : \theta < \theta_B$), and its decision rule δ_{AL}. The associated confidence set is

$$C_U(\mathbf{x}, 1 - \alpha/2) = \{\theta_B : \delta_{AL}(\mathbf{x}, \alpha/2; \theta_B) = 0\}, \tag{2.9}$$

and let the smallest interval that contains it be $(-\infty, \theta_U(\mathbf{x}, 1 - \alpha/2)]$. Then a *central* $100(1 - \alpha)\%$ confidence interval is the two-sided interval that is the intersection of both one-sided intervals, giving

$$C_c(\mathbf{x}, 1 - \alpha) = [\theta_L(x, 1 - \alpha/2), \theta_U(x, 1 - \alpha/2)]. \tag{2.10}$$

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

A central confidence interval created this way is valid, since

$$\Pr[\theta \in C_c(\mathbf{X}, 1-\alpha)|\theta] \geq 1-\alpha.$$

Now we return to the Fisher's exact test example.

Example 2.5 *(continued) Now consider the confidence interval created by inverting the central Fisher's exact test, the test created from two one-sided Fisher's exact tests (unlike the two-sided test, for each pair of the one-sided hypotheses, there is only one version of a one-sided Fisher's exact test). Again, using the data,* $7/262$ *from one group and* $30/494$ *from the other, we get a p-value of* $p = 0.0518$ *and a 95% central confidence interval on the odds ratio of* $(0.155, 1.006)$.

2.3.3 Compatibility

On page 14 we defined compatibility between p-value functions and confidence intervals, where the p-value was based on a series of point null hypotheses. Here we define compatibility more formally so that it can apply to other types of tests and p-values.

Consider first a single hypothesis test using the formal definition given on page 13). Let the hypothesis test be $\{\delta, A\}$, where $A = \{\mathcal{X}, \mathcal{P}_0, \mathcal{P}_1\}$ is the set of assumptions defining the sample space, $\mathcal{X}$, the null hypothesis model or models, $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\}$, and the alternative hypothesis models, $\mathcal{P}_1 = \{P_\theta : \theta \in \Theta_1\}$. Define the (nonrandomized) decision rule in terms of a p-value, so that $\delta(\mathbf{x}, \alpha; \Theta_0) = I(p(\mathbf{x}, \Theta_0) \leq \alpha)$. Let $C_S(\mathbf{x}, 1-\alpha)$ be a confidence set procedure.

Definition 2.1 Compatibility of test and confidence set: *A hypothesis test* $\{\delta, A\}$ *is compatible with a confidence set procedure* $C_S(\mathbf{x}, 1-\alpha)$ *when, for any* $\mathbf{x}$, $\delta(\mathbf{x}, \alpha; \Theta_0) = 1$ *(or equivalently* $p(\mathbf{x}, \Theta_0) \leq \alpha$*) if and only if* $\theta \notin \Theta_0$ *for all* $\theta \in C_S(\mathbf{x}, 1-\alpha)$*. The definition is the same if the confidence set is a confidence interval.*

Now consider a series of hypothesis tests. Let t be the index for each specific hypothesis test and let the associated null and alternative parameter spaces be $\Theta_0(t)$ and $\Theta_1(t)$. For example, a series of alternative-is-greater hypotheses could be represented by parameter spaces $\Theta_0(t) = \{\theta : \theta \leq t\}$ and $\Theta_1(t) = \{\theta : \theta > t\}$. Let the associated p-value function for the series be $p(\mathbf{x}, \Theta_0(t))$. Then analogously to Definition 2.1, the p-value function is *compatible* with a confidence set procedure $C_S(\mathbf{x}, 1-\alpha)$ when for any t and $\mathbf{x}$, $p(\mathbf{x}, \Theta_0(t)) \leq \alpha$ if and only if $\theta \notin \Theta_0(t)$ for all $\theta \in C_S(\mathbf{x}, 1-\alpha)$.

A closely related property of a p-value function is *coherence*. A p-value function $p(\mathbf{x}, \Theta_0(t))$ is coherent if for every $\mathbf{x} \in \mathcal{X}$, then $p(\mathbf{x}, \Theta_0(s)) \leq p(\mathbf{x}, \Theta_0(t))$ if $\Theta_0(s) \subseteq \Theta_0(t)$ (see Röhmel (2005) or Fay and Hunsberger (2021)).

