```
t.ci <- t.test(STAR$g4reading[STAR$classtype == 1],
               STAR$g4reading[STAR$classtype == 2])
t.ci

##
##  Welch Two Sample t-test
##
## data:STAR$g4reading[STAR$classtype==1] and STAR$g4reading[STAR$classtype == 2]
## t = 1.3195, df = 1541.2, p-value = 0.1872
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.703591  8.706055
## sample estimates:
## mean of x mean of y
##  723.3912  719.8900
```

The degrees of freedom are calculated as 1541.2. Because the size of our sample is not too small, the resulting confidence interval is only slightly wider than the one based on the normal approximation reported above.

7.2 Hypothesis Testing

In section 6.1.5, we presented an analysis of Arnold Schwarzenegger's 2009 veto message to the California legislature, and showed that the particular order of words in his message was highly unlikely to be a consequence of coincidence alone. This was done by examining the likelihood of observing the event that actually happened under a particular probability model. In section 6.6.3, a similar method was used to detect election fraud in Russia, where we generated hypothetical election results and compared them with the actual election outcome to investigate whether the latter was anomalous. In this section, we formalize this logic and introduce a general principle of statistical *hypothesis testing* that underlies such analysis. This principle enables us to determine whether or not the occurrence of an observed event is likely to be due to chance alone.

7.2.1 TEA-TASTING EXPERIMENT

In his classic book *The Design of Experiments*, Ronald Fisher introduced the idea of a statistical hypothesis test. During an afternoon tea party at the University of Cambridge, a lady declared that tea tastes different depending on whether the tea is poured into the milk or the milk is poured into the tea. Fisher examined this claim by using a randomized experiment in which 8 identical cups were prepared and 4 were randomly selected for milk to be poured into the tea. For the remaining 4 cups, the milk was poured first. The lady was then asked to identify, for each cup, whether the tea or the milk had been poured first. To everyone's surprise, the lady correctly classified all the cups. Did this happen by luck or did the lady actually possess the ability to detect the order, as she claimed?

Table 7.2. Tea-Tasting Experiment.

Cups	*Lady's guess*	*Actual order*	*Scenarios*			···
1	M	M	T	T	T	
2	T	T	T	T	M	
3	T	T	T	T	M	
4	M	M	T	M	M	
5	M	M	M	M	T	
6	T	T	M	M	T	
7	T	T	M	T	M	
8	M	M	M	M	T	
Number of correct guesses		8	4	6	2	···

Note: "M" and "T" represent two scenarios, "milk is poured first" and "tea is poured first," respectively. Under the hypothesis that the lady has no ability to distinguish the order in which milk and tea were poured into each cup, her guess will be identical regardless of which cups had milk/tea poured first.

To analyze this randomized experiment, we draw on potential outcomes as explained in chapter 2. For each of the 8 cups, we consider two potential guesses given by the lady, which may or may not depend on whether milk or tea was actually poured into the cup first. If we hypothesize that the lady had no ability to distinguish whether milk or tea was poured into the cup first, then her guess should not depend on the actual order in which milk and tea were poured. In other words, under this hypothesis, the two potential outcomes should be identical. Recall the *fundamental problem of causal inference*, which states that only one of the two potential outcomes can be observed. Here, the hypothesis that the lady possesses no ability to distinguish the two types of tea with milk reveals her responses under counterfactual scenarios.

Fisher's analysis proceeds under this hypothesis and involves computing the number of correctly guessed cups under every possible assignment combination. As discussed in section 7.1.1, this experiment is an example of *complete randomization*, where the number of observations assigned to each condition is fixed a priori. In contrast, *simple randomization* would randomize each cup independently without such a constraint. Table 7.2 illustrates Fisher's method. The second column of the table shows the lady's actual guess for each cup, which is identical to the true order (third column) in which milk and tea were poured into the cup. In the remaining columns, we show three arbitrarily selected combinations of assigning 4 cups to "milk first" and the other 4 to "tea first." Although these counterfactual assignment combinations did not occur in the actual experiment, we can compute the number of correctly guessed cups under each scenario with the aforementioned hypothesis that the lady lacks the ability to distinguish between the two types of tea with milk and thus different assignments do not affect the lady's guess. This is done by simply comparing the lady's guess (second column), which is assumed to remain unchanged, with each counterfactual assignment. For example, if the cups had received the assignments in the fifth column of the table, then the number of correctly classified cups would have been 6.

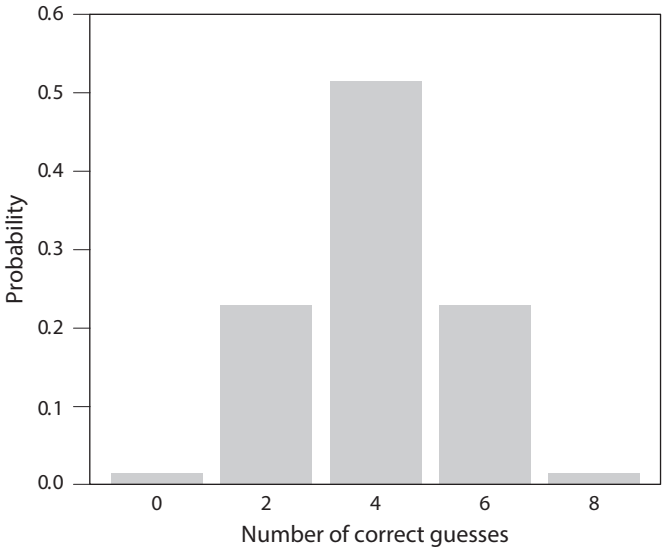

Figure 7.2. Sampling Distribution for the Tea-Tasting Experiment. The bar plot shows the distribution of the number of correctly classified cups.

Under this setup, the key question concerns the likelihood that the lady would have classified all 8 cups correctly if she had not had the ability to distinguish the taste difference. Since each assignment combination is equally likely in this randomized experiment, we can compute the probability of perfect classification by counting the number of ways in which we assign 4 cups to the "milk first" condition and the remaining 4 cups to the "tea first" condition (see equation (6.1)). The number of combinations is given by ${}_8C_4 = 8!/(4! \times (8-4)!) = 70$ because 4 cups out of 8 were selected to have tea poured in first. Thus, under the assumption that the lady has no ability to distinguish the taste difference, the probability that she guesses all cups correctly is 1/70, or approximately 0.01, which is quite small. We conclude from this analysis that the lady's perfect classification is unlikely to have occurred due to chance alone.

Moreover, as shown in figure 7.2, we can characterize the exact distribution of the number of correctly specified cups over all possible assignment combinations. How is this distribution derived? First, there is only one assignment combination, presented as the actual order in the third column of the table, that makes the lady's guesses a set of perfect classifications. Similarly, there is one assignment combination that makes all of her guesses incorrect. In this experiment, the number of ways in which the lady guesses 2 cups correctly is equivalent to the product of two things: the number of ways in which the lady correctly classifies one of the 4 "milk first" conditions and the number of ways in which the lady incorrectly classifies 3 of them. We can compute this as ${}_4C_1 \times {}_4C_3 = 16$. The same calculation applies to the number of assignment combinations that leads to 6 correctly classified cups. Similarly, we can compute the number of combinations that lead to 4 correctly classified cups, which is given by ${}_4C_2 \times {}_4C_2 = 36$. Finally, because by design the number of cups assigned to each condition is equal, there is no instance where the number of correctly classified cups

is odd. Below, we compute the probability of each event by using the `choose()` function, which enables us to compute combinations.

```
## truth: enumerate the number of assignment combinations
true <- c(choose(4, 0) * choose(4, 4),
          choose(4, 1) * choose(4, 3),
          choose(4, 2) * choose(4, 2),
          choose(4, 3) * choose(4, 1),
          choose(4, 4) * choose(4, 0))
true
```

```
## [1]  1 16 36 16  1
```

```
## compute probability: divide it by the total number of events
true <- true / sum(true)
## number of correctly classified cups as labels
names(true) <- c(0, 2, 4, 6, 8)
true
```

```
##          0          2          4          6          8
## 0.01428571 0.22857143 0.51428571 0.22857143 0.01428571
```

