Agenda: #64F AP Stats Ch20A

Objectives :

1) Learn how confidence intervals and hypotheses tests associate.

CW1) (15 min)

Read intro and understand scenarios of Florida motorcycle lows.

CW2) (20min)

Hypotheses Test activities

CW3) (15min)

Confidence intervals activities

CW4) (5min)

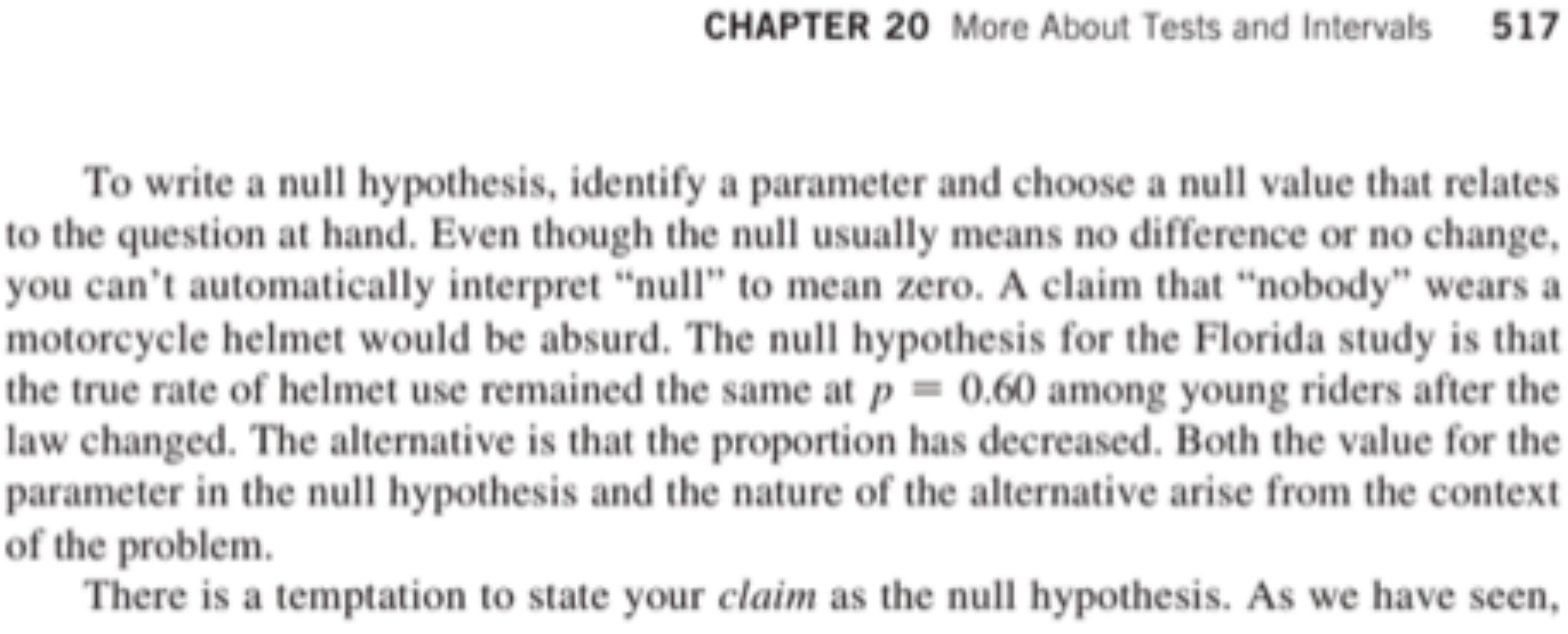
Closing Activities

chapter 20 More About Tests and Intervals

In 2000 Florida changed its motorcycle helmet law. No longer are riders 21 and older required to wear helmets. Under the new law, those under 21 still must wear helmets, but a report by the Preusser Group (www.preussergroup.com) suggests that helmet use may have declined in this group, too.

It isn't practical to survey young motorcycle riders. (For example, how can you construct a sampling frame? If you contacted licensed riders, would they admit to riding illegally without a helmet?) The researchers adopted a different strategy. Police reports of motorcycle accidents record whether the rider wore a helmet and give the rider's age. Before the change in the helmet law, 60% of youths involved in a motorcycle accident had been wearing their helmets. The Preusser study looked at accident reports during 2001–2003, the three years following the law change, considering these riders to be a representative sample of the larger population. They observed 781 young riders who were involved in accidents. Of these, 396 (or 50.7%) were wearing helmets. Is this evidence of a decline in helmet-wearing, or just the natural fluctuation of such statistics?

