Chapter 22

What Is a Test of Significance?

Lecture Slides

Case Study: What Is a Test of Significance? 1

A February 5, 2015, article in *Inside Higher Education* reported that college students today are spending less time socializing face-to-face and more time socializing via social media.

What was the basis for this finding?

Case Study: What Is a Test of Significance? 2

Every year since 1985, the Higher Education Research Institute at UCLA has conducted a survey of college freshmen.

The 2014 survey involved a random sample of 153,015 of the more than 1.6 million first-time, full-time freshmen students at 227 of the nation’s baccalaureate colleges and universities.

In the the 2014 survey, the percent of students who reported spending 6 or more hours per week interacting in online social networks was 27.2%.

Case Study: What Is a Test of Significance? 3

27.2% in 2014 is a slight increase from 26.9% in 2013.

Students also reported spending less time socializing with friends in person.

In 2014, 18% of students (an all-time low) reported spending at least 16 hours per week socializing with friends.

This was down from 2013 when 20.1% of students reported spending this same amount of time per week socializing with friends.

Case Study: What Is a Test of Significance? 4

Further, in 2014, 38.8% of students (an all-time high) reported dedicating 5 hours per week or less to socializing; in 2013, this figure was 36.3%.

The *Inside Higher Education* article included a quote from the current director of a foundation that works to prevent suicide among students. He noted that time spent on social media is replacing time spent hanging out in person with friends, but even though the ways students are interacting is different, they are still finding ways to connect with each other.

Case Study: What Is a Test of Significance? 5

The sample size in the 2013 survey was 165,743, so the findings reported come from two very large samples.

Although different percents were reported in 2013 and 2014, changes in the percents are small.

Could it be that the difference between the two samples is just due to the luck of the draw in randomly choosing the respondents?

Case Study: What Is a Test of Significance? 6

In this chapter, we discuss methods, called tests of significance, that help us decide whether an observed difference can plausibly be attributed to chance.

By the end of this chapter, you will know how to interpret such tests and whether the differences in the Higher Education Research Institute's surveys of college freshmen can be plausibly attributed to chance.

The Reasoning of Statistical Tests of Significance 1

The local hot-shot playground basketball player claims to make 80% of his free throws.

"Show me," you say. $H_0 : p = 0.8$

He shoots 20 free throws and makes 8 of them.

$\hat{p} = 8/20 = 0.4$

"Aha," you conclude, "if he makes 80%, he would almost never make as few as 8 of 20. So I don't believe his claim."

$H_a : p < 0.8$

The Reasoning of Statistical Tests of Significance 2

That’s the reasoning of statistical tests of significance at the playground level: an outcome that is very unlikely if a claim is true is good evidence that the claim is not true.

Statistical inference uses data from a sample to draw conclusions about a population.

So, statistical tests deal with claims about a population.

The Reasoning of Statistical Tests of Significance 3

Statistical tests ask if sample data give good evidence against a claim.

A statistical test says, “If we took many samples and the claim were true, we would rarely get a result like this.”

To get a numerical measure of how strong the sample evidence is, replace the vague term *rarely* with a probability.

Example: Is the Coffee Fresh? 1

People of taste are supposed to prefer fresh-brewed coffee to the instant variety.

A skeptic claims that coffee drinkers can’t tell the difference.

Let’s do an experiment to test this claim.

Each of 50 subjects tastes two unmarked cups of coffee and says which he or she prefers. One cup in each pair contains instant coffee; the other, fresh-brewed coffee.

Example: Is the Coffee Fresh? 2

The statistic that records the result of our experiment is the proportion $\hat{p}$ of the sample who say they like the fresh-brewed coffee better.