As done in chapter 6, we can also approximate this distribution using Monte Carlo simulations. We generate 1000 hypothetical experiments to approximate the sampling distribution of the number of correctly classified cups. To do this, we use the `sample()` function and *sample without replacement* 8 elements from a vector of 4 M's and 4 T's. This is equivalent to randomly assigning 4 cups to the "milk first" condition and the remaining 4 to the "tea first" condition. We then compute the fraction of trials that yield a certain number of correctly specified cups. The following code chunk shows this simulation approach. We find that the differences between the simulation results and the analytical answers are quite small.

```
## simulations
sims <- 1000
## lady's guess: M stands for "milk first," T stands for "tea first"
guess <- c("M", "T", "T", "M", "M", "T", "T", "M")
correct <- rep(NA, sims) # place holder for number of correct guesses
for (i in 1:sims) {
    ## randomize which cups get milk/tea first
    cups <- sample(c(rep("T", 4), rep("M", 4)), replace = FALSE)
    correct[i] <- sum(guess == cups) # number of correct guesses
}
## estimated probability for each number of correct guesses
prop.table(table(correct))
```

```
## correct
##     0     2     4     6     8
## 0.015 0.227 0.500 0.248 0.010
```

```
## comparison with analytical answers; the differences are small
prop.table(table(correct)) - true
```

```
## correct
##             0             2             4             6
##  0.0007142857 -0.0015714286 -0.0142857143  0.0194285714
##             8
## -0.0042857143
```

The major advantage of Fisher's analysis is that the inference is solely based on the randomization of treatment assignment. Such inference is called *randomization inference*. Methods based on randomization inference typically do not require a strong assumption about the data-generating process because researchers control the randomization of treatment assignment, which alone serves as the basis of inference.

7.2.2 THE GENERAL FRAMEWORK

The tea-tasting experiment described above illustrates a general framework called statistical hypothesis testing. Statistical hypothesis testing is based on probabilistic *proof by contradiction*. Proof by contradiction is a general strategy of mathematical proof in which one demonstrates that assuming the contrary of what we would like to prove leads to a logical contradiction. For example, consider the proposition that there is no smallest positive *rational number*. To prove this proposition, we assume that the conclusion is false. That is, suppose that there exists a smallest positive rational number a. Recall that any rational number can be expressed as the fraction of two integers: $a = p/q > 0$ where both the numerator p and the nonzero denominator q are positive integers. But, for example, $b = a/2$ is smaller than a, and yet b is also a rational number. This contradicts the hypothesis that a is the smallest positive rational number.

In the case of statistical hypothesis testing, we can never reject a hypothesis with 100% certainty. Consequently, we use a probabilistic version of proof by contradiction. We begin by assuming a hypothesis we would like to eventually refute. This hypothesis is called a *null hypothesis*, often denoted by H_0. In the current application, the null hypothesis is that the lady has no ability to tell whether milk or tea is poured first into a cup. This is an example of *sharp null hypothesis* because all potential outcomes for each observation are determined, and therefore known, under this hypothesis. In contrast, we will later consider a nonsharp null hypothesis, which fixes the *average* potential outcome rather than every potential outcome.

Second, we choose a *test statistic*, which is some function of observed data. For the tea-tasting experiment, the test statistic is the number of correctly specified cups. Next, under the null hypothesis, we derive the *sampling distribution* of the test statistic, which is given in figure 7.2 for our application. This distribution is also called the *reference distribution*. Finally, we ask whether the observed value of the test statistic

Table 7.3. Type I and Type II Errors in Hypothesis Testing.

	Reject H_0	Retain H_0
H_0 is true	**type I error**	correct
H_0 is false	correct	**type II error**

Note: H_0 represents the null hypothesis.

is likely to occur under the reference distribution. In the current experiment, the number of correctly classified cups is observed to be 8. If 8 is likely under the reference distribution, we retain the null hypothesis. If it is unlikely, then we reject the null hypothesis.

In this textbook, we prefer to use phrases such as "fail to reject the null hypothesis" and "retain the null hypothesis" instead of "accept the null hypothesis." Philosophical views on this issue differ, but we adopt a perspective that failure to reject the null hypothesis is evidence for some degree of consistency between the data and the hypothesis, but does not necessarily indicate the correctness of the null hypothesis. Others, however, argue that the failure to reject the null hypothesis implies acceptance of the hypothesis. Regardless of one's stance on this issue, statistical hypothesis testing provides empirical support for scientific theories.

How should we quantify the degree to which the observed value of the test statistic is unlikely to occur under the null hypothesis? We use the *p-value* for this purpose. The p-value can be understood as the probability that under the null hypothesis, we observe a value of the test statistic at least as extreme as the one we actually observed. A smaller p-value provides stronger evidence against the null hypothesis. Importantly, the p-value does not represent the probability that the null hypothesis is true. This probability is actually either 1 or 0 because the null hypothesis is either true or false, though researchers do not know which.

In order to decide whether or not to reject the null hypothesis, we must specify the *level of test* α (as explained later, this α is the same as the confidence level α for confidence intervals discussed earlier). If the p-value is less than or equal to α, then we reject the null hypothesis. The level of test represents the probability of false rejection if the null hypothesis is true. This error is called *type I error*. Typically, we would like the level of test to be low. Commonly used values of α are 0.05 and 0.01.

Table 7.3 shows two types of errors in hypothesis testing. While researchers can specify the degree of type I error by choosing the level of test α, it is not possible to directly control *type II error*, which results when researchers retain a false null hypothesis. Notably, there is a clear trade-off between type I and type II errors in that minimizing type I error usually increases the risk of type II error. As an extreme example, suppose that we never reject the null hypothesis. Under this scenario, the probability of type I error is 0 if the null hypothesis is true, but the probability of type II error is 1 if the null hypothesis is false.

In the case of the tea-tasting experiment, the test statistic is the number of correctly classified cups. Since the observed value of this test statistic was 8, which is the most extreme value, the p-value equals the probability that the number of correct guesses is 8 or $1/70 \approx 0.014$. If the lady correctly classified 6 cups instead of 8, two values are at

least as extreme as the observed value: 6 and 8. Therefore, in this case, the p-value is $({}_4C_0 \times {}_4C_4 + {}_4C_1 \times {}_4C_3)/70 = (1 + 16)/70 \approx 0.243$.

These p-values are *one-sided p-values* (or *one-tailed p-values*) because they consider only the values of the test statistic that are greater than or equal to the observed value. Under this one-sided *alternative hypothesis*, which is the complement of the null hypothesis, we ignore an extreme response on the other side, such as classifying all 8 cups incorrectly. In contrast, if we specify a two-sided alternative hypothesis, then computing the *two-sided p-value* (or *two-tailed p-value*) requires consideration of extreme values on both sides. If the reference distribution is symmetric, then the two-sided p-value is twice as great as the one-sided value. In the tea-tasting experiment, the two-sided p-value is $2/70 \approx 0.029$. If the lady had correctly guessed 6 cups, then the two-sided p-value is $2 \times (1 + 16)/70 \approx 0.486$.

While the framework described here is applicable to any statistical hypothesis testing, the particular hypothesis testing procedure used for the tea-tasting experiment is called *Fisher's exact test*. As explained earlier, this test is an example of randomization inference, where the validity of the test can be justified based on the randomization of treatment assignment.