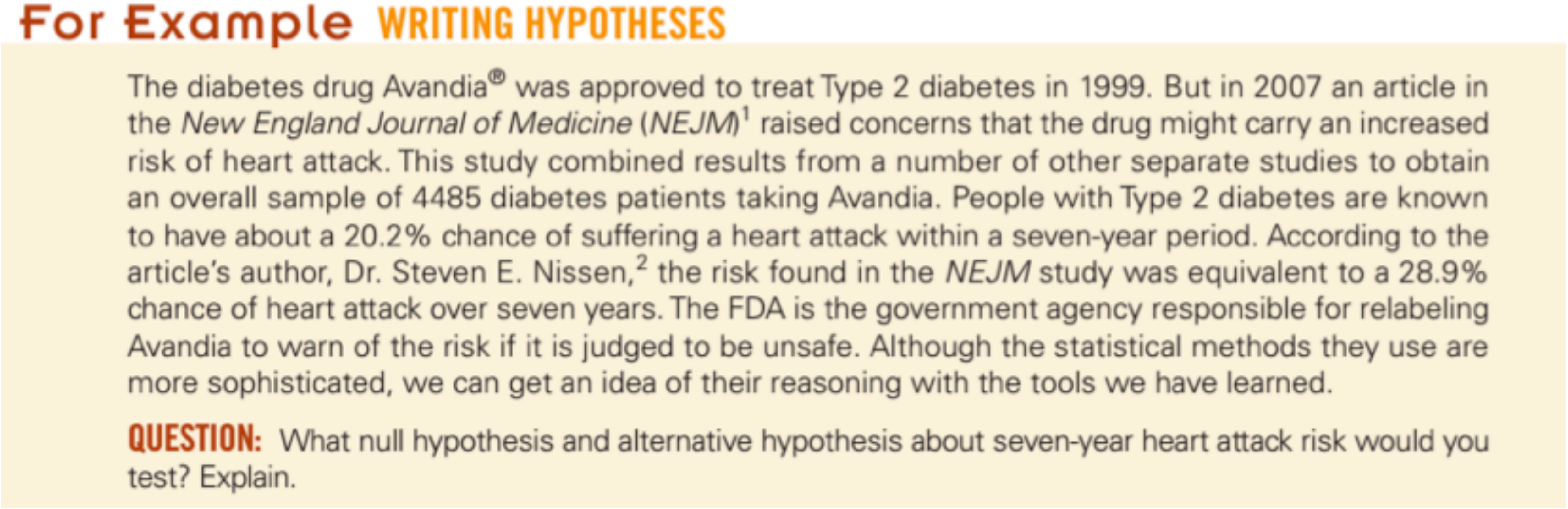
How do we choose the null hypothesis? The appropriate null arises directly from the context of the problem. It is dictated, not by the data, but by the situation. One good way to identify both the null and alternative hypotheses is to think about why the study is being done and what we hope to learn from the test. Typical null hypotheses might be that the proportion of patients recovering after receiving a new drug is the same as we would expect of patients receiving a placebo or that the mean strength attained by athletes training with new equipment is the same as with the old equipment. The alternative hypotheses would be that the new drug cures a higher proportion of patients or that the new equipment results in a greater mean strength.



To write a null hypothesis, identify a parameter and choose a null value that relates to the question at hand. Even though the null usually means no difference or no change, you can't automatically interpret "null" to mean zero. A claim that "nobody" wears a motorcycle helmet would be absurd. The null hypothesis for the Florida study is that the true rate of helmet use remained the same at $p = 0.60$ among young riders after the law changed. The alternative is that the proportion has decreased. Both the value for the parameter in the null hypothesis and the nature of the alternative arise from the context of the problem.

There is a temptation to state your *claim* as the null hypothesis. As we have seen,

Zero In on the Null



For Example WRITING HYPOTHESES

The diabetes drug Avandia® was approved to treat Type 2 diabetes in 1999. But in 2007 an article in the *New England Journal of Medicine* (*NEJM*)[1] raised concerns that the drug might carry an increased risk of heart attack. This study combined results from a number of other separate studies to obtain an overall sample of 4485 diabetes patients taking Avandia. People with Type 2 diabetes are known to have about a 20.2% chance of suffering a heart attack within a seven-year period. According to the article's author, Dr. Steven E. Nissen,[2] the risk found in the *NEJM* study was equivalent to a 28.9% chance of heart attack over seven years. The FDA is the government agency responsible for relabeling Avandia to warn of the risk if it is judged to be unsafe. Although the statistical methods they use are more sophisticated, we can get an idea of their reasoning with the tools we have learned.

QUESTION: What null hypothesis and alternative hypothesis about seven-year heart attack risk would you test? Explain.

How to Think About P-Values

A P-value is a conditional probability. It tells us the probability of getting results at least as unusual as the observed statistic, *given* that the null hypothesis is true. We can write $\text{P-value} = P(\text{observed statistic value [or even more extreme]} \mid H_0)$.

Writing the P-value this way helps to make clear that the P-value is *not* the probability that the null hypothesis is true. It is a probability about the data. Let's say that again:

The P-value is not the probability that the null hypothesis is true.

The P-value is not even the conditional probability that the null hypothesis is true given the data. We would write that probability as $P(\mathrm{H}_0 \mid \text{observed statistic value})$. This is a conditional probability but in reverse. It would be nice to know this probability, but we can't. As we saw in Chapter 14, reversing the order in a conditional probability is difficult, and the results can be counterintuitive.

