Print this page

4.3 DETERMINING STATISTICAL SIGNIFICANCE

Statistical Decisions

In previous sections we have seen how to set up null and alternative hypotheses corresponding to a question of interest, collect sample data, and calculate a p-value using a randomization distribution. We know that a very small p-value means that such a sample is very unlikely to occur by random chance alone and provides strong evidence against the null hypothesis, H_0, in favor of the alternative, H_a. If the evidence is strong enough against the null hypothesis, we can *reject the null hypothesis* in favor of the alternative. On the other hand, if the data are reasonably likely to occur when the null hypothesis is true, we *do not reject the null hypothesis*.

When Making a Formal Decision in a Statistical Test Based on Sample Data:

Reject H_0 if a sample so extreme is unlikely when H_0 is true. This means we have found evidence to support H_a.

Do not reject H_0 if a sample is not too extreme when H_0 is true. This means the test is inconclusive, and either H_0 or H_a may be true.

In either case, be sure to interpret the decision in the context of the question of interest.

Notice that the formal decision is generally made in terms of whether or not we reject the null hypothesis: Reject H_0 or do not reject H_0. If the data are significant, we reject H_0. If the data are not significant, we do not reject H_0. When the sample is not significant, we do not say that we “accept H_0.” Finding a lack of convincing evidence against the null hypothesis should not be confused with finding strong evidence *for* the null hypothesis. In fact, in a hypothesis test, the conclusion is never that we have found evidence for the null hypothesis. The next example illustrates this point.

Example 4.21

Walking Elephants

Suppose that we have a mystery animal named X and consider the hypotheses

$$H_0: \text{ X is an elephant}$$
$$H_a: \text{ X is not an elephant}$$

What conclusion would you draw from each of the following pieces of evidence?

(a) X has four legs.

(b) X walks on two legs.

Solution

(a) It is not at all unusual for an elephant to have four legs, so that evidence would certainly not lead to rejecting this null hypothesis. However, we do not “Accept H_0” and we do not conclude that X *must* be an elephant. Rather we say that the data do not provide significant evidence against H_0 and we cannot determine whether X is or is not an elephant.

(b) While it is not impossible for an elephant to walk on two legs (for example, you might think of trained circus elephants), it is certainly very uncommon. So “walking on two legs” would be sufficient evidence to reject H_0 and conclude X is probably not an elephant.

Michael Edwards/Stone/GettyImages, Inc.

An elephant standing on two legs

If we reject H_0, we have found evidence for the alternative hypothesis. If we fail to reject H_0, we have not found evidence of anything. These are the only two possible outcomes of a formal hypothesis test. Again, we never find evidence *for* the null hypothesis. Furthermore, even if we reject the null hypothesis, we never conclude that our sample statistic is the true value of the parameter—it has simply provided evidence to reject the null hypothesis claim.

Example 4.22

In Data 4.5, a company is testing whether chicken meat from a supplier has an average arsenic level higher than 80 ppb. The hypotheses are

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

where μ is the mean arsenic level in chicken from this supplier.

(a) If the null hypothesis is rejected, what can the company conclude?

(b) If the null hypothesis is not rejected, what can the company conclude?

Solution

(a) If the null hypothesis is rejected, the company has found evidence that the average level of arsenic in chickens from that supplier is greater than 80, and the company should stop buying chicken from that supplier.
(b) If the null hypothesis of $\mu=80$ is not rejected, the company cannot conclude anything significant from this sample about the average level of arsenic in chickens from that supplier. The company would not have sufficient evidence to cancel its relationship with the supplier, since the arsenic level may or may not be greater than 80 ppb.

Practice Problems 4.3H

How Small is Small Enough? The Significance Level

You're probably wondering, *how small does a p-value have to be for us to reject* H_0? If we agree that a p-value of 0.0001 is clearly strong enough evidence to reject H_0 and a p-value of 0.50 provides insufficient evidence to make such a conclusion, there must be some point between 0.0001 and 0.50 where we cross the threshold between statistical significance and random chance. That point, measuring when something becomes rare enough to be called "unusual," might vary a lot from person to person. We should agree in advance on a reasonable cutoff point. Statisticians call this cutoff point the *significance level* of a test and usually denote it with the Greek letter α (alpha). For example, if $\alpha=0.05$ we say we are doing a 5% test and will reject the null hypothesis if the p-value for the sample is smaller than 0.05. Often, shorthand notation such as $P<0.05$ is used to indicate that the p-value is less than 0.05, which means the results are significant at a 5% level.

Significance Level

The **significance level**, α, for a test of hypotheses is a boundary below which we conclude that a p-value shows statistically significant evidence against the null hypothesis.

Common significance levels are $\alpha=0.05$, $\alpha=0.01$, or $\alpha=0.10$.

Given a specific significance level, α, the formal decision in a statistical test, based on comparing the p-value from a sample to α, is very straightforward.

Formal Statistical Decision Based on a Significance Level

Given a significance level α and the p-value from a sample, we:

Reject H_0 if the $\text{p-value} < \alpha$.

Do not reject H_0 if the $\text{p-value} \geq \alpha$.

Example 4.23

Dogs and Owners: The Conclusion!

In Section 4.2 we construct a randomization distribution to see how unusual it is to see 16 or more correct matches among 25 sets of dog-owner pairs under a null hypothesis of no dog-owner resemblance ($H_0 : p = 0.5$ vs $H_a : p > 0.5$). The estimated p-value from the randomization distribution is 0.1145. Using a 5% significance level, what decision do we make? Does the decision change if we use $\alpha = 0.10$ or $\alpha = 0.01$?

Solution ▶

Since the p-value of 0.1145 is greater than the significance level of $\alpha = 0.05$, we do not reject H_0 at a 5% level. This sample does not provide convincing evidence that dogs tend to resemble their owners. Since the p-value is also more than both 0.10 and 0.01, the decision and interpretation are the same for those significance levels.

Notice that we always follow up the formal decision (reject H_0 or do not reject H_0) with a statement that interprets the decision in the context of the data situation. This is an important step in addressing any statistical question that involves real data.

