Agenda: #61F AP Stats Ch19B

Objectives : Learn what testing hypotheses

1) Last lesson's conclusion with different significant values (15min)

2) Reading Page 495~499 30 min (15min)

we stop one by one and answer questions.

3) By reading alternative section,

3) Conclusion (House Keeping- Project) 10min

P-Values: Are We Surprised?

Beyond a Reasonable Doubt We ask whether the data were unlikely beyond a reasonable doubt. We've just calculated that probability. The probability that the observed statistic value (or an even more extreme value) could occur if the null model were true—in this case, 0.067—is the P-value.

The fundamental step in our reasoning is the question "Are these data surprising, given the null hypothesis?" The key calculation is to determine exactly how likely the data we observed would be if the null hypothesis were a true model of the world. Specifically, we want to find the *probability* of seeing data like these (or something even more extreme) *given* that the null hypothesis is true. This probability tells us how surprised we'd be to see the data we collected if the null hypothesis is true. It's so important that it gets a special name: it's called the **P-value**.[1]

When the P-value *is* small enough, it says that we are very surprised. It means that it's very unlikely we'd observe data like these if our null hypothesis were true. The model we started with (the null hypothesis) and the data we collected are at odds with each other, so we have to make a choice. Either the null hypothesis is correct and we've just seen something remarkable, or the null hypothesis is wrong, and we were wrong to use it as the basis for computing our P-value. On the other hand, if you believe in data more than in assumptions, then, given that choice, you should reject the null hypothesis.

When the P-value is high, we haven't seen anything unlikely or surprising at all. Events that have a high probability of happening happen often. The data are consistent with the model from the null hypothesis, and we have no reason to reject the null hypothesis. But many other similar hypotheses could also account for the data we've seen, so *we haven't proven that the null hypothesis is true*. The most we can say is that it doesn't appear to be false. Formally, we "fail to reject" the null hypothesis. That's a pretty weak conclusion, but it's all we can do with a high P-value.

■ NOTATION ALERT

We have many P's to keep straight. We use an uppercase P for probabilities, as in $P(\mathbf{A})$, and for the special probability we care about in hypothesis testing, the P-value.

We use lowercase p to denote our model's underlying proportion parameter and $\hat{p}$ to denote our observed proportion statistic.

What to Do with an "Innocent" Defendant

If the evidence is not strong enough to reject the defendant's presumption of innocence, what verdict does the jury return? They say "not guilty." Notice that they do not say that the defendant is innocent. All they say is that they have not seen sufficient evidence to convict, to reject innocence. The defendant may, in fact, be innocent, but the jury has no way to be sure.

Said statistically, the jury's null hypothesis is H_0: innocent defendant. If the evidence is too unlikely given this assumption—if the P-value is too small—the jury rejects the null hypothesis and finds the defendant guilty. But—and this is an important distinction—if there is *insufficient evidence* to convict the defendant, the jury does not decide that H_0 is true and declare the defendant innocent. Juries can only *fail to reject* the null hypothesis and declare the defendant "not guilty."

In the same way, if the data are not particularly unlikely under the assumption that the null hypothesis is true, then the most we can do is to "fail to reject" our null hypothesis. We never declare the null hypothesis to be true (or "accept" the null), because we simply do not know whether it's true or not. (After all, more evidence may come along later.)

In the trial, the burden of proof is on the prosecution. In a hypothesis test, the burden of proof is on the unusual claim. The null hypothesis is the ordinary state of affairs, so it's the alternative to the null hypothesis that we consider unusual and for which we must marshal evidence.

In the trial, the burden of proof is on the prosecution. In a hypothesis test, the burden of proof is on the unusual claim. The null hypothesis is the ordinary state of affairs, so it's the alternative to the null hypothesis that we consider unusual and for which we must marshal evidence.

Imagine a clinical trial testing the effectiveness of a new headache remedy. In Chapter 12, we saw the value of comparing such treatments to a placebo. The null hypothesis, then, is that the new treatment is no more effective than the placebo. This is important because some patients will improve even when administered the placebo treatment. If we use only six people to test the drug, the results are likely *not to be clear* and we'll be unable to reject the hypothesis. Does this mean the drug doesn't work? Of course not. It simply means that we don't have enough evidence to reject our assumption. That's why we don't start by assuming that the drug *is more effective*. If we were to do that, then we could test just a few people, find that the results aren't clear, and claim that since we've been unable to reject our original assumption the drug must be effective. The FDA is unlikely to be impressed by that argument.