Fisher's exact test can be implemented in R using the `fisher.test()` function. The main input of this function is a 2×2 contingency table in matrix form, where the rows and columns represent a binary treatment assignment variable and a binary outcome variable, respectively. Here, as examples, we create tables for the tea-tasting experiment: one case with all 8 cups correctly classified and the other case with 6 out of 8 cups correctly classified. In each table, rows represent actual assignments and columns provide reported guesses with the diagonal elements corresponding to the correct guesses.

```
## all correct
x <- matrix(c(4, 0, 0, 4), byrow = TRUE, ncol = 2, nrow = 2)
## 6 correct
y <- matrix(c(3, 1, 1, 3), byrow = TRUE, ncol = 2, nrow = 2)
## "M" milk first, "T" tea first
rownames(x) <- colnames(x) <- rownames(y)<- colnames(y) <- c("M", "T")
x

##   M T
## M 4 0
## T 0 4

y

##   M T
## M 3 1
## T 1 3
```

We can specify an alternative hypothesis by setting the `alternative` argument to `"two.sided"` (default), `"greater"`, or `"less"`. In the following code chunk, we conduct Fisher's exact test with one-sided and two-sided alternatives. We confirm that

the p-values obtained from the `fisher.test()` function are identical to those we calculated on our own.

```
## one-sided test for 8 correct guesses
fisher.test(x, alternative = "greater")

##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.01429
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  2.003768      Inf
## sample estimates:
## odds ratio
##        Inf

## two-sided test for 6 correct guesses
fisher.test(y)

##
##  Fisher's Exact Test for Count Data
##
## data:  y
## p-value = 0.4857
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.2117329 621.9337505
## sample estimates:
## odds ratio
##   6.408309
```

We now summarize the general procedure of statistical hypothesis testing.

In general, **statistical hypothesis testing** consists of the following five steps:

1. Specify a **null hypothesis** and an alternative hypothesis.
2. Choose a test statistic and the **level of test** α.
3. Derive the **reference distribution**, which refers to the sampling distribution of the test statistic under the null hypothesis.
4. Compute the p-**value**, either one-sided or two-sided depending on the alternative hypothesis.
5. Reject the null hypothesis if the p-value is less than or equal to α. Otherwise, retain the null hypothesis (i.e., fail to reject the null hypothesis).

While statistical hypothesis testing is a principled way to quantify uncertainty, the methodology has an important disadvantage. In particular, it forces researchers to make a binary decision about whether to reject the null hypothesis. In many situations, however, we are not interested in the null hypothesis itself. In fact, we may believe that the null hypothesis never strictly holds true. Instead, it could be more fruitful to quantify the degree to which the observed data deviate from the null hypothesis. In the tea-tasting experiment, we may wish to measure the extent to which the lady can taste the difference rather than simply whether or not she possesses any ability in this regard. While the p-value represents the degree to which empirical evidence refutes the null hypothesis, it does not directly correspond to the substantive quantity of interest. In other words, while hypothesis testing can determine *statistical significance*, it often fails to provide a direct measure of *scientific significance*.

7.2.3 ONE-SAMPLE TESTS

Using the general principle of statistical hypothesis testing we have introduced, a variety of hypothesis tests can be developed. We consider one-sample and two-sample tests, which are among the most commonly used tests. *One-sample tests* of means are used to examine the null hypothesis that the population mean equals a specific value. *Two-sample tests*, on the other hand, are based on the null hypothesis that the means of two populations equal each other. Two-sample tests are particularly useful when analyzing randomized controlled trials, enabling researchers to investigate whether or not the observed difference in average outcomes between the treatment and control groups is likely to arise by random chance alone. These tests are used more frequently than Fisher's exact test, described earlier, because they do not rely on the sharp null hypothesis that no unit is affected by the treatment. Instead, two-sample tests concern whether treatment influences an outcome *on average*.

We start, as an example of one-sample tests, with a reanalysis of the sample surveys given in section 7.1.4. Suppose that our null hypothesis is that in the population exactly half of voters support Obama and the other half do not, i.e., $H_0 : p = 0.5$. Let an alternative hypothesis be that Obama's support rate is not 0.5, i.e., $H_1 : p \neq 0.5$. Now, suppose that we conduct a simple random sample and interview 1018 selected individuals, $n = 1018$. In this sample, 550 of them express support for Obama whereas the other individuals do not. This implies that the sample proportion of Obama's supporters is 54%, i.e., $\overline{X}_n = 550/1018$. Clearly, the sample proportion differs from the hypothesized proportion, 0.5, but is this difference statistically significant? Is the difference within the sampling error? Statistical hypothesis testing can answer this question.

We follow the general procedure of hypothesis testing laid out in section 7.2.2. Since the null and alternative hypotheses are defined above, we next choose a test statistic and the level of the test. We use the sample proportion $\overline{X}_n$ as our test statistic and set $\alpha = 0.05$. We then derive the sampling distribution of this test statistic under the null hypothesis. Following the discussion in section 7.1.3 and utilizing equation (7.12), we use the central limit theorem to approximate the reference distribution of $\overline{X}_n$ as $\mathcal{N}(0.5, 0.5(1 - 0.5)/1018)$, where the variance is computed using the formula $\mathbb{V}(X)/n = p(1 - p)/n$. Note that this variance of the reference distribution is constructed using Obama's support rate under the null hypothesis, i.e., $p = 0.5$.

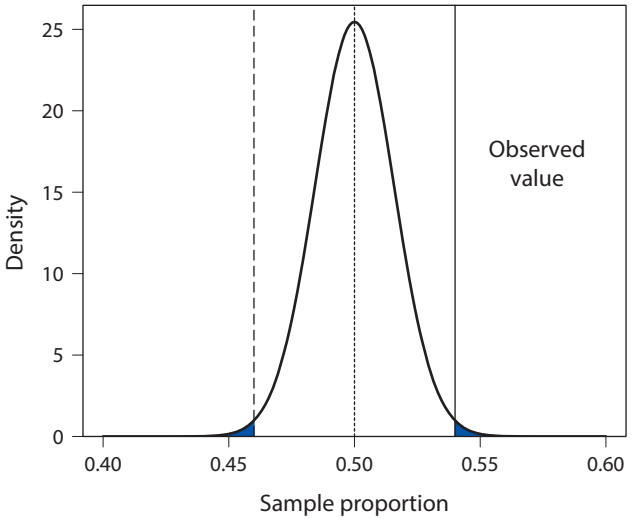

Figure 7.3. One-Sided and Two-Sided p-Values. The density curve represents the reference distribution under the null hypothesis that the population proportion is 0.5. The observed value is indicated by the solid vertical line. The two-sided p-value equals the sum of the two blue shaded areas under the curve, whereas the one-sided p-value is equal to the one of the two blue areas under the curve (depending on the alternative hypothesis).

Under this setup, the *two-sided p-value*, corresponding to our null and alternative hypotheses, can be computed as the probability that under the null hypothesis we observe a value more extreme than the observed value, i.e., $\overline{X}_n = 550/1018$. Figure 7.3 shows this graphically where a more extreme value is indicated by any value either above the observed value (solid line approximately at 0.54) or below its symmetric value (dotted line approximately at 0.46). Thus, the two-sided p-value equals the sum of the two blue shaded areas under the density curve. We use the `pnorm()` function to calculate each area where the argument `lower.tail` needs to be set to `FALSE` in order to compute the upper blue area in the figure.

```
n <- 1018
x.bar <- 550 / n
se <- sqrt(0.5 * 0.5 / n) # standard deviation of sampling distribution
## upper blue area in the figure
upper <- pnorm(x.bar, mean = 0.5, sd = se, lower.tail = FALSE)
## lower blue area in the figure; identical to the upper area
lower <- pnorm(0.5 - (x.bar - 0.5), mean = 0.5, sd = se)
## two-sided p-value
upper + lower
```

```
## [1] 0.01016866
```

In this particular case, since both the upper and lower shaded areas have the same area (because the normal distribution is symmetric around its mean), we can simply

double one of the areas to obtain the two-sided p-value. Note that this may not work in other cases where the reference distribution is not symmetric.

```
2 * upper
## [1] 0.01016866
```

If, on the other hand, our alternative hypothesis is $p > 0.5$ rather than $p \neq 0.5$, then we must compute the one-sided p-value. In this case, there is no need to consider the possibility of an extremely small value because the alternative hypothesis specifies p to be greater than the null value. Hence, the one-sided p-value is given by the blue area under the curve above the observed value in the figure.

```
## one-sided p-value
upper
## [1] 0.005084332
```

Regardless of whether we use the one-sided or two-sided p-value, we reject the null hypothesis that Obama's support in the population is exactly 50%. We conclude that the 4 percentage point difference we observe is unlikely to arise due to chance alone.

