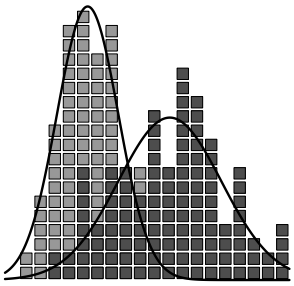
Chapter 4

Hypothesis Testing



Obtaining a point estimate of an unknown parameter using the techniques in Chapter 2, and an associated interval estimate using the techniques in Chapter 3, is one type of statistical inference. Using these methods, the experimenter obtains a best guess for the unknown parameter (the point estimate) and a measure of the precision of the point estimate (the interval estimate). This approach is adequate in some settings, but inadequate in other settings.

There is a second broad type of statistical inference known as *hypothesis testing* which is introduced in this chapter. Hypothesis testing is a formal approach that has been devised by statisticians for testing whether data supports or refutes a conjecture (that is, a hypothesis) concerning an unknown population parameter. Over the years it has become the de facto standard method for drawing conclusions from a data set. Hypothesis testing contains more elements than point and interval estimation; these elements will be defined and illustrated in Section 4.1.

4.1 Elements of Hypothesis Testing

Hypothesis testing is used to answer a question that is posed concerning an unknown parameter. We begin with the definition of a hypothesis.

Definition 4.1 A *hypothesis* is a statement concerning an unknown parameter.

As a simple illustration, consider a coin with probability p of turning up heads when flipped, where p is an unknown but fixed parameter satisfying $0 < p < 1$. The hypothesis associated with the coin being fair is

$$p = \frac{1}{2}.$$

The hypothesis associated with the coin being biased is

$$p \neq \frac{1}{2}.$$

The hypothesis associated with the coin being biased in favor of turning up heads is

$$p > \frac{1}{2}.$$

These hypotheses could be tested with a coin flipping experiment.

Now consider a somewhat more complex example. A researcher has developed a new drug that allegedly extends the average survival time of a patient with liver cancer more than the drug currently in use. Let μ_X be the population mean survival time of a liver cancer patient who uses the current drug. Let μ_Y be the population mean survival time of a liver cancer patient who uses the new drug. The hypothesis that the two drugs perform equally well in terms of the population mean survival time of a liver cancer patient is

$$\mu_X = \mu_Y .$$

The hypothesis that the new drug performs better than the current drug in terms of the population mean survival time is

$$\mu_X < \mu_Y .$$

This second hypothesis is often known as the *research worker's hypothesis* because this is the hypothesis that the experimenter is trying to show. Hypothesis testing parallels the scientific method. The two pillars of the scientific method are theory and experimentation. Formulating a hypothesis, as we have in the liver cancer drug example, corresponds to the development of a theory. In order to test the hypothesis, or theory, an experiment is conducted. In the case of the two drugs, the survival times $X_1, X_2, \ldots, X_n$ of n liver cancer patients taking the current drug are collected; the survival times $Y_1, Y_2, \ldots, Y_m$ of m liver cancer patients taking the new drug are also collected. Care should be taken in the way that the patients are placed into the two groups so that the survival times in the two groups are not influenced by other factors such as age, gender, or the progress of the disease. This process is known to statisticians as *stratified sampling*. In addition, hypothesis testing in medicine often uses a *double-blind experiment* in which both the patient and the physician are unaware of whether the current drug or the new drug is being used. This procedure is used to ward off the *placebo effect*, which can bias the results of the experiment. A hypothesis test can then be conducted to assess the effectiveness of the new drug relative to the effectiveness of the current drug.

The time-tested approach in hypothesis testing is to test one hypothesis, known as the null hypothesis, versus a second hypothesis, known as the alternative hypothesis. These two competing hypotheses are defined next.

Definition 4.2 A *hypothesis test* is a statistical inference technique for deciding between two competing propositions concerning an unknown parameter. The first proposition, known as the *null hypothesis*, is denoted by H_0. The second proposition, known as the *alternative hypothesis*, is denoted by H_1. The decision between H_0 and H_1 is based on a set of data values and is stated as either "reject H_0" or "fail to reject H_0."

Stating the null and alternative hypothesis is one of the first steps in conducting a hypothesis test. Some further details associated with H_0 and H_1 are given below.

