Print this page

4.2 MEASURING EVIDENCE WITH P-VALUES

What Is a P-value?

In Section 4.1 we learned that evidence against the null hypothesis is measured by examining how extreme sample results would be, if the null hypothesis were true. This leads us to one of the most important ideas of statistical inference: the *p-value* of the sample. The p-value gives us a formal way to measure the strength of evidence a sample provides against the null hypothesis and in support of the alternative hypothesis.

The P-value

The **p-value** of the sample data in a statistical test is the probability, when the null hypothesis is true, of obtaining a sample as extreme as (or more extreme than) the observed sample.

The smaller the p-value, the stronger the statistical evidence is against the null hypothesis and in favor of the alternative.

There are various ways to calculate p-values. In this chapter we'll take an approach similar to the bootstrapping procedures of Chapter 3 and calculate the p-value by generating lots of simulated samples. In Chapter 3, we use *bootstrap samples* to show the distribution of sample statistics if we resample from the original sample to approximate a sampling distribution for the population. Here, we are interested in the sort of statistics we observe *if we assume the null hypothesis is true*. Thus, when testing hypotheses, we simulate samples in a way that is consistent with the null hypothesis. We call these *randomization samples*.

For each simulated sample, we calculate the statistic of interest. We collect the values of the statistic for many randomization samples to generate a *randomization distribution*. This distribution approximates a sampling distribution from a population where the null hypothesis holds. If the statistic from the original sample lies in a typical part of that distribution, we do not find it to be significant. On the other hand, if the statistic for our original sample lies in an extreme, unlikely part of the randomization distribution, usually out in one of the tails, we have statistically significant evidence against H_0 and in support of H_a.

Randomization Distribution

Simulate many samples using a random process that matches the way the original data were collected and that *assumes the null hypothesis is true*. Collect the values of a sample statistic for each sample to create a **randomization distribution**.

Assess the significance of the *original* sample by determining where its sample statistic lies in the

randomization distribution.

P-values from Randomization Distributions

By now you are probably curious, *do* dogs really resemble their owners? Recall that in Data 4.1 we saw that 16 out of 25 dogs were matched correctly with their owners. Just how extreme is this result?

In Data 4.1 we are interested in testing $H_0: p=0.5$ vs $H_a: p>0.5$, where p is the proportion of correct matches between dogs and owners. In the actual sample of 25 trials, we found 16 matches, giving a sample proportion of $\hat{p} = 0.64$. How unlikely is it to see 16 or more matches out of 25 tries if participants are just guessing at random? One way to assess the chance of this happening is to simulate lots of samples of 25 trials when $p=0.5$ and keep track of how often 16 or more matches occur. While it is impractical to repeat the experimental procedure many times (lots of sets of 25 dog-owner pairs), we could easily simulate the results with an equivalent process that ensures $p=0.5$. For example, we could flip a fair coin 25 times and count the number of heads, then repeat the process several thousand times. That might also be fairly time consuming (and produce a very sore flipping thumb), so we generally use computer simulations instead. For example, Figure 4.7 shows a histogram of the number of heads (correct dog-owner matches) in each of 10,000 sets of 25 simulated coin flips where $p=0.5$.

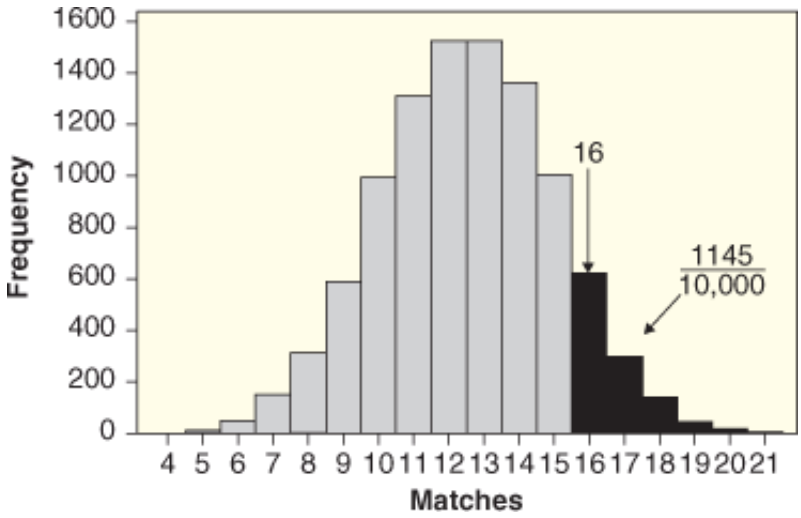

Figure 4.7 ***Randomization distribution of matches in 10,000 sets of 25 trials when*** **p=0.5**

Example 4.13

Explain, using the definition of a p-value, how we can find the p-value from the randomization distribution in Figure 4.7. What does the p-value tell us about the likelihood of dogs looking like their owners?

Solution

A p-value is the probability, when the null hypothesis is true, of obtaining a sample as extreme as the observed sample. In this case, the null hypothesis ($p=0.5$) represents random guessing. The randomization distribution in Figure 4.7 shows many simulated samples where we assume the null

hypothesis is true. To find the p-value, we want to find the proportion of these simulated samples that are as extreme as the observed sample. Among these 10,000 simulated samples of size 25, we find 1145 cases with a number of matches greater than or equal to the 16 that were observed in the original dog-owner study (see the region in the upper tail of Figure 4.7). Thus, for the dog-owner data, we have

$$\text{p-value} = \frac{1145}{10{,}000} = 0.1145$$

This value, 0.1145, estimates the probability of seeing results as extreme as 16 if people are randomly guessing. Even if there were no dog-owner resemblances, we would expect to get 16 or more correct at least 11% of the time. This is not so unusual, so we expect that the original sample is not statistically significant. We do not have sufficient evidence to conclude that dogs tend to resemble their owners.

You may be wondering why we compute the probability of *at least* 16 heads, rather than exactly 16 heads. In many situations, there are many possible outcomes, and the probability of getting any single outcome (even a typical one) will be very small. For example, consider flipping 100 coins. Getting 50 heads is the most likely answer, and certainly 50 heads should not be considered atypical, but there is actually quite a small probability of getting *exactly* 50 heads. For this reason, we always measure the probability of getting a result as extreme as that observed.

Note that the value of 0.1145 (from our simulation of dog-owner matches) is only an approximation of the p-value. If we had created another set of 10,000 simulated results, we might see a slightly different number of simulations with 16 or more matches. For example, two other simulations of 10,000 samples are shown in Figure 4.8 and yield p-values[13]Technically, these are called “empirical” or “estimated” p-values, but we will often refer to them simply as the p-values from each of the randomization distributions. of 0.1137 and 0.1162. While these p-values differ a bit from simulation to simulation, the basic conclusion that results so extreme occur in slightly more than 1 out of every 10 samples is the same in every instance. As in Chapter 3, we observe a fairly regular, consistent pattern in the general shapes of each of these randomization distributions. They each peak near 12 or 13 matches (as we would expect with $n=25$ and $p=0.5$), are relatively symmetric bell shapes, and display roughly the same variability. As we will see in later chapters, these features are all quite predictable and can be exploited in certain situations to compute p-values very efficiently.

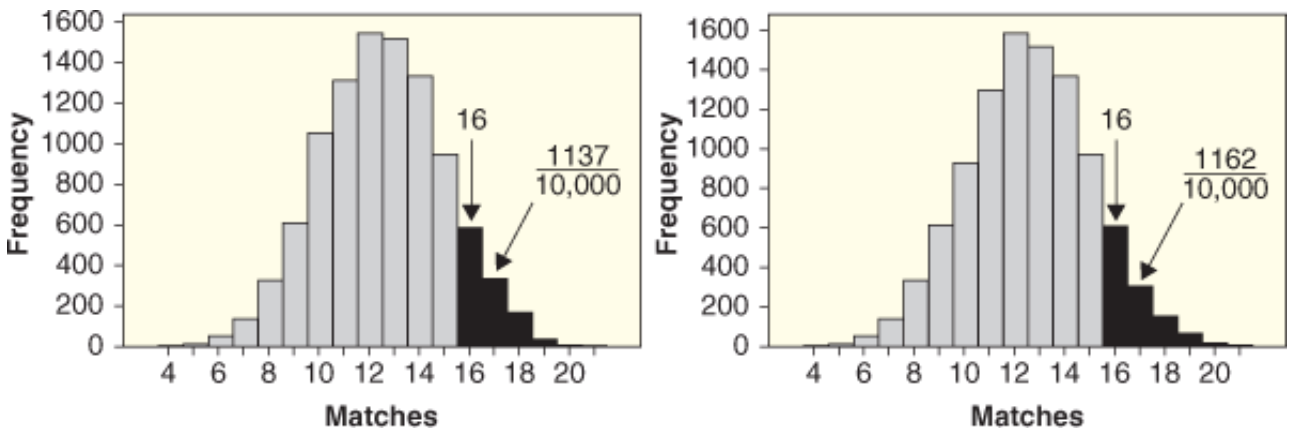

Figure 4.8 ***Two more sets of 10,000 simulations of dog-owner matches with $n=25$ and $p=0.5$***

Example 4.14

Flip a coin 25 times and count the number of heads. Where does your point fall on the randomization distribution?