Example 4.24

In Example 4.15 we consider other possible outcomes for number of matches in the dog-owner experiment. For each of these situations, what is the decision when doing a 5% test of $H_0 : p = 0.5$ vs $H_a : p > 0.5$?

(a) 19 correct matches with $\text{p-value} = 0.0075$

(b) 15 correct matches with $\text{p-value} = 0.215$

Solution ▶

(a) With 19 correct matches out of 25 trials, the p-value of 0.0075 is less than $\alpha = 0.05$. If we saw 19 correct matches out of 25, the results would be statistically significant, providing sufficient evidence to reject H_0, and conclude that dogs tend to resemble their owners.

(b) Finding only 15 correct matches in the 25 trials would be even less significant than the 16 matches in the original study. The p-value of 0.215 is not even close to being less than $\alpha = 0.05$ so we do not reject H_0. If we saw only 15 correct matches, we would have insufficient evidence to say that a dog-owner resemblance exists.

We can visualize the significance level α as a portion of the total area in a randomization distribution. Figure 4.22 shows the randomization distribution for dog-owner matches. Notice that if we use $\alpha=0.05$, we decide to reject H_0 for all sample values in the 5% of area in the upper tail (17 or more matches out of 25 trials). If instead we use $\alpha=0.01$, we only reject H_0 for sample values in the upper 1% of area (about 19 or more matches). The samples in the extreme tail that lead to rejecting H_0 at a significance level of α take up a proportion of area equal to α. Since this is a one-tailed test, the whole area is in one tail. For a two-tailed test, the area of α would be split evenly between the two tails.

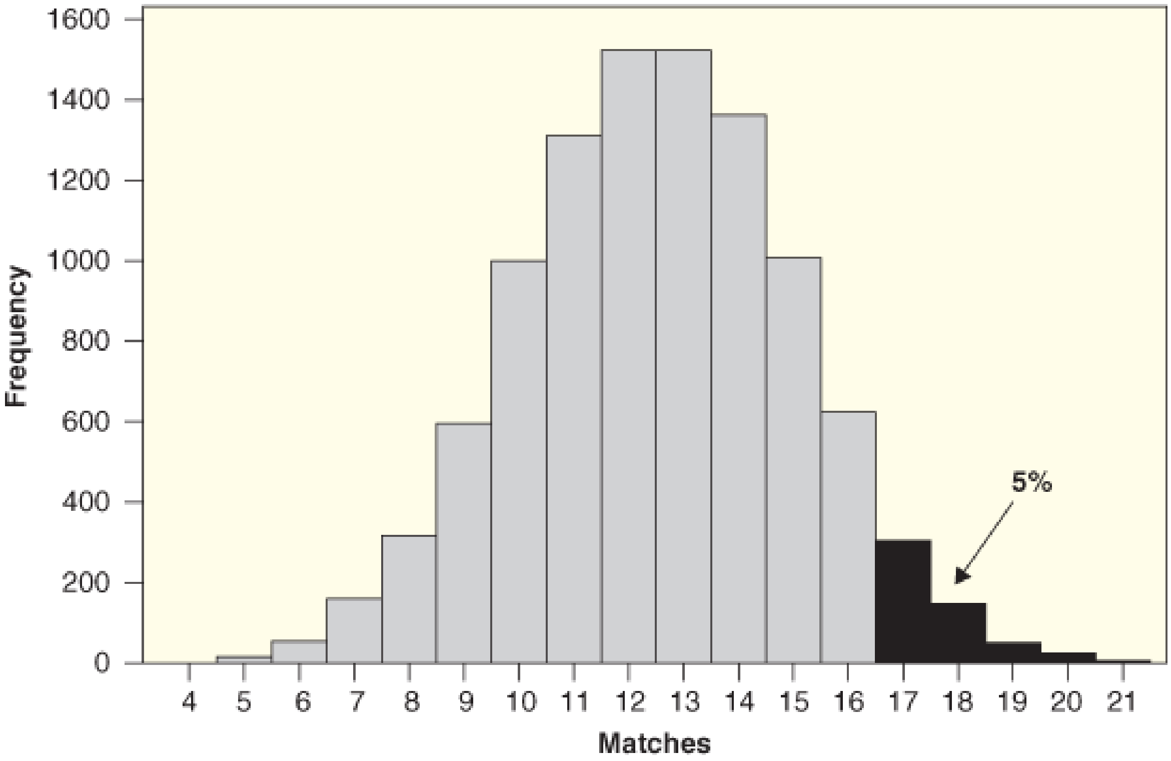

Figure 4.22 ***5% significance level as proportion of total area in a randomization distribution***

How do we decide on a significance level in the first place? Many research areas have standard accepted values of α for their fields. In other cases, researchers are allowed to choose α based on what makes the most sense for their particular situation, although it is important that α is chosen *before* the experiment takes place. In still other cases, researchers may report results of a test based on which of a set of commonly used α values give significant results. For example, a claim that “results are significant at a 10% level” may be viewed as somewhat less strong than “results are significant at a 1% level.” However, without knowing the p-value, we can never know for sure which gives stronger evidence. In all cases, providing the p-value itself is preferred since then a reader can more accurately assess the strength of the evidence.

Practice Problems 4.3I

Type I and Type II Errors

Practice Problems 4.3J

Formal hypothesis testing produces one of two possible generic decisions (ignoring context): “reject H_0” or “do not reject H_0.” In reality, the claims about the population described by H_0 and H_a might be either true or false. Perhaps dogs really do tend to resemble owners (H_a), or maybe this phenomenon doesn't exist at all (H_0) and people are just guessing. When we make a formal decision to “reject H_0,” we generally are accepting some risk that H_0 might actually be true. For example, we may have been unlucky and stumbled upon one of those “1 in a 1000” samples that are very rare to see when H_0 holds but still are not impossible. This is an example of what we call a *Type I error*: rejecting a true H_0. The other possible error to make in a statistical test is to fail to reject H_0 when it is false and the alternative H_a is actually true. We call this a *Type II error*: failing to reject a false H_0. See Table 4.7.

Table 4.7 ***Possible errors in a formal statistical decision***

	Reject H_0	**Do not reject H_0**
H_0 is true	Type I error	No error
H_0 is false	No error	Type II error

Example 4.25

Describe the consequences of making Type I and Type II errors in each case.