We find that 36 of our 50 subjects choose the fresh coffee. That is,

$$\hat{p} = \frac{36}{50} = 0.72 \; or \; 72\%$$

Example: Is the Coffee Fresh? 3

To make a point, let's compare our outcome $\hat{p}$ = 0.72 with another possible result. If only 28 of the 50 subjects like the fresh coffee better than instant coffee, the sample proportion is

$$\hat{p} = \frac{28}{50} = 0.56\ or\ 56\%$$

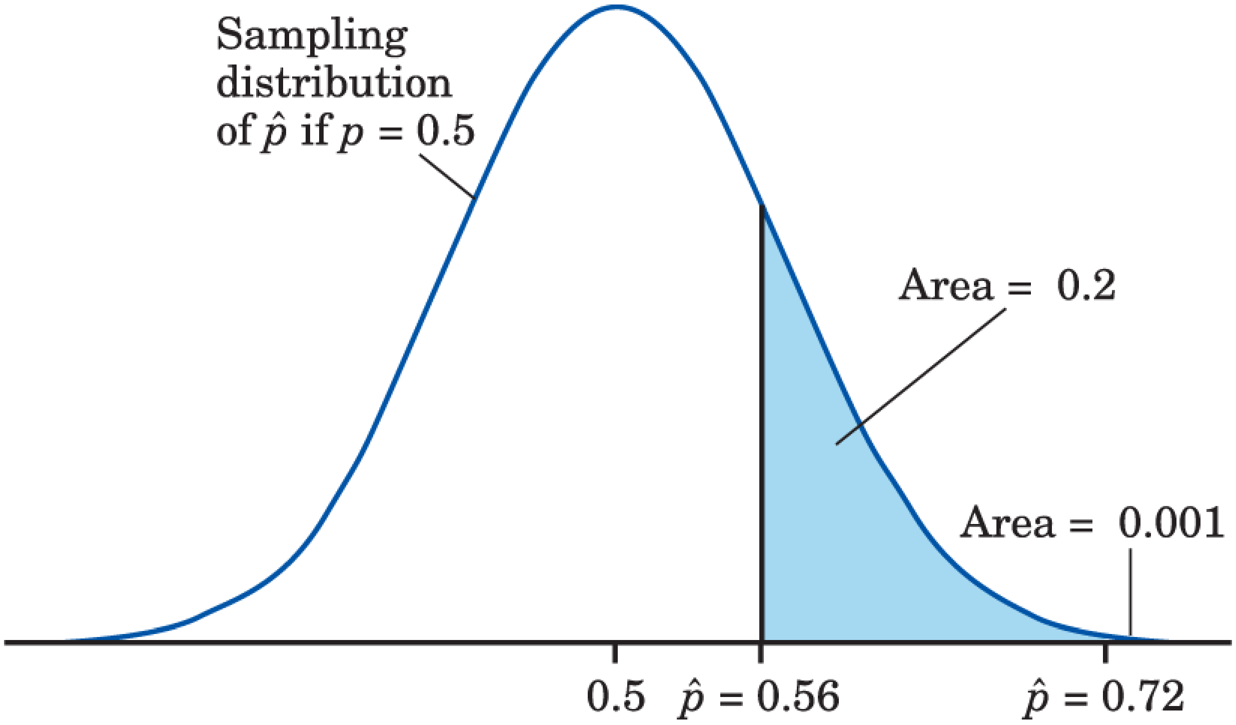
Surely 72% is stronger evidence against the skeptic's claim than 56%. But how much stronger?

Example: Is the Coffee Fresh? 4

- **The claim**. The skeptic claims that coffee drinkers can't tell fresh from instant, so that only half will choose fresh-brewed coffee. That is, he claims that the population proportion *p* is only 0.5. Suppose for the sake of argument that this claim is true. $H_0 : p = 0.5$
- **The sampling distribution**. If the claim *p* = 0.5 were true and we tested many random samples of 50 coffee drinkers, the sample proportion $\hat{p}$ would vary from sample to sample according to (approximately) the Normal distribution with mean = *p* = 0.5 and

standard deviation = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{50}} = 0.0707$

Example: Is the Coffee Fresh? 5



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Is the Coffee Fresh? 6

- **The data**. Place the sample proportion $\hat{p}$ on the sampling distribution. You see in Figure 22.1 that $\hat{p}$ = 0.56 isn’t an unusual value, but that $\hat{p}$ = 0.72 is unusual. We would rarely get 72% of a sample of 50 coffee drinkers preferring fresh-brewed coffee if only 50% of all coffee drinkers felt that way. So, the sample data do give evidence against the claim. → $H_0: p = 0.5$

- **The probability**. We can measure the strength of the evidence against the claim by a probability. What is the probability that a sample gives a $\hat{p}$ this large or larger if the truth about the population is that *p* = 0.5?