Solution

Most of you will get numbers in the "typical" part of the distribution, but some of you may get an atypical number out in one of the tails. Probably none of you will get less than 4 heads or more than 21 heads, since none of our 10,000 simulated values are that extreme. About 11% of you will get at least 16 heads.

Example 4.15

(a) Use Figure 4.7 or 4.8 to determine which has a smaller p-value: sample results of 15 correct matches or 19 correct matches?

(b) It turns out that the p-value for 15 correct matches is about 0.2151, while the p-value for 19 correct matches is about 0.0075. Interpret each of these as a proportion of the total area in the randomization histogram in Figure 4.7.

(c) Interpret each of the p-values from part (b) in terms of the probability of the results happening by random chance.

(d) Which of the p-values from part (b), 0.2151 or 0.0075, provides the strongest evidence against the null hypothesis and in support of the alternative hypothesis?

Solution

(a) The part of the tail past 15 is much larger than the part past 19, so the p-value will be smaller for 19 correct matches.

(b) The fact that the p-value for 15 is 0.2151 tells us that the area in the tail past 15 is about 21.5% of the total area under the distribution. See Figure 4.9(a). Likewise, since the p-value for 19 is 0.0075, the area in the tail past 19 is a very small percentage of the total area, as we see in Figure 4.9(b).

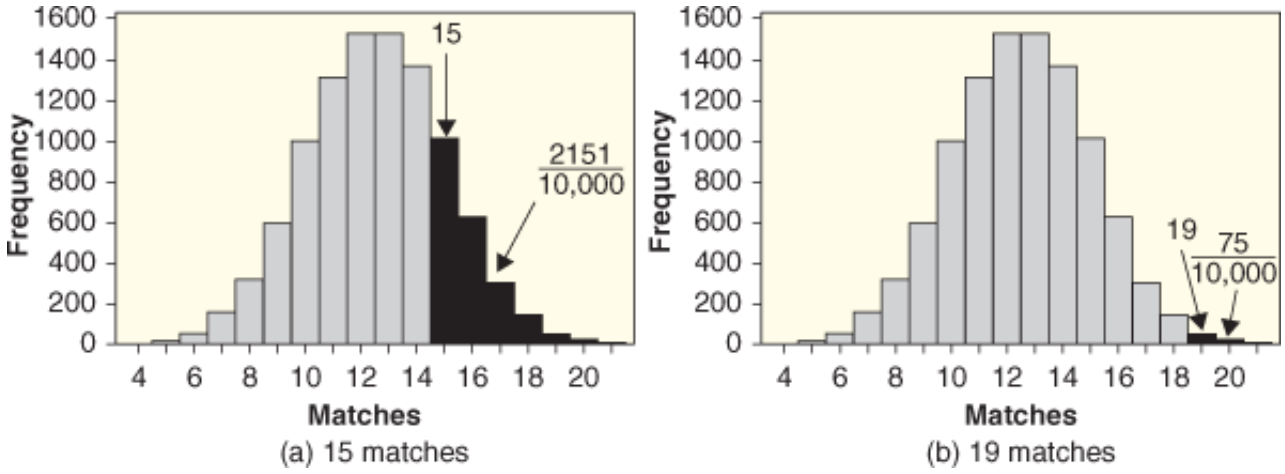

Figure 4.9 ***Empirical p-values for different correct dog-owner matches in n=25 trials and p=0.5***

(c) The p-value tells us the probability of the sample results (or ones more extreme) happening by random chance *if the null hypothesis is true.* The p-value for 15 tells us that, if people are just randomly guessing, they will get 15 or more correct matches about 21.5% of the time. The p-value for 19 tells us that, if people are just randomly guessing, they will get 19 or more correct matches only about 0.75% of the time, or only 7 or 8 times out of 1000. If people are guessing, getting at least 15 correct out of 25 is relatively common whereas getting at least 19 correct is very unlikely.

(d) We know that *the smaller the p-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis*. The smaller p-value of 0.0075 provides the strongest evidence that dogs resemble their owners. (This makes sense, since 19 correct matches provides greater evidence than 15 correct matches out of 25.)

Example 4.16

Use Figure 4.7 to estimate the p-value for sample results of 22 correct matches out of 25. What is the conclusion of the test in this case? Do you get the same result if you use either of the distributions in Figure 4.8?

Solution

The randomization distribution in Figure 4.7 does not even extend out as far as 22, so in all 10,000 simulations, there are no results as extreme as 22. The same result is evident in both distributions in Figure 4.8. Although it is *possible* to get 22 or even all 25 correct just by guessing, it is *very* unlikely. The p-value is approximately zero, which shows very strong evidence against the null hypothesis and in support of the alternative hypothesis. If 22 dogs had been correctly matched with their owners in our experiment, we would have had strong evidence that dogs do resemble their owners.

Practice Problems 4.2E

DATA 4.6 **Finger Tapping and Caffeine**

© Hakan Dere/iStockphoto Photodisc/Getty Images,Inc.

Does caffeine facilitate rapid movements?

Many people feel they need a cup of coffee or other source of caffeine to "get going" in the morning. The effects of caffeine on the body have been extensively studied. In one experiment,[14]Hand, A.J., Daly, F., Lund, A.D., McConway, K.J., and Ostrowski, E., *Handbook of Small Data Sets*, Chapman and Hall, London, 1994, p. 40. researchers trained a sample of male college students to tap their fingers at a rapid rate. The sample was then divided at random into two groups of 10 students each. Each student drank the equivalent of about two cups of coffee, which included about 200 mg of caffeine for the students in one group but was decaffeinated coffee for the second group. After a 2-hour period, each student was tested to measure finger tapping rate (taps per minute). The students did not know whether or not their drinks included caffeine and the person measuring the tap rates was also unaware of the groups. This was a double-blind experiment with only the statistician analyzing the data having information linking the group membership to the observed tap rates. (Think back to Chapter 1: Why is this important?) The goal of the experiment was to determine whether caffeine produces an increase in the average tap rate. The finger-tapping rates measured in this experiment are summarized in Table 4.4 and stored in **CaffeineTaps**.

Table 4.4 ***Finger tap rates for subjects with and without caffeine***

Caffeine	246	248	250	252	248	250	246	248	245	250	Mean = 248.3
No caffeine	242	245	244	248	247	248	242	244	246	242	Mean = 244.8

Example 4.17

State null and alternative hypotheses for the finger-tapping experiment.

Solution

We are dealing with quantitative data and are interested in the average tap rate for the two different groups. Appropriate parameters to consider are μ_c and μ_n, the mean tap rates for populations of male students with and without caffeine, respectively. The researchers are looking for a difference in means in a particular direction (higher mean tap rate for the caffeine group) so the hypotheses are

$$H_0: \ \mu_c = \mu_n$$
$$H_a: \ \mu_c > \mu_n$$

We see that, in the sample data, the mean tap rate of the caffeine group is higher than that of the no-caffeine group. The key question is whether that difference is statistically significant or could have occurred merely due to the random assignment of students to the two groups. Figure 4.10 shows comparative dotplots of the finger tap rates for the two samples, with arrows at the two sample means. Note that everyone in the caffeine group has a tap rate above the mean of the no-caffeine group, while none of the rates from the no-caffeine group are above the mean of the caffeine group. This would tend to support the alternative hypothesis that caffeine increases the average tap rate. On the other hand, there is considerable overlap between the two distributions, so perhaps it was just random chance that happened to assign several of the slowest tappers to the no-caffeine group and the fastest tappers to the group with caffeine.