- Some authors use the notation H_a or H_A for the alternative hypothesis. The more traditional H_1 is used here because of some helpful notational advantages.
- The null hypothesis has an inherently different nature than the alternative hypothesis because it is assumed to be true until enough evidence in the data causes it to be rejected. The null hypothesis is hardly ever true in any practical problem; it is almost always an approximation.
- The null hypothesis H_0 tends to be "no effect," "no difference," or "no change." The alternative hypothesis, on the other hand, is often the hoped-for conclusion of the test.
- The language used in the decision associated with the conclusion of the hypothesis test, that is, "reject H_0" or "fail to reject H_0," is important. There is a subtle but important difference

between the conclusions "fail to reject H_0" and "accept H_0." The words "fail to reject H_0" imply that there is not enough statistical evidence in the sample to reject H_0 in favor of H_1. The null hypothesis H_0 has not been confirmed—there just simply isn't enough evidence in the data to reject it.

- Some authors refer to H_0 and H_1 as "statistical hypotheses."
- In the examples considered in this chapter, the data values will always be a random sample.

The first example of a hypothesis test is used to decide whether an unknown parameter θ assumes one value or another. To keep the mathematics simple, only a single observation, denoted by X, is collected to decide between the null and alternative hypotheses.

Example 4.1 A single observation X is collected to test the hypothesis

$$H_0 : \theta = 1$$

versus

$$H_1 : \theta = 3,$$

where θ is a positive unknown parameter from the Rayleigh population distribution described by the probability density function

$$f(x) = 2\theta^{-2} x e^{-(x/\theta)^2} \qquad x > 0.$$

Conduct a hypothesis test associated with the data value $x = 2.2$.

The null and alternative hypotheses completely define the population probability distribution of X. The population probability density functions associated with H_0 and H_1 are graphed in Figure 4.1. Let x be a realization of the random variable X. Because of the support of the population probability distribution, x will always be positive. It is clear from Figure 4.1 that a smaller value of x is evidence in favor of H_0 and a larger value of x is evidence in favor of H_1. Now consider the particular value of x collected in this case, that is, $x = 2.2$. The data value is out in the far right-hand tail of the

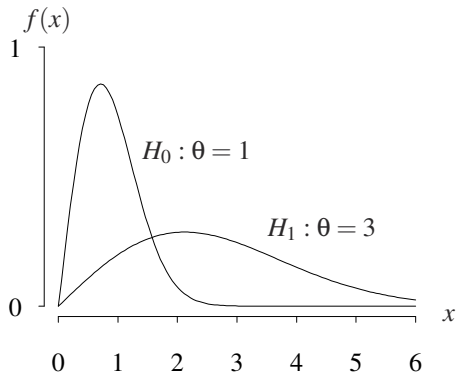


Figure 4.1: Population probability density functions for $\theta = 1$ and $\theta = 3$.

probability density function associated with $\theta = 1$, but in the center of the probability density function associated with $\theta = 3$, so it is reasonable to conclude that H_0 should be rejected in this case. The probability distribution associated with $\theta = 1$ is unlikely to have produced $x = 2.2$. This conclusion was drawn in a rather intuitive and unscientific manner by appealing to Figure 4.1, so the next example establishes a cutoff value for x for deciding between H_0 and H_1.

The null and alternative hypotheses in the previous example completely defined the probability density function of X. A hypothesis that completely defines a probability distribution is known as a simple hypothesis, which is defined next.

Definition 4.3 A hypothesis that completely specifies a probability distribution is called a *simple hypothesis*. A hypothesis that does not completely specify a probability distribution is called a *composite hypothesis*.

The next example considers a simple null hypothesis and a simple alternative hypothesis to test a hypothesis concerning the parameter θ from a special case of the beta distribution.

Example 4.2 A single observation X is collected to test the simple null hypothesis

$$H_0 : \theta = 1/2$$

versus the simple alternative hypothesis

$$H_1 : \theta = 3,$$

where θ is a positive unknown parameter from the continuous population distribution described by the probability density function

$$f(x) = \theta x^{\theta-1} \qquad 0 < x < 1.$$

Analyze an appropriate decision rule based on the single data value.