Example: Is the Coffee Fresh? 7

- **The probability (cont'd)**. If $\hat{p}$ = 0.56, this probability is the shaded area under the Normal curve in Figure 22.1. This area is 0.20. Our sample actually gave $\hat{p}$ = 0.72. The probability of getting a sample outcome this large is only 0.001, an area too small to see in Figure 22.1. An outcome that would occur just by chance in 20% of all samples is not strong evidence against the claim. But an outcome that would happen only 1 in 1000 times is good evidence.

The Reasoning of Statistical Tests of Significance 4

There are two possible explanations of the fact that 72% of our subjects prefer fresh to instant coffee:

1. The skeptic is correct (p = 0.5), and by bad luck a very unlikely outcome occurred.
2. The population proportion favoring fresh coffee is greater than 0.5, so the sample outcome is about what would be expected. We cannot be certain that Explanation 1 is untrue. Our taste test results could be due to chance alone. But the probability that such a result would occur by chance is so small (0.001) that we are quite confident that Explanation 2 is right.

Hypotheses and *P*-values 1

In most studies, we hope to show that some definite effect is present in the population.

In Example 1, we suspect that a majority of coffee drinkers prefer fresh-brewed coffee.

A statistical test begins by supposing for the sake of argument that the effect we seek is *not* present. We then look for evidence against this supposition and in favor of the effect we hope to find.

The first step in a test of significance is to state a claim that we will try to find evidence *against*.

Hypotheses and *P*-values 2

Null Hypothesis (H_O)

The claim being tested in a statistical test is called the **null hypothesis**. The test is designed to assess the strength of the evidence against the null hypothesis. Usually, the null hypothesis is a statement of “no effect” or “no difference.”

Hypotheses and *P*-values 3

The term *null hypothesis* is abbreviated H_0 and is read as H-nought, H-oh, and sometimes even H-null.

It is a statement about the population and so must be stated in terms of a population parameter.

In Example 1, the parameter is the proportion *p* of all coffee drinkers who prefer fresh to instant coffee. The null hypothesis is:

$$H_0 : p = 0.5$$

Hypotheses and *P*-values 4

The statement we hope or suspect is true instead of H_0 is called the **alternative hypothesis** and is abbreviated H_a.

In Example 1, the alternative hypothesis is that a majority of the population favor fresh coffee. In terms of the population parameter, this is:

$$H_a: p > 0.5$$

Hypotheses and *P*-values 5

A significance test looks for evidence against the null hypothesis and in favor of the alternative hypothesis.

The evidence is strong if the outcome we observe would rarely occur if the null hypothesis is true but is more probable if the alternative hypothesis is true.

For example, it would be surprising to find 36 of 50 subjects favoring fresh coffee if, in fact, only half of the population feel this way.

How surprising?

Hypotheses and *P*-values 6

A significance test answers this question by giving a probability: the probability of getting an outcome at least as far as the actually observed outcome from what we would expect when H_0 is true.

What counts as “far from what we would expect” depends on H_a as well as H_0.

In the taste test, the probability we want is the probability that 36 or more of 50 subjects favor fresh coffee. If the null hypothesis $p = 0.5$ is true, this probability is very small (0.001). That’s good evidence that the null hypothesis is not true.