When using the normal distribution as the reference distribution, researchers often use the z-score to standardize the test statistic by subtracting its mean and dividing it by its standard deviation. Once this transformation is made, the reference distribution becomes the standard normal distribution. That is, if we use μ_0 to denote the hypothesized mean under the null hypothesis, we have the following result so long as the sample size is sufficiently large (due to the central limit theorem):

$$\frac{\overline{X}_n - \mu_0}{\text{standard error of } \overline{X}_n} \sim \mathcal{N}(0, 1). \tag{7.19}$$

Note that this transformation does not change the outcome of the hypothesis testing conducted above. In fact, the p-value will be identical with or without this transformation. However, one can easily compare the z-score with the critical values shown in table 7.1 in order to determine whether to reject the null hypothesis without computing the p-value. For example, under the two-sided alternative hypothesis, if the z-score is greater than 1.96, then we reject the null hypothesis. We now show, using the current example, that we obtain the same p-value as above.

```
z.score <- (x.bar - 0.5) / se
z.score
## [1] 2.57004
pnorm(z.score, lower.tail = FALSE) # one-sided p-value
## [1] 0.005084332
```

```
2 * pnorm(z.score, lower.tail = FALSE) # two-sided p-value
## [1] 0.01016866
```

This test, which is based on the z-score of the sample mean, is called the *one-sample z-test*. Although we used this test for a Bernoulli random variable in this example, the test can be applied to a wide range of nonbinary random variables so long as the sample size is sufficiently large and the central limit theorem is applicable. For nonbinary random variables, we will use the sample variance to estimate the standard error. If the random variable X is distributed according to the normal distribution, then the same test statistic, i.e., the z-score of the sample mean, follows the t-distribution with $n-1$ degrees of freedom instead of the standard normal distribution. This *one-sample t-test* is more conservative than the one-sample z-test, meaning that the former gives a greater p-value than the latter. Some researchers prefer conservative inference and hence use the one-sample t-test rather than the one-sample z-test.

Suppose that $\{X_1, X_2, \ldots, X_n\}$ are n independently and identically distributed random variables with mean μ and variance σ^2. The **one-sample z-test** consists of the following components:

1. Null hypothesis that the population mean μ is equal to a prespecified value μ_0: $H_0 : \mu = \mu_0$
2. Alternative hypothesis: $H_1 : \mu \neq \mu_0$ (two-sided), $H_1 : \mu > \mu_0$ (one-sided), or $H_1 : \mu < \mu_0$ (one-sided)
3. Test statistic (z-statistic): $Z_n = (\overline{X}_n - \mu_0)/\sqrt{\hat{\sigma}^2/n}$, where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (sample mean)
4. Reference distribution: $Z_n \sim \mathcal{N}(0, 1)$ when n is large
5. Variance: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ (sample variance) or $\hat{\sigma}^2 = \mu_0(1-\mu_0)$ if X is a Bernoulli random variable
6. p-value: $\Phi(-|Z_n|)$ (one-sided) and $2\Phi(-|Z_n|)$ (two-sided), where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution

If X is normally distributed, the same test statistic Z_n is called the t-statistic and follows the t-distribution with $n-1$ degrees of freedom. The p-value will be based on the cumulative distribution of this t-distribution. This is called the **one-sample t-test**, which is more conservative than the one-sample z-test.

There exists a general one-to-one relationship between confidence intervals and hypothesis tests. Compare equation (7.19) with equation (7.15). The difference is that the unknown population mean $\mathbb{E}(X)$ in the former is replaced with the hypothesized population mean μ_0 in the latter. Note that under a null hypothesis the hypothesized mean μ_0 represents the actual population mean. This suggests that we reject a null hypothesis $H_0 : \mu = \mu_0$ using the α-level two-sided test if and only if the $(1-\alpha)\times 100\%$

confidence interval does not contain μ_0. We can confirm this result using the current example by checking that 0.5 is contained in the 99% confidence interval (since we reject the null hypothesis when $\alpha = 0.1$) but not in the 95% confidence interval (we fail to reject the null when $\alpha = 0.05$).

```
## 99% confidence interval contains 0.5
c(x.bar - qnorm(0.995) * se, x.bar + qnorm(0.995) * se)

## [1] 0.4999093 0.5806408

## 95% confidence interval does not contain 0.5
c(x.bar - qnorm(0.975) * se, x.bar + qnorm(0.975) * se)

## [1] 0.5095605 0.5709896
```

It turns out that this one-to-one relationship between confidence intervals and hypothesis testing holds in general. Many researchers, however, prefer to report confidence intervals rather than p-values because the former also contain information about the magnitude of effects, quantifying *scientific significance* as well as *statistical significance*.

We conducted the one-sample z-test for sample proportion "by hand" above in order to illustrate the underlying idea. However, R has the `prop.test()` function, which enables us to conduct this test in a single line of R code. For the one-sample test of sample proportion like the one above, the function takes the number of successes as the main argument `x` and the number of trials as the argument `n`. In addition, one can specify the success probability under the null hypothesis as `p`, as well as the alternative hypothesis (`"two.sided"` for the two-sided alternative hypothesis, and either `"less"` or `"greater"` for the one-sided alternative hypothesis). The default confidence level is 95%, which we can change with the `conf.level` argument.

Finally, the `correct` argument determines whether a continuity correction should be applied in order to improve the approximation (the default is `TRUE`). This correction is generally recommended, especially when the sample size is small because the binomial distribution, which is a discrete distribution, is approximated by a continuous distribution, i.e., the normal distribution. We first show that `prop.test()` without a continuity correction gives a result identical to the one obtained earlier. We then show the result based on the continuity correction.

```
## no continuity correction to get the same p-value as above
prop.test(550, n = n, p = 0.5, correct = FALSE)

##
##  1-sample proportions test without continuity
##  correction
##
## data:  550 out of n, null probability 0.5
## X-squared = 6.6051, df = 1, p-value = 0.01017
## alternative hypothesis: true p is not equal to 0.5
```

```
## 95 percent confidence interval:
##  0.5095661 0.5706812
## sample estimates:
##       p
## 0.540275
```

```
## with continuity correction
prop.test(550, n = n, p = 0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  550 out of n, null probability 0.5
## X-squared = 6.445, df = 1, p-value = 0.01113
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5090744 0.5711680
## sample estimates:
##       p
## 0.540275
```

The `prop.test()` function also conveniently yields confidence intervals. Note that the standard error used for confidence intervals is different from the standard error used for hypothesis testing. This is because the latter standard error is derived under the null hypothesis $\sqrt{p(1-p)/n}$, whereas the standard error for confidence intervals is computed using the estimated proportion, $\sqrt{\overline{X}_n(1-\overline{X}_n)/n}$. To illustrate a different level of confidence intervals, we can compute 99% confidence intervals using the `conf.level` argument.

```
prop.test(550, n = n, p = 0.5, conf.level = 0.99)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  550 out of n, null probability 0.5
## X-squared = 6.445, df = 1, p-value = 0.01113
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.4994182 0.5806040
## sample estimates:
##       p
## 0.540275
```

As another example, we revisit the analysis of the STAR project given in section 7.1.5. We first conduct a one-sample t-test just for illustration. Suppose that we test the null hypothesis that the population mean test score is 710, i.e., $H_0 : \mu = 710$.

