**Section 5.3: Hypothesis Testing for a Proportion**



STAT 1201
Introduction to Probability and Statistics

# 5.3 Hypothesis testing for a proportion

The following question comes from a book written by Hans Rosling, Anna Rosling Rönnlund, and Ola Rosling called *Factfulness*:

> *How many of the world's 1 year old children today have been vaccinated against some disease:*
>
> *a. 20%*
> *b. 50%*
> *c. 80%*

Write down what your answer (or guess), and when you're ready, find the answer in the footnote.[13]

In this section, we'll be exploring how people with a 4-year college degree perform on this and other world health questions as we learn about hypothesis tests, which are a framework used to rigorously evaluate competing ideas and claims.

## 5.3.1 Hypothesis testing framework

We're interested in understanding how much people know about world health and development. If we take a multiple choice world health question, then we might like to understand if

**$H_0$:** People never learn these particular topics and their responses are simply equivalent to random guesses.

**$H_A$:** People have knowledge that helps them do better than random guessing, or perhaps, they have false knowledge that leads them to actually do worse than random guessing.

These competing ideas are called **hypotheses**. We call $H_0$ the null hypothesis and $H_A$ the alternative hypothesis. When there is a subscript 0 like in $H_0$, data scientists pronounce it as "nought" (e.g. $H_0$ is pronounced "H-nought").

**NULL AND ALTERNATIVE HYPOTHESES**

The **null hypothesis ($H_0$)** often represents a skeptical perspective or a claim to be tested. The **alternative hypothesis ($H_A$)** represents an alternative claim under consideration and is often represented by a range of possible parameter values.

Our job as data scientists is to play the role of a skeptic: before we buy into the alternative hypothesis, we need to see strong supporting evidence.

The null hypothesis often represents a skeptical position or a perspective of "no difference". In our first example, we'll consider whether the typical person does any different than random guessing on Roslings' question about infant vaccinations.

The alternative hypothesis generally represents a new or stronger perspective. In the case of the question about infant vaccinations, it would certainly be interesting to learn whether people do better than random guessing, since that would mean that the typical person knows something about world health statistics. It would also be very interesting if we learned that people do *worse* than random guessing, which would suggest people believe incorrect information about world health.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative. The hallmarks of hypothesis testing are also found in the US court system.

[13] The correct answer is (c): 80% of the world's 1 year olds have been vaccinated against some disease.

**GUIDED PRACTICE 5.17**

G A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?[14]

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true.* Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

When considering Roslings' question about infant vaccination, the null hypothesis represents the notion that the people we will be considering – college-educated adults – are as accurate as random guessing. That is, the proportion $p$ of respondents who pick the correct answer, that 80% of 1 year olds have been vaccinated against some disease, is about 33.3% (or 1-in-3 if wanting to be perfectly precise). The alternative hypothesis is that this proportion is something other than 33.3%. While it's helpful to write these hypotheses in words, it can be useful to write them using mathematical notation:

$H_0$**:** $p = 0.333$

$H_A$**:** $p \neq 0.333$

In this hypothesis setup, we want to make a conclusion about the population parameter $p$. The value we are comparing the parameter to is called the **null value**, which in this case is 0.333. It's common to label the null value with the same symbol as the parameter but with a subscript '0'. That is, in this case, the null value is $p_0 = 0.333$ (pronounced "p-nought equals 0.333").

**EXAMPLE 5.18**

It may seem impossible that the proportion of people who get the correct answer is *exactly* 33.3%. If we don't believe the null hypothesis, should we simply reject it?

---

E No. While we may not buy into the notion that the proportion is exactly 33.3%, the hypothesis testing framework requires that there be strong evidence before we reject the null hypothesis and conclude something more interesting.

After all, even if we don't believe the proportion is *exactly* 33.3%, that doesn't really tell us anything useful! We would still be stuck with the original question: do people do better or worse than random guessing on Roslings' question? Without data that strongly points in one direction or the other, it is both uninteresting and pointless to reject $H_0$.