Hypotheses and *P*-values 7

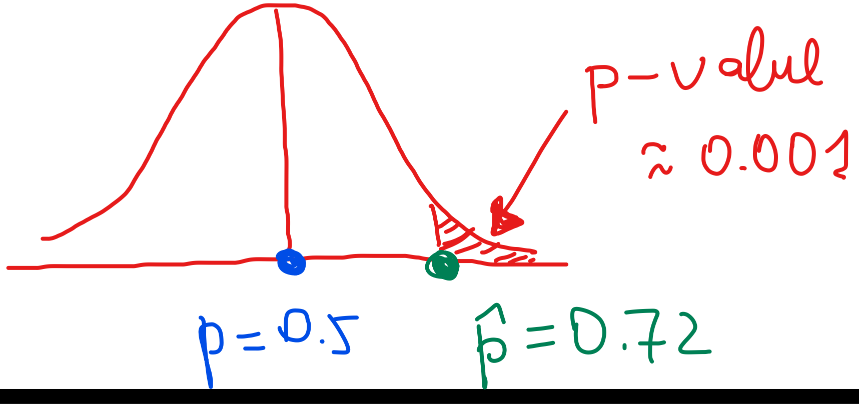
The probability, computed assuming that H_0 is true, that the sample outcome would be as extreme or more extreme than the actually observed outcome is called the ***P*-value** of the test.

The smaller the *P*-value is, the stronger is the evidence against H_0 provided by the data.

$H_0: p = 0.5$

$H_a: p > 0.5$

Evidence $\hat{p} = 0.72$



Hypotheses and *P*-values 8

In practice, most statistical tests are carried out by computer software that calculates the *P*-value for us.

It is usual to report the *P*-value in describing the results of studies in many fields.

You should, therefore, understand what *P*-values say even if you don’t do statistical tests yourself, just as you should understand what *95% confidence* means even if you don’t calculate your own confidence intervals.

Example: Working through College 1

Do college students work too much? According to a 2015 report from the Georgetown University Center on Education and the Workforce, 70% of college students in the United States have a full- or part-time job while enrolled in college.

If we express 70% as a proportion, this means the claimed population proportion is *p* = 0.7.

$H_0 : p = 0.7$

Example: Working through College 2

An administrator from a local college questions the accuracy of this claim. In particular, she believes the true proportion of students at her college who have full- or part-time jobs while enrolled in college is different from 0.7. $H_a: p \neq 0.7$

The administrator is able to survey a random sample of 325 students from her college, and she finds that 238 of these students have full- or part-time jobs.

Example: Working through College 3

The sample proportion is $\hat{p} = \frac{238}{325} = 0.732.$

That’s a bit more than 70%.

Is this evidence that the true population proportion is something different than 0.7?

This is a job for a significance test.

Example: Working through College 4

The Hypotheses

The null hypothesis says that the population proportion is 0.7 (p = 0.7).

The administrator believes that this value is incorrect, but she does not theorize ahead of time that the true value is higher than or lower than 0.7. She just believes the true population proportion is something different than 0.7, so the alternative hypothesis is just “the population proportion is not 0.7.” The two hypotheses are:

$$H_0 : p = 0.7$$
$$H_a : p \neq 0.7$$

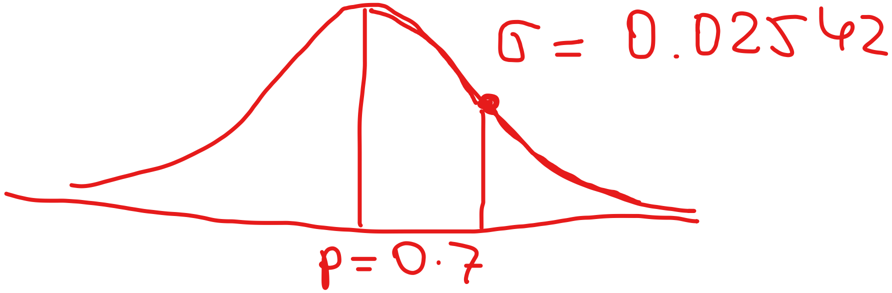
Example: Working through College 5

The sampling distribution

If the null hypothesis is true, the sample proportion of heads has approximately the Normal distribution with mean = *p* = 0.7 and

standard deviation = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.7(1-0.7)}{325}} = 0.02542$