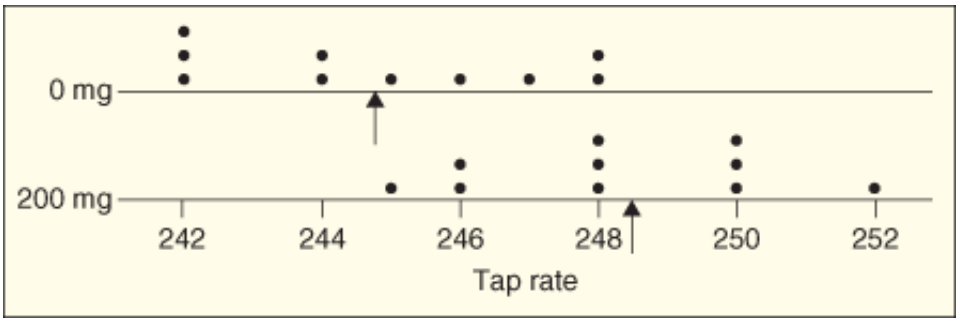

Figure 4.10 ***Dotplots of tap rates for groups with and without caffeine***

The sample statistic of interest is the difference in the sample means. For these data, the observed difference in the sample means is

$$D = \bar{x}_c - \bar{x}_n = 248.3 - 244.8 = 3.5$$

To determine whether this difference is statistically significant, we need to find the chance of a difference as large as D=3.5 occurring if caffeine really has no effect on tap rates. In other words, we need to find the p-value.

We generate a randomization distribution by assuming the null hypothesis is true. In this case, the null hypothesis is μ_c=μ_n or, more generally, that caffeine has no effect on tap rate. This assumption means that a person's tap rate would be the same whether the person is assigned to the caffeine group or the no-caffeine group. Any of the values observed in the caffeine group could just as easily have come from the no-caffeine group and vice versa if a different random assignment had been made at the start of the

experiment.

To create the randomization distribution by assuming H_0 is true, then, we randomly assign the 20 observed values to the two groups and compute D, the difference in means, for each such randomly generated assignment. For example, the data in Table 4.5 show the same 20 tap rates after a new random assignment into the two groups. Now the difference in the sample means is $\bar{x}_c - \bar{x}_n = 246.8 - 246.3 = 0.5$.

Table 4.5 ***Random assignment of tap rates to groups***

Caffeine	244	250	248	246	248	245	246	247	248	246	Mean = 246.8
No-caffeine	250	244	252	248	242	250	242	245	242	248	Mean = 246.3

We can imagine putting all 20 sample values on index cards, shuffling the deck and repeatedly dealing them at random into two piles of 10 values each. Each such random deal represents a different random assignment of subjects to the two experimental groups and, if the null hypothesis (no effect due to caffeine) is true, gives a plausible value for the difference in the two sample means. If we repeat this process many times, we obtain a randomization distribution of plausible differences and can see where our actual observed difference, D=3.5, falls.

Figure 4.11 shows a dotplot of the differences in means based on a computer simulation of 1000 random assignments of these 20 tap rate values into two groups of size 10. Since we started by assuming that H_0:μ_c=μ_n is true (which implies that μ_c−μ_n=0), it is no surprise that the distribution of the randomization differences in means is centered approximately at zero and symmetric.

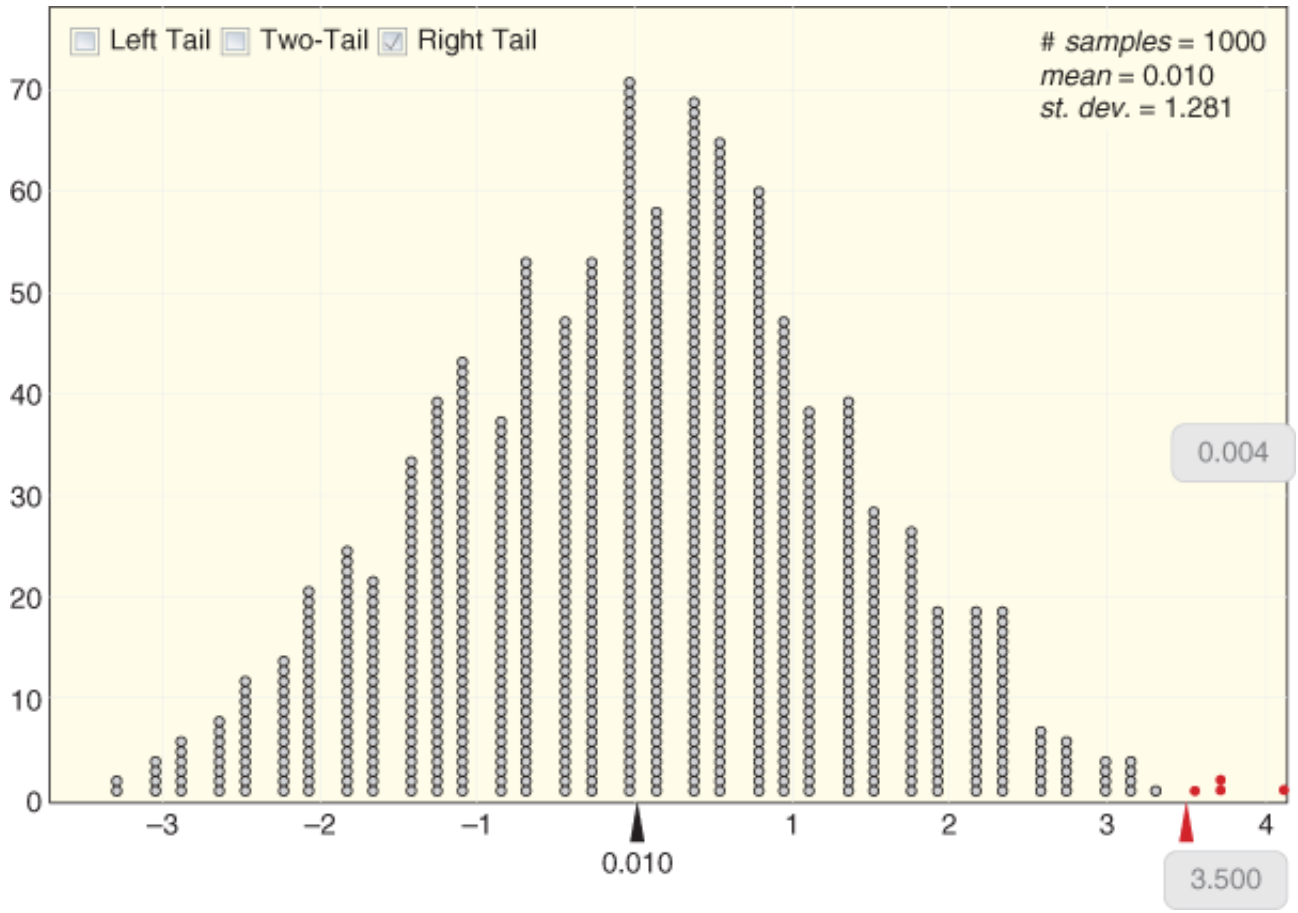

Figure 4.11 ***Distribution of differences in finger tap means for 1000 randomizations***

Example 4.18

How unlikely is the original observed difference $D=3.5$? Use Figure 4.11 to find and interpret the p-value. Is there evidence that caffeine increases tap rate?

Solution

We see in Figure 4.11 that the observed difference $D=3.5$ is very far in the upper "tail" of the distribution—certainly not in a typical location. In fact, only four of the simulated group assignments produced a difference in means equal to or larger than what was observed in the actual experiment. We have

$$\text{p-value} = \frac{4}{1000} = 0.004$$

This provides very strong evidence that the experimental results are statistically significant; it is very unlikely to see a difference this extreme if we simply assign two groups at random assuming caffeine has no effect on the tap rates. Since this was an experiment with the researchers randomly assigning the values of the explanatory variable (caffeine, no-caffeine), it is appropriate to infer causation and we have strong evidence that caffeine does increase tap rates.

P-values and the Alternative Hypothesis

As we have seen, we create the randomization distribution by assuming the null hypothesis is true. Does the alternative hypothesis play any role in computing a p-value? It turns out it does. In some tests, the alternative hypothesis specifies a particular direction (greater than or less than). We refer to these as *one-tailed* or *one-sided* tests since we seek evidence in just one direction from the null value. In other cases, we are only looking to see if there is a difference without specifying in advance in which direction it might lie. These are called *two-tailed* or *two-sided* tests. Whether a test is one-sided or two-sided is determined by the alternative hypothesis (H_a), which in turn is derived directly from the question of interest. The definition of "more extreme" when computing a p-value is dependent on whether the alternative hypothesis is one-tailed or two-tailed, as the next example illustrates.