The null and alternative hypotheses are both simple because they completely define the population probability distribution of X. The population probability density functions associated with H_0 and H_1 are graphed in Figure 4.2. Let x be a realization of the random variable X. Because of the support of the population probability distribution, x will always fall in the interval $0 < x < 1$. It is clear from Figure 4.2 that a smaller value of x is evidence in favor of H_0 and a larger value of x is evidence in favor of H_1. Using terminology that will be defined formally subsequently, x is known as the *test statistic*. Since we do not have any criteria yet for selecting which values of x correspond to the two decisions, let's arbitrarily choose the cut-off value of $c = 0.64$. If $x > 0.64$, then we reject H_0; if $x \leq 0.64$, then we fail to reject H_0. Using terminology that will be defined formally subsequently, $c = 0.64$ is known as the *critical value* for the hypothesis test and $\{x | x > 0.64\}$ is known as the *critical region* for the hypothesis test. It is clear from Figure 4.2 that this is not a perfect hypothesis test in the sense that it will not always provide the correct conclusion with respect to which of the two distributions produced the data value. In that light, it is useful to calculate the probability that the test will draw the wrong conclusion. There are two types of errors that can be made. The first, known as a *Type I error*, corresponds to rejecting H_0 when H_0 is true. The second, known as

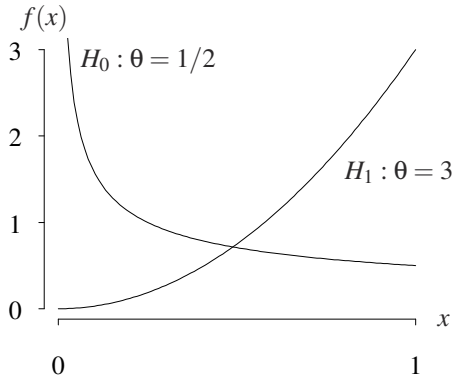


Figure 4.2: Population probability density functions for $\theta = 1/2$ and $\theta = 3$.

a *Type II error*, corresponds to failing to reject H_0 when H_1 is true. For this particular hypothesis test, the probability of a Type I error is

$$\begin{aligned}
\alpha &= P(\text{Type I error}) \\
&= P(\text{reject } H_0 \,|\, H_0 \text{ true}) \\
&= P(X > 0.64 \,|\, \theta = 1/2) \\
&= \int_{0.64}^{1} \frac{1}{2} x^{-1/2}\, dx \\
&= \left[\sqrt{x}\,\right]_{0.64}^{1} \\
&= 0.2,
\end{aligned}$$

as illustrated in Figure 4.3. The curve is the probability density function of X when H_0 is true, that is, when $\theta = 1/2$. The shaded area is the probability that X exceeds 0.64, which is the probability of a Type I error. The probability of a Type II error, on the other hand, is

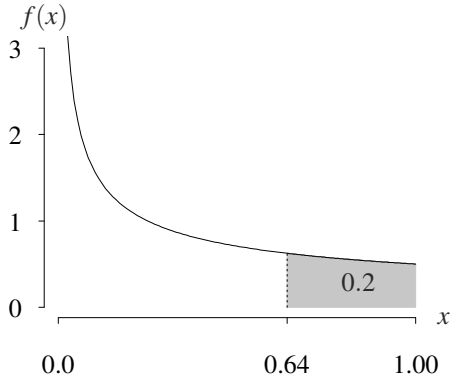


Figure 4.3: The probability of a Type I error $\alpha = P(X > 0.64 \,|\, \theta = 1/2) = 0.2$.

$$\begin{aligned}
\beta &= P(\text{Type II error}) \\
&= P(\text{fail to reject } H_0 \,|\, H_1 \text{ true}) \\
&= P(X \leq 0.64 \,|\, \theta = 3) \\
&= \int_0^{0.64} 3x^2\,dx \\
&= \left[x^3\right]_0^{0.64} \\
&\cong 0.2621,
\end{aligned}$$

as illustrated in Figure 4.4. The curve is the probability density function of X when H_1 is true, that is, when $\theta = 3$. The shaded area is the probability that X is less than or equal to 0.64, which is the probability of a Type II error.