**GUIDED PRACTICE 5.19**

G Another example of a real-world hypothesis testing situation is evaluating whether a new drug is better or worse than an existing drug at treating a particular disease. What should we use for the null and alternative hypotheses in this case?[15]

[14] The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

[15] The null hypothesis ($H_0$) in this case is the declaration of *no difference*: the drugs are equally effective. The alternative hypothesis ($H_A$) is that the new drug performs differently than the original, i.e. it could perform better or worse.

### 5.3.2 Testing hypotheses using confidence intervals

We will use the `rosling_responses` data set to evaluate the hypothesis test evaluating whether college-educated adults who get the question about infant vaccination correct is different from 33.3%. This data set summarizes the answers of 50 college-educated adults. Of these 50 adults, 24% of respondents got the question correct that 80% of 1 year olds have been vaccinated against some disease.

Up until now, our discussion has been philosophical. However, now that we have data, we might ask ourselves: does the data provide strong evidence that the proportion of all college-educated adults who would answer this question correctly is different than 33.3%?

We learned in Section 5.1 that there is fluctuation from one sample to another, and it is unlikely that our sample proportion, $\hat{p}$, will exactly equal $p$, but we want to make a conclusion about $p$. We have a nagging concern: is this deviation of 24% from 33.3% simply due to chance, or does the data provide strong evidence that the population proportion is different from 33.3%?

In Section 5.2, we learned how to quantify the uncertainty in our estimate using confidence intervals. The same method for measuring variability can be useful for the hypothesis test.

**EXAMPLE 5.20**

Check whether it is reasonable to construct a confidence interval for $p$ using the sample data, and if so, construct a 95% confidence interval.

---

The conditions are met for $\hat{p}$ to be approximately normal: the data come from a simple random sample (satisfies independence), and $n\hat{p} = 12$ and $n(1-\hat{p}) = 38$ are both at least 10 (success-failure condition).

(E) To construct the confidence interval, we will need to identify the point estimate ($\hat{p} = 0.24$), the critical value for the 95% confidence level ($z^\star = 1.96$), and the standard error of $\hat{p}$ ($SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = 0.060$). With those pieces, the confidence interval for $p$ can be constructed:

$$\hat{p} \pm z^\star \times SE_{\hat{p}}$$
$$0.24 \pm 1.96 \times 0.060$$
$$(0.122, 0.358)$$

We are 95% confident that the proportion of all college-educated adults to correctly answer this particular question about infant vaccination is between 12.2% and 35.8%.

Because the null value in the hypothesis test is $p_0 = 0.333$, which falls within the range of plausible values from the confidence interval, we cannot say the null value is implausible.[16] That is, the data do not provide sufficient evidence to reject the notion that the performance of college-educated adults was different than random guessing, and we do not reject the null hypothesis, $H_0$.

**EXAMPLE 5.21**

Explain why we cannot conclude that college-educated adults simply guessed on the infant vaccination question.

---

(E) While we failed to reject $H_0$, that does not necessarily mean the null hypothesis is true. Perhaps there was an actual difference, but we were not able to detect it with the relatively small sample of 50.

**DOUBLE NEGATIVES CAN SOMETIMES BE USED IN STATISTICS**

In many statistical explanations, we use double negatives. For instance, we might say that the null hypothesis is *not implausible* or we *failed to reject* the null hypothesis. Double negatives are used to communicate that while we are not rejecting a position, we are also not saying it is correct.

[16] Arguably this method is slightly imprecise. As we'll see in a few pages, the standard error is often computed slightly differently in the context of a hypothesis test for a proportion.