We use the `t.test()` function where we specify the null value μ_0 using the `mu` argument. The other arguments such as `alternative` and `conf.level` work in the exact same way as for the `prop.test()` function. We use the reading test score for our analysis and conduct a two-sided one-sample t-test. As the result below shows, we retain, at the 0.05 level, the null hypothesis that the population mean of test score is 710. The resulting p-value is small, leading to the rejection of the null hypothesis.

```
## two-sided one-sample t-test
t.test(STAR$g4reading, mu = 710)

##
##  One Sample t-test
##
## data:  STAR$g4reading
## t = 10.407, df = 2352, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 710
## 95 percent confidence interval:
##  719.1284 723.3671
## sample estimates:
## mean of x
##  721.2478
```

7.2.4 TWO-SAMPLE TESTS

We now move to a more realistic analysis of the STAR project. When analyzing randomized controlled trials like this, researchers often conduct a statistical hypothesis test with the null hypothesis that the population average treatment effect (PATE) is zero, i.e., $H_0 : \mathbb{E}(Y_i(1) - Y_i(0)) = 0$ with a two-sided alternative hypothesis given by $H_1 : \mathbb{E}(Y_i(1) - Y_i(0)) \neq 0$. If we assume that the PATE cannot be negative, then we employ a one-sided alternative hypothesis, $H_1 : \mathbb{E}(Y_i(1) - Y_i(0)) > 0$. In contrast, if we assume that the PATE cannot be positive, we set $H_1 : \mathbb{E}(Y_i(1) - Y_i(0)) < 0$. In this application, we would like to test whether or not the PATE of small class size on the grade-four reading score (relative to regular class size) is zero.

To test this null hypothesis, we use the difference-in-means estimator as a test statistic. More generally, beyond randomized controlled trials, we can use the two-sample tests based on the difference-in-means estimator to investigate the null hypothesis that the means are equal between these two populations. What is the reference distribution of this test statistic? We can approximate it by appealing to the central limit theorem as in section 7.1.5. The theorem implies that the sample means of the treatment and control groups have a normal distribution. Therefore, under the null hypothesis of equal means between the two populations, the difference between these two sample means is also normally distributed with mean zero. Furthermore, the z-score of the difference in sample means follows the standard normal distribution. We can use this fact to conduct the *two-sample z-test* (see equation (7.18) for the expression of standard error, which serves as the denominator of the test statistic). As in the one-sample tests,

if the outcomes are assumed to be normally distributed, the *two-sample t-test* can be used, which yields a more conservative inference.

Suppose that $\{X_1, X_2, \ldots, X_{n_0}\}$ represent n_0 independently and identically distributed random variables with mean μ_0 and variance σ_0^2. Similarly, $\{Y_1, Y_2, \ldots, Y_{n_1}\}$ represent n_1 independently and identically distributed random variables with mean μ_1 and variance σ_1^2. The **two-sample** z**-test** of sample means consists of the following components:

1. Null hypothesis that two populations have the same mean: $H_0 : \mu_0 = \mu_1$
2. Alternative hypothesis: $H_1 : \mu_0 \neq \mu_1$ (two-sided), $H_1 : \mu_0 > \mu_1$ (one-sided), or $H_1 : \mu_0 < \mu_1$ (one-sided)
3. Test statistic (z-statistic): $Z_n = (\overline{Y}_{n_1} - \overline{X}_{n_0}) / \sqrt{\frac{1}{n_1}\hat{\sigma}_1^2 + \frac{1}{n_0}\hat{\sigma}_0^2}$
4. Reference distribution: $Z_n \sim \mathcal{N}(0, 1)$ when n_0 and n_1 are large
5. Variance: $\hat{\sigma}_0^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(X_i - \overline{X}_{n_0})^2$ and $\hat{\sigma}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(Y_i - \overline{Y}_{n_1})^2$ (sample variances) or $\hat{\sigma}_0^2 = \hat{\sigma}_1^2 = \hat{p}(1 - \hat{p})$ with $\hat{p} = \frac{n_0}{n_0 + n_1}\overline{X}_{n_0} + \frac{n_1}{n_0 + n_1}\overline{Y}_{n_1}$ if X and Y are Bernoulli random variables
6. p-value: $\Phi(-|Z_n|)$ (one-sided) and $2\Phi(-|Z_n|)$ (two-sided), where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution

If X and Y are normally distributed, the same test statistic Z_n is called the t-statistic and follows the t-distribution. The p-value will be based on the cumulative distribution of this t-distribution. This is called the **two-sample** t**-test**, which is more conservative than the one-sample z-test.

Recall from section 7.1.5 that the estimated PATE is stored as an R object `ate.est` whereas its standard error is given by the R object `ate.se`. Using these objects, we compute the one-sided and two-sided p-values as follows.

```
## one-sided p-value
pnorm(-abs(ate.est), mean = 0, sd = ate.se)

## [1] 0.09350361

## two-sided p-value
2 * pnorm(-abs(ate.est), mean = 0, sd = ate.se)

## [1] 0.1870072
```

Since this p-value is much greater than the typical threshold of 5%, we cannot reject the hypothesis that the average treatment effect of small class size on the fourth-grade reading test score is zero.

The hypothesis test conducted above is based on the large sample approximation because we relied upon the central limit theorem to derive the reference distribution. Similar to the discussion in section 7.1.5, if we assume that the outcome variable

is normally distributed, then we could use the t-distribution instead of the normal distribution to conduct a hypothesis test. As a test statistic, we use the z-score for the difference-in-means estimator, which is called the t-statistic in the case of this two-sample t-test. Unlike the one-sample example discussed in section 7.1.5, however, the degrees of freedom must be approximated for the *two-sample t-test*. Because the t-distribution generally has heavier tails than the normal distribution, the *t-test* is more conservative and hence is often preferred even when the outcome variable may not be normally distributed.

In R, we can conduct a two-sample t-test using the `t.test()` function as we did for a one-sample t-test. For the two-sample t-test, the function takes two vectors, each of which contains data for one of the two groups. We can specify the difference between the means of the two groups, or the PATE in this application, under the null hypothesis via the `mu` argument. The default value for this argument is zero, which is what we would like to use in the current example.

```
## testing the null of zero average treatment effect
t.test(STAR$g4reading[STAR$classtype == 1],
       STAR$g4reading[STAR$classtype == 2])

##
##  Welch Two Sample t-test
##
## data:STAR$g4reading[STAR$classtype==1] and STAR$g4reading[STAR$classtype == 2]
## t = 1.3195, df = 1541.2, p-value = 0.1872
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.703591  8.706055
## sample estimates:
## mean of x mean of y
##  723.3912  719.8900
```

The output displays the value of the t-statistic as well as the p-value and the degrees of freedom for Student's t-distribution used for the test. Since the p-value is greater than the standard threshold of $\alpha = 0.05$, we fail to reject the null hypothesis that the average treatment effect of small class size on the fourth-grade reading score is zero. As in the case of `prop.test()`, the output of the `t.test()` function contains the confidence interval for the corresponding level. As expected from the use of the t-distribution, this confidence interval is slightly wider than the confidence interval based on the normal approximation we obtained in section 7.1.5. The confidence interval also contains zero, which is consistent with the fact that we fail to reject the null hypothesis of zero average treatment effect.

As another application of hypothesis tests, we reanalyze the labor market discrimination experiment described in section 2.1. In this experiment, fictitious résumés of job applicants were sent to potential employers. Researchers randomly assigned stereotypically African-American or Caucasian names to each résumé and examined whether or not the callback rate depended on the race of the applicant. The data set we analyze is contained in the CSV file `resume.csv`. The names and descriptions

of variables in this data set are given in table 2.1. The outcome variable of interest is `call`, which indicates whether or not each résumé received a callback. The treatment variable is the race of the applicant, `race`, and we focus on the comparison between black-sounding and white-sounding names.