2.3.4 Three-Decision Rules

For many situations, we want to know not just whether $\theta = \theta_B$, but if it is not equal, we want to know if $\theta > \theta_B$ or $\theta < \theta_B$. For example, suppose θ represents the difference in means of

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

the response from a randomized trial of a new treatment compared to the standard treatment, and $\theta_B = 0$ represents equal means. At the design stage, we do not know if standard is better than new or vice versa, so it will be difficult to choose one of the one-sided hypothesis tests in advance. In this case we consider the *three-decision rule*, where the three decisions are: (1) fail to reject $H_0 : \theta = \theta_B$; (2) reject $H_0 : \theta = \theta_B$ and conclude $\theta > \theta_B$; or (3) reject $H_0 : \theta = \theta_B$ and conclude $\theta < \theta_B$.

The three-decision rule can be tested by using a central confidence interval together with its associated two-sided p-value, we can then combine the two-sided hypothesis with both one-sided hypotheses to create a three-decision rule. It works like this. Let p_c be the p-value created from doubling the minimum of the one-sided p-values,

$$p_c(\mathbf{x}, \theta_B) = \min\{1, 2p_{AG}(\mathbf{x}, \theta_B), 2p_{AL}(\mathbf{x}, \theta_B)\}, \tag{2.11}$$

where $p_{AG}(\mathbf{x}, \theta_B)$ is the p-value associated with the one-sided null hypothesis $H_0 : \theta \leq \theta_B$ (AG represents the alternative-is-greater hypotheses). If $p_c \leq \alpha$, we reject $H_0 : \theta = \theta_B$ at level α. Alternatively, we say we reject $H_0 : \theta = \theta_B$ with a two-sided p-value equal p_c, which implies that we would reject for any $p_c \leq \alpha$. Then if the $100(1-\alpha)\%$ central confidence interval, say (θ_L, θ_U), is compatible with p_c, this means the following: either $\theta_B < \theta_L$ in which case we reject $H_0 : \theta \leq \theta_B$ with p-value $p_c/2 = p_{AG}$, or $\theta_U < \theta_B$ in which case we reject $H_0 : \theta \geq \theta_B$ with p-value $p_c/2 = p_{AL}$.

2.4 Properties of Hypothesis Tests

It is useful to present the Type I and II errors in Table 2.1.

A Type I error is when we reject the null hypothesis when the null is true, while a Type II error is when we fail to reject the null hypothesis when the alternative is true.

Consider first the Type I error. Ideally, we design our decision rules to be valid, meaning the rate of Type I errors under any probability model in the null hypothesis is less than or equal to the α-level, α. Rejecting the null hypothesis is known as finding a *significant effect* (regardless of the truth, which is not known). It can be difficult to develop a decision rule that is valid for all sample sizes. In some cases, we can get approximate validity. One important property for these kinds of decision rules is that you have asymptotic validity, so that the approximation gets better as you collect more data. We give more details about asymptotics in Section 3.7 and Chapter 10.

Now consider the Type II error, failing to reject the null hypothesis when it is false. When we do not reject the null, we often use the phrase "fail to reject the null" (or equivalently "fail

		Truth	
		Null	*Alternative*
Decision	*Reject Null*	Type I Error	
	Fail to Reject Null		Type II Error

Table 2.1 *Tabular description of Type I and Type II errors*

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

to find a significant effect") to emphasize that we cannot make a strong statement regarding either hypothesis. We can fail to reject the null because either (1) the null hypothesis is true, or (2) even though the null hypothesis is false, the data do not appear extreme under the null perhaps due to randomness or not enough data. A common mistake is to interpret failure to reject the null as showing that the null is true, when it often means that there are insufficient data to say anything practically meaningful about the hypotheses.

Let $\beta(\theta)$ be the Type II error rate when the probability parameter vector is θ. We want $\beta(\theta)$ to be small for some reasonably likely values of $\theta \in \Theta_1$. Typically, instead of talking about Type II error rates, we talk about *power*, the probability of rejecting the null given θ. The implications for minimizing Type II error rates and maximizing power are the same, since the power under θ is equal to $1 - \beta(\theta)$. For applied work, we typically require that the power be large, about 80% or 90%, under a reasonably likely alternative, or an alternative that is based on minimum clinical significance. If the power is small, then it is likely that at the end of the study we will fail to reject the null hypothesis, which may not give us much new scientific information.

Most reasonable hypothesis tests are *consistent* for some $\theta \in \Theta_1$, meaning that the power, $1 - \beta(\theta)$ increases to 1 as the sample size approaches infinity. For these hypotheses, we can determine the smallest sample size, say n, such that the power is at least the target power (say 90%) for a specific θ. This is an important part of designing a study (see Chapter 20).