**GUIDED PRACTICE 5.22**

Let's move onto a second question posed by the Roslings:

> *There are 2 billion children in the world today aged 0-15 years old, how many children will there be in year 2100 according to the United Nations?*
>
> *a. 4 billion.*
> *b. 3 billion.*
> *c. 2 billion.*

G

Set up appropriate hypotheses to evaluate whether college-educated adults are better than random guessing on this question. Also, see if you can guess the correct answer before checking the answer in the footnote![17]

**GUIDED PRACTICE 5.23**

G

This time we took a larger sample of 228 college-educated adults, 34 (14.9%) selected the correct answer to the question in Guided Practice 5.22: 2 billion. Can we model the sample proportion using a normal distribution and construct a confidence interval?[18]

**EXAMPLE 5.24**

Compute a 95% confidence interval for the fraction of college-educated adults who answered the children-in-2100 question correctly, and evaluate the hypotheses in Guided Practice 5.22.

---

To compute the standard error, we'll again use $\hat{p}$ in place of $p$ for the calculation:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.149(1-0.149)}{228}} = 0.024$$

In Guided Practice 5.23, we found that $\hat{p}$ can be modeled using a normal distribution, which ensures a 95% confidence interval may be accurately constructed as

E

$$\hat{p} \pm z^\star \times SE \quad \rightarrow \quad 0.149 \pm 1.96 \times 0.024 \quad \rightarrow \quad (0.103, 0.195)$$

Because the null value, $p_0 = 0.333$, is not in the confidence interval, a population proportion of 0.333 is implausible and we reject the null hypothesis. That is, the data provide statistically significant evidence that the actual proportion of college adults who get the children-in-2100 question correct is different from random guessing. Because the entire 95% confidence interval is below 0.333, we can conclude college-educated adults do *worse* than random guessing on this question.

One subtle consideration is that we used a 95% confidence interval. What if we had used a 99% confidence level? Or even a 99.9% confidence level? It's possible to come to a different conclusion if using a different confidence level. Therefore, when we make a conclusion based on confidence interval, we should also be sure it is clear what confidence level we used.

The worse-than-random performance on this last question is not a fluke: there are many such world health questions where people do worse than random guessing. In general, the answers suggest that people tend to be more pessimistic about progress than reality suggests. This topic is discussed in much greater detail in the Roslings' book, *Factfulness*.

---

[17] The appropriate hypotheses are:
$H_0$: the proportion who get the answer correct is the same as random guessing: 1-in-3, or $p = 0.333$.
$H_A$: the proportion who get the answer correct is different than random guessing, $p \neq 0.333$.
The correct answer to the question is 2 billion. While the world population is projected to increase, the average age is also expected to rise. That is, the majority of the population growth will happen in older age groups, meaning people are projected to live longer in the future across much of the world.

[18] We check both conditions, which are satisfied, so it is reasonable to use a normal distribution for $\hat{p}$:
**Independence.** Since the data are from a simple random sample, the observations are independent.
**Success-failure.** We'll use $\hat{p}$ in place of $p$ to check: $n\hat{p} = 34$ and $n(1-\hat{p}) = 194$. Both are greater than 10, so the success-failure condition is satisfied.

### 5.3.3 Decision errors

Hypothesis tests are not flawless: we can make an incorrect decision in a statistical hypothesis test based on the data. For example, in the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free. One key distinction with statistical hypothesis tests is that we have the tools necessary to probabilistically quantify how often we make errors in our conclusions.

Recall that there are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios, which are summarized in Figure 5.8.

| | | **Test conclusion** | |
|---|---|---|---|
| | | do not reject $H_0$ | reject $H_0$ in favor of $H_A$ |
| **Truth** | $H_0$ true | okay | **Type 1 Error** |
| | $H_A$ true | **Type 2 Error** | okay |

Figure 5.8: Four different scenarios for hypothesis tests.

A **Type 1 Error** is rejecting the null hypothesis when $H_0$ is actually true. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

**GUIDED PRACTICE 5.25**

G In a US court, the defendant is either innocent ($H_0$) or guilty ($H_A$). What does a Type 1 Error represent in this context? What does a Type 2 Error represent? Figure 5.8 may be useful.[19]

**EXAMPLE 5.26**

How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?

E To lower the Type 1 Error rate, we might raise our standard for conviction from "beyond a reasonable doubt" to "beyond a conceivable doubt" so fewer people would be wrongly convicted. However, this would also make it more difficult to convict the people who are actually guilty, so we would make more Type 2 Errors.