(a) In the dogs-owners experiment where we test $H_0{:}p=0.5$ vs $H_a{:}p>0.5$

(b) In Example 4.21 where we have a mystery animal named X and test H_0: X is an elephant vs H_a: X is not an elephant

Solution

(a) A Type I error is to reject a true H_0. In the dog-owner study, a Type I error is to conclude that dogs do resemble their owners when actually there is no relationship between appearances of dogs and owners.

A Type II error is to fail to reject a false H_0. In this case, a Type II error means the test based on our sample data does not convince us that dogs look like their owners when dogs actually do tend to resemble their owners.

(b) If we see evidence (perhaps that X walks on two legs) that is so rare we conclude that X is not an elephant and it turns out that X is an elephant (perhaps trained in a circus), we have made a Type I error.

For a Type II error, we might find evidence (perhaps having four legs) that is not unusual for an elephant, so we do not reject H_0 and then discover that X is actually a giraffe.

If our results are significant and we reject H_0, there is usually no way of knowing whether we are correct or whether we have made a Type I error. If our results are insignificant and we fail to reject H_0, we could be correct or we could have made a Type II error. While we can never rule out these possibilities entirely, we do have some control over the chance of making these errors.

While we wish to avoid both types of errors, in practice we have to accept some trade-off between them. We could reduce the chance of making a Type I error by making it very hard to reject H_0, but then we would probably make Type II errors more often. On the other hand, if we routinely reject H_0, we would rarely be guilty of a Type II error, but we would end up rejecting too many H_0's that were actually true. Remember that at the outset we set up our hypotheses with H_0 representing the "status quo" and only reject it (in favor of H_a) when there is *convincing* evidence against it. For this reason, we are generally more concerned with keeping the chances of making a Type I error relatively small, even if it sometimes means we accept a larger chance of making a Type II error.

How can we reduce the chance of making a Type I error? In other words, how can we make it harder to reject H_0 when it is actually true? One key is to think about the significance level, α. The randomization distribution represents what we expect to see if the null hypothesis is true, and as we see in Figure 4.22, the proportion of samples that lead to rejecting H_0 is equal to α. If we make α smaller, fewer samples would be that extreme, meaning we would reject H_0 less often. The smaller we make the significance level α, the less likely we are to make a Type I error when H_0 is true.

Choosing a Significance Level

The significance level, α, represents the tolerable probability of making a Type I error.

If the consequences of a Type I error are severe (for example, approving a new drug that is potentially dangerous), we might use a very small α (perhaps even $\alpha=0.005$). However, remember that using a very small α also increases the likelihood that we make a Type II error when the alternative H_a is true. For this reason we usually use the common significance levels of 5%, 10%, or 1%.

Example 4.26

Analogy to Law

It is often helpful to think of significance tests as similar to cases in a court of law. For each italicized word or phrase below, give the analogy in a statistical test.

(a) A person is *innocent* until proven *guilty*.

(b) The *evidence* provided must indicate the suspect's guilt beyond a *reasonable doubt*.

(c) There are two types of errors a jury can make:

- *Releasing a guilty person*
- *Convicting an innocent person*

Solution

(a) “Innocent” is the null hypothesis, H_0 (the status quo that we assume to be the situation until we see convincing evidence to the contrary). “Guilty” represents the alternative hypothesis, H_a (the claim that instigates the trial).

(b) The “evidence” is the data from the sample and its p-value. The “reasonable doubt” corresponds to the significance level, α. We reject the claim of innocence (H_0) and determine the suspect is guilty (H_a) when the evidence (p-value) is very unlikely (less than α) to occur if the suspect is really innocent.

(c) “Releasing a guilty person” corresponds to a Type II error, since we fail to find evidence to reject a false H_0. “Convicting an innocent person” corresponds to a Type I error, since we (incorrectly) find evidence in the data to reject a true H_0. As in our legal system, we are usually more worried about a Type I error (convicting an innocent person) than about a Type II error (releasing a guilty person). Also as in our legal system, there is a trade-off between the two kinds of errors when we test hypotheses.

In medical terms we often think of a Type I error as a “false positive”—a test that indicates a patient has an illness when actually none is present, and a Type II error as a “false negative”—a test that fails to detect an actual illness.

Practice Problems 4.3K

Less Formal Statistical Decisions

Practice Problems 4.3L

Classical hypothesis testing requires a formal decision to “Reject H_0” or “Do not reject H_0” depending on whether or not the p-value is less than the desired significance level. The general idea for a 5% test is illustrated in Figure 4.23. If the p-value is less than 5%, we reject the null in favor of the alternative; otherwise we find the evidence insufficient to discard the null hypothesis.

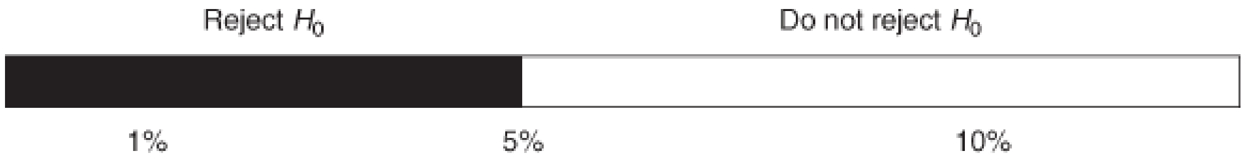

Figure 4.23 ***Formal decision rule for p-values with a 5% significance test***

In Data 4.2 we describe an experiment to see if smiles have an effect on leniency in a disciplinary action. Participants viewed a photograph of the “suspect” who either was smiling or had a neutral expression. There were 34 participants in each group who made disciplinary decisions that were interpreted to give a leniency score (on a 10-point scale) for each case. The data are stored in **Smiles** and we are interested in whether smiling makes a difference (in either direction) on the leniency scores. Letting μ_s and μ_n be the mean leniency scores for smiling and neutral suspects in general, we test the hypotheses

$$H_0\text{: } \mu_s = \mu_n$$
$$H_a\text{: } \mu_s \neq \mu_n$$

For the sample data, we find $\bar{x}_s = 4.91$ and $\bar{x}_n = 4.12$, so the difference in sample means is $D=4.91-4.12=0.79$. The randomization distribution in Figure 4.24 shows the results of the differences in sample means for 1000 simulations where the 34 “smile” and “neutral” labels were randomly assigned to the 68 leniency scores. There are 23 values in the upper tail of the 1000 simulations that are larger than the original sample difference of $D=0.79$.