We test the null hypothesis that the probability of receiving a callback is the same between résumés with black-sounding names and those with white-sounding names. We use the `prop.test()` function to implement the two-sample z-test. The input is a table whose columns represent the counts of successes and failures and rows represent the two groups to be compared. We will use a one-sided test because résumés with black-sounding names are hypothesized to receive fewer callbacks.

```
resume <- read.csv("resume.csv")
## organize the data in tables
x <- table(resume$race, resume$call)
x

##
##              0    1
##   black 2278  157
##   white 2200  235

## one-sided test
prop.test(x, alternative = "greater")

##
##  2-sample test for equality of proportions with
##  continuity correction
##
## data:  x
## X-squared = 16.449, df = 1, p-value = 2.499e-05
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.01881967 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.9355236 0.9034908
```

Thus, the result supports the alternative hypothesis that résumés with white-sounding names are more likely to receive callbacks than those with black-sounding names. It is instructive to directly compute this p-value without using the `prop.test()` function. Under the null hypothesis of equal proportions between the two groups, i.e., $H_0 : \mu_0 = \mu_1$, the standard error of the difference-in-means (or more accurately difference-in-proportions) estimator can be computed as

$$\sqrt{\frac{\widehat{\mathbb{V}(X)}}{n_0} + \frac{\widehat{\mathbb{V}(Y)}}{n_1}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_0} + \frac{\hat{p}(1-\hat{p})}{n_1}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}, \quad (7.20)$$

where X and Y are the outcome variables for the résumés with black-sounding and white-sounding names, respectively, n_0 and n_1 are sample sizes, and $\hat{p} = \frac{1}{n_0+n_1}(\sum_{i=1}^{n_0} X_i + \sum_{i=1}^{n_1} Y_i)$ is the overall sample proportion. We use the same estimate $\hat{p}(1-\hat{p})$ for the variances of X and Y because under the null hypothesis of identical proportions, their variances, which are based on the proportions, are also identical.

```
## sample size
n0 <- sum(resume$race == "black")
n1 <- sum(resume$race == "white")
## sample proportions
p <- mean(resume$call) # overall
p0 <- mean(resume$call[resume$race == "black"]) # black
p1 <- mean(resume$call[resume$race == "white"]) # white
## point estimate
est <- p1 - p0
est
```

```
## [1] 0.03203285
```

```
## standard error
se <- sqrt(p * (1 - p) * (1 / n0 + 1 / n1))
se
```

```
## [1] 0.007796894
```

```
## z-statistic
zstat <- est / se
zstat
```

```
## [1] 4.108412
```

```
## one-sided p-value
pnorm(-abs(zstat))
```

```
## [1] 1.991943e-05
```

The exact same p-value can be obtained using the `prop.test()` function without a continuity correction.

```
prop.test(x, alternative = "greater", correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without
##  continuity correction
##
## data:  x
## X-squared = 16.879, df = 1, p-value = 1.992e-05
## alternative hypothesis: greater
## 95 percent confidence interval:
```

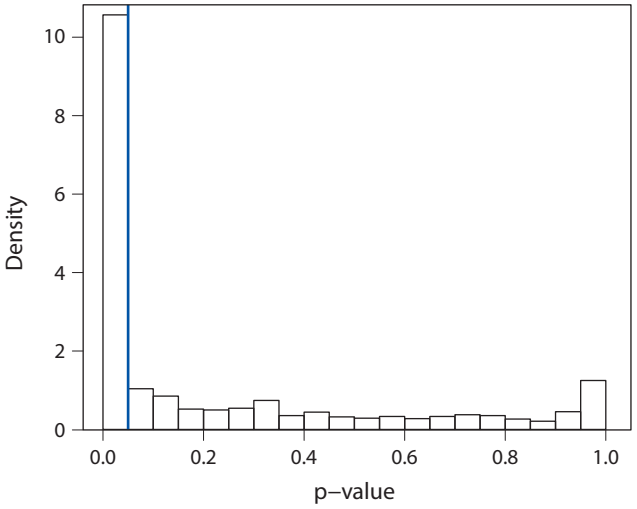

Figure 7.4. The Distribution of p-Values for Hypothesis Tests Published in Two Leading Political Science Journals.

```
##  0.01923035 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.9355236 0.9034908
```

7.2.5 PITFALLS OF HYPOTHESIS TESTING

Since Fisher's tea-tasting experiment, hypothesis testing has been extensively used in the scientific community to determine whether or not empirical findings are statistically significant. Statistical hypothesis testing represents a rigorous methodology to draw a conclusion in the presence of uncertainty. However, the prevalent use of hypothesis testing also leads to *publication bias* because only statistically significant results, and especially the ones that are surprising to the scientific community, tend to be published. In many social science journals, the α-level of 5% is regarded as the cutoff that determines whether empirical findings are statistically significant or not. As a result, researchers tend to submit their papers to journals only when their empirical results have p-values smaller than this 5% threshold. In addition, journals may also be more likely to publish statistically significant results than nonsignificant results. This is problematic because even if the null hypothesis is true, researchers have a 5% chance of obtaining a p-value less than 5%.

In one study, two researchers examined more than 100 articles published in the two leading political science journals over a decade or so.[2] The researchers collected the p-values for the hypotheses tested in those articles. Figure 7.4 shows that a majority of reported findings have p-values less than or equal to the 5% threshold, which is

[2] Alan Gerber and Neil Malhotra (2008) "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals." *Quarterly Journal of Political Science*, vol. 3, no. 3, pp. 313–326.

(a) Paul the Octopus (b) Mani the Parakeet

Figure 7.5. Two Animal Oracles that Correctly Predicted the Outcomes of Soccer Matches. Sources: (a) Reuters/Wolfgang Rattay. (b) AP Images/Joan Leong.

indicated by the blue vertical line. In addition, there appears to be a discontinuous jump at the threshold, suggesting that journals are publishing more empirical results that are just below the threshold than results just above it.

Another important pitfall regarding hypothesis testing is *multiple testing*. Recall that statistical hypothesis testing is probabilistic. We never know with 100% certainty whether the null hypothesis is true. Instead, as explained earlier, we typically have type I and type II errors when conducting hypothesis tests (see table 7.3). Multiple testing problems refer to the possibility of *false discoveries* when testing multiple hypotheses.

To see this, suppose that a researcher tests 10 hypotheses when, unbeknown to the researcher, all of these hypotheses are in fact false. What is the probability that the researcher rejects at least one null hypothesis using 5% as the threshold? If we assume independence among these hypotheses tests, we can compute this probability as

$$\begin{aligned} P(\text{reject at least one hypothesis}) &= 1 - P(\text{reject no hypothesis}) \\ &= 1 - 0.95^{10} \approx 0.40. \end{aligned}$$

The second equality follows because the probability of not rejecting the null hypothesis when the null hypothesis is true is $1 - \alpha = 0.95$ and we assume independence among these 10 hypothesis tests. Thus, the researcher has a 40% chance of making at least one false discovery. The lesson here is that if we conduct many hypothesis tests, we are likely to falsely find statistically significant results.

To illustrate the multiple testing problem, consider "Paul the Octopus" shown in figure 7.5a. This octopus in a German aquarium attracted media attention during the 2010 World Cup soccer tournament by correctly predicting all seven matches involving Germany, as well as the outcome of the final match between the Netherlands and Spain. Paul predicted by choosing to enter one of two containers with a country flag as shown in the figure. Given this data, we can conduct a hypothesis test with the null hypothesis that Paul does not possess any ability to predict soccer matches. Under this null hypothesis, Paul randomly guesses a winner out of two countries in question. What is the probability that Paul correctly predicts the outcomes of all 8 matches? Since Paul has a 50% chance of correctly predicting each match, this one-sided p-value is equal

to $1/2^8 \approx 0.004$. This value is well below the usual 5% threshold and hence can be considered statistically significant.