**GUIDED PRACTICE 5.27**

G How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[20]

Exercises 5.25-5.27 provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject $H_0$ unless we have strong evidence. But what precisely does *strong evidence* mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject $H_0$ more than 5% of the time. This corresponds to a **significance level** of 0.05. That is, if the null hypothesis is true, the significance level indicates how often the data lead us to incorrectly reject $H_0$. We often write the significance level using $\alpha$ (the Greek letter *alpha*): $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 5.3.5.

[19] If the court makes a Type 1 Error, this means the defendant is innocent ($H_0$ true) but wrongly convicted. Note that a Type 1 Error is only possible if we've rejected the null hypothesis.

A Type 2 Error means the court failed to reject $H_0$ (i.e. failed to convict the person) when she was in fact guilty ($H_A$ true). Note that a Type 2 Error is only possible if we have failed to reject the null hypothesis.

[20] To lower the Type 2 Error rate, we want to convict more guilty people. We could lower the standards for conviction from "beyond a reasonable doubt" to "beyond a little doubt". Lowering the bar for guilt will also result in more wrongful convictions, raising the Type 1 Error rate.

If we use a 95% confidence interval to evaluate a hypothesis test and the null hypothesis happens to be true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter. This happens about 5% of the time (2.5% in each tail). Similarly, using a 99% confidence interval to evaluate a hypothesis is equivalent to a significance level of $\alpha = 0.01$.

A confidence interval is very helpful in determining whether or not to reject the null hypothesis. However, the confidence interval approach isn't always sustainable. In several sections, we will encounter situations where a confidence interval cannot be constructed. For example, if we wanted to evaluate the hypothesis that several proportions are equal, it isn't clear how to construct and compare many confidence intervals altogether.

Next we will introduce a statistic called the *p-value* to help us expand our statistical toolkit, which will enable us to both better understand the strength of evidence and work in more complex data scenarios in later sections.

### 5.3.4 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative hypothesis. Statistical hypothesis testing typically uses the p-value method rather than making a decision based on confidence intervals.

**P-VALUE**

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true. We typically use a summary statistic of the data, in this section the sample proportion, to help compute the p-value and evaluate the hypotheses.

**EXAMPLE 5.28**

Pew Research asked a random sample of 1000 American adults whether they supported the increased usage of coal to produce energy. Set up hypotheses to evaluate whether a majority of American adults support or oppose the increased usage of coal.

(E) The uninteresting result is that there is no majority either way: half of Americans support and the other half oppose expanding the use of coal to produce energy. The alternative hypothesis would be that there is a majority support or oppose (though we do not known which one!) expanding the use of coal. If $p$ represents the proportion supporting, then we can write the hypotheses as

$H_0$**:** $p = 0.5$

$H_A$**:** $p \neq 0.5$

In this case, the null value is $p_0 = 0.5$.

When evaluating hypotheses for proportions using the p-value method, we will slightly modify how we check the success-failure condition and compute the standard error for the single proportion case. These changes aren't dramatic, but pay close attention to how we use the null value, $p_0$.

**EXAMPLE 5.29**

Pew Research's sample show that 37% of American adults support increased usage of coal. We now wonder, does 37% represent a real difference from the null hypothesis of 50%? What would the sampling distribution of $\hat{p}$ look like if the null hypothesis were true?

---

If the null hypothesis were true, the population proportion would be the null value, 0.5. We previously learned that the sampling distribution of $\hat{p}$ will be normal when two conditions are met:

**Independence.** The poll was based on a simple random sample, so independence is satisfied.

**Success-failure.** Based on the poll's sample size of $n = 1000$, the success-failure condition is met, since

$$np \overset{H_0}{=} 1000 \times 0.5 = 500 \qquad n(1-p) \overset{H_0}{=} 1000 \times (1-0.5) = 500$$

(E) are both at least 10. Note that the success-failure condition was checked using the null value, $p_0 = 0.5$; this is the first procedural difference from confidence intervals.