To see if a coin is biased, we might flip the coin 10 times and count the number of heads. For a fair coin, we expect the proportion of heads to be $p=0.5$ so we have $H_0: p=0.5$. A randomization distribution for the number of heads in 10 flips of a coin with $p=0.5$ using 1000 simulations is shown in Figure 4.12.

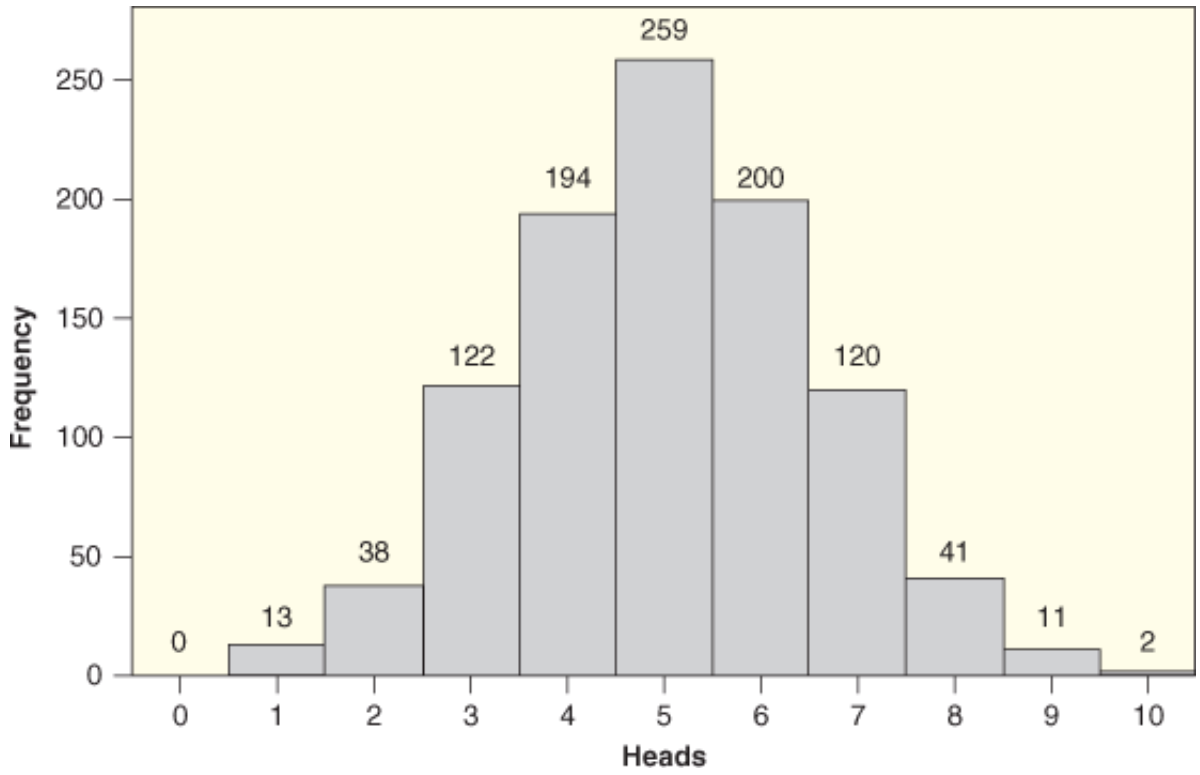

Figure 4.12 ***Randomization distribution of number of heads in 1000 sets of n=10 simulated coin flips***

Example 4.19

If the alternative hypothesis for the coin-flipping data is

$$H_a: \quad p > 0.5$$

use Figure 4.12 to estimate the p-value for each of the following observed numbers of heads in 10 flips:

(a) 8 heads

(b) 6 heads

(c) 4 heads

Solution

(a) We see in Figure 4.12 that, in our simulated coin flips, we obtained 8 heads 41 times, 9 heads 11 times, and all 10 heads twice. In other words, our results were greater than or equal to 8 on $41+11+2=54$ of 1000 simulations, so the estimated p-value is $54/1000=0.054$. Since the alternative hypothesis is one tailed, only looking for evidence that $p>0.5$, we only care about results that are more extreme in that direction.

(b) In Figure 4.12, we see that the simulated results were greater than or equal to 6 heads in $200+120+41+11+2=374$ times out of the 1000 simulations, so the empirical p-value is $374/1000=0.374$.

(c) Again, because of the alternative hypothesis, we only care about results more extreme to the right of our observed value of 4 heads. We see in Figure 4.12 that this includes $194+259+200+120+41+11+2=827$ times out of 1000, so the p-value is $827/1000=0.827$. This p-value is very large and provides no evidence at all for the alternative hypothesis. As we expect, seeing 4 heads in 10 flips of a coin gives no evidence that the coin is biased in a way that will produce too many heads.

Example 4.20

Now assume the alternative hypothesis for the coin-flipping data is

$$H_a: \quad p \neq 0.5$$

and estimate the p-value if we see 8 heads in 10 flips.

Solution

Using Figure 4.12 as in the example above, we again see that there are 54 simulated samples (out of 1000) that produced 8 or more heads. But now, with the two-tailed alternative $H_a: p \neq 0.5$, results "as extreme as" 8 heads include both tails of the distribution: At least 8 heads, or at least 8 tails (2 or fewer heads). To calculate the probability of results as extreme as 8, we need to measure the area in both tails. In Figure 4.13 we see that $0+13+38=51$ of the 1000 simulations had 2 or fewer heads in the 10 flips and $41+11+2=54$ had 8 or more heads. Thus one estimate of the p-value would be $(51+54)/1000=105/1000=0.105$. You can see why these are called *two-tailed* tests since we use counts of the extreme values from both tails of the randomization distribution.

Another way to estimate the p-value for a two-tailed test is to just double the one-tail p-value. This gives an empirical p-value of $2(54)/1000=0.108$, which is very close to what we obtain by adding the two tails separately. The results are similar in this example because the randomization distribution is fairly symmetric around the expected value (when H_0 is true) of five heads.

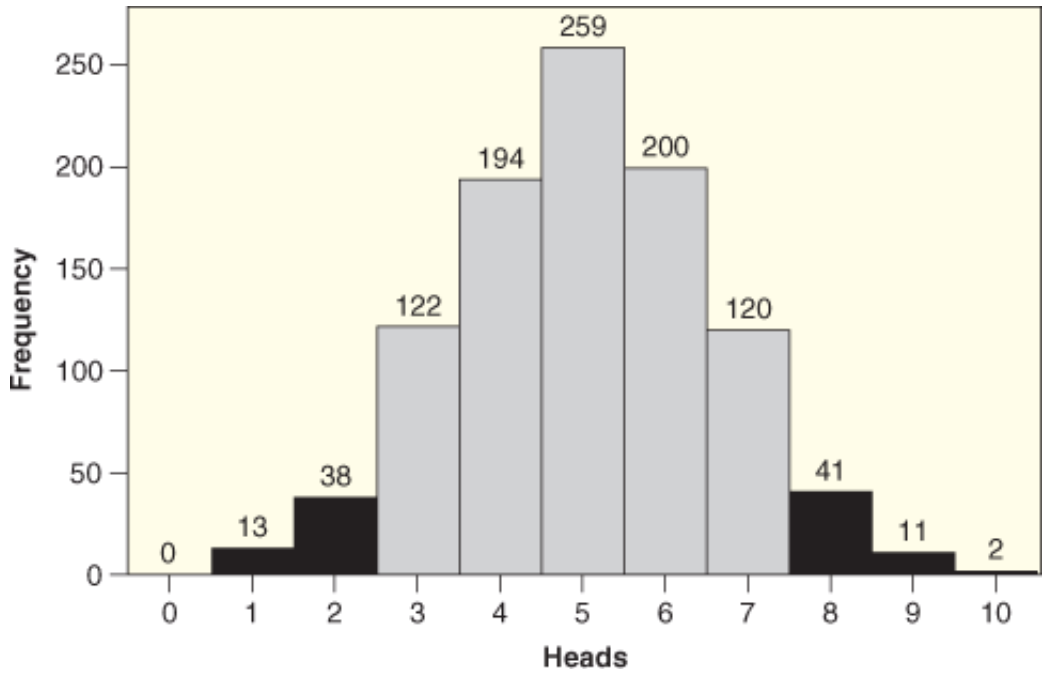

Figure 4.13 ***Two-tailed p-value for 8 heads in 10 flips when $H_0: p=0.5$***

To measure p-values for two-tailed tests, we need to account for both tails of the distribution. Although many of the randomization distributions we encounter will be symmetric, that is not always the case (see Exercise 4.64). For this reason, we usually estimate a p-value for a two-tailed test based on a randomization distribution using the second method illustrated above: Find the proportion of randomization samples with statistics at or beyond what was observed in our original sample and double that value to account for the other tail. That gives a consistent interpretation of "as extreme as" to mean "as unlikely as" rather than just "as far away as," since in asymmetric cases it may be difficult to determine from what center point to measure distance.