However, the problem of multiple testing suggests that if we have many animals predict soccer matches, we are likely to find an animal that appears to be prophetic. During the same world cup, another animal, "Mani the Parakeet" shown in figure 7.5b, was reported to have a similar oracle ability. The parakeet correctly predicted only 6 out of 8 matches. Each time, he selected one of two pieces of paper with his beak and flipped it to reveal a winner, without viewing country flags as Paul did. Since no scientific theory suggests animals can possess such predictive ability, we may conclude that Paul and Mani represent false discoveries due to the problem of multiple testing. Although beyond the scope of this book, statisticians have developed various methods that make appropriate adjustments for multiple testing.

The **multiple testing problem** is that conducting many hypothesis tests is likely to result in false discoveries, i.e., incorrect rejection of null hypotheses.

7.2.6 POWER ANALYSIS

Another problem of hypothesis testing is that null hypotheses are often not interesting. For example, who would believe that the small class in the STAR study has *exactly* zero average causal effect on students' test scores as assumed under the null hypothesis? The effect size might be small, but it is hard to imagine that it is exactly zero. A related problem is that failure to reject the null hypothesis does not necessarily mean that the null hypothesis is true. Failure to reject the null may arise because data are not informative about the null hypothesis. For example, if the sample size is too small, then even if the true average treatment effect is not zero, researchers may fail to reject the null hypothesis of zero average effect because the standard error is too large.

We use *power analysis* in order to formalize the degree of informativeness of data in hypothesis tests. The *power* of a statistical hypothesis test is defined as one minus the probability of *type II error*:

$$\text{power} = 1 - P(\text{type II error}).$$

Recall from the discussion in section 7.2.2 that type II error occurs when researchers retain a false null hypothesis. Therefore, we would like to maximize the power of a statistical hypothesis test so that we can detect departure from the null hypothesis as much as possible.

Power analysis is often used to determine the smallest sample size necessary to estimate the parameter with enough precision that its observed value is distinguishable from the parameter value assumed under the null hypothesis. This is typically done as part of research planning in order to inform data collection. In sample surveys, for example, researchers wish to know the number of people they must interview in order to reject the null hypothesis of an exact tie in support level when one candidate is ahead of the other by a prespecified degree (see also the discussion in section 7.1.4). Moreover, experimentalists use power analysis to compute the number of observations necessary

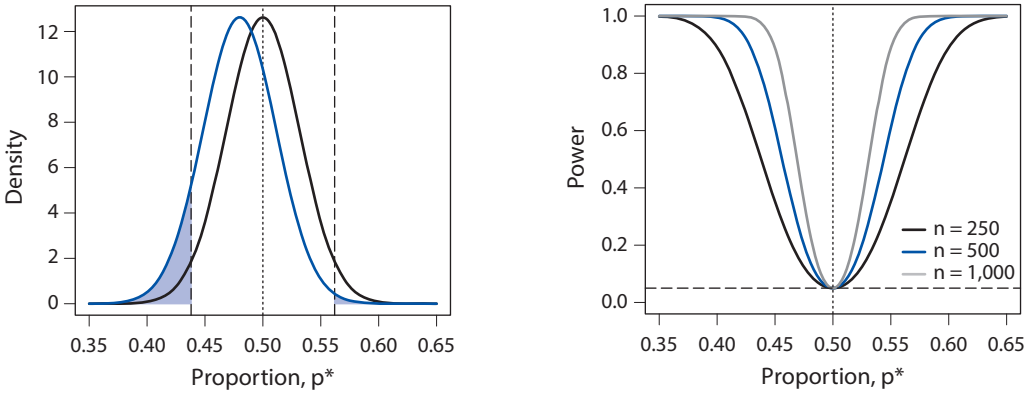

Figure 7.6. Illustration of Power Analysis. In the left-hand plot, the solid black line represents the sampling distribution of sample proportion under the null hypothesis $p = 0.5$ (vertical dotted line). The blue solid line represents the sampling distribution of the test statistic under a hypothetical data-generating process, which has mean 0.48. The sum of the two blue shaded areas equals the power of this statistical test when the significance level is $\alpha = 0.05$. The vertical dashed lines represent thresholds, above or below which the null hypothesis will be rejected. The right-hand plot displays the power function under the same setting with three different sample sizes.

to reject the null hypothesis of zero average treatment effect when the effect is actually not zero. As a result, power analysis is often required for research grant applications in order to justify the budget that researchers are requesting.

Again, we use survey sampling as an example. Suppose that we wish to find out how many respondents we must interview to be able to reject the null hypothesis that the support level for Obama, denoted by p, is exactly 50% when the true support level is at least 2 percentage points away from an exact tie, i.e., 48% or less, or 52% or greater. That is, 2 percentage points is the smallest deviation from the null hypothesis we would like to detect with a high probability. Further assume that we will use the sample proportion as the test statistic, and that the significance level is set to $\alpha = 0.05$ with a two-sided alternative hypothesis.

To compute the power, we need to consider two sampling distributions of the test statistic. The first is the sampling distribution under the null distribution. We have already derived the large sample approximation of this sampling distribution earlier: $\mathcal{N}(p, p(1-p)/n)$, where p is the null value of the population proportion. In our application, $p = 0.5$. The second is the sampling distribution under a hypothetical data-generating process. In the current case, this distribution is approximated by $\mathcal{N}(p^*, p^*(1-p^*)/n)$ via the *central limit theorem*, where p^* is either less than or equal to 0.48 or greater than or equal to 0.52.

The left-hand plot of figure 7.6 graphically illustrates the mechanics of power analysis in this case. In the plot, the two sampling distributions of the sample proportion, one centered around 0.5 under the null hypothesis (black solid line) and the other centered around 0.48 under a hypothetical data-generating process (blue solid line), are shown. We choose 0.48 as the mean value under the hypothetical data-generating process because any distribution with a mean less than this value would

result in greater statistical power, which is the probability of correctly rejecting the null, and hence would require a smaller sample size. For the meantime, we set the sample size n to 250.

Under this setting, we compute the power of the statistical test, which is the probability of rejecting the null hypothesis. To do this, we first derive the thresholds that determine the rejection region. As shown in section 7.2.3, the threshold is equal to the null value p_0 plus or minus the product of the standard error and critical value $z_{\alpha/2}$, i.e., $p_0 \pm z_{\alpha/2} \times \text{standard error}$, where in the current setting $p_0 = 0.5$ and $z_{\alpha/2} \approx 1.96$. In the left-hand plot of the figure, these thresholds are denoted by black dashed lines and we reject the null hypothesis if an observed value is more extreme than they are.

We use the probability distribution indicated by the blue solid line in the figure when computing the probability of rejection under the hypothetical data-generating process. That is, the power of the test equals the sum of the two blue shaded areas in the figure, one large area below the lower threshold and the other small area above the upper threshold. Formally, it is given by

$$\text{power} = P(\overline{X}_n < p - z_{\alpha/2} \times \text{standard error}) + P(\overline{X}_n > p + z_{\alpha/2} \times \text{standard error}).$$

In this equation, the sample proportion $\overline{X}_n$ is assumed to be approximately distributed according to $\mathcal{N}(p^*, p^*(1-p^*)/n)$, where in the current application p^* is set to 0.48. We can compute the power of a test in R as follows.

```
## set the parameters
n <- 250
p.star <- 0.48 # data-generating process
p <- 0.5 # null value
alpha <- 0.05
## critical value
cr.value <- qnorm(1 - alpha / 2)
## standard errors under the hypothetical data-generating process
se.star <- sqrt(p.star * (1 - p.star) / n)
## standard error under the null
se <- sqrt(p * (1 - p) / n)
## power
pnorm(p - cr.value * se, mean = p.star, sd = se.star) +
    pnorm(p + cr.value * se, mean = p.star, sd = se.star, lower.tail = FALSE)

## [1] 0.09673114
```

Under these conditions, the power of the test is only 10%. We can examine how the power of this test changes as a function of the sample size and hypothetical data-generating process. The right-hand plot of figure 7.6 presents the *power function*, where the horizontal axis represents the population proportion under the hypothetical data-generating process and each line indicates a different sample size. We observe that the power of a statistical test increases as the sample size becomes greater and the true population proportion p^* shifts away from the null value $p = 0.5$.