We can find the P-value, $P(\text{observed statistic value} \mid \mathrm{H}_0)$, because H_0 gives the parameter values that we need to calculate the required probability. But there's no direct way to find $P(\mathrm{H}_0 \mid \text{observed statistic value})$.[3] As tempting as it may be to say that a P-value of 0.03 means there's a 3% chance that the null hypothesis is true, that just isn't right. All we can say is that, given the null hypothesis, there's a 3% chance of observing the statistic value that we have actually observed (or one more unlike the null value).

What to Do with a Small P-Value

We know that a small P-value means that the result we just observed is unlikely to occur if the null hypothesis is true. So we have evidence against the null hypothesis. An even smaller P-value implies stronger evidence against the null hypothesis, but it doesn't mean that the null hypothesis is "less true" (see "How Guilty Is the Suspect" on page 520).

How small the P-value has to be for you to reject the null hypothesis depends on a lot of things, not all of which can be precisely quantified. Your belief in the null hypothesis will influence your decision. Your trust in the data, in the experimental method if the data come from a planned experiment, in the survey protocol if the data come from a designed survey, all influence your decision. The P-value should serve as a measure of the strength of the evidence against the null hypothesis, but should never serve as a hard and fast rule for decisions. You have to take that responsibility on yourself.

As a review, let's look at the helmet law example from the chapter opener. Did helmet wearing among young riders decrease after the law allowed older riders to ride without helmets? What is the evidence?

Step-by-Step Example ANOTHER ONE-PROPORTION *z*-TEST

Question: Has helmet use in Florida declined among riders under the age of 21 subsequent to the change in the helmet laws?

THINK ➡ **Plan** State the problem and discuss the variables and the W's.

Hypotheses The null hypothesis is established by the rate set before the change in the law. The study was concerned with safety, so they'll want to know of any decline in helmet use, making this a lower-tail test.

I want to know whether the rate of helmet wearing among Florida's motorcycle riders under the age of 21 decreased after the law changed to allow older riders to go without helmets. The proportion before the law was passed was 60% so I'll use that as my null hypothesis value. The alternative is one-sided because I'm interested only in seeing if the rate decreased. I have data from accident records showing 396 of 781 young riders were wearing helmets.

$$H_0: p = 0.60$$

$$H_A: p < 0.60$$

SHOW ➡ **Model** Check the conditions.

✓ **Independence Assumption:** The data are for riders involved in accidents during a three-year period. Individuals are independent of one another.

✗ **Randomization Condition:** No randomization was applied, but we are considering these riders involved in accidents to be a representative sample of all riders. We should take care in generalizing our conclusions.

✓ **10% Condition:** These 781 riders are a small sample of a larger population of all young motorcycle riders.

✓ **Success/Failure Condition:** We'd expect $np = 781(0.6) = 468.6$ helmeted riders and $nq = 781(0.4) = 312.4$ non-helmeted. Both are at least 10.

Specify the sampling distribution model and name the test.

The conditions are satisfied, so I can use a Normal model and perform a one-proportion *z*-test.

SHOW ➡ Mechanics Find the standard deviation of the sampling model using the hypothesized proportion.

There were 396 helmet wearers among the 781 accident victims.

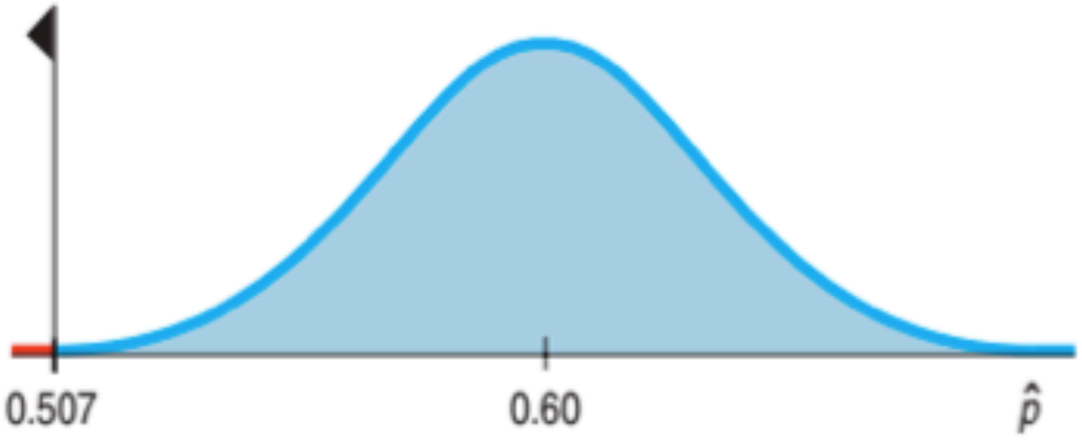
$$\hat{p} = \frac{396}{781} = 0.507$$

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.60)(0.40)}{781}} = 0.0175$$

Find the *z*-score for the observed proportion.