Ideally, for any set of assumptions, A, we want to pick the decision rule that is most powerful among all possible decision rules for any $\theta \in \Theta_1$. The resulting test, $\{\delta, A\}$, is called the *uniformly most powerful* (UMP) test. For many sets of assumptions, there does not exist a UMP test. In this case, when choosing between two decision rules (assuming both are valid), one would choose the one that has larger power under the alternative model of interest for some fixed sample size. Alternatively, one could fix $\theta \in \Theta_1$ and compare the minimum sample size needed to achieve a specific power, $1 - \beta(\theta)$, between the two decision rules. For example, if N_1 and N_2 represent those sample sizes for decision rules δ_1 and δ_2, then the relative efficiency of δ_2 with respect to δ_1 is defined as the ratio, $r = N_1/N_2$. In other words, δ_2 will require r times the sample size of δ_1 to get the same power (given θ).

Another way to choose between decision rules is to choose the one that has larger relative efficiency asymptotically. Since most tests are consistent for all $\theta \in \Theta_1$, the way to measure the asymptotic relative efficiency (ARE) is to consider a sequence of parameters, θ_n, that approach the boundary between the null and alternative as n goes to infinity. In most cases, the ARE is useful because it does not depend on the power or on which θ in the alternative is chosen.

Two properties of decision rules that are useful for applied work are the very closely related properties of invariance and equivariance. Invariance under a set of transformations is when the decision rule does not change after applying the transformation to the data. For example, one might want a decision rule to give the same result if you make a monotonic transformation of the responses, such as taking the log of the response. Equivariance is when the decision rule changes appropriately if you transform the data and the hypotheses in the same way. For example, consider a binomial response, where x represents the number of successes with parameters n and θ. Let $\delta(x, \alpha; \theta_0)$ be a decision rule testing the null $H_0 : \theta = \theta_0$. Suppose that instead you measure failures, $n - x$. Then the decision rule would be equivariant to switching the successes and failures if $\delta(x, \alpha; \theta_0) = \delta(n - x, \alpha; 1 - \theta_0)$ for all

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

x and θ_0. Another type of equivariance is palindromic equivariance, defined as equivariant to systematically switching the order of the responses (and parameters) so that the largest response (or parameter) becomes the smallest, the next largest becomes the next smallest, and so on. In much of the literature, no distinction is made between invariance and equivariance, and both are known as invariance. For example, "invariant to monotonic transformations" usually means "equivariant to monotonic transformations," and this is usually clear from the context. In the rest of this book, we usually use the term invariance to include equivariance.

Robustness is an important idea for applied work. A test is robust if the results do not change drastically for minor violations of the assumptions. For example, if we perform a one-sample t-test on data that is only approximately normally distributed, the results will often be reasonable. For some types of approximately normal data, however, the t-test may have poor power. For example, if the population is roughly normal except for a few extreme outliers, the t-test will have very poor power for testing the mean (see Chapter 5). Another way to improve the robustness is to change your hypotheses and base your tests on parameters that are not highly dependent on a small percentage of outliers. For example, you might test inferences about the median rather than the mean. These robustness ideas will be emphasized throughout this book.

Other properties, not emphasized much in this book are mentioned in Section 2.6.

2.5 Summary

A hypothesis test has several components:

- A set of possible data values, known as the sample space or support;
- A set of probability models that is partitioned into two models (the null and alternative hypotheses); and
- A decision rule, which determines whether the null hypothesis is rejected. It is a function of the data and the α-level.

An α-level, α, is set by the researcher. If the probability of the decision rule rejecting the null hypothesis is always less than or equal to α, given that the null is true, then the rule is said to be valid. Another useful quantity is the p-value which is the smallest α such that we would reject the null hypothesis for a given data set at level α (and additionally we would reject for all $\alpha^* > \alpha$). We desire valid decision rules, but sometimes settle for approximately valid rules.

We can create central confidence intervals by inverting two one-sided hypothesis tests, and these intervals are good for making directional inferences (i.e., making three-decision rules).

The Type I error rate is the probability of rejecting the null hypothesis when it is true. The Type II error rate is the probability of failing to reject the null when the alternative is true. Failing to reject the null hypothesis often occurs when there are insufficient data to say anything meaningful about the hypothesis, so it is important not to declare that the null is true. Power is the probability of rejecting the null hypothesis, usually under an alternative hypothesis model. For applied work, we typically require that power be large, such as 80 or 90%, whenever a reasonably likely alternative is true; otherwise, the research could provide little new information.