Under the null $H_0: p = 0.7$, we have:



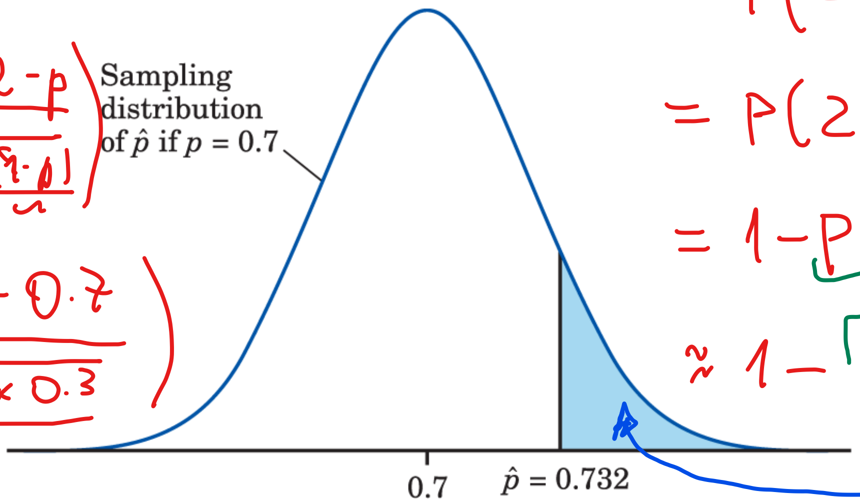
Example: Working through College 6

The data. Figure 22.2 shows this sampling distribution with the sample outcome $\hat{p}$ = 0.732 marked. The picture already suggests that this is not an unlikely outcome that would give strong evidence against the claim that p = 0.7.

$$P(\hat{p} \geq 0.732)$$

$$= P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{0.732 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$\approx P\left(Z \geq \frac{0.732 - 0.7}{\sqrt{\frac{0.7 \times 0.3}{325}}}\right)$$



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

$$= P\left(Z \geq \frac{0.032}{0.02542}\right)$$

$$= P(Z \geq 1.258851)$$

$$= 1 - \underbrace{P(Z < 1.258851)}$$

$$\approx 1 - \overbrace{0.8962}$$

$$= \boxed{0.1038}$$

See more extensive table B on OWL (Final exam resources)

Example: Working through College 7

The *P*-value. How unlikely is an outcome as far from 0.7 as $\hat{p} = 0.732$?

Because the alternative hypothesis allows *p* to lie on either side of 0.7, values of $\hat{p}$ far from 0.7 in either direction provide evidence against H_0 and in favor of H_a. The *P*-value is, therefore, the probability that the observed $\hat{p}$ lies as far from 0.7 in either direction as the observed $\hat{p} = 0.732$.

Figure 22.3 shows this probability as area under the Normal curve. It is $P = 0.19$.

We get $P \approx 0.2$. See next.