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.507 - 0.60}{0.0175} = -5.31$$

Make a picture. Sketch a Normal model centered at the hypothesized helmet rate of 60%. This is a lower-tail test, so shade the region to the left of the observed rate.



Given this *z*-score, the P-value is obviously very low.

The observed helmet rate is 5.31 standard deviations below the former rate. The corresponding P-value is less than 0.001.

Conclusion Link the P-value to your decision about the null hypothesis, and then state your conclusion in context.

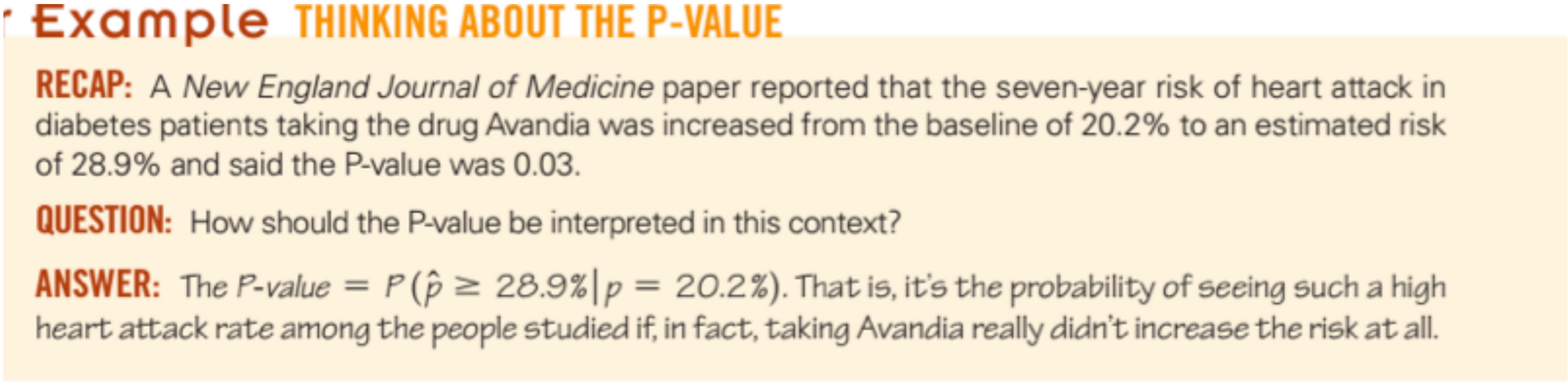
The very small P-value says that if the true rate of helmet-wearing among riders under 21 were still 60%, the probability of observing a rate no higher than 50.7% in a sample like this is less than 1 chance in 1000, so I reject the null hypothesis. There is strong evidence that there has been a decline in helmet use among riders under 21.

The P-value in the helmet example is quite small—less than 0.001. That's strong evidence to suggest that the rate has decreased since the law was changed. But it doesn't say that it was "a lot lower." To answer that question, you'd need to construct a confidence interval:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.507 \pm 1.96(0.0175) = (0.472, 0.542)$$

(using 95% confidence).

There is strong evidence that the rate is no longer 60%, but the small P-value by itself says nothing about how much lower the rate might be. The confidence interval provides that information; the rate seems to be closer to 50% now. Whether a change from 60% to 50% makes an important difference in safety is a judgment that depends on the situation, but not on the P-value. Not coincidentally, on July 1, 2008, Florida required a motorcycle "endorsement" for all motorcycle riders. For riders under 21, that requires a motorcycle safety course. Although only about 70% of motorcycle riders are endorsed, the percentage of unendorsed riders involved in crashes dropped considerably after 2008.[4]



r Example THINKING ABOUT THE P-VALUE

RECAP: A *New England Journal of Medicine* paper reported that the seven-year risk of heart attack in diabetes patients taking the drug Avandia was increased from the baseline of 20.2% to an estimated risk of 28.9% and said the P-value was 0.03.

QUESTION: How should the P-value be interpreted in this context?

ANSWER: The *P-value* $= P(\hat{p} \geq 28.9\% \mid p = 20.2\%)$. That is, it's the probability of seeing such a high heart attack rate among the people studied if, in fact, taking Avandia really didn't increase the risk at all.

What to Do with a High P-Value

Therapeutic touch (TT), taught in many schools of nursing, is a therapy in which the practitioner moves her hands near, but does not touch, a patient in an attempt to manipulate a "human energy field." Therapeutic touch practitioners believe that by adjusting this field they can promote healing. However, no instrument has ever detected a human energy field, and no experiment has ever shown that TT practitioners can detect such a field.

In 1998, the *Journal of the American Medical Association* published a paper reporting work by a then nine-year-old girl.[5] She had performed a simple experiment in which she challenged 15 TT practitioners to detect whether her unseen hand was hovering over their left or right hand (selected by the flip of a coin).

The practitioners "warmed up" with a period during which they could see the experimenter's hand, and each said that they could detect the girl's human energy field. Then a screen was placed so that the practitioners could not see the girl's hand, and they attempted 10 trials each. Overall, of 150 trials, the TT practitioners were successful only 70 times—a success proportion of 46.7%.