It is worth considering the impact of the critical value $c = 0.64$ on α and β. What is the impact of increasing c? Appealing to Figures 4.3 and 4.4, this would result in a smaller α, but a larger β. On the flip side, what is the impact of decreasing c? Again appealing to Figures 4.2 and 4.3, this would result in a larger α, but a smaller β. This see-saw relationship between α and β is a characteristic of hypothesis testing.

What if the α and β values calculated in this example are unacceptably high? As was the case with confidence intervals that were unacceptably wide, more data values could be collected in order to bring both of these probabilities down.

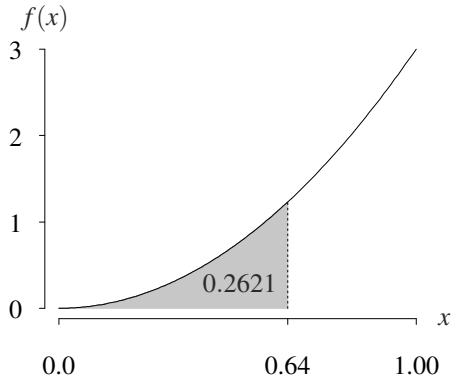


Figure 4.4: The probability of a Type II error is $\beta = P(X \leq 0.64 \,|\, \theta = 3) \cong 0.2621$.

Several new terms used in the previous example are now formally defined. A hypothesis test concerns an unknown parameter θ defined on the parameter space Ω. In the previous example, the parameter space was

$$\Omega = \{\theta \,|\, \theta > 0\}.$$

The null and alternative hypotheses can be written generically as

$$H_0 : \theta \in \Omega_0$$

and

$$H_1 : \theta \in \Omega_1,$$

where Ω_0 and Ω_1 are subsets of Ω. In the previous example, $\Omega_0 = \{\theta \,|\, \theta = 1/2\}$ and $\Omega_1 = \{\theta \,|\, \theta = 3\}$ are sets containing just a single point because the null hypothesis and the alternative hypothesis are both simple hypotheses. The set of all data values that lead to rejecting H_0 is known as the critical region, which is defined next.

Definition 4.4 For a hypothesis test of a null hypothesis H_0 versus an alternative hypothesis H_1 based on the data values $X_1, X_2, \ldots, X_n$, the set of data values that lead to rejecting H_0 is known as the *critical region* C.

The critical region is sometimes referred to as the *rejection region*. In the previous example, the critical region was $C = \{X \,|\, X > 0.64\}$. So in general, the decision rule for the test is:

- if $(X_1, X_2, \ldots, X_n) \in C$, then reject H_0,
- if $(X_1, X_2, \ldots, X_n) \notin C$, then fail to reject H_0.

The data values $X_1, X_2, \ldots, X_n$ assume values in an n-dimensional support, so specifying C can be tedious. The usual approach in hypothesis testing is to define a test statistic that is a function of the data values and an associated critical value to implicitly define the critical region C.

Definition 4.5 A *test statistic* is a real-valued function of the data values $X_1, X_2, \ldots, X_n$ that is used in a hypothesis test. The *critical value(s)* associated with the test statistic are cutoff value(s) that determine the boundary of the critical region.

The test statistic in the previous example was x, an experimental value of the random variable X, and the critical value was $c = 0.64$. Determining an appropriate test statistic for a hypothesis test is a nontrivial matter that will be postponed for later in the chapter. For the next few examples that follow, the test statistic will be given in the problem statement. The two types of errors that were introduced in the previous example, along with their associated probabilities, are defined next.

Definition 4.6 Consider a hypothesis test of a simple null hypothesis H_0 versus a simple alternative hypothesis H_1. Rejecting H_0 when H_0 is true is a *Type I error*. Failing to reject H_0 when H_1 is true is a *Type II error*. The probability of a Type I error, also known as the *significance level*, is

$$\alpha = P(\text{Type I error}) = P\left(\text{reject } H_0 \,|\, H_0 \text{ true}\right).$$

The probability of a Type II error is

$$\beta = P(\text{Type II error}) = P\left(\text{fail to reject } H_0 \,|\, H_1 \text{ true}\right).$$

Since α and β are probabilities that are associated with errors, every effort should be made to keep them as small as possible subject to the cost and time constraints associated with data collection. The pre-specified, constant significance level α goes by several aliases:

- level of significance,
- size of the critical region,
- power of the test when H_0 is true,
- the probability of committing a Type I error.