Estimating a P-value from a Randomization Distribution

For a one-tailed alternative: Find the proportion of randomization samples that equal or exceed the original statistic in the direction (tail) indicated by the alternative hypothesis.

For a two-tailed alternative: Find the proportion of randomization samples in the *smaller* tail at or beyond the original statistic and then double the proportion to account for the other tail.

Take care when applying this method for a two-tailed test to always use the proportion in the smaller tail. For example, in the coin flip example, if we saw 4 heads and were doing a two-tailed test, the estimated p-value based on the randomization distribution in Figure 4.12 would be $2(0+13+38+122+194)/1000=2(367)/1000=0.734$. Note that doubling the *upper* tail p-value for 4 heads from Example 4.19(c) would give $2(0.827)=1.654$, which is impossible. A p-value can never be larger than 1!

Practice Problems 4.2F

Just as different bootstrap distributions gave slightly different confidence intervals, different randomization distributions will give slightly different p-values. Different simulations yield slightly different counts and p-value estimates which are similar, but not identical. Our goal in constructing the randomization distribution is to get an idea of whether the sample data are unusual if the null hypothesis is true, and variation in the third decimal place of the p-value is not something to worry about. However, just as with confidence intervals, if we do care about accuracy even in the third decimal place, we can simply increase the number of simulated randomizations.

Practice Problems 4.2G

How small does the p-value need to be for us to consider the sample to be statistically significant? That is the topic we consider in the next section. At that point you will be equipped to use the information shown in the randomization distribution to make a decision concerning the competing claims in the null and alternative hypotheses. After Section 4.3 we will return to the question of constructing randomization distributions in Section 4.4 to see how to use *StatKey* or other technology to create randomization distributions and to illustrate more examples for testing different types of hypotheses.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Interpret a p-value as the probability of results as extreme as the observed results happening by random chance, if the null hypothesis is true
- Estimate a p-value from a randomization distribution
- Connect the definition of a p-value to the motivation behind a randomization distribution
- Comparatively quantify strength of evidence using p-values
- Distinguish between one-tailed and two-tailed tests in estimating p-values

Exercises for Section 4.2

SKILL BUILDER 1

In Exercises 4.41 to 4.44, two p-values are given. Which one provides the strongest evidence against H_0?

4.41 p-value=0.90 or p-value=0.08

ANSWER ⊕

WORKED SOLUTION ⊕

4.42 p-value=0.04 or p-value=0.62

4.43 p-value=0.007 or p-value=0.13

ANSWER ⊕

WORKED SOLUTION ⊕

4.44 p-value=0.02 or p-value=0.0008

SKILL BUILDER 2

In Exercises 4.45 to 4.47, a randomization distribution based on 1000 simulated samples is given along with the relevant null and alternative hypotheses. Which p-value most closely matches the observed statistic?

4.45 Figure 4.14 shows a randomization distribution for testing H_0:μ=50 vs H_a:μ>50. In each case, use the distribution to decide which value is closer to the p-value for the observed sample mean.

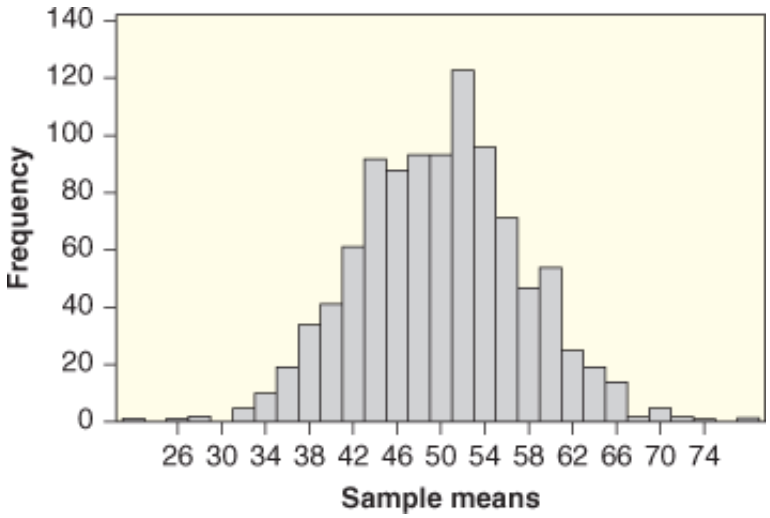

Figure 4.14 ***Randomization distribution for Exercise 4.45***

(a) The p-value for $\bar{x} = 68$ is closest to: 0.01 or 0.25?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) The p-value for $\bar{x} = 54$ is closest to: 0.10 or 0.30?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) The p-value for $\bar{x} = 63$ is closest to: 0.05 or 0.50?

ANSWER ⊕

WORKED SOLUTION ⊕

4.46 Figure 4.15 shows a randomization distribution for testing H_0:p=0.3 vs H_a:p<0.3. In each case, use the distribution to decide which value is closer to the p-value for the observed sample proportion.

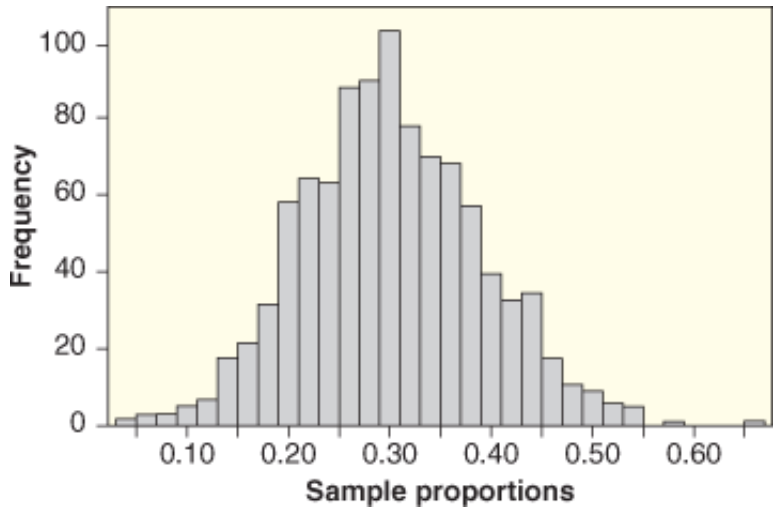

Figure 4.15 ***Randomization distribution for Exercise 4.46***

(a) The p-value for $\hat{p} = 0.25$ is closest to: 0.001 or 0.30?

(b) The p-value for $\hat{p} = 0.15$ is closest to: 0.04 or 0.40?

(c) The p-value for $\hat{p} = 0.35$ is closest to: 0.30 or 0.70?

4.47 Figure 4.16 shows a randomization distribution for testing H_0:μ_1=μ_2 vs H_a:$\mu_1 \neq \mu_2$. The statistic used for each sample is $D = \bar{x}_1 - \bar{x}_2$. In each case, use the distribution to decide which value is closer to the p-value for the observed difference in sample means.

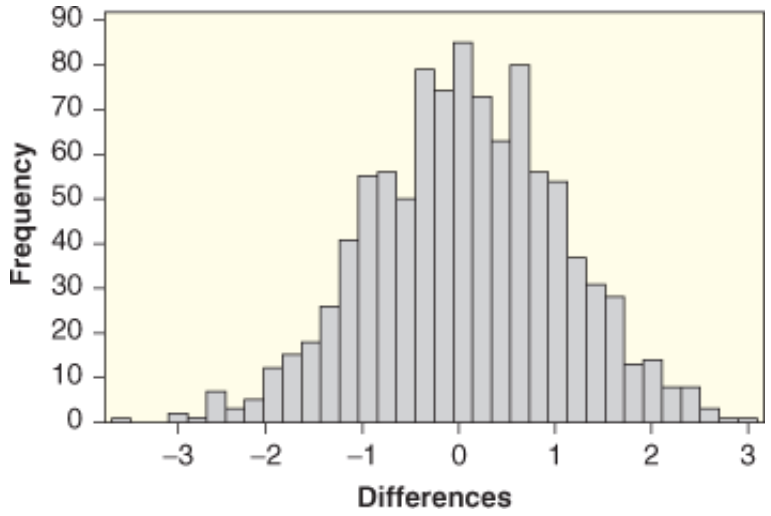

Figure 4.16 ***Randomization distribution for Exercises 4.47 to 4.51***

(a) The p-value for $D = \bar{x}_1 - \bar{x}_2 = -2.9$ is closest to: 0.01 or 0.250?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) The p-value for $D = \bar{x}_1 - \bar{x}_2 = 1.2$ is closest to: 0.30 or 0.60?