The null hypothesis here is that the TT practitioners were just guessing. If that were the case, since the hand was chosen using a coin flip, the practitioners would guess correctly 50% of the time. So the null hypothesis is that $p = 0.5$ and the alternative that they could actually detect a human energy field is (one-sided) $p > 0.5$.

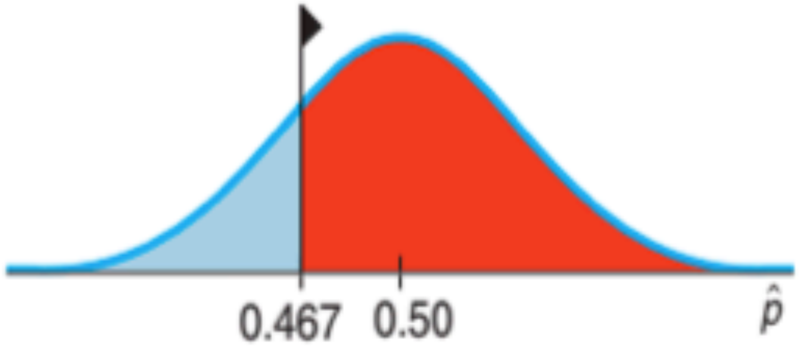
What would constitute evidence that they weren't guessing? Certainly, a very high proportion of correct guesses out of 150 would convince most people. Exactly how high the proportion of correct guesses has to be for you to reject the null hypothesis depends on how small a P-value you need to be convinced (which, in turn, depends on how often you're willing to make mistakes—a topic we'll discuss later in the chapter).

But let's look again at the TT practitioners' proportion. Does it provide any evidence that they weren't guessing? The proportion of correct guesses is 46.7%—that's *less* than the hypothesized value, not greater! When we find $SD(\hat{p}) = 0.041$ (or 4.1%) we can see that 46.7% is almost 1 SD *below* the hypothesized proportion:

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.5)(0.5)}{150}} \approx 0.041$$

The observed proportion, $\hat{p}$, is 0.467.

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.467 - 0.5}{0.041} = -0.805$$



The observed success rate is 0.805 standard deviations below the hypothesized mean.

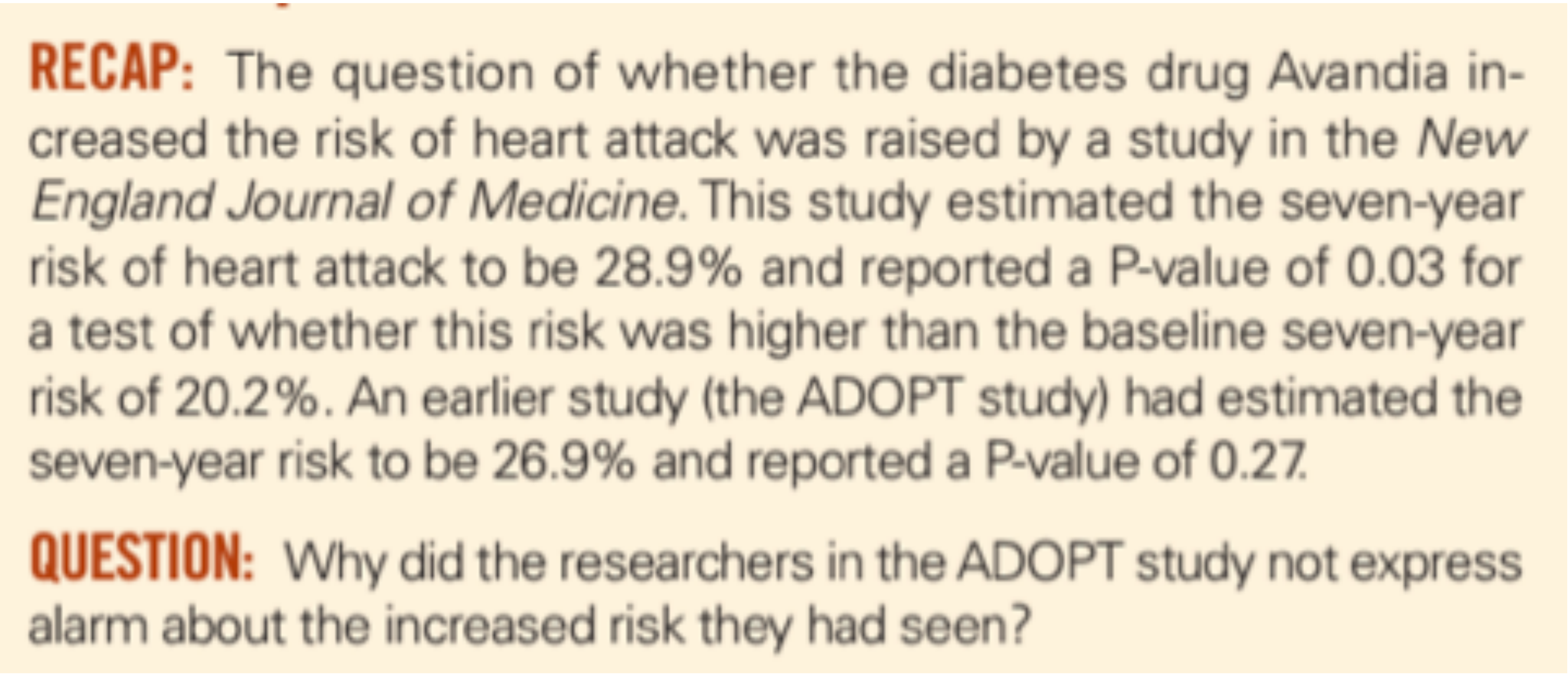
$$\text{P-value} = p(z > -0.805) = 0.790$$