The above specific example illustrates the main principle of power analysis. We summarize the general procedure below.

Power is defined as the probability of rejecting the null hypothesis when the null hypothesis is false, which is equal to one minus the probability of type II error. **Power analysis** consists of the following steps:

1. Select the settings of the statistical hypothesis test you plan to use. This includes the specification of the test statistic, null and alternative hypotheses, and significance level.
2. Choose the population parameter value under a hypothetical data-generating process.
3. Compute the probability of rejecting the null hypothesis under this data-generating process with a given sample size.

One can then vary the sample size to examine how the power of the test changes to decide the **sample size** necessary for the desired level of power.

The power analysis can be conducted in a similar manner for two-sample tests. Consider the two-sample test of proportions, which can be used to analyze a randomized experiment with a binary outcome variable. The test statistic is the difference in sample proportion between the treatment and control groups, $\overline{Y}_{n_1} - \overline{X}_{n_0}$. Under the null hypothesis that this difference in the population, or the population average treatment effect (PATE), is equal to zero, the sampling distribution of the test statistic is given by $\mathcal{N}(0, p(1-p)(1/n_1 + 1/n_0))$, where p is the overall population proportion (see equation (7.20)), which is equal to the weighted average of the proportions in the two groups, $p = (n_0 p_0 + n_1 p_1)/(n_0 + n_1)$. To compute the power of the statistical test in this case, we must specify the population proportion separately for the treatment and control groups, p_1^* and p_0^*, under a hypothetical data-generating process. Then, the sampling distribution of the test statistic under this data-generating process is given by $\mathcal{N}(p_1^* - p_0^*, p_1^*(1-p_1^*)/n_1 + p_0^*(1-p_0^*)/n_0)$. Using this information, we can compute the probability of rejecting the null.

As an example, consider the résumé experiment analyzed in section 2.1. Suppose that we plan to send out 500 résumés with black-sounding names and another 500 résumés with white-sounding names. Further, assume that we expect the callback rate to be around 5% for black names and 10% for white names.

```
## parameters
n1 <- 500
n0 <- 500
p1.star <- 0.05
p0.star <- 0.1
```

To compute the power of this statistical test, we first compute the overall callback rate as a weighted average of callback rates of the two groups, where the weights are

their sample size. We then compute the standard error under the null hypothesis, i.e., standard error $= \sqrt{p(1-p)(1/n_0 + 1/n_1)}$, as well as under the hypothetical data-generating process, i.e., standard error$^* = \sqrt{p_1^*(1-p_1^*)/n_1 + p_0^*(1-p_0^*)/n_0}$.

```
## overall callback rate as a weighted average
p <- (n1 * p1.star + n0 * p0.star) / (n1 + n0)
## standard error under the null
se <- sqrt(p * (1 - p) * (1 / n1 + 1 / n0))
## standard error under the hypothetical data-generating process
se.star <- sqrt(p1.star * (1 - p1.star) / n1 + p0.star * (1 - p0.star) / n0)
```

We can now compute the power by calculating the probability that the difference in two proportions, $\overline{Y}_n - \overline{X}_n$, takes a value either less than $-z_{\alpha/2} \times$ standard error or greater than $-z_{\alpha/2} \times$ standard error*, under the hypothetical data-generating process.

```
pnorm(-cr.value * se, mean = p1.star - p0.star, sd = se.star) +
    pnorm(cr.value * se, mean = p1.star - p0.star,
          sd = se.star, lower.tail = FALSE)
```

```
## [1] 0.85228
```

While for illustration we computed the power by hand, we can use the `power.prop.test()` function available in R. This function, which is applicable to the two-sample test for proportions, can either compute the power given a set of parameters or determine a parameter value given a target power level. The arguments of this function include the sample size per group (`n`), population proportions for two groups (`p1.star` and `p2.star`), significance level (`sig.level`), and power (`power`). Note that the function assumes the two groups have an identical sample size, i.e., $n_0 = n_1$. To compute the power, we set `power = NULL` (default). The following syntax gives a result identical to what we computed above.

```
power.prop.test(n = 500, p1 = 0.05, p2 = 0.1, sig.level = 0.05)
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 500
##              p1 = 0.05
##              p2 = 0.1
##       sig.level = 0.05
##           power = 0.8522797
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The `power.prop.test()` function also enables sample size calculation by simply setting the `power` argument to a desired level and setting `n` to `NULL` (default). For example, if we want to know, under the same conditions as above, the minimum sample size necessary to obtain a 90% level of power, we use the following R syntax. The result below implies that we need at least 582 observations per group in order to achieve this power.

```
power.prop.test(p1 = 0.05, p2 = 0.1, sig.level = 0.05, power = 0.9)
##
##      Two-sample comparison of proportions power calculation
##
##               n = 581.0821
##              p1 = 0.05
##              p2 = 0.1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

For continuous variables, we can conduct a power analysis based on *Student's t-test*, introduced in section 7.2.4. The logic is exactly the same as that described above for one-sample and two-sample tests of proportions. The `power.t.test()` function can perform a power analysis where the `type` argument specifies a two-sample (`"two.sample"`) or one-sample (`"one.sample"`) test. For a one-sample *t*-test, we must specify the mean `delta` and standard deviation `sd` of a normal random variable under a hypothetical data-generating process. For a two-sample *t*-test, the function assumes that the standard deviation and sample size are identical for the two groups. We, therefore, specify the true difference-in-means `delta` under a hypothetical data-generating process as well as a standard deviation `sd`. Finally, the function assumes the null hypothesis that the mean is zero for a one-sample test and the mean difference is zero for a two-sample test. If the null value is not zero, then one simply has to adjust the hypothetical data-generating process by subtracting that value from the true mean (or mean difference).

Below, we present two examples of using the `power.t.test()` function. The first is the power calculation for a one-sample test with a true mean of 0.25 and standard deviation of 1. The sample size is 100. Recall that the assumed mean value under the null hypothesis is zero.

```
power.t.test(n = 100, delta = 0.25, sd = 1, type = "one.sample")
##
##      One-sample t test power calculation
##
```

```
##               n = 100
##           delta = 0.25
##              sd = 1
##       sig.level = 0.05
##           power = 0.6969757
##     alternative = two.sided
```

Under this setting, the power is calculated to be 70%. What is the sample size we need to have a power of 0.9 under the same setting? We can answer this question by specifying the `power` argument in the `power.t.test()` function while leaving the `n` argument unspecified.

```
power.t.test(power = 0.9, delta = 0.25, sd = 1, type = "one.sample")
```

```
##
##      One-sample t test power calculation
##
##               n = 170.0511
##           delta = 0.25
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

The minimum sample size for obtaining a power of 0.9 or greater is 171. The second example is the sample size calculation for a one-sided two-sample test with a true mean difference of 0.25 and standard deviation of 1. We set the desired power to be 90%.

```
power.t.test(delta = 0.25, sd = 1, type = "two.sample",
             alternative = "one.sided", power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##               n = 274.7222
##           delta = 0.25
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

The result shows that we need a minimum of 275 observations per group to achieve a power of 90% under this setting.