ANSWER ⊕

WORKED SOLUTION ⊕

SKILL BUILDER 3

Exercises 4.48 to 4.51 also refer to Figure 4.16, which shows a randomization distribution for hypotheses H_0:μ_1=μ_2 vs H_a:$\mu_1 \neq \mu_2$. The statistic used for each sample is $D = \bar{x}_1 - \bar{x}_2$. Answer parts (a) and (b) using the two possible sample results given in each exercise.

(a) For each D-value, sketch a smooth curve to roughly approximate the distribution in Figure 4.16, mark the D-value on the horizontal axis, and shade in the proportion of area corresponding to the p-value.

(b) Which sample provides the strongest evidence against H_0? Why?

4.48 D=2.8 or D=1.3

4.49 D=0.7 or D=−1.3

ANSWER ⊕

WORKED SOLUTION ⊕

4.50 $\bar{x}_1 = 17.3, \bar{x}_2 = 18.7$ or $\bar{x}_1 = 19.0, \bar{x}_2 = 15.4$

4.51 $\bar{x}_1 = 95.7, \bar{x}_2 = 93.5$ or $\bar{x}_1 = 94.1, \bar{x}_2 = 96.3$

ANSWER ⊕

WORKED SOLUTION ⊕

4.52 Arsenic in Chicken

Data 4.5 discusses a test to determine if the mean level of arsenic in chicken meat is above 80 ppb. If a restaurant chain finds significant evidence that the mean arsenic level is above 80, the chain will stop using that supplier of chicken meat. The hypotheses are

$$H_0: \ \mu = 80$$
$$H_a: \ \mu > 80$$

where μ represents the mean arsenic level in all chicken meat from that supplier. Samples from two different suppliers are analyzed, and the resulting p-values are given:

Sample from Supplier A: p-value is 0.0003

Sample from Supplier B: p-value is 0.3500

(a) Interpret each p-value in terms of the probability of the results happening by random chance.

(b) Which p-value shows stronger evidence for the alternative hypothesis? What does this mean in terms of arsenic and chickens?

(c) Which supplier, A or B, should the chain get chickens from in order to avoid too high a level of arsenic?

4.53 Multiple Sclerosis and Sunlight

It is believed that sunlight offers some protection against multiple sclerosis (MS) since the disease is rare near the equator and more prevalent at high latitudes. What is it about sunlight that offers this protection? To find out, researchers[15]Seppa, N., "*Sunlight May Cut MS Risk by Itself*," *Science News*,

April 24, 2010, p. 9, reporting on a study in *Proceedings of the National Academy of Science*, March 22, 2010. injected mice with proteins that induce a condition in mice comparable to MS in humans. The control mice got only the injection, while a second group of mice were exposed to UV light before and after the injection, and a third group of mice received vitamin D supplements before and after the injection. In the test comparing UV light to the control group, evidence was found that the mice exposed to UV suppressed the MS-like disease significantly better than the control mice. In the test comparing mice getting vitamin D supplements to the control group, the mice given the vitamin D did not fare significantly better than the control group. If the p-values for the two tests are 0.472 and 0.002, which p-value goes with which test?

ANSWER ⊕

WORKED SOLUTION ⊕

4.54 Dogs and Owners

The data for the 10,000 simulated dog-owner matches shown in Figure 4.7 are given in Table 4.6. We are testing H_0:p=0.5 (random guessing) vs H_a:p>0.5 (evidence of a dog-owner resemblance).

Table 4.6 ***Data for Figure 4.7 on simulated numbers of correct matches in 25 trials***

Correct matches	4	5	6	7	8	9	10	11	12
Frequency	1	17	54	148	341	599	972	1302	1549
Correct matches	13	14	15	16	17	18	19	20	21
Frequency	1551	1344	977	612	322	142	51	14	4

(a) Use the data in the table to verify that the p-value for the observed statistic of 16 correct matches is 0.1145.

(b) Use the data to calculate a p-value for an observed statistic of 20 correct matches.

(c) Use the data to calculate a p-value for an observed statistic of 14 correct matches.

(d) Which of the three p-values in parts (a) to (c) gives the strongest evidence against H_0?

(e) If any of the p-values in parts (a) to (c) indicate statistical significance, which one would it be?

4.55 Finger Tapping and Caffeine

In Data 4.6 we look at finger-tapping rates to see if ingesting caffeine increases average tap rate. Letting μ_c and μ_n represent the average tap rate of people who have had coffee with caffeine and without caffeine, respectively, the null and alternative hypotheses are

$$H_0: \mu_c = \mu_n$$
$$H_a: \mu_c > \mu_n$$

(a) Sketch a smooth curve that roughly approximates the distribution in Figure 4.17 and shade in the proportion of area corresponding to the p-value for a difference in average sample tap rates of $D = \bar{x}_c - \bar{x}_n = 1.6$. Which of the following values is closest to the p-value: 0.60, 0.45, 0.11, or 0.03?

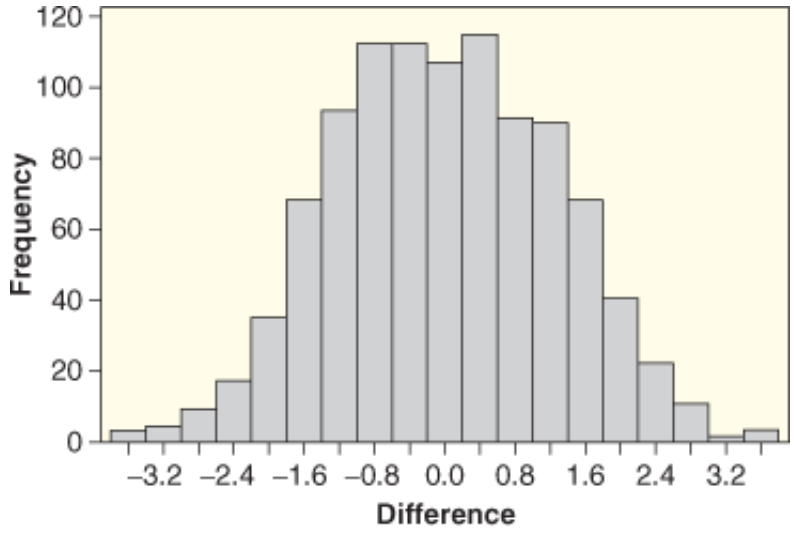

Figure 4.17 ***Distribution of differences in means for 1000 randomizations when*** $\mu_c=\mu_n$

ANSWER

WORKED SOLUTION

(b) On another sketch of the distribution in Figure 4.17, shade in the proportion of area corresponding to the p-value for a difference in average sample tap rates of $D=\bar{x}_c-\bar{x}_n=2.4$. Which of the following values is closest to the p-value: 0.60, 0.45, 0.11, or 0.03?

ANSWER

WORKED SOLUTION

(c) Which of the results given in parts (a) and (b) provides the strongest evidence that caffeine increases average finger-tapping rate? Why?

ANSWER

WORKED SOLUTION

4.56 Influencing Voters: Is a Phone Call Better Than a Flyer?

Exercise 4.38 describes a study to investigate whether a recorded phone call is more effective than a flyer in persuading voters to vote for a particular candidate. The response variable is the proportion of voters planning to vote for the candidate, with p_c and p_f representing the proportions for the two methods (receiving a phone call and receiving a flyer, respectively.) The sample statistic of interest is $D=\hat{p}_c-\hat{p}_f$. We are testing H_0:p_c=p_f vs H_a:p_c>p_f. A randomization distribution for this test is shown in Figure 4.18.