If the practitioners had been highly successful, we would have seen a low P-value. In that case, we would then have concluded that they could actually detect a human energy field.

But that's not what happened. What we observed was a $\hat{p} = 0.467$ success rate. The P-value for this proportion is greater than 0.5 because the observed value is on the "wrong" side of the null hypothesis value. To convince us, the practitioners should be doing better than guessing, not worse!

Obviously, we won't be rejecting the null hypothesis; for us to reject it, the P-value would have to be quite small.

Big P-values just mean that what we've observed isn't surprising. That is, the results are in line with our assumption that the null hypothesis models the world, so we have no reason to reject it. A big P-value doesn't prove that the null hypothesis is true, but it certainly offers no evidence that it's *not* true. When we see a large P-value, all we can say is that we "don't reject the null hypothesis."

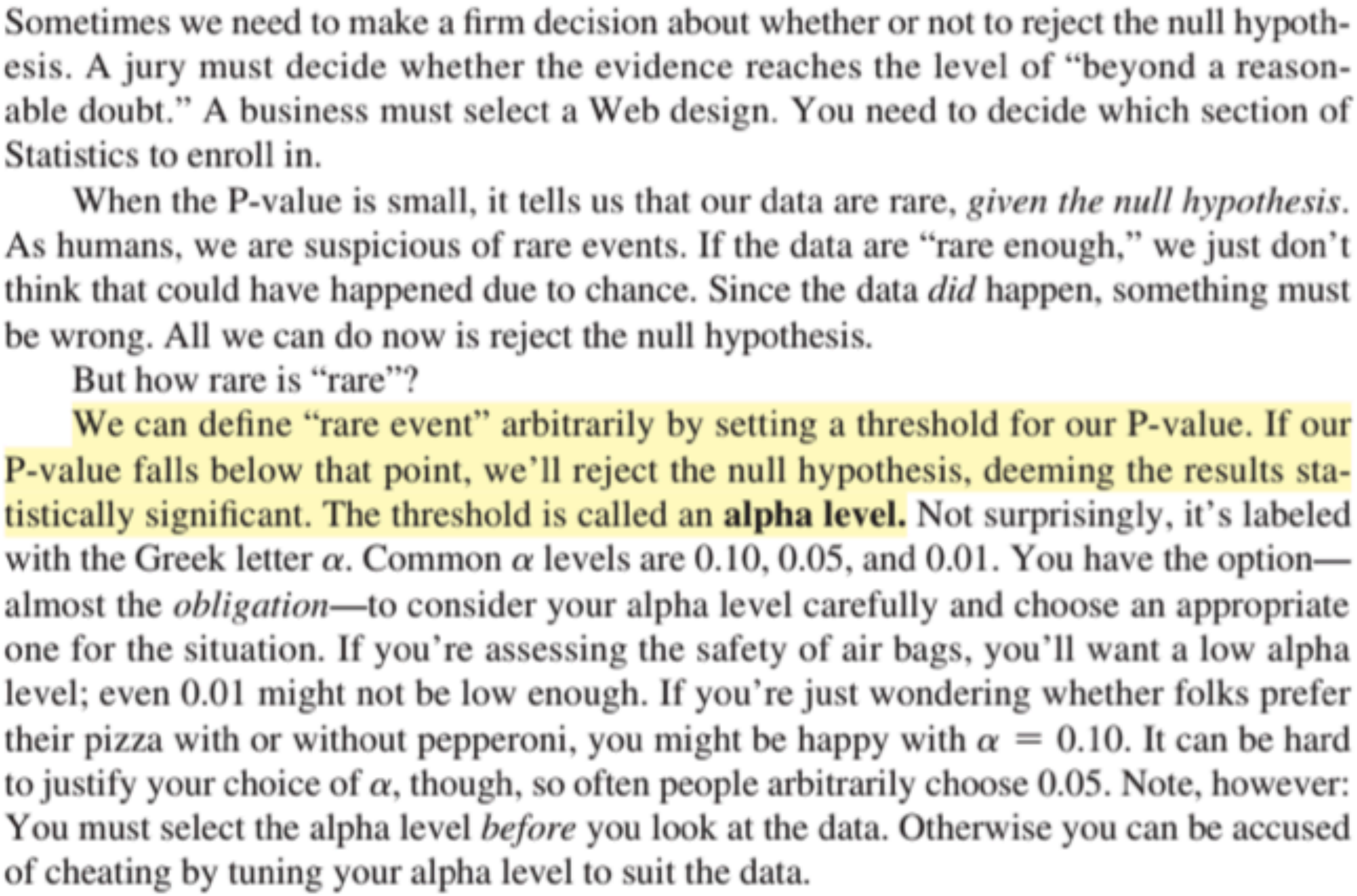


RECAP: The question of whether the diabetes drug Avandia increased the risk of heart attack was raised by a study in the *New England Journal of Medicine*. This study estimated the seven-year risk of heart attack to be 28.9% and reported a P-value of 0.03 for a test of whether this risk was higher than the baseline seven-year risk of 20.2%. An earlier study (the ADOPT study) had estimated the seven-year risk to be 26.9% and reported a P-value of 0.27.

QUESTION: Why did the researchers in the ADOPT study not express alarm about the increased risk they had seen?

ANSWER: A P-value of 0.27 means that a heart attack rate at least as high as the one they observed could be expected in 27% of similar experiments even if, in fact, there were no increased risk from taking Avandia. That's not remarkable enough to reject the null hypothesis. In other words, the ADOPT study wasn't convincing.

Alpha Levels



Sometimes we need to make a firm decision about whether or not to reject the null hypothesis. A jury must decide whether the evidence reaches the level of “beyond a reasonable doubt.” A business must select a Web design. You need to decide which section of Statistics to enroll in.