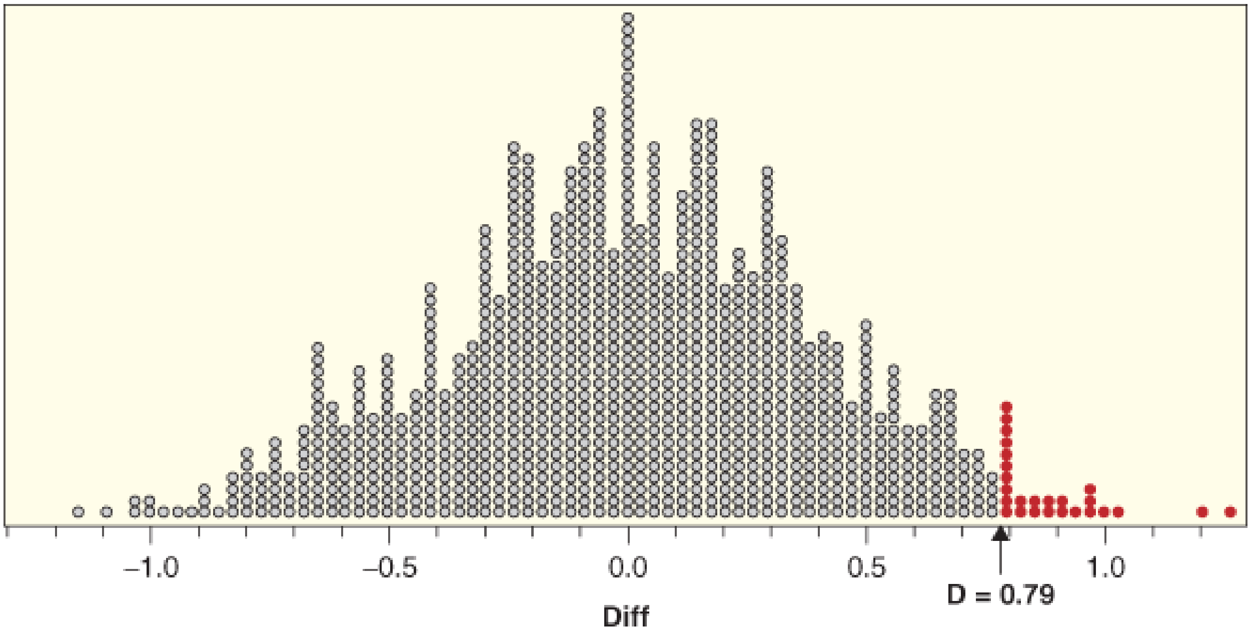

Figure 4.24 ***Randomization distribution of differences in leniency means, $D = \bar{x}_s - \bar{x}_n$***

Example 4.27

(a) Use the randomization distribution and the information above to estimate a p-value in the smiling and leniency experiment. Use a 5% significance level to reach a decision.

(b) If we change the score for just one of the participants in the smiling and leniency experiment by a single point, either less lenient for someone in the smile group or more lenient for someone in the neutral group, the difference in means becomes $D=0.76$ and four *new* points in the randomization distribution would exceed this difference. Repeat part (a) for this value of D.

Solution

(a) Since we are doing a two-tailed test with 23 out of 1000 simulated differences more extreme than $D=0.79$ the estimated p-value is $2 \cdot 23/1000=0.046$. This p-value is less than the significance level of $\alpha=0.05$ so we reject H_0 and conclude that the difference in mean leniency scores is more than we expect to see by random chance alone. Based on these data we conclude that smiling makes a difference and we expect more leniency, on average, to be awarded to smiling suspects. If you go before a disciplinary panel, you should smile!

(b) The randomization distribution in Figure 4.24 has $23+4=27$ cases above $D=0.76$, which produces a p-value of $2 \cdot 27/1000 = 0.054$. This p-value is not less than 5%, so we do not reject H_0 and thus conclude that we do not have sufficient evidence to show that smiling makes a difference in the amount of leniency. If you go before a disciplinary panel, it may not matter whether you smile or maintain a neutral expression.

Notice in Example 4.27 that changing just one person's score by a single point dramatically changes the conclusion of the test. One of the drawbacks of the classical approach to hypothesis testing is that it forces us to make very black-white decisions. We either reject the null or don't reject it. In some situations we might feel more comfortable with a less prescriptive decision. We might be “pretty sure” that H_0 should be rejected or find some, but not entirely convincing, evidence against it. For this reason we sometimes interpret a p-value less formally by merely indicating the strength of evidence it shows against the null hypothesis. For example, the p-values of 0.046 and 0.054 in Example 4.27 might both be interpreted as showing moderate but not very strong evidence that smiling helps increase leniency.

Figure 4.25 gives a schematic representation of a less formal way to interpret p-values as strength of evidence against a null hypothesis. Contrast this with the formal decision rule shown in Figure 4.23. Which way is right? They both have their merits. As we continue studying significance testing, keep both approaches in mind so that you can make a concrete decision for a given significance level but also interpret any p-value as a measure of strength of evidence.

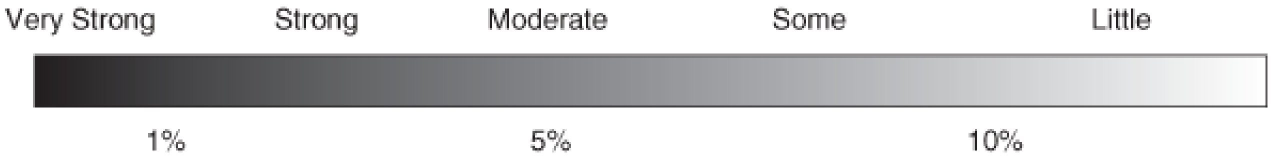

Figure 4.25 *Informal strengths of evidence against H_0*

Practice Problems 4.3M

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Make a formal decision in a hypothesis test by comparing a p-value to a given significance level
- State the conclusion to a hypothesis test in context
- Interpret Type I and Type II errors in hypothesis tests
- Recognize a significance level as measuring the tolerable chance of making a Type I error
- Make a less formal decision that reflects the strength of evidence in a p-value

Exercises for Section 4.3

SKILL BUILDER 1

Exercises 4.67 to 4.70 give a p-value. State the conclusion of the test based on this p-value in terms of "Reject H_0" or "Do not reject H_0" if we use a 5% significance level.