Don’t “Accept” the Null Hypothesis

Think about the null hypothesis that H_0: All swans are white. Does collecting a sample of 100 white swans prove the null hypothesis? The data are *consistent* with this hypothesis and seem to lend support to it, but they don’t *prove* it. In fact, all we can do is disprove the null hypothesis—for example, by finding just one non-white swan.

The Reasoning of Hypothesis Testing

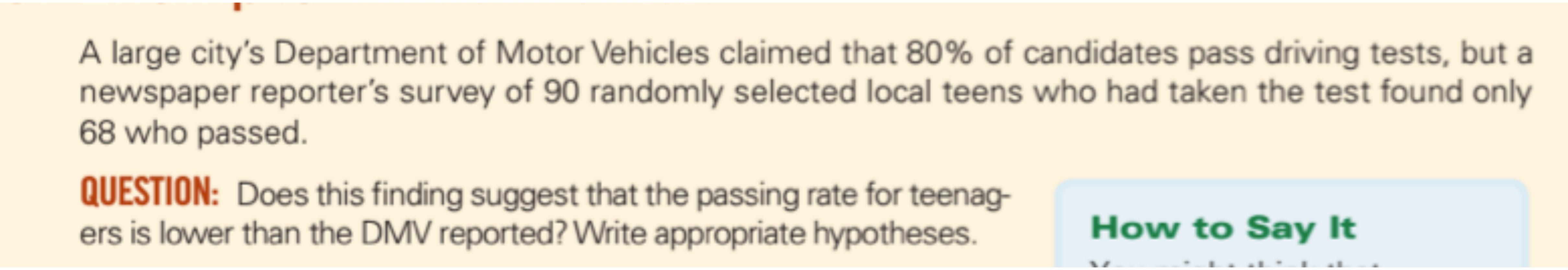
> "The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."
>
> —*Sir Ronald Fisher,*
> The Design of Experiments

Hypothesis tests follow a carefully structured path. To avoid getting lost as we navigate down it, we divide that path into four distinct sections.

1. Hypotheses

First, we state the null hypothesis. That's usually the skeptical claim that nothing's different. Are we considering a (New! Improved!) possibly better method? The null hypothesis says, "Oh yeah? Convince me!" To convert a skeptic, we must pile up enough evidence against the null hypothesis that we can reasonably reject it.

In statistical hypothesis testing, hypotheses are almost always about model parameters. To assess how unusual our data may be, we need a null model. The null hypothesis specifies a particular parameter value to use in our model. In the usual shorthand, we write H_0: *parameter* = *hypothesized value*. The alternative hypothesis, H_A, contains the values of the parameter we consider plausible when we reject the null.



A large city's Department of Motor Vehicles claimed that 80% of candidates pass driving tests, but a newspaper reporter's survey of 90 randomly selected local teens who had taken the test found only 68 who passed.

QUESTION: Does this finding suggest that the passing rate for teenagers is lower than the DMV reported? Write appropriate hypotheses.

How to Say It

For Example WRITING HYPOTHESES

A large city's Department of Motor Vehicles claimed that 80% of candidates pass driving tests, but a newspaper reporter's survey of 90 randomly selected local teens who had taken the test found only 68 who passed.

QUESTION: Does this finding suggest that the passing rate for teenagers is lower than the DMV reported? Write appropriate hypotheses.

ANSWER: *I'll assume that the passing rate for teenagers is the same as the DMV's overall rate of 80%, unless there's strong evidence that it's lower.*

$$H_0: p = 0.80$$

$$H_A: p < 0.80$$

How to Say It

You might think that the 0 in H_0 should be pronounced as "zero" or "0," but it's actually pronounced "naught" as in "all is for naught."

2. Model

To plan a statistical hypothesis test, specify the *model* you will use to test the null hypothesis and the parameter of interest. Of course, all models require assumptions, so you will need to state them and check any corresponding conditions.

Your Model step should end with a statement such as

Because the conditions are satisfied, I can model the sampling distribution of the sample proportion with a Normal model.

Watch out, though. Your Model step could end with

Because the conditions are not satisfied, I can't proceed with the test.

If that's the case, stop and reconsider.

Each test in the book has a name that you should include in your report. We'll see many tests in the chapters that follow. Some will be about more than one sample, some will involve statistics other than proportions, and some will use models other than the Normal (and so will not use z-scores). The test about proportions is called a **one-proportion z-test**.[2]

A S ***Activity:* Was the Observed Outcome Unlikely?** Complete the test you started in the first activity for this chapter. The narration explains the steps of the hypothesis test.