When the P-value is small, it tells us that our data are rare, *given the null hypothesis*. As humans, we are suspicious of rare events. If the data are “rare enough,” we just don’t think that could have happened due to chance. Since the data *did* happen, something must be wrong. All we can do now is reject the null hypothesis.

But how rare is “rare”?

We can define “rare event” arbitrarily by setting a threshold for our P-value. If our P-value falls below that point, we’ll reject the null hypothesis, deeming the results statistically significant. The threshold is called an **alpha level.** Not surprisingly, it’s labeled with the Greek letter α. Common α levels are 0.10, 0.05, and 0.01. You have the option—almost the *obligation*—to consider your alpha level carefully and choose an appropriate one for the situation. If you’re assessing the safety of air bags, you’ll want a low alpha level; even 0.01 might not be low enough. If you’re just wondering whether folks prefer their pizza with or without pepperoni, you might be happy with $\alpha = 0.10$. It can be hard to justify your choice of α, though, so often people arbitrarily choose 0.05. Note, however: You must select the alpha level *before* you look at the data. Otherwise you can be accused of cheating by tuning your alpha level to suit the data.

The alpha level is also called the **significance level.** When we reject the null hypothesis, we say that the test is “significant at that level.” For example, we might say that we reject the null hypothesis “at the 5% level of significance.”

What can you say if the P-value does not fall below α?

When you have not found sufficient evidence to reject the null according to the standard you have established, you should say that “The data have failed to provide sufficient evidence to reject the null hypothesis.” Don’t say that you “accept the null hypothesis.” You certainly haven’t proven or established it; it was merely assumed to begin with. Say that you’ve failed to reject it.

The automatic nature of the reject/fail-to-reject decision when we use an alpha level may make you uncomfortable. If your P-value falls just slightly above your alpha level, you’re not allowed to reject the null. Yet a P-value just barely below the alpha level leads to rejection. If this bothers you, you’re in good company. Many statisticians think it better to report the P-value than to base a decision on an arbitrary alpha level.