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

The following properties are desirable when selecting among possible decision rules: high power, invariance (i.e., the inferences are unchanged with certain transformations to the data), and robustness (i.e., the test performs well even with some violations of assumptions).

2.6 Extensions and Bibliographic Notes

Lehmann and Romano (2005) give a comprehensive theoretical treatment of statistical hypothesis tests. Their text has an emphasis on optimal tests (e.g., uniformly most powerful tests). For example, it covers the Neyman–Pearson lemma, that defines the most powerful test for comparing simple null and alternative hypotheses (a simple hypothesis is one that has only one element, e.g., the simple null has Θ_0 that is one parameter not a set of parameters). Further, their text defines an unbiased test: a test where the power to reject is at least as large as the size for all $\theta \in \Theta_1$. By restricting the set of tests to unbiased tests, one can sometimes find a test that is uniformly most powerful among those unbiased tests; this is called the *UMP unbiased test*. Similarly, we can consider an invariance property, and find the UMP tests among those with that invariance property.

2.7 Notes

N1 **Set notation:** A set is a collection of objects. For example, the even numbers greater than 0 and less than 10 are a set, $\{2, 4, 6, 8\}$. We can write the same set in the following way: $\{x : x \text{ is even and } x > 0 \text{ and } x < 10\}$. This notation is read as, the collection of all xs such that (the ":" means such that) x meets all three conditions mentioned. This notation can be useful if you are defining a set in relation to some function. For example, $\{x : f(x) > 5\}$, means all xs such that $f(x) > 5$. If $\mathcal{X}$ is a set, the notation $x \in \mathcal{X}$ means x is in that set.

N2 **Randomized decision rules:** An important theoretical use of randomized decision rules is for getting hypothesis tests with sizes (*size* is defined below Equation 2.1) exactly equal to α when the data are discrete. Consider Example 2.1, a binomial with $n = 10$. If we set

$$\delta(x, 0.05) = \begin{cases} 1 & \text{if } x \geq 9 \\ \frac{0.05 - \Pr[X \geq 9 | \theta = .5]}{\Pr[X = 8 | \theta = .5]} = 0.893 & \text{if } x = 8 \\ 0 & \text{if } x \leq 7, \end{cases}$$

then the size will be exactly equal to 0.05. The problem is that if $x = 8$, then sometimes we reject the null hypothesis but sometimes we do not, and whether we reject or not does not depend on the data. So two analysts could do the proper analysis on the same data and come up with different answers. This makes many applied statisticians uncomfortable, especially since the randomness is obviously coming from the random number generator not the data. Sometimes applied statisticians will use methods that use random number generators to implement a method with many samples (for example the use of bootstrap sampling, or Monte Carlo sampling), but in those cases the randomness is added to aid in calculations and the probability of getting very different answers by two different implementations can be kept small by using a large number of samples in the calculation.

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

N3 **Setting α-level:** Ebola is a very serious disease, and as of the beginning of 2015 there had not been a vaccine that had been tested for efficacy in humans. There was a very large outbreak of Ebola in West Africa that started in 2014, and a study was quickly designed to test some of the best vaccines in development. During the protocol development, some questioned about whether to use an α-level of 0.05 for a two-sided test. If the vaccine showed promise, but the significance was between 0.05 and 0.10, it might make sense to stop the study and start supplying the vaccine to the population if the outbreak was still continuing at that point. In this case, the Type I error (calling a useless vaccine effective) might not be as big a problem as a Type II error (failing to detect a significant effect on an effective vaccine).

N4 **Monotonic decision rules:** Formally, a monotonic decision rule has $\delta(\mathbf{x}, \alpha_1) \leq \delta(\mathbf{x}, \alpha_2)$ for all $\mathbf{x}$ and $\alpha_1 < \alpha_2$. For some simple examples that show nonmonotonic decision rules are possible (but not necessarily practical) see (Lehmann and Romano, 2005, Problems 3.17, 3.58). One reason nonmonotonic decision rules have been proposed is to increase power for unbiased (see Section 2.6) tests, but this can lead to some strange behaviour (Perlman and Wu, 1999).