If the null hypothesis were true, the sampling distribution indicates that a sample proportion based on $n = 1000$ observations would be normally distributed. Next, we can compute the standard error, where we will again use the null value $p_0 = 0.5$ in the calculation:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \overset{H_0}{=} \sqrt{\frac{0.5 \times (1-0.5)}{1000}} = 0.016$$

This marks the other procedural difference from confidence intervals: since the sampling distribution is determined under the null proportion, the null value $p_0$ was used for the proportion in the calculation rather than $\hat{p}$.

Ultimately, if the null hypothesis were true, then the sample proportion should follow a normal distribution with mean 0.5 and a standard error of 0.016. This distribution is shown in Figure 5.9.

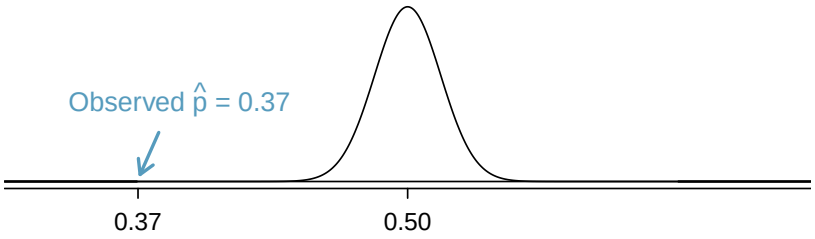



Figure 5.9: If the null hypothesis were true, this normal distribution describes the distribution of $\hat{p}$.

---

**CHECKING SUCCESS-FAILURE AND COMPUTING $SE_{\hat{p}}$ FOR A HYPOTHESIS TEST**

When using the p-value method to evaluate a hypothesis test, we check the conditions for $\hat{p}$ and construct the standard error using the null value, $p_0$, instead of using the sample proportion.

In a hypothesis test with a p-value, we are supposing the null hypothesis is true, which is a different mindset than when we compute a confidence interval. This is why we use $p_0$ instead of $\hat{p}$ when we check conditions and compute the standard error in this context.

---

When we identify the sampling distribution under the null hypothesis, it has a special name: the **null distribution**. The p-value represents the probability of the observed $\hat{p}$, or a $\hat{p}$ that is more extreme, if the null hypothesis were true. To find the p-value, we generally find the null distribution, and then we find a tail area in that distribution corresponding to our point estimate.

**EXAMPLE 5.30**

If the null hypothesis were true, determine the chance of finding $\hat{p}$ at least as far into the tails as 0.37 under the null distribution, which is a normal distribution with mean $\mu = 0.5$ and $SE = 0.016$.

---

This is a normal probability problem where $x = 0.37$. First, we draw a simple graph to represent the situation, similar to what is shown in Figure 5.9. Since $\hat{p}$ is so far out in the tail, we know the tail area is going to be very small. To find it, we start by computing the Z-score using the mean of 0.5 and the standard error of 0.016:

$$Z = \frac{0.37 - 0.5}{0.016} = -8.125$$

(E)

We can use software to find the tail area: $2.2 \times 10^{-16}$ (0.00000000000000022). If using the normal probability table in Appendix C.1, we'd find that $Z = -8.125$ is off the table, so we would use the smallest area listed: 0.0002.

The potential $\hat{p}$'s in the upper tail beyond 0.63, which are shown in Figure 5.10, also represent observations at least as extreme as the observed value of 0.37. To account for these values that are also more extreme under the hypothesis setup, we double the lower tail to get an estimate of the p-value: $4.4 \times 10^{-16}$ (or if using the table method, 0.0004).

The p-value represents the probability of observing such an extreme sample proportion by chance, if the null hypothesis were true.