One-Proportion *z*-Test

The conditions for the one-proportion *z*-test are the same as for the one-proportion *z*-interval. We test the hypothesis H_0: $p = p_0$ using the statistic $z = \frac{(\hat{p} - p_0)}{SD(\hat{p})}$. We use the hypothesized proportion to find the standard deviation, $SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$.

When the conditions are met and the null hypothesis is true, this statistic follows the standard Normal model, so we can use that model to obtain a P-value.

For Example CHECKING THE CONDITIONS

RECAP: A large city's DMV claimed that 80% of candidates pass driving tests. A reporter has results from a survey of 90 randomly selected local teens who had taken the test.

For Example CHECKING THE CONDITIONS

RECAP: A large city's DMV claimed that 80% of candidates pass driving tests. A reporter has results from a survey of 90 randomly selected local teens who had taken the test.

QUESTION: Are the conditions for inference satisfied?

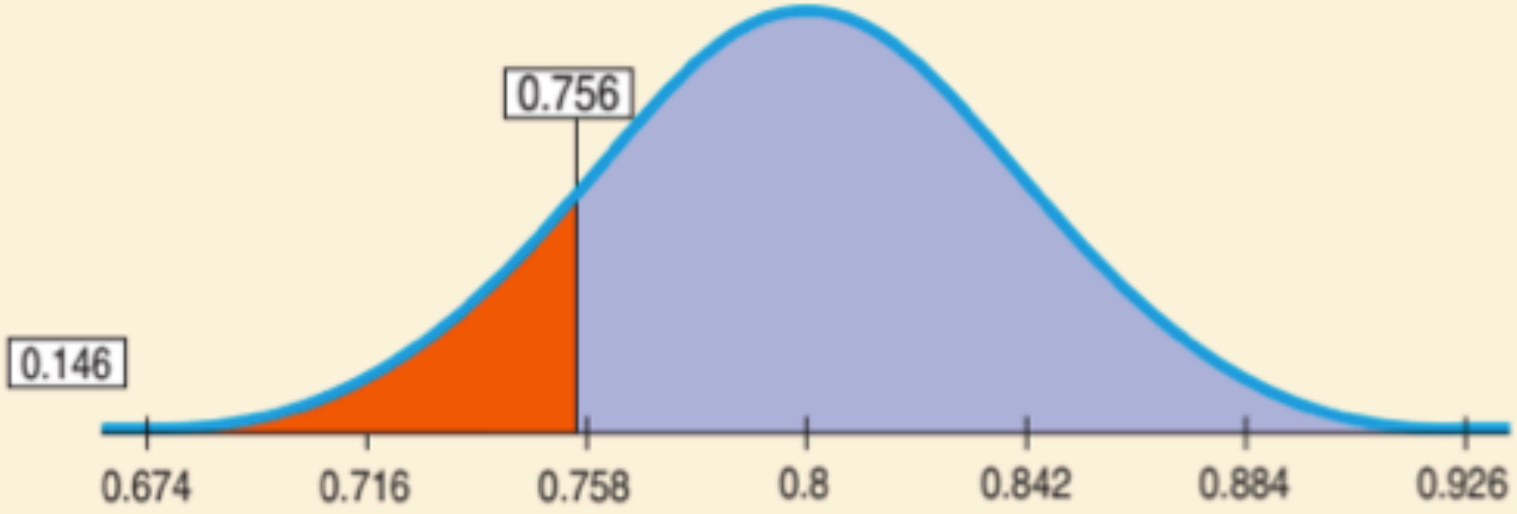
- ✓ **Randomization Condition:** The 90 teens surveyed were a random sample of local teenage driving candidates.
- ✓ **10% Condition:** 90 is fewer than 10% of the teenagers who take driving tests in a large city.
- ✓ **Success/Failure Condition:** We expect $np_0 = 90(0.80) = 72$ successes and $nq_0 = 90(0.20) = 18$ failures. Both are at least 10.