N5 **Surprisal:** For cases when the p-value can be interpreted as the probability of observing equal or more extreme data under the null, a transformation has been proposed: using the surprisal, $s = -\log_2(p)$. The surprisal gives a way to envision small probabilities. An $s = 6$ is the probability associated with seeing six heads in a row from fair coin tosses. A p-value of $p = 0.05$ corresponds to $s = 4.3$ which is close to seeing four heads in a row (see Greenland, 2017).

N6 **Two-sample permutation test:** Here is how to show that a two-sample permutation test is valid under the assumptions of Example 2.4. First consider conditioning on the order statistics of $\mathbf{y}$: the values of $\mathbf{y}$ ordered from smallest to largest. This is simply a way of conditioning on the 10 values of $\mathbf{y}$, but not on the original order of the indexes associated with them. This conditioning makes the calculation of the p-values tractable. Let $\mathbf{y}_o$ be the vector of order statistics, and $\mathbf{Y}_o$ the associated vector of random variables. For clarity, we write the constant c as $c(\mathbf{y}_o)$ to emphasize that it depends on $\mathbf{y}_o$. Then unconditionally we have,

$$\int \Pr[T(\mathbf{X}) \geq c(\mathbf{y}_o) | \mathbf{Y}_o = \mathbf{y}_o] \Pr[\mathbf{Y}_o = \mathbf{y}_o],$$

where the integration is over all the possible values of $\mathbf{y}_o$ (and $T(\mathbf{x})$ is a function of the data like the difference in means or difference in mean ranks). We do not need to worry about the details of the integration because we solve for $c(\mathbf{y}_o)$ using the methods of Example 2.3. To see this, note that when y_o is fixed, then the only possible values of $\mathbf{y}$ are the permutations of the indices for the y_i. And it turns out that permuting the indices of the y_i and keeping z_i fixed, gives the same possibilities as keeping the y_i fixed and permuting the z_i. But because there are 5 ones and 5 twos in z_i, we can group the 10! different permutations of the indices of z_i into 252 groups of 5! 5! permutations, where within each group all 5! 5! permutations give the same $\mathbf{z}$. Thus, we can solve for $c(y_o)$ so that $\Pr[T(\mathbf{X}) \geq c(\mathbf{y}_o) | \mathbf{Y}_o = \mathbf{y}_o] \leq \alpha$ for all $\mathbf{y}_o$. So since each conditional term is $\leq \alpha$, when we integrate over the set of possible values of $\mathbf{y}_o$, this integral is also $\leq \alpha$. Mathematically,

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press

$$\Pr[T(\mathbf{X}) \geq c(\mathbf{y}_o)] \leq \int \alpha \Pr[\mathbf{Y}_o = \mathbf{y}_o] = \alpha.$$

N7 **Fisher–Irwin test and central Fisher exact tests in R:** The default in fisher.test in R (at least for version 4.0.4 or earlier), is to use the Fisher–Irwin two-sided p-value, but to invert the central Fisher's exact test to get the confidence intervals (hence, those confidence intervals are neither matching nor compatible with the Fisher–Irwin p-values). To get matching confidence intervals use the exact2x2 R package. The Fisher–Irwin test cannot get compatible confidence intervals but can get matching confidence intervals. See Fay (2010a) or Fay and Hunsberger (2021) for discussion of the Fisher's exact tests and their matching confidence intervals. For the general issue of matching confidence intervals see Fay (2010b). For another example, see Problem P2.

2.8 Problems

P1 Show that for a valid decision rule, for any $\theta \in \Theta_0$, the probability of failing to reject the null hypothesis at level α is at least $1 - \alpha$. Hint: start with the definition of a valid decision rule, Equation 2.1. Multiply the expression by -1 and add 1 to both sides, then use $\Pr[\delta = 0] = 1 - \Pr[\delta = 1]$ to get

$$\min_{\theta \in \Theta_0} \Pr[\delta(\mathbf{X}, \alpha) = 0 | \theta] \geq 1 - \alpha.$$

P2 **Matching confidence interval not compatible:** For an example of a confidence set derived from inverting a one-sided hypothesis test that does not give an interval, consider the one-sided exact unconditional test for the two-sample binomial problem using the score method (see Section 10.7). Using the exact2x2 R package explore this. For example, testing 130/248 against 76/170, and test different null hypotheses, $H_0 : \beta \geq \beta_0$, for $\beta_0 = 0.02, 0.024$, or 0.0.26. Are the p-values monotonic? If they are not what does this mean in terms of confidence sets possibly being confidence intervals? (Fay and Hunsberger, 2021).

https://doi.org/10.1017/9781108528825.003 Published online by Cambridge University Press