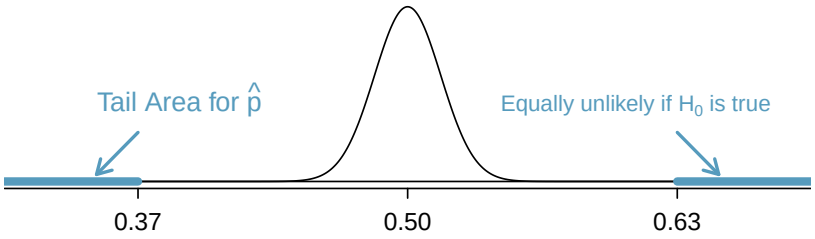



Figure 5.10: If $H_0$ were true, then the values above 0.63 are just as unlikely as values below 0.37.

**EXAMPLE 5.31**

How should we evaluate the hypotheses using the p-value of $4.4 \times 10^{-16}$? Use the standard significance level of $\alpha = 0.05$.

---

If the null hypothesis were true, there's only an incredibly small chance of observing such an extreme deviation of $\hat{p}$ from 0.5. This means one of the following must be true:

(E)

1. The null hypothesis is true, and we just happened to get observe something so extreme that only happens about once in every 23 quadrillion times (1 quadrillion = 1 million × 1 billion).
2. The alternative hypothesis is true, which would be consistent with observing a sample proportion far from 0.5.

The first scenario is laughably improbable, while the second scenario seems much more plausible.

Formally, when we evaluate a hypothesis test, we compare the p-value to the significance level, which in this case is $\alpha = 0.05$. Since the p-value is less than $\alpha$, we reject the null hypothesis. That is, the data provide strong evidence against $H_0$. The data indicate the direction of the difference: a majority of Americans do not support expanding the use of coal-powered energy.

**COMPARE THE P-VALUE TO $\alpha$ TO EVALUATE $H_0$**

When the p-value is less than the significance level, $\alpha$, reject $H_0$. We would report a conclusion that the data provide strong evidence supporting the alternative hypothesis.

When the p-value is greater than $\alpha$, do not reject $H_0$, and report that we do not have sufficient evidence to reject the null hypothesis.

In either case, it is important to describe the conclusion in the context of the data.

**GUIDED PRACTICE 5.32**

(G) Do a majority of Americans support or oppose nuclear arms reduction? Set up hypotheses to evaluate this question.[21]

**EXAMPLE 5.33**

A simple random sample of 1028 US adults in March 2013 show that 56% support nuclear arms reduction. Does this provide convincing evidence that a majority of Americans supported nuclear arms reduction at the 5% significance level?

First, check conditions:

**Independence.** The poll was of a simple random sample of US adults, meaning the observations are independent.

**Success-failure.** In a one-proportion hypothesis test, this condition is checked using the null proportion, which is $p_0 = 0.5$ in this context: $np_0 = n(1 - p_0) = 1028 \times 0.5 = 514 \geq 10$.

With these conditions verified, we can model $\hat{p}$ using a normal model.

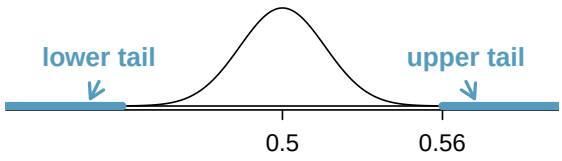
Next the standard error can be computed. The null value $p_0$ is used again here, because this is a hypothesis test for a single proportion.

(E)
$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1028}} = 0.0156$$

Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.56 - 0.50}{0.0156} = 3.85$$

It's generally helpful to draw null distribution and the tail areas of interest for computing the p-value:




The upper tail area is about 0.0001, and we double this tail area to get the p-value: 0.0002. Because the p-value is smaller than 0.05, we reject $H_0$. The poll provides convincing evidence that a majority of Americans supported nuclear arms reduction efforts in March 2013.

[21] We would like to understand if a majority supports or opposes, or ultimately, if there is no difference. If $p$ is the proportion of Americans who support nuclear arms reduction, then $H_0$: $p = 0.50$ and $H_A$: $p \neq 0.50$.