ANSWER: *The conditions are satisfied, so it's okay to use a Normal model and perform a one-proportion z-test.*

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.8)(0.2)}{90}} \approx 0.042$$

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.756 - 0.800}{0.042} \approx -1.05$$

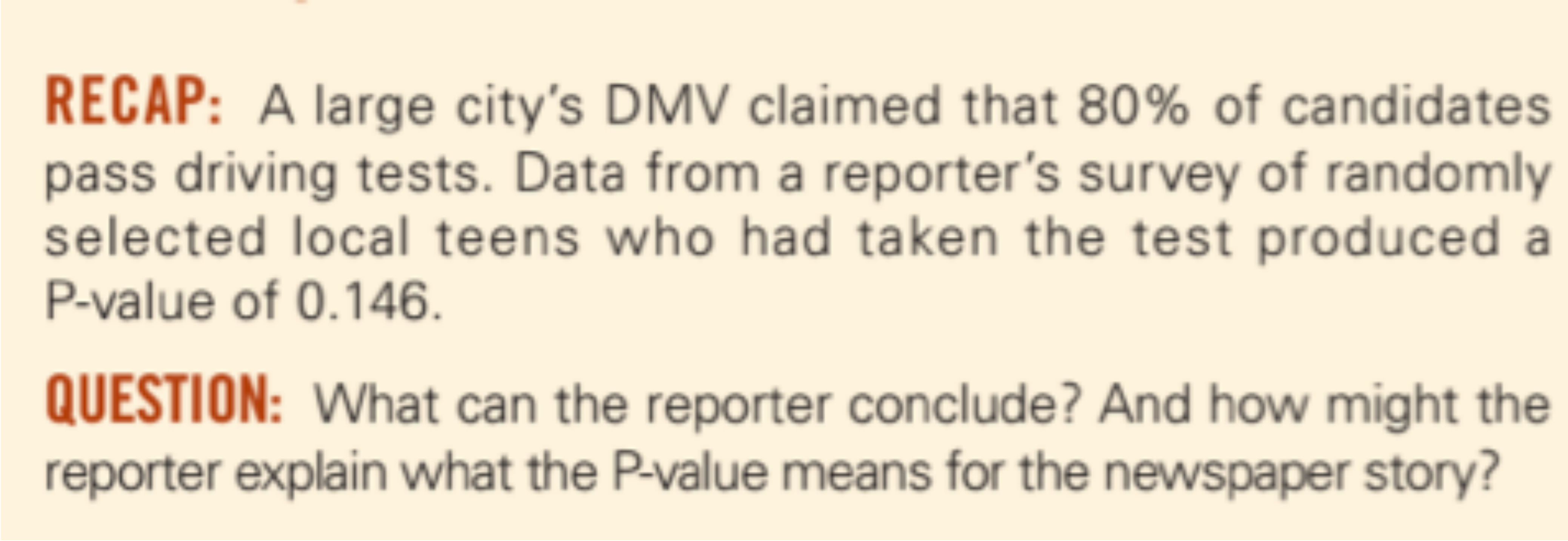
$$\text{P-value} = P(z < -1.05) = 0.146$$



4. Conclusion

The conclusion in a hypothesis test is always a statement about the null hypothesis. The conclusion must state either that we reject or that we fail to reject the null hypothesis. And, as always, the conclusion should be stated in context.

Your conclusion about the null hypothesis should never be the end of a testing procedure. Often, there are actions to take or policies to change. In our ingot example, management must decide whether to continue the changes proposed by the engineers. The decision always includes the practical consideration of whether the new method is worth the cost. Suppose management decides to reject the null hypothesis of 20% cracking in favor of the alternative that the percentage has been reduced. They must still evaluate how much the cracking rate has been reduced and how much it cost to accomplish the reduction. The *size of the effect* is always a concern when we test hypotheses. A good way to look at the **effect size** is to examine a confidence interval.



RECAP: A large city's DMV claimed that 80% of candidates pass driving tests. Data from a reporter's survey of randomly selected local teens who had taken the test produced a P-value of 0.146.

QUESTION: What can the reporter conclude? And how might the reporter explain what the P-value means for the newspaper story?

RECAP: A large city's DMV claimed that 80% of candidates pass driving tests. Data from a reporter's survey of randomly selected local teens who had taken the test produced a P-value of 0.146.

QUESTION: What can the reporter conclude? And how might the reporter explain what the P-value means for the newspaper story?

ANSWER: *Because the P-value of 0.146 is so large, I fail to reject the null hypothesis. These survey data do not provide sufficient evidence to convince us that the passing rate for teenagers taking the driving test is lower than 80%.*

If the passing rate for teenage driving candidates were actually 80%, we'd expect to see success rates this low in about 1 in 7 (14.6%) samples of this size. This seems too likely to happen just by chance to assert that the DMV's stated success rate does not apply to teens.