As was the case with confidence intervals, popular values of α are 0.1, 0.05, 0.01. The probability $1-\beta$, which represents

$$1-\beta = P(\text{reject } H_0 \,|\, H_1 \text{ true}),$$

is known as the *power* of the test. Table 4.1 shows the four outcomes to any hypothesis test. The columns of the 2×2 table denote the true state of nature with respect to H_0 and H_1; the rows of the table denote the decision to the hypothesis test that is made by the experimenter. The associated probabilities are given in parentheses. Two of the four cases correspond to a correct decision, and the other two correspond to an error.

		Reality	
		H_0 true	H_1 true
Decision	Fail to reject H_0	Correct $(1-\alpha)$	Type II error (β)
	Reject H_0	Type I error (α)	Correct $(1-\beta)$

Table 4.1: Four outcomes of a hypothesis test.

Most of the elements of a hypothesis test have now been defined. This leads us to the following algorithm, or method, for conducting a hypothesis test via the following steps.

1. Develop a hunch, or theory, concerning a problem of interest.
2. Translate the theory into a question concerning an unknown parameter θ.
3. State the null hypothesis H_0 in terms of θ.
4. State the alternative hypothesis H_1 in terms of θ.
5. Determine the significance level α of the test.
6. Determine the sample size n.
7. Determine a test statistic, and derive its probability distribution when H_0 is true.
8. Determine the associated critical value(s).
9. Determine the associated critical region.
10. Collect the data values $x_1, x_2, \ldots, x_n$.
11. Compute the observed value of the test statistic.
12. Decide between H_0 and H_1 based on whether the test statistic falls in the critical region.

Many of the elements of a statistical test, for example the null hypothesis, the alternative hypothesis, the significance level, the sample size, and the test statistic, are all defined *before* any data is collected. To define these elements after the data is collected could result in manipulation of the final conclusion. Looking at the data prior to formulating the steps in a hypothesis test is known as "data snooping" and is considered a violation of the process.

The choice of the level of significance α is important in a hypothesis test because it determines the critical value c, which is the key value determining the final outcome of the hypothesis test. The

value of the level of significance α should reflect the consequences of rejecting H_0 when H_0 is true. The more dire the consequences (for example, loss of life), the smaller α should be.

The final step in the process will result in "reject H_0" or "fail to reject H_0," but should also be translated back into the language associated with the problem of interest from the first step. For example, a hypothesis test might result in the conclusion that "the coin is not fair" or "the new drug prolongs the population mean survival time for liver cancer patients."

In the previous example, we defined a test statistic X, a critical value $c = 0.64$, and an associated critical region $C = \{X \,|\, X > 0.64\}$, which led to the calculation of α, namely $\alpha = 0.2$. In the next example, we collect $n = 2$ data values and set α initially, as in the hypothesis testing algorithm, to determine a critical value c.

Example 4.3 Let X_1 and X_2 be a random sample of size $n = 2$ from a $N(\theta, 1)$ population, where the population mean θ is an unknown parameter. Use the test statistic $X_1 + X_2$ to test the simple null hypothesis

$$H_0 : \theta = 5$$

against the simple alternative hypothesis

$$H_1 : \theta = 3$$

at $\alpha = 0.1$. Find the critical value c and the probability of a Type II error β.

The population variance has been set to the constant $\sigma^2 = 1$ in the problem statement so that there is just one unknown parameter θ. Since X_1 and X_2 are independent $N(\theta, 1)$ random variables, small values of the test statistic $X_1 + X_2$ are evidence in favor of rejecting H_0. The decision rule for the hypothesis test is to reject H_0 if $X_1 + X_2 < c$. In other words, the critical region is $C = \{(X_1, X_2) \,|\, X_1 + X_2 < c\}$. The first step is to find the critical value c that provides a threshold for the value of $X_1 + X_2$ to reject H_0. Since the sum of two independent normal random variables is also normally distributed,

- $X_1 + X_2 \sim N(10, 2)$ under H_0, and
- $X_1 + X_2 \sim N(6, 2)$ under H_1.