**HYPOTHESIS TESTING FOR A SINGLE PROPORTION**

Once you've determined a one-proportion hypothesis test is the correct procedure, there are four steps to completing the test:

**Prepare.** Identify the parameter of interest, list hypotheses, identify the significance level, and identify $\hat{p}$ and $n$.

**Check.** Verify conditions to ensure $\hat{p}$ is nearly normal under $H_0$. For one-proportion hypothesis tests, use the null value to check the success-failure condition.

**Calculate.** If the conditions hold, compute the standard error, again using $p_0$, compute the Z-score, and identify the p-value.

**Conclude.** Evaluate the hypothesis test by comparing the p-value to $\alpha$, and provide a conclusion in the context of the problem.

### 5.3.5 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is $\alpha = 0.05$. However, it can be helpful to adjust the significance level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring $H_A$ before we would reject $H_0$.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we might choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject $H_0$ when the alternative hypothesis is actually true.

Additionally, if the cost of collecting data is small relative to the cost of a Type 2 Error, then it may also be a good strategy to collect more data. Under this strategy, the Type 2 Error can be reduced while not affecting the Type 1 Error rate. Of course, collecting extra data is often costly, so there is typically a cost-benefit analysis to be considered.

**EXAMPLE 5.34**

A car manufacturer is considering switching to a new, higher quality piece of equipment that constructs vehicle door hinges. They figure that they will save money in the long run if this new machine produces hinges that have flaws less than 0.2% of the time. However, if the hinges are flawed more than 0.2% of the time, they wouldn't get a good enough return-on-investment from the new piece of equipment, and they would lose money. Is there good reason to modify the significance level in such a hypothesis test?

(E)

The null hypothesis would be that the rate of flawed hinges is 0.2%, while the alternative is that it the rate is different than 0.2%. This decision is just one of many that have a marginal impact on the car and company. A significance level of 0.05 seems reasonable since neither a Type 1 or Type 2 Error should be dangerous or (relatively) much more expensive.

**EXAMPLE 5.35**

The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not door hinges. If the durability of these safety components is shown to be better than the current supplier, they will switch manufacturers. Is there good reason to modify the significance level in such an evaluation?

(E)

The null hypothesis would be that the suppliers' parts are equally reliable. Because safety is involved, the car company should be eager to switch to the slightly more expensive manufacturer (reject $H_0$), even if the evidence of increased safety is only moderately strong. A slightly larger significance level, such as $\alpha = 0.10$, might be appropriate.

**GUIDED PRACTICE 5.36**

(G) A part inside of a machine is very expensive to replace. However, the machine usually functions properly even if this part is broken, so the part is replaced only if we are extremely certain it is broken based on a series of measurements. Identify appropriate hypotheses for this test (in plain language) and suggest an appropriate significance level.[22]

**WHY IS 0.05 THE DEFAULT?**

The $\alpha = 0.05$ threshold is most common. But why? Maybe the standard level should be smaller, or perhaps larger. If you're a little puzzled, you're reading with an extra critical eye – good job! We've made a 5-minute task to help clarify *why 0.05*:

www.openintro.org/why05

### 5.3.6 Statistical significance versus practical significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample. Sometimes researchers will take such large samples that even the slightest difference is detected, even differences where there is no practical value. In such cases, we still say the difference is **statistically significant**, but it is not **practically significant**. For example, an online experiment might identify that placing additional ads on a movie review website statistically significantly increases viewership of a TV show by 0.001%, but this increase might not have any practical value.

One role of a data scientist in conducting a study often includes planning the size of the study. The data scientist might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain other information, such as a very rough estimate of the true proportion $p$, so that she could roughly estimate the standard error. From here, she can suggest a sample size that is sufficiently large that, if there is a real difference that is meaningful, we could detect it. While larger sample sizes may still be used, these calculations are especially helpful when considering costs or potential risks, such as possible health impacts to volunteers in a medical study.