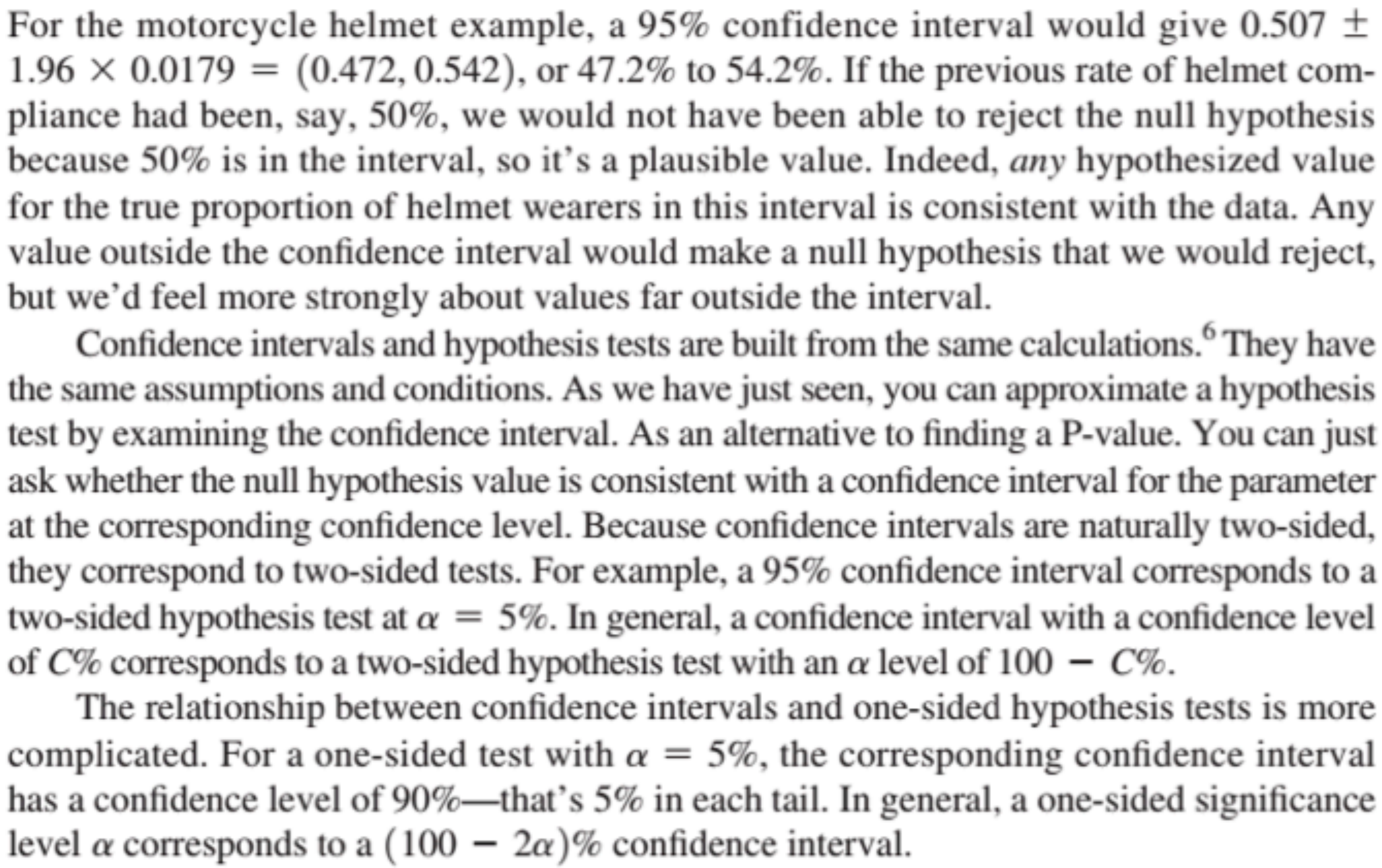
Practical vs. Statistical Significance

What do we mean when we say that a test is statistically significant? All we mean is that the test statistic had a P-value lower than our alpha level. Don't be lulled into thinking that statistical significance carries with it any sense of practical importance or impact.

For large samples, even small, unimportant ("insignificant") deviations from the null hypothesis can be statistically significant. On the other hand, if the sample is not large enough, even large financially or scientifically "significant" differences may not be statistically significant.

It's good practice to report the magnitude of the difference between the observed statistic value and the null hypothesis value (in the data units) along with the P-value on which we base statistical significance.

Confidence Intervals and Hypothesis Tests



For the motorcycle helmet example, a 95% confidence interval would give $0.507 \pm 1.96 \times 0.0179 = (0.472, 0.542)$, or 47.2% to 54.2%. If the previous rate of helmet compliance had been, say, 50%, we would not have been able to reject the null hypothesis because 50% is in the interval, so it's a plausible value. Indeed, *any* hypothesized value for the true proportion of helmet wearers in this interval is consistent with the data. Any value outside the confidence interval would make a null hypothesis that we would reject, but we'd feel more strongly about values far outside the interval.

Confidence intervals and hypothesis tests are built from the same calculations.[6] They have the same assumptions and conditions. As we have just seen, you can approximate a hypothesis test by examining the confidence interval. As an alternative to finding a P-value. You can just ask whether the null hypothesis value is consistent with a confidence interval for the parameter at the corresponding confidence level. Because confidence intervals are naturally two-sided, they correspond to two-sided tests. For example, a 95% confidence interval corresponds to a two-sided hypothesis test at $\alpha = 5\%$. In general, a confidence interval with a confidence level of $C\%$ corresponds to a two-sided hypothesis test with an α level of $100 - C\%$.

The relationship between confidence intervals and one-sided hypothesis tests is more complicated. For a one-sided test with $\alpha = 5\%$, the corresponding confidence interval has a confidence level of 90%—that's 5% in each tail. In general, a one-sided significance level α corresponds to a $(100 - 2\alpha)\%$ confidence interval.

or Example MAKING A DECISION BASED ON A CONFIDENCE INTERVAL

RECAP: The baseline seven-year risk of heart attacks for diabetics is 20.2%. In 2007 a *NEJM* study reported a 95% confidence interval equivalent to 20.8% to 40.0% for the risk among patients taking the diabetes drug Avandia.

QUESTION: What did this confidence interval suggest to the FDA about the safety of the drug?

ANSWER: The FDA could be 95% confident that the interval from 20.8% to 40.0% included the true risk of heart attack for diabetes patients taking Avandia. Because the lower limit of this interval was higher than the baseline risk of 20.2%, there was evidence of an increased risk.