4.67 p-value$=0.0007$

ANSWER ⊕

WORKED SOLUTION ⊕

4.68 p-value$=0.0320$

4.69 p-value$=0.2531$

ANSWER ⊕

WORKED SOLUTION ⊕

4.70 p-value$=0.1145$

SKILL BUILDER 2

In Exercises 4.71 to 4.74, using the p-value given, are the results significant at a 10% level? At a 5% level? At a 1% level?

4.71 p-value$=0.0320$

ANSWER ⊕

WORKED SOLUTION ⊕

4.72 p-value$=0.2800$

4.73 p-value$=0.008$

ANSWER ⊕

WORKED SOLUTION ⊕

4.74 p-value$=0.0621$

SKILL BUILDER 3

In Exercises 4.75 and 4.76, match the four p-values with the appropriate conclusion:

(a) The evidence against the null hypothesis is significant, but only at the 10% level.

(b) The evidence against the null and in favor of the alternative is very strong.

(c) There is not enough evidence to reject the null hypothesis, even at the 10% level.

(d) The result is significant at a 5% level but not at a 1% level.

4.75

(a) 0.0875

ANSWER ⊕

WORKED SOLUTION ⊕

(b) 0.5457

ANSWER ⊕

WORKED SOLUTION ⊕

(c) 0.0217

ANSWER ⊕

WORKED SOLUTION ⊕

(d) 0.00003

ANSWER ⊕

WORKED SOLUTION ⊕

4.76

(a) 0.00008

(b) 0.0571

(c) 0.0368

(d) 0.1753

4.77 Significance Levels

Test A is described in a journal article as being significant with "$P<.01$"; Test B in the same article is described as being significant with "$P<.10$." Using only this information, which test would you suspect provides stronger evidence for its alternative hypothesis?

ANSWER ⊕

WORKED SOLUTION ⊕

4.78 Interpreting a P-value

In each case, indicate whether the statement is a proper interpretation of what a p-value measures.

(a) The probability the null hypothesis H_0 is true.

(b) The probability that the alternative hypothesis H_a is true.

(c) The probability of seeing data as extreme as the sample, when the null hypothesis H_0 is true.

(d) The probability of making a Type I error if the null hypothesis H_0 is true.

(e) The probability of making a Type II error if the alternative hypothesis H_a is true.

4.79 Divorce Opinions and Gender

In Data 4.4, we introduce the results of a May 2010 Gallup poll of 1029 U.S. adults. When asked if they view divorce as "morally acceptable," 71% of the men and 67% of the women in the sample responded yes. In the test for a difference in proportions, a randomization distribution gives a p-value of 0.165. Does this indicate a significant difference between men and women in how they view divorce?

ANSWER ⊕

WORKED SOLUTION ⊕

4.80 Red Wine and Weight Loss

Resveratrol, an ingredient in red wine and grapes, has been shown to promote weight loss in rodents. A recent study[19]BioMed Central. "*Lemurs Lose Weight with 'Life-Extending' Supplement Resveratrol*," *ScienceDaily*, June 22, 2010. investigates whether the same phenomenon holds true in primates. The grey mouse lemur, a primate, demonstrates seasonal spontaneous obesity in preparation for winter, doubling its body mass. A sample of six lemurs had their resting metabolic rate, body mass gain, food intake, and locomotor activity measured for one week prior to resveratrol supplementation (to serve as a baseline) and then the four indicators were measured again after treatment with a resveratrol supplement for four weeks. Some p-values for tests comparing the mean differences in these variables (before vs after treatment) are given below. In parts (a) to (d), state the conclusion of the test using a 5% significance level, and interpret the conclusion in context.

(a) In a test to see if mean resting metabolic rate is higher after treatment, $p=0.013$.

(b) In a test to see if mean body mass gain is lower after treatment, $p=0.007$

(c) In a test to see if mean food intake is affected by the treatment, $p=0.035$.

(d) In a test to see if mean locomotor activity is affected by the treatment, $p=0.980$

(e) In which test is the strongest evidence found? The weakest?

(f) How do your answers to parts (a) to (d) change if the researchers make their conclusions using a stricter 1% significance level?

(g) For each p-value, give an informal conclusion in the context of the problem describing the level of evidence for the result.

(h) The sample only included six lemurs. Do you think that we can generalize to the population of all lemurs that body mass gain is lower on average after four weeks of a resveratrol supplement? Why or why not?

4.81 Euchre

Exercise 4.40 describes an ongoing game of Euchre, in which the game continues until one of the two teams is deemed to be *statistically significantly* better than the other team. Which significance level, 5% or 1%, will make the game last longer?

ANSWER ⊕

WORKED SOLUTION ⊕

4.82 Sleep or Caffeine for Memory?

The consumption of caffeine to benefit alertness is a common activity practiced by 90% of adults in North America. Often caffeine is used in order to replace the need for sleep. One recent study[20]Mednick, S., Cai, D., Kanady, J., and Drummond, S., "*Comparing the Benefits of Caffeine, Naps and Placebo on Verbal, Motor and Perceptual Memory*," *Behavioural Brain Research*, 2008; 193: 79-86. compares students' ability to recall memorized information after either the consumption of caffeine or a brief sleep. A random sample of 35 adults (between the ages of 18 and 39) were randomly divided into three groups and verbally given a list of 24 words to memorize. During a break, one of the groups takes a nap for an hour and a half, another group is kept awake and then given a caffeine pill an hour prior to testing, and the third group is given a placebo. The response variable of interest is the number of words participants are able to recall following the break. The summary statistics for the three groups are in Table 4.8. We are interested in testing whether there is evidence of a difference in average recall ability between any two of the treatments. Thus we have three possible tests between different pairs of groups: Sleep vs Caffeine, Sleep vs Placebo, and Caffeine vs Placebo.