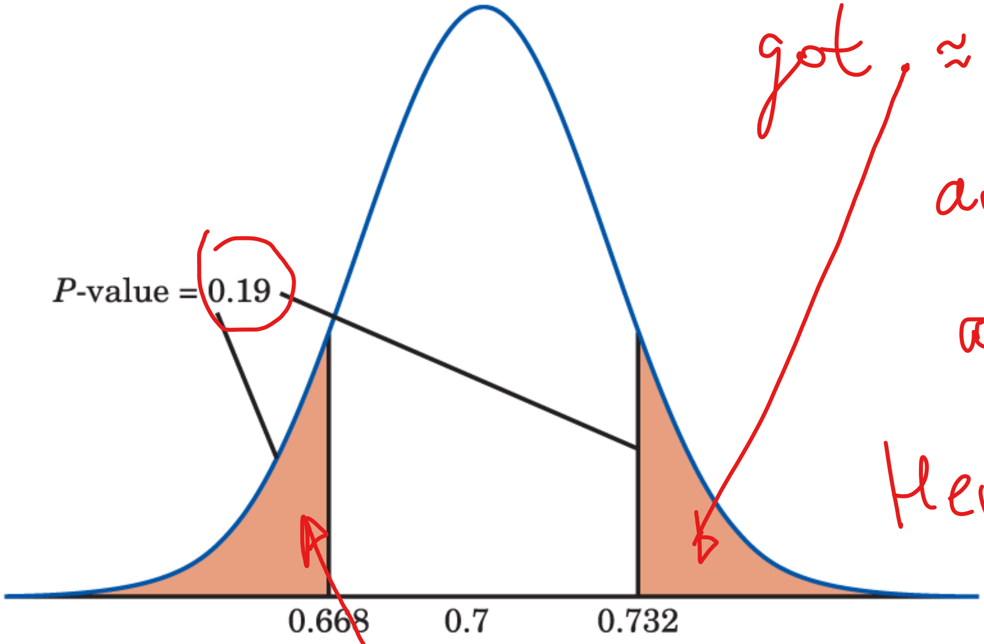
Example: Working through College 8

The *P*-value.

On slide 32, we got ≈ 0.1

and so is also ≈ 0.1.

Hence, the two areas give ≈ 0.2 (≈ 0.19)



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Working through College 9

The Conclusion

If the true population proportion is 0.7, the probability we would obtain a sample proportion this far or farther from 0.7 is 0.19.

We therefore cannot reject the claim that 70% of college students work full- or part-time while attending college.

Hypotheses and *P*-values 9

The alternative H_a: $p > 0.5$ in Example 1 is a one-sided alternative because the effect we seek evidence for says that the population proportion is greater than one-half.

The alternative H_a: $p \neq 0.7$ in Example 2 is a two-sided alternative because we ask whether or not the proportion differs from 0.7. Whether the alternative is one-sided or two-sided determines whether sample results that are extreme in one direction or in both directions count as evidence against H_0 in favor of H_a.

Statistical Significance 1

We can decide in advance how much evidence against H_0 we will insist on.

!!!

The way to do this is to say, before any data are collected, how small a *P*-value we require.

The decisive value of *P* is called the **significance level**.

Statistical Significance 2

It is usual to write the significance level as α, the Greek letter alpha.

If we choose $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would happen no more than 5% of the time (one time in 20) when H_0 is true.

If we choose $\alpha = 0.01$, we are insisting on stronger evidence against H_0, evidence so strong that it would appear only 1% of the time (one time in 100) if H_0 is, in fact, true.

Statistical Significance 3

If the *P*-value is as small or smaller than α, we say that the data are **statistically significant at level α**.

Statistical Significance 4

Significant in the statistical sense does not mean “important.” It means simply “not likely to happen just by chance.”

Now we have attached a number to statistical significance to say what *not likely* means.

You will often see significance at level 0.01 expressed by the statement, “The results were significant ($P < 0.01$).” Here, *P* stands for the *P*-value.

Very important to give P-value.

Statistical Significance 5

One traditional level of significance to use is 0.05.

The origins of this appear to trace back to the British statistician and geneticist Sir Ronald A. Fisher.

Fisher once wrote that it was convenient to consider sample statistics that are two or more standard deviations away from the mean as being significant.

Of course, we don’t have to make use of traditional levels of significance such as 5% and 1%.

Usually specified by the "regulator" (e.g. Health Canada)

Statistical Significance 6

The *P*-value is more informative because it allows us to assess significance at any level we choose.

A result with *P* = 0.03 is significant at the α = 0.05 level but not significant at the α = 0.01 level.