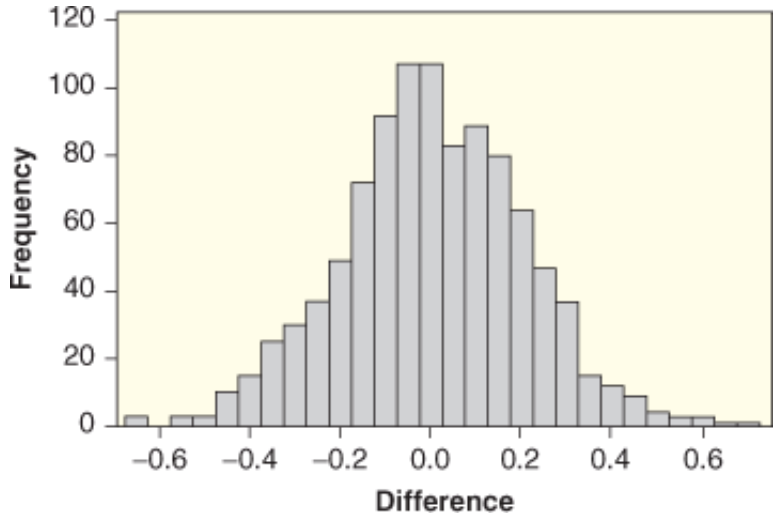

Figure 4.18 ***Randomization distribution using n = 1000 for testing*** $H_0: p_c = p_f$

(a) Sketch a smooth curve that roughly approximates the distribution in Figure 4.18 and shade in the proportion of the area corresponding to the p-value for the sample statistic $D = 0.3$.

(b) Four possible sample statistics are given, along with four possible p-values. Match the statistics with the p-values:

Statistics: 0.1, 0.3, 0.5, 0.7

P-values: 0.012, 0.001, 0.365, 0.085

(c) Interpret the p-value 0.001 in terms of the probability of the results happening by random chance.

(d) Of the four p-values given in part (b), which provides the strongest evidence that a phone call is more effective?

4.57 Influencing Voters: Is There a Difference in Effectiveness between a Phone Call and a Flyer?

Exercise 4.37 describes a study to investigate which method, a recorded phone call or a flyer, is more effective in persuading voters to vote for a particular candidate. Since in this case, the alternative hypothesis is not specified in a particular direction, the hypotheses are $H_0: p_c = p_f$ vs $H_a: p_c \neq p_f$. All else is as in Exercise 4.56, including the randomization distribution shown in Figure 4.18.

(a) Sketch smooth curves that roughly approximate the distribution in Figure 4.18 and shade in the proportion of the area corresponding to the p-value for each of $D = 0.2$ and $D = -0.4$.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Two possible sample statistics are given below, along with several possible p-values. Select the most accurate p-value for each sample statistic.

Statistics: $D = 0.2, D = -0.4$

P-values: 0.008, 0.066, 0.150, 0.392, 0.842

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Of all five p-values given in part (b), which provides the strongest evidence that the methods are not equally effective?

ANSWER ⊕

WORKED SOLUTION ⊕

4.58 Colonoscopy, Anyone?

A colonoscopy is a screening test for colon cancer, recommended as a routine test for adults over age 50. A new study[16]Zauber, et al., "*Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths*," *New England Journal of Medicine*, 2012; 366: 687-696. provides the best evidence yet that this test saves lives. The proportion of people with colon polyps expected to die from colon cancer is 0.01. A sample of 2602 people who had polyps removed during a colonoscopy were followed for 20 years, and 12 of them died from colon cancer. Does this provide evidence that the proportion of people who die from colon cancer after having polyps removed in a colonoscopy is significantly less than the expected proportion (without a colonoscopy) of 0.01?

(a) What are the null and alternative hypotheses?

(b) What is the sample proportion?

(c) Figure 4.19 shows a randomization distribution for this test. Use the fact that there are 1000 dots in the distribution to find the p-value. Explain your reasoning.

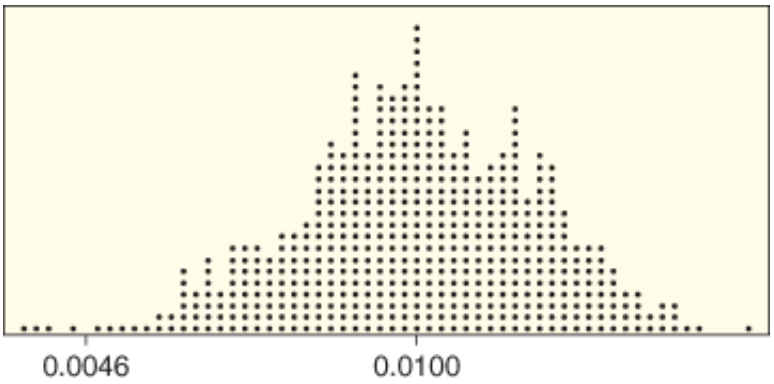

Figure 4.19 ***Randomization distribution for 1000 samples testing effectiveness of colonoscopies***

(d) Does the p-value appear to show significant evidence that colonoscopies save lives?

4.59 Measuring the Impact of Great Teachers

An education study in Tennessee in the 1980s (known as Project Star) randomly assigned 12,000 students to kindergarten classes, with the result that all classes had fairly similar socioeconomic mixes of students.[17]Leonhardt, D., "*The Case for $320,000 Kindergarten Teachers*," *The New York Times*, July 27, 2010, reporting on a study by R. Chetty, a Harvard economist, and his colleagues. The students are now about 30 years old, and the study is ongoing. In each case below, assume that we are conducting a test to compare performance of students taught by outstanding kindergarten teachers with performance of students taught by mediocre kindergarten teachers. What does the quoted information tell us about

whether the p-value is relatively *large* or relatively *small* in a test for the indicated effect?

(a) On the tests at the end of the kindergarten school year, "some classes did far better than others. The differences were too big to be explained by randomness."

ANSWER

WORKED SOLUTION

(b) By junior high and high school, the effect appears to be gone: "Children who had excellent early schooling do little better on tests than similar children who did not."

ANSWER

WORKED SOLUTION

(c) The newest results, reported in July 2010 by economist Chetty, show that the effects seem to re-emerge in adulthood. The students who were in a classroom that made significant gains in kindergarten were significantly "more likely to go to college, … less likely to become single parents, … more likely to be saving for retirement, … Perhaps most striking, they were earning more." (Economists Chetty and Saez estimate that a standout kindergarten teacher is worth about $320,000 a year in increased future earnings of one class of students. If you had an outstanding grade-school teacher, consider sending a thank you note!)

ANSWER

WORKED SOLUTION

4.60 Smiles and Leniency

Data 4.2 describes an experiment to study the effects of smiling on leniency in judging students accused of cheating. The full data are in **Smiles**. In Example 4.2 we consider hypotheses $H_0: \mu_s = \mu_n$ vs $H_a: \mu_s > \mu_n$ to test if the data provide evidence that average leniency score is higher for smiling students (μ_s) than for students with a neutral expression (μ_n). A dotplot for the difference in sample means based on 1000 random assignments of leniency scores from the original sample to smile and neutral groups is shown in Figure 4.20.