[22] Here the null hypothesis is that the part is not broken, and the alternative is that it is broken. If we don't have sufficient evidence to reject $H_0$, we would not replace the part. It sounds like failing to fix the part if it is broken ($H_0$ false, $H_A$ true) is not very problematic, and replacing the part is expensive. Thus, we should require very strong evidence against $H_0$ before we replace the part. Choose a small significance level, such as $\alpha = 0.01$.

### 5.3.7 One-sided hypothesis tests (special topic)

So far we've only considered what are called **two-sided hypothesis tests**, where we care about detecting whether $p$ is either above or below some null value $p_0$. There is a second type of hypothesis test called a **one-sided hypothesis test**. For a one-sided hypothesis test, the hypotheses take one of the following forms:

1. There's only value in detecting if the population parameter is *less than* some value $p_0$. In this case, the alternative hypothesis is written as $p < p_0$ for some null value $p_0$.

2. There's only value in detecting if the population parameter is *more than* some value $p_0$: In this case, the alternative hypothesis is written as $p > p_0$.

While we adjust the form of the alternative hypothesis, we continue to write the null hypothesis using an equals-sign in the one-sided hypothesis test case.

In the entire hypothesis testing procedure, there is only one difference in evaluating a one-sided hypothesis test vs a two-sided hypothesis test: how to compute the p-value. In a one-sided hypothesis test, we compute the p-value as the tail area in the *direction of the alternative hypothesis only*, meaning it is represented by a single tail area. Herein lies the reason why one-sided tests are sometimes interesting: if we don't have to double the tail area to get the p-value, then the p-value is smaller and the level of evidence required to identify an interesting finding in the direction of the alternative hypothesis goes down. However, one-sided tests aren't all sunshine and rainbows: the heavy price paid is that any interesting findings in the opposite direction must be disregarded.

**EXAMPLE 5.37**

In Section 1.1, we encountered an example where doctors were interested in determining whether stents would help people who had a high risk of stroke. The researchers believed the stents would help. Unfortunately, the data showed the opposite: patients who received stents actually did worse. Why was using a two-sided test so important in this context?

(E)

Before the study, researchers had reason to believe that stents would help patients since existing research suggested stents helped in patients with heart attacks. It would surely have been tempting to use a one-sided test in this situation, and had they done this, they would have limited their ability to identify potential harm to patients.

Example 5.37 highlights that using a one-sided hypothesis creates a risk of overlooking data supporting the opposite conclusion. We could have made a similar error when reviewing the Roslings' question data this section; if we had a pre-conceived notion that college-educated people wouldn't do worse than random guessing and so used a one-sided test, we would have missed the really interesting finding that many people have incorrect knowledge about global public health.

When might a one-sided test be appropriate to use? *Very rarely.* Should you ever find yourself considering using a one-sided test, carefully answer the following question:

> *What would I, or others, conclude if the data happens to go clearly in the opposite direction than my alternative hypothesis?*

If you or others would find any value in making a conclusion about the data that goes in the opposite direction of a one-sided test, then a two-sided hypothesis test should actually be used. These considerations can be subtle, so exercise caution. We will only apply two-sided tests in the rest of this book.

**EXAMPLE 5.38**

Why can't we simply run a one-sided test that goes in the direction of the data?

---

We've been building a careful framework that controls for the Type 1 Error, which is the significance level $\alpha$ in a hypothesis test. We'll use the $\alpha = 0.05$ below to keep things simple.

Imagine we could pick the one-sided test after we saw the data. What will go wrong?

(E)

- If $\hat{p}$ is *smaller* than the null value, then a one-sided test where $p < p_0$ would mean that any observation in the *lower* 5% tail of the null distribution would lead to us rejecting $H_0$.
- If $\hat{p}$ is *larger* than the null value, then a one-sided test where $p > p_0$ would mean that any observation in the *upper* 5% tail of the null distribution would lead to us rejecting $H_0$.

Then if $H_0$ were true, there's a 10% chance of being in one of the two tails, so our testing error is actually $\alpha = 0.10$, not 0.05. That is, not being careful about when to use one-sided tests effectively undermines the methods we're working so hard to develop and utilize.