Table 4.8 ***Effect of sleep and caffeine on memory***

			Standard
Group	**Sample Size**	**Mean**	**Deviation**
Sleep	12	15.25	3.3
Caffeine	12	12.25	3.5
Placebo	11	13.70	3.0

(a) In the test comparing the sleep group to the caffeine group, the p-value is 0.003. What is the conclusion of the test? In the sample, which group had better recall ability? According to the test results, do you think sleep is really better than caffeine for recall ability?

(b) In the test comparing the sleep group to the placebo group, the p-value is 0.06. What is the conclusion of the test using a 5% significance level? Using a 10% significance level? How strong is the evidence of a difference in mean recall ability between these two treatments?

(c) In the test comparing the caffeine group to the placebo group, the p-value is 0.22. What is the conclusion of the test? In the sample, which group had better recall ability? According to the test results, would we be justified in concluding that caffeine impairs recall ability?

(d) According to this study, what should you do before an exam that asks you to recall information?

4.83 **Price and Marketing**

How influenced are consumers by price and marketing? If something costs more, do our expectations lead us to believe it is better? Because expectations play such a large role in reality, can a product that costs more (but is in reality identical) actually be more effective? Baba Shiv, a neuroeconomist at Stanford, conducted a study[21]Shiv, B., Carmon, Z. and Ariely D., "*Placebo Effects of Marketing Actions: Consumers May Get What They Pay For*," *Journal of Marketing Research*, 2005; 42: 383-393. involving 204 undergraduates. In the study, all students consumed a popular energy drink which claims on its packaging to increase mental acuity. The students were then asked to solve a series of puzzles. The students were charged either regular price ($1.89) for the drink or a discount price ($0.89). The students receiving the discount price were told that they were able to buy the drink at a discount since the drinks had been purchased in bulk. The authors of the study describe the results: "the number of puzzles solved was lower in the reduced-price condition ($M = 4.2$) than in the regular-price condition ($M = 5.8$) ... $p <$

0.0001.”

(a) What can you conclude from the study? How strong is the evidence for the conclusion?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) These results have been replicated in many similar studies. As Jonah Lehrer tells us: “According to Shiv, a kind of placebo effect is at work. Since we expect cheaper goods to be less effective, they generally are less effective, even if they are identical to more expensive products. This is why brand-name aspirin works better than generic aspirin and why Coke tastes better than cheaper colas, even if most consumers can't tell the difference in blind taste tests.”[22]Lehrer, J., “*Grape Expectations: What Wine Can Tell Us About the Nature of Reality*,” *The Boston Globe*, February 28, 2008. Discuss the implications of this research in marketing and pricing.

ANSWER ⊕

WORKED SOLUTION ⊕

4.84 Mercury Levels in Fish

Figure 4.26 shows a scatterplot of the acidity (pH) for a sample of $n=53$ Florida lakes vs the average mercury level (ppm) found in fish taken from each lake. The full dataset is introduced in Data 2.4 and is available in **FloridaLakes**. There appears to be a negative trend in the scatterplot, and we wish to test whether there is significant evidence of a negative association between pH and mercury levels.

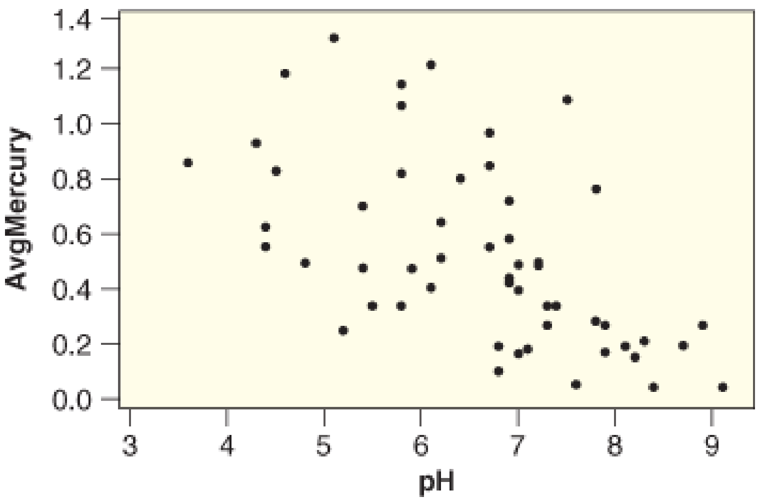

Figure 4.26 ***Water pH vs mercury levels of fish in Florida lakes***

(a) What are the null and alternative hypotheses?

(b) For these data, a statistical software package produces the following output:

$$r=-0.575 \quad p\text{-}value=0.000017$$

Use the p-value to give the conclusion of the test. Include an assessment of the strength of the evidence and state your result in terms of rejecting or failing to reject H_0 *and* in terms of pH and mercury.

(c) Is this convincing evidence that low pH *causes* the average mercury level in fish to increase? Why or why not?

4.85 Penalty Shots in Soccer

A recent article noted that it may be possible to accurately predict which way a penalty-shot kicker in soccer will direct his shot.[23]"*A Penalty Kicker's Cues*," *The Week*, July 16, 2010, p. 21. The study finds that certain types of body language by a soccer player—called "tells"—can be accurately read to predict whether the ball will go left or right. For a given body movement leading up to the kick, the question is whether there is strong evidence that the proportion of kicks that go right is significantly different from one-half.

(a) What are the null and alternative hypotheses in this situation?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) If sample results for one type of body movement give a p-value of 0.3184, what is the conclusion of the test? Should a goalie learn to distinguish this movement?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) If sample results for a different type of body movement give a p-value of 0.0006, what is the conclusion of the test? Should a goalie learn to distinguish this movement?