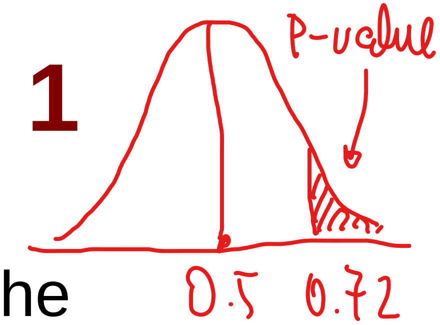
Nonetheless, the traditional significance levels are widely accepted guidelines for "how much evidence is enough." We might say that *P* < 0.10 indicates "some evidence" against the null hypothesis, *P* < 0.05 is "moderate evidence," and *P* < 0.01 is "strong evidence."

Calculating *P*-values 1

Finding the *P*-values we gave in Examples 1 and 2 requires doing Normal distribution calculations using Table B of Normal percentiles.

In practice, software does the calculation for us, but here is an example that shows how to use Table B.

Example: Tasting Coffee 1



The hypotheses. In Example 1, we want to test the hypotheses H_0 : $p = 0.5$ H_a: $p > 0.5$. Here, p is the proportion of the population of all coffee drinkers who prefer fresh coffee to instant coffee.

The sampling distribution. If the null hypothesis is true, so that $p = 0.5$, we saw in Example 1 that $\hat{p}$ follows a Normal distribution with mean 0.5 and standard deviation 0.0707.

The data. A sample of 50 people found that 36 preferred fresh coffee. The sample proportion is $\hat{p} = 0.72$.

Example: Tasting Coffee 2

The *P*-value. The alternative hypothesis is one-sided on the high side. So, the *P*-value is the probability of getting an outcome at least as large as 0.72. Figure 22.1 displays this probability as an area under the Normal sampling distribution curve. To find any Normal curve probability, move to the standard scale. When we convert a sample statistic to a standard score when conducting a statistical test of significance, the standard score is commonly referred to as a test statistic.

$$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \longleftarrow \text{test statistic}$$

$H_0: p = p_0 \ (=0.5)$

Example: Tasting Coffee 3

The *P*-value. The test statistic for the outcome $\hat{p}$ = 0.72 is

$$\textit{standard score} = \frac{\textit{observation} - \textit{mean}}{\textit{standard deviation}}$$

$$= \frac{0.72 - 0.5}{0.0707} = 3.1$$

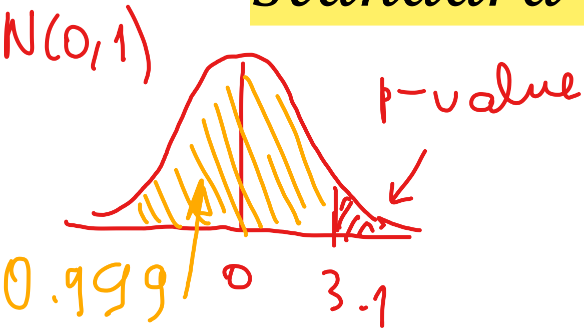


Table B says that standard score 3.1 is the 99.9 percentile of a Normal distribution. That is, the area under a Normal curve to the left of 3.1 (in the standard scale) is 0.999. The area to the right is therefore 0.001, and that is our *P*-value.

Example: Tasting Coffee 4

The Conclusion

The small *P*-value means that these data provide very strong evidence that a majority of the population prefers fresh coffee.

Calculating *P*-values 2

Test Statistic

When conducting a statistical test of significance, the standard score that is computed based on the sample data is commonly referred to as a **test statistic**.

Tests for a Population Mean

The reasoning that leads to significance tests for hypotheses about a population mean μ follows the reasoning that leads to tests about a population proportion *p*.

The big idea is to use the sampling distribution that the sample mean $\bar{x}$ would have if the null hypothesis were true.

Locate the value of $\bar{x}$ from your data on this distribution, and see if it is unlikely. A value of $\bar{x}$ that would rarely appear if H_0 were true is evidence that H_0 is not true.