The phrase "under H_0" is equivalent to "when H_0 is true." Thus, the critical value c is the 10th percentile of a $N(10, 2)$ random variable because $\alpha = 0.1$. This can be calculated with the R command

```
qnorm(0.1, 10, sqrt(2))
```

which returns $c = 8.1876$. So the decision rule for the hypothesis test is to reject H_0 whenever $X_1 + X_2 < 8.1876$. The geometry behind this choice of critical value is shown in Figure 4.5. The curve is the probability density function of a $N(10, 2)$ random variable, which is the probability distribution of $X_1 + X_2$ under H_0. The shaded area under the curve to the left of the critical value $c = 8.1876$ is the level of significance $\alpha = 0.1$. We now turn to the calculation of β. Using the definition of a Type II error,

$$\begin{aligned} \beta &= P(\text{Type II error}) \\ &= P(\text{fail to reject } H_0 \,|\, H_1 \text{ true}) \\ &= P(X_1 + X_2 \geq 8.1876 \,|\, \theta = 3) \\ &= 0.0609. \end{aligned}$$

This probability was calculated with the following R statements.

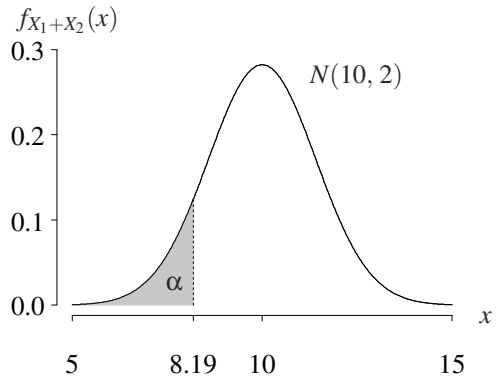


Figure 4.5: The sampling distribution of $X_1 + X_2$ when $\theta = 5$ with $\alpha = 0.1$.

```
critical.value = qnorm(0.1, 10, sqrt(2))
1 - pnorm(critical.value, 6, sqrt(2))
```

The probability of a Type I error α and the probability of a Type II error β are shown in Figure 4.6. The area under the $N(10, 2)$ probability density function to the left of $c = 8.1876$ is $\alpha = P(\text{Type I error}) = 0.1$. The area under the $N(6, 2)$ probability density function to the right of $c = 8.1876$ is $\beta = P(\text{Type II error}) = 0.0609$. The see-saw relationship between α and β is again apparent from this figure as c moves along the horizontal axis. As the critical value $c = 8.1876$ moves to the left, α decreases and β increases. As the critical value $c = 8.1876$ moves to the right, α increases and β decreases. The only way to decrease both α and β simultaneously is to increase the sample size n.

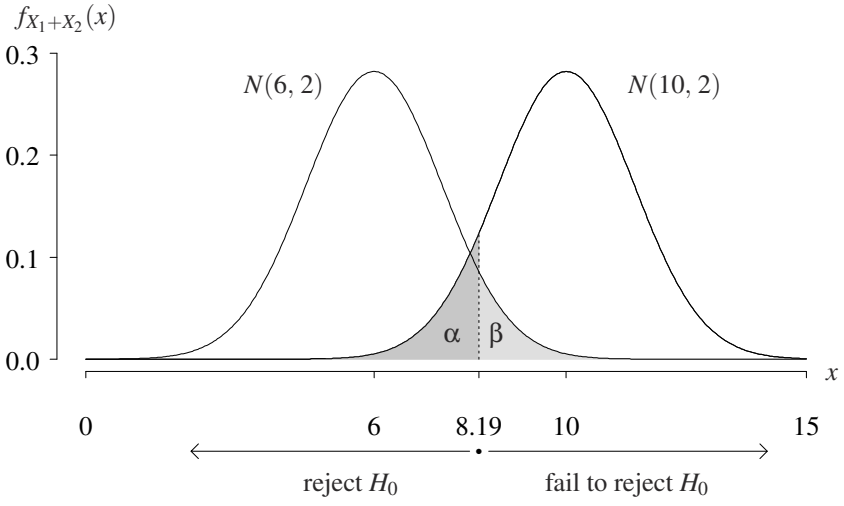


Figure 4.6: The sampling distribution of $X_1 + X_2$ when $\theta = 3$ and $\theta = 5$.