ANSWER ⊕

WORKED SOLUTION ⊕

4.86 Radiation from Cell Phones and Brain Activity

Does heavy cell phone use affect brain activity? There is some concern about possible negative effects of radiofrequency signals delivered to the brain. In a randomized matched-pairs study,[24]Volkow, et al., "*Effects of Cell Phone Radiofrequency Signal Exposure on Brain Glucose Metabolism*," *Journal of the American Medical Association*, 2011; 305(8): 808-813. 47 healthy participants had cell phones placed on the left and right ears. Brain glucose metabolism (a measure of brain activity) was measured for all participants under two conditions: with one cell phone turned on for 50 minutes (the "on" condition) and with both cell phones off (the "off" condition). The amplitude of radiofrequency waves emitted by the cell phones during the "on" condition was also measured.

(a) Is this an experiment or an observational study? Explain what it means to say that this was a "matched-pairs" study.

(b) How was randomization likely used in the study? Why did participants have cell phones on their ears during the "off" condition?

(c) The investigators were interested in seeing whether average brain glucose metabolism was different based on whether the cell phones were turned on or off. State the null and alternative hypotheses for this test.

(d) The p-value for the test in part (c) is 0.004. State the conclusion of this test in context.

(e) The investigators were also interested in seeing if brain glucose metabolism was significantly correlated with the amplitude of the radiofrequency waves. What graph might we use to visualize this relationship?

(f) State the null and alternative hypotheses for the test in part (e).

(g) The article states that the p-value for the test in part (e) satisfies $p<0.001$. State the conclusion of this test in context.

4.87 ADHD and Pesticides

In Exercise 4.16, we describe an observational study investigating a possible relationship between exposure to organophosphate pesticides as measured in urinary metabolites (DAP) and diagnosis of ADHD (attention-deficit/hyperactivity disorder). In reporting the results of this study, the authors[25]Bouchard, M., Bellinger, D., Wright, R., and Weisskopf, M., "*Attention-Deficit/Hyperactivity Disorder and Urinary Metabolites of Organophosphate Pesticides*," *Pediatrics*, 2010; 125: e1270-e1277. make the following statements:

- "The threshold for statistical significance was set at $P<.05$."
- "The odds of meeting the … criteria for ADHD increased with the urinary concentrations of total DAP metabolites."
- "The association was statistically significant."

(a) What can we conclude about the p-value obtained in analyzing the data?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Based on these statements, can we distinguish whether the evidence of association is very strong vs moderately strong? Why or why not?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Can we conclude that exposure to pesticides is related to the likelihood of an ADHD diagnosis?

ANSWER ⊕

WORKED SOLUTION ⊕

(d) Can we conclude that exposure to pesticides *causes* more cases of ADHD? Why or why not?

ANSWER ⊕

WORKED SOLUTION ⊕

4.88 Diabetes and Pollution

Diabetes tends to be more prevalent in urban populations, but why this is so is not fully understood. A recent study[26]Data recreated from information in Sun et al., "*Ambient Air Pollution Exaggerates Adipose Inflammation and Insulin Resistance in a Mouse Model of Diet-Induced Obesity*," *Journal of the American Heart Association*, 2009; 119(4): 538-546. on mice was designed to investigate the link between diabetes and air pollution. The study involved 28 mice, with 14 randomly selected to have filtered air pumped into their cage while the other 14 breathed particulate matter that simulated common air pollution. The response variable is the amount of insulin resistance each mouse had after 24 weeks. Higher insulin resistance indicates a greater risk for developing diabetes.

(a) Is this an observational study or randomized experiment?

(b) If we are interested in whether there is a difference in mean insulin resistance between the two groups, what are the null and alternative hypotheses?

(c) The difference in sample means from the original sample is $D=\bar{x}_{FA}-\bar{x}_{PM}=1.8-6.2=-4.4$. Figure 4.27 shows 1000 random assignments of insulin-resistant scores from the original sample to each of the two groups. Is it likely we will reject the null hypothesis?

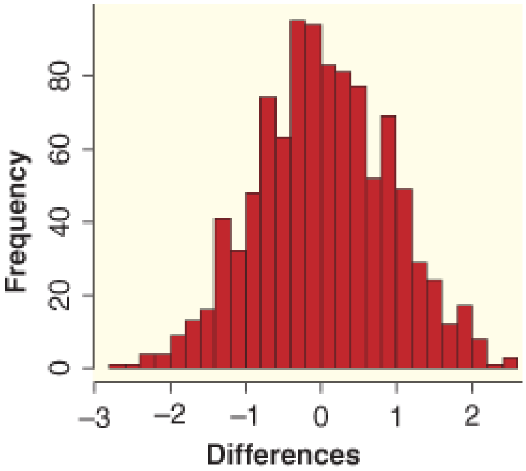

Figure 4.27 ***Randomization distribution for 1000 simulations with $H_0:\mu_{FA}=\mu_{PM}$***

(d) What is the p-value?

(e) State the conclusion of the test in context.

4.89 Beer and Mosquitoes

Does consuming beer attract mosquitoes? Exercise 4.17 discusses an experiment done in Africa testing possible ways to reduce the spread of malaria by mosquitoes. In the experiment, 43 volunteers were randomly assigned to consume either a liter of beer or a liter of water, and the attractiveness to mosquitoes of each volunteer was measured. The experiment was designed to test whether beer consumption increases mosquito attraction. The report[27]Lefvre, T., et al., "*Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes*." *PLoS ONE*, 2010; 5(3): e9546. states that "Beer consumption, as opposed to water consumption, significantly increased the activation … of *An. gambiae* [mosquitoes] … ($P<0.001$)."

(a) Is this convincing evidence that consuming beer is associated with higher mosquito attraction? Why or why not?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) How strong is the evidence for the result? Explain.

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Based on these results, is it reasonable to conclude that consuming beer *causes* an increase in mosquito attraction? Why or why not?

ANSWER ⊕

WORKED SOLUTION ⊕

4.90 Exercise and the Brain

Exercise 4.19 describes a study investigating the effects of exercise on cognitive function.[28]Gobeske, K., et al., "*BMP Signaling Mediates Effects of Exercise on Hippocampal Neurogenesis and Cognition in Mice*," *PLoS One*, 2009; 4(10): e7506. Separate groups of mice were exposed to running wheels for 0, 2, 4, 7, or 10 days. Cognitive function was measured by Y-maze performance. The study was testing whether exercise improves brain function, whether exercise reduces levels of BMP (a protein which makes the brain slower and less nimble), and whether exercise increases the levels of noggin (which improves the brain's ability). For each of the results quoted in parts a, b, and c, interpret the information about the p-value in terms of evidence for the effect.