$H_0 : \mu = \mu_0$

$H_a : \mu < \mu_0$ (or $\mu > \mu_0$) (or $\mu \neq \mu_0$)

Example: Length of Human Pregnancies 1

Although women are typically given a delivery date that is calculated as 280 days after the onset of their last menstrual period, only 4% of women deliver babies at 280 days.

$H_0 : \mu = 280$

A more likely average time from ovulation to birth may be much less than 280 days.

$H_a : \mu < 280$

To test this theory, a random sample of 95 women with healthy pregnancies is monitored from ovulation to birth.

n sample size

Example: Length of Human Pregnancies 2

The mean length of pregnancy is found to be $\bar{x} = 275$ days, with a standard deviation of $s = 10$ days.

Is this sample result good evidence that the mean length of a healthy pregnancy for all women is less than 280 days?

The hypotheses. The researcher’s claim is that the mean length of pregnancy is less than 280 days. The hypotheses are

$$H_0 : \mu = 280 \quad H_a: \mu < 280$$

Example: Length of Human Pregnancies 3

The sampling distribution. If the null hypothesis is true, the sample mean $\bar{x}$ has approximately the Normal distribution with mean μ = 280 and standard error $\frac{s}{\sqrt{n}} = \frac{10}{\sqrt{95}} = 1.03$

comes from H_0

comes from data

We once again use the sample standard deviation *s* in place of the unknown population standard deviation σ.

Example: Length of Human Pregnancies 4

The data. The researcher's sample gave $\bar{x}$ = 275. The standard score, or test statistic, for this outcome is

$$standard\ score = \frac{observation - mean}{standard\ error}$$

$5/\sqrt{n}$

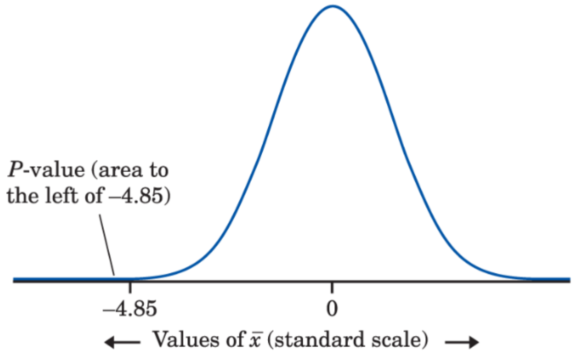
$$= \frac{275 - 280}{\boxed{1.03}} = -4.85$$

That is, the sample result is about 4.85 standard errors below the mean we would expect if, on the average, healthy pregnancies lasted for 280 days.

Example: Length of Human Pregnancies 5

The *P*-value

The *P*-value for our one-sided test is the area to the left of −4.85 under the Normal curve. The area to the left of −3.4 in Table B is 0.0003. Because −4.85 is smaller than −3.4, we know that the area to its left is smaller than 0.0003. Thus, our *P*-value is smaller than 0.0003.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Length of Human Pregnancies 6

The Conclusion

A *P*-value of less than 0.0003 is strong evidence that the mean length of a healthy human pregnancy is below the commonly reported length of 280 days.

Statistics in Summary 1

- A confidence interval estimates an unknown parameter. A **test of significance** assesses the evidence for some claim about the value of an unknown parameter.

- In practice, the purpose of a statistical test is to answer the question, “Could the effect we see in the sample just be an accident due to chance, or is it good evidence that the effect is really there in the population?”

Statistics in Summary 2

- Significance tests answer this question by giving the probability that a sample effect as large as the one we see in this sample would arise just by chance. This probability is ***P*-value**. A small *P*-value says that our outcome is unlikely to happen just by chance.

- To set up a test, state a **null hypothesis** that says the effect you seek is not present in the population. The **alternative hypothesis** says that the effect is present.

Statistics in Summary 3

- The *P*-value is the probability, calculated taking the null hypothesis to be true, of an outcome as extreme in the direction specified by the alternative hypothesis as the actually observed outcome.

- A sample result is **statistically significant at the 5% level (or at the 0.05 level)** if it would occur just by chance no more than 5% of the time in repeated samples.