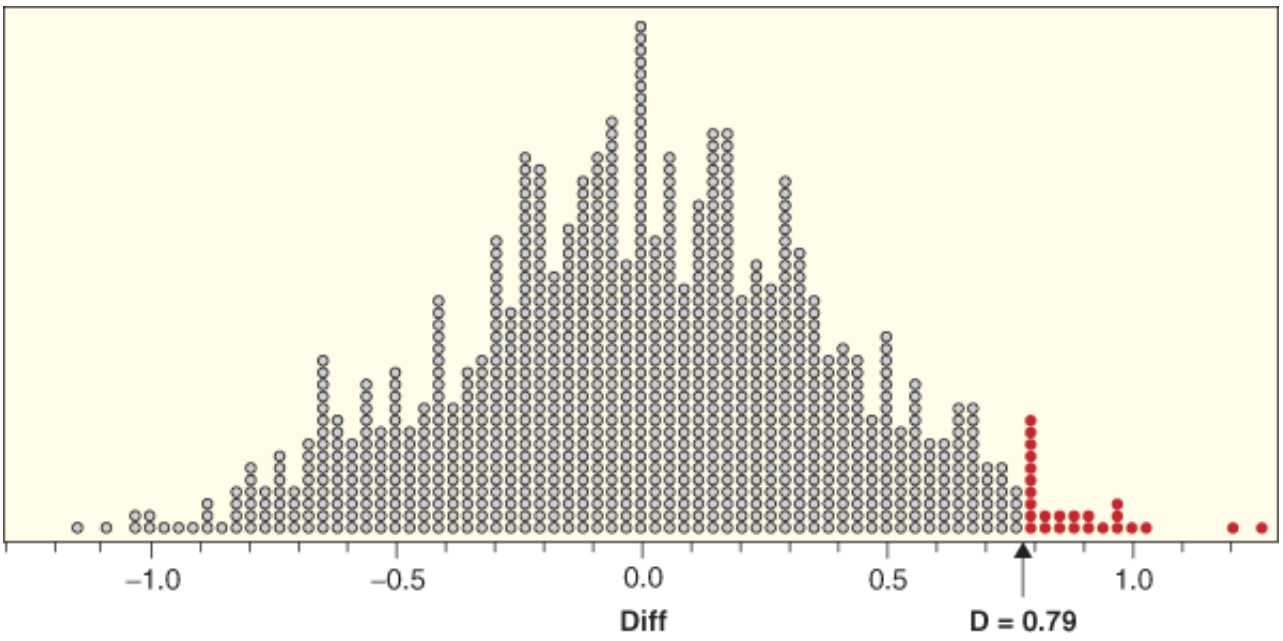

Figure 4.20 ***Randomization distribution for 1000 samples testing H_0:μ_s=μ_n using Smiles data***

(a) The difference in sample means for the original sample is $D = \bar{x}_s - \bar{x}_n = 4.91 - 4.12 = 0.79$ (as shown in Figure 4.20). What is the p-value for the one-tailed test?

Hint: There are 27 dots in the tail beyond 0.79.

(b) In Example 4.3 we consider the test with a two-tailed alternative, H_0:μ_s=μ_n vs H_a:$\mu_s \neq \mu_n$, where we make no assumption in advance on whether smiling helps or discourages leniency. How would the randomization distribution in Figure 4.20 change for this test? How would the p-value change?

4.61 Definition of a P-value

Using the definition of a p-value, explain why the area in the tail of a randomization distribution is used to compute a p-value.

ANSWER ⊕

WORKED SOLUTION ⊕

4.62 Classroom Games

Two professors[18]Dufewenberg, M. and Swarthout, J.T., "*Play to Learn? An Experiment*," from a working paper, at http://econ.arizona.edu/docs/Working_Papers/2009/Econ-WP-09-03.pdf. at the University of Arizona were interested in whether having students actually play a game would help them analyze theoretical properties of the game. The professors performed an experiment in which students played one of two games before coming to a class where both games were discussed. Students were randomly assigned to which of the two games they played, which we'll call Game 1 and Game 2. On a later exam, students were asked to solve problems involving both games, with Question 1 referring to Game 1 and Question 2 referring to Game 2. When comparing the performance of the two groups on the exam question related to Game 1, they suspected that the mean for students who had played Game 1 (μ_1) would be higher than the mean for the other students μ_2, so they considered the hypotheses

$H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 > \mu_2$.

(a) The paper states: *"test of difference in means results in a p-value of 0.7619."* Do you think this provides sufficient evidence to conclude that playing Game 1 helped student performance on that exam question? Explain.

(b) If they were to repeat this experiment 1000 times, and there really is no effect from playing the game, roughly how many times would you expect the results to be as extreme as those observed in the actual study?

(c) When testing a difference in mean performance between the two groups on exam Question 2 related to Game 2 (so now the alternative is reversed to be $H_a: \mu_1 < \mu_2$ where μ_1 and μ_2 represent the mean on Question 2 for the respective groups), they computed a p-value of 0.5490. Explain what it means (in the context of this problem) for both p-values to be greater than 0.5.

4.63 Classroom Games: Is One Question Harder?

Exercise 4.62 describes an experiment involving playing games in class. One concern in the experiment is that the exam question related to Game 1 might be a lot easier or harder than the question for Game 2. In fact, when they compared the mean performance of all students on Question 1 to Question 2 (using a two-tailed test for a difference in means), they report a p-value equal to 0.0012.

(a) If you were to repeat this experiment 1000 times, and there really is no difference in the difficulty of the questions, how often would you expect the means to be as different as observed in the actual study?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Do you think this p-value indicates that there is a difference in the average difficulty of the two questions? Why or why not?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Based on the information given, can you tell which (if either) of the two questions is easier?

ANSWER ⊕

WORKED SOLUTION ⊕

4.64 What Is Your Lucky Number?

Thirty students are asked to choose a random number between 0 and 9, inclusive, to create a dataset of $n = 30$ digits. If the numbers are truly random, we would expect about three 0's, three 1's, three 2's, and so on. If the dataset includes eight 7's, how unusual is that? If we look exclusively at the number of 7's, we expect the proportion of 7's to be 0.1 (since there are 10 possible numbers) and the number of 7's to be 3 in a sample of size 30. We are testing $H_0: p = 0.1$ vs $H_a: p \neq 0.1$, where p is the proportion of 7's. We can generate the randomization distribution by generating 1000 sets of 30 random digits and recording $X =$ the number of 7's in each simulated sample. See Figure 4.21.

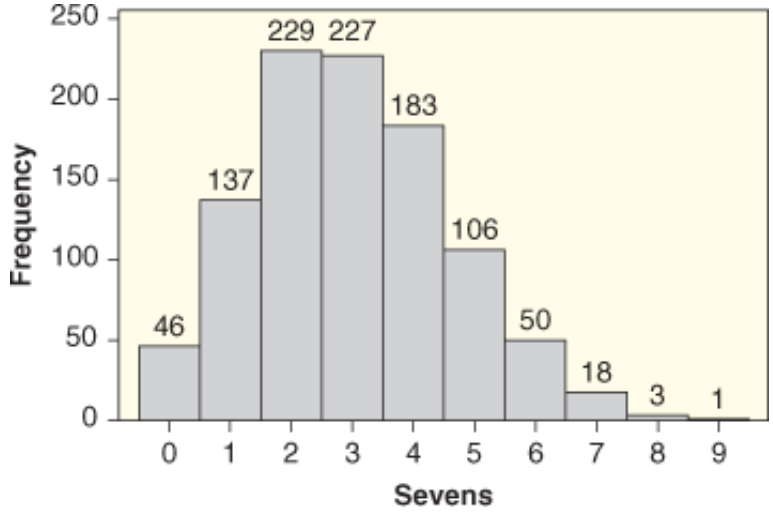

Figure 4.21 ***Randomization distribution for 1000 samples of number of 7's in 30 digits when*** $H_0: p=0.1$

(a) Notice that this randomization distribution is *not* symmetric. This is a two-tailed test, so we need to consider both "tails." How far is $X=8$ from the expected value of 3? What number would be equally far out on the other side? Explain why it is better in this situation to double the observed one-tailed p-value rather than to add the exact values on both sides.

(b) What is the p-value for the observed statistic of $X=8$ sevens when doing the two-tailed test?

(c) The randomization distribution in Figure 4.21 would apply to any digit (not just 7's) if the null hypothesis is $H_0: p=0.1$. Suppose we want to test if students tend to avoid choosing zero when picking a random digit. If we now let p be the proportion of 0's all students choose, the alternative would be $H_a: p<0.1$. What is the smallest p-value we could get using the randomization distribution in Figure 4.21? What would have to happen in the sample of digits from 30 students for this p-value to occur?

4.65 Rolling Dice

You roll a die 60 times and record the sample proportion of fives, and you want to test whether the die is biased to give more fives than a fair die would ordinarily give. To find the p-value for your sample data, you create a randomization distribution of proportions of fives in many simulated samples of size 60 with a fair die.

(a) State the null and alternative hypotheses.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Where will the center of the distribution be? Why?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Give an example of a sample proportion for which the number of 5's obtained is *less* than what you would expect in a fair die.

ANSWER ⊕

WORKED SOLUTION

(d) Will your answer to part (c) lie on the left or the right of the center of the randomization distribution?

ANSWER

WORKED SOLUTION

(e) To find the p-value for your answer to part (c), would you look at the left, right, or both tails?

ANSWER

WORKED SOLUTION

(f) For your answer in part (c), can you say anything about the size of the p-value?

ANSWER

WORKED SOLUTION

4.66 Determining Statistical Significance

How small would a p-value have to be in order for you to consider results statistically significant? Explain. (There is no correct answer! This is just asking for your personal opinion. We'll study this in more detail in the next section.)