(a) "Exercise improved Y-maze performance in most mice by the 7th day of exposure, with further increases after 10 days for all mice tested ($p<.01$)."

(b) "After only two days of running, BMP … was reduced … and it remained decreased for all subsequent time-points ($p<.01$)."

(c) "Levels of noggin … did not change until 4 days, but had increased 1.5-fold by 7-10 days of exercise ($p<.001$)."

(d) Which of the tests appears to show the strongest statistical effect?

(e) What (if anything) can we conclude about the effects of exercise on mice?

4.91 Translating Information to Other Significance Levels

Suppose in a two-tailed test of $H_0{:}\rho=0$ vs $H_a{:}\rho\neq0$, we reject H_0 when using a 5% significance level. Which of the conclusions below (if any) would also definitely be valid for the same data? Explain your reasoning in each case.

(a) Reject $H_0{:}\rho=0$ in favor of $H_a{:}\rho\neq0$ at a 1% significance level.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Reject $H_0{:}\rho=0$ in favor of $H_a{:}\rho\neq0$ at a 10% significance level.

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Reject $H_0{:}\rho=0$ in favor of the one-tail alternative, $H_a{:}\rho>0$, at a 5% significance level, assuming the sample correlation is positive.

ANSWER ⊕

WORKED SOLUTION ⊕

4.92 Flaxseed and Omega-3

Exercise 4.29 describes a company that advertises that its milled flaxseed contains, on average, at least 3800 mg of ALNA, the primary omega-3 fatty acid in flaxseed, per tablespoon. In each case below, which of the standard significance levels, 1% or 5% or 10%, makes the most sense for that situation?

(a) The company plans to conduct a test just to double-check that its claim is correct. The company is eager to find evidence that the average amount per tablespoon is greater than 3800 (their alternative hypothesis) and is not really worried about making a mistake. The test is internal to the company and there are unlikely to be any real consequences either way.

(b) Suppose, instead, that a consumer organization plans to conduct a test to see if there is evidence *against* the claim that the product contains at least 3800 mg per tablespoon. If the organization finds evidence that the advertising claim is false, it will file a lawsuit against the flaxseed company. The organization wants to be very sure that the evidence is strong, since there could be very serious consequences if the company is sued incorrectly.

SELECTING A SIGNIFICANCE LEVEL

For each situation described in Exercises 4.93 to 4.98, indicate whether it makes more sense to use a relatively large significance level (such as $\alpha=0.10$) or a relatively small significance level (such as $\alpha=0.01$).

4.93 Testing a new drug with potentially dangerous side effects to see if it is significantly better than the drug currently in use. If it is found to be more effective, it will be prescribed to millions of people.

ANSWER ⊕

WORKED SOLUTION ⊕

4.94 Using your statistics class as a sample to see if there is evidence of a difference between male and female students in how many hours are spent watching television per week.

4.95 Using a sample of 10 games each to see if your average score at Wii bowling is significantly more than your friend's average score.

ANSWER ⊕

WORKED SOLUTION ⊕

4.96 Testing to see if a well-known company is lying in its advertising. If there is evidence that the company is lying, the Federal Trade Commission will file a lawsuit against them.

4.97 Testing to see whether taking a vitamin supplement each day has significant health benefits. There are no (known) harmful side effects of the supplement.

ANSWER ⊕

WORKED SOLUTION ⊕

4.98 A pharmaceutical company is testing to see whether its new drug is significantly better than the existing drug on the market. It is more expensive than the existing drug. Which significance level would the company prefer? Which significance level would the consumer prefer?

TYPE I AND TYPE II ERRORS

For each situation given in Exercises 4.99 to 4.103, describe what it means in that context to make a Type I and Type II error. Personally, which do you feel is a worse error to make in the given situation?

4.99 The situation described in Exercise 4.93

ANSWER ⊕

WORKED SOLUTION ⊕

4.100 The situation described in Exercise 4.94

4.101 The situation described in Exercise 4.95

ANSWER ⊕

WORKED SOLUTION ⊕

4.102 The situation described in Exercise 4.96

4.103 The situation described in Exercise 4.97

ANSWER ⊕

WORKED SOLUTION ⊕

4.104 Influencing Voters

Exercise 4.38 describes a possible study to see if there is evidence that a recorded phone call is more effective than a mailed flyer in getting voters to support a certain candidate. The study assumes a significance level of $\alpha=0.05$.

(a) What is the conclusion in the context of this study if the p-value for the test is 0.027?

(b) In the conclusion in part (a), which type of error are we possibly making: Type I or Type II? Describe what that type of error means in this situation.

(c) What is the conclusion if the p-value for the test is 0.18?

(d) In the conclusion in part (c), which type of error are we possibly making: Type I or Type II? Describe what that type of error means in this situation.

4.105 Significant and Insignificant Results

(a) If we are conducting a statistical test and determine that our sample shows significant results, there are two possible realities: We are right in our conclusion or we are wrong. In each case, describe the situation in terms of hypotheses and/or errors.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) If we are conducting a statistical test and determine that our sample shows insignificant results, there are two possible realities: We are right in our conclusion or we are wrong. In each case, describe the situation in terms of hypotheses and/or errors.

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Explain why we generally won't ever know which of the realities (in either case) is correct.

ANSWER ⊕

WORKED SOLUTION ⊕

4.106 Classroom Games

Exercise 4.62 describes a situation in which game theory students are randomly assigned to play either Game 1 or Game 2, and then are given an exam containing questions on both games. Two one-tailed tests were conducted: one testing whether students who played Game 1 did better than students who played Game 2 on the question about Game 1, and one testing whether students who played Game 2 did better than students who played Game 1 on the question about Game 2. The p-values were 0.762 and 0.549, respectively. The p-values greater than 0.5 mean that, in the sample, the students who played the *opposite* game did better on each question. What does this study tell us about possible effects of actually playing a game and answering a theoretical question about it? Explain.