Hypothesis Testing for a Proportion

August 21, 2019

Hypothesis Testing Framework

Suppose we're interested in examining how people perform on a multiple choice question related to world health. We might like to understand if

$\boldsymbol{H_0}$: People never learn these topics and their responses are random guesses.

$\boldsymbol{H_A}$: People have knowledge that helps them do better than random guessing, or perhaps have false knowledge that leads them to do worse than random guessing.

Hypotheses

We talked briefly about hypothesis before! Recall that

- H_0 is the **null hypothesis**.
- H_A is the **alternative hypothesis**.

Hypotheses

- The null hypothesis represents a skeptical perspective or a perspective of "no difference". This is the claim to be tested.
- The alternative hypothesis is some new, alternate claim. It is often represented by a range of possible values.

We will define these more precisely as we go.

Hypotheses

Let's return to our example about a world health question.

- Suppose there are 4 possible answers and only 1 correct answer.
- The responses being random guesses corresponds to

$$H_0 : p = \frac{1}{4}$$

- The responses relating to some knowledge (whether correct or incorrect) corresponds to

$$H_A : p \neq \frac{1}{4}$$

Hypotheses

The alternative hypothesis usually represents a new or stronger perspective.

- It would be interesting to know that people know something about world health (if in fact $p > 1/4$).
- It would also be interesting to know if people have misleading information about world health (if in fact $p < 1/4$).

Hypothesis Testing

The hypothesis testing framework is very general!

- Any time someone makes a claim that's difficult to believe, we start by being skeptical.
- If enough evidence is presented to support that claim, we may reject our skeptical position and change our minds.

Example: Juries

A jury on a criminal case makes two possible decisions: innocent or guilty.

In principle, the US court system operates under "innocent until proven guilty".

How might we set this up in a formal hypothesis framework?

Example: Juries

If a person is innocent until proven guilty, our default assumption should be that the person is innocent:

$$\boldsymbol{H_0}: \text{ the defendant is innocent.}$$

We should be skeptical of the claim that a person is guilty, concluding guilt only if we are convinced beyond a reasonable doubt:

$$\boldsymbol{H_A}: \text{ the defendant is guilty.}$$

Example: Juries

- Crucially, even if we aren't convinced that a person is innocent, we may still fail to convict.
- That is, we may fail to convict because we are unsure.
- This is because a jury's decision is based on our being overwhelmingly convinced of *guilt*, not of innocence.
- The prosecutor may fail to provide enough evidence to convince us of guilt, but that doesn't necessarily mean that the defendant is innocent.

Hypothesis Testing

The jury framework is a lot like hypothesis testing:

- We may find sufficient evidence to reject the null hypothesis.
- We may also not find sufficient evidence to reject the null hypothesis.
- However, even if we lack this evidence, we typically do not accept the null hypothesis as true.
- Failing to find sufficient evidence for the alternative hypothesis does not necessarily mean that the null hypothesis is true!

Hypotheses

Let's return to our example about a world health question.

- Recall that
$$H_0 : p = \frac{1}{4}$$
- and
$$H_A : p \neq \frac{1}{4}.$$

The Null Value

- In this setting, we want to know something about the population parameter p.
- We compare this to the value 0.25, called the **null value**.
- We denote the null value by p_0 ("p-nought"). Here, $p_0 = 0.25$.

Example

- It may seem impossible that the proportion of people who get the right answer is *exactly* chance level ($p = 0.25$).
- However, recall that our framework requires that there be strong evidence in order to reject this notion.
- We are not trying to conclude that $p = 0.25$ (we don't tend to conclude the null hypothesis).
- If the proportion is 0.2501 rather than exactly 0.25, we haven't really learned anything interesting.

Hypothesis Testing Using Confidence Intervals

We will use the `Rosling responses` data set to evaluate the hypothesis test evaluating whether college-educated adults get a question about infant vaccination correct.

The question posed is: *How many of the world's 1 year old children today have been vaccinated against some disease?*

1. 20%
2. 50%
3. 80%

Example

- We want to know if the proportion of college-educated adults who get the question correct is different from 33.3%.
- The data set summarizes the answers of 50 college-educated adults.
- Of these 50 adults, 24% of respondents got the question correct (80% of 1 year olds have been vaccinated against some disease).

Example

Now that we have data, we might wonder if the data provide strong evidence that the proportion of college-educated adults is different than 33.3%.

- We know that there is fluctuation from one sample to another.
- We also know that it is unlikely that $\hat{p}$ will exactly equal p.
- Still, we want to draw a conclusion about p.

Example

We need to know if our sample statistic $\hat{p} = 0.24$

- suggests that the true proportion is something other than $p = 0.333$

OR

- if this deviation is due to random chance.

We know how to quantify the uncertainty in our estimate using confidence intervals. How can we apply this concept to hypothesis tests?

Example

Construct a 95% confidence interval for p using the `Rosling responses` data.

Example

First we need to confirm that the Central Limit Theorem applies to this data.

$$n\hat{p} = 50 \times 0.24 = 12 \geq 10$$

and

$$n(1 - \hat{p}) = 50 \times 0.76 = 38 \geq 10$$

The success-failure condition holds, so we can move on to building our interval.

Example

- The point estimate is $\hat{p} = 0.24$.
- $\alpha = 1 - 0.95 = 0.05$
- The critical value is $z_{0.05/2} = z_{0.025} = 1.96$
- The standard error is

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.060$$

Example

Then

$$\hat{p} \pm z_{\alpha/2} \times SE_{\hat{p}}$$
$$0.24 \pm 1.96 \times 0.060$$

which is the interval $(0.122, 0.358)$.

We can be 95% confident that the proportion of college-educated adults to correctly answer the infant vaccination question is between 12.2% and 35.8%.

Hypothesis Testing Using Confidence Intervals

So we have a confidence interval... now what?

Our interval is $(0.122, 0.358)$.

- We are interested in the null value $p_0 = 0.333$.
- Notice that $p_0 = 0.333$ falls within our interval.
- Therefore $p_0 = 0.333$ is in our range of plausible values.

Since $p_0 = 0.333$ is one of our plausible values, we cannot say that the null value is implausible.

Example

Note that we cannot make the claim that college-educated adults simply guess on this question!

- Failing to reject H_0 is not the same thing as concluding H_0.
- There are still lots of other plausible values that are different from $p_0 = 0.333$!
- It is possible that there is a difference that we were unable to detect with this particular study.

Double Negatives in Statistics

- We use a lot of double negatives when talking about hypotheses.
- We might say things like
  - "the null hypothesis is not implausible"
  - "we failed to reject the null hypothesis"
- We use these to say that we are not rejecting, but are also not accepting, the null.

Hypothesis Testing Using Confidence Intervals

Essentially, if p_0 is within the interval $\hat{p} \pm MoE$, then we do not reject the null hypothesis.

If p_0 is *not* within the interval $\hat{p} \pm MoE$, then we reject the null hypothesis and conclude the alternative.

Decision Errors

- It is entirely possible that we make the right conclusion based on our data... but the wrong conclusion based on the true (unknown) parameter!
- In our criminal court example, sometimes people are wrongly convicted. Other times, guilty people are not convicted at all.
- Unlike in the courts, statistics gives us the tools to quantify how often we make these sorts of errors.

Decision Errors

- There are two competing hypotheses: null and alternative.
- In a hypothesis test, we make some statement about which might be true.
- There are four possible scenarios. We can
  1. Reject H_0 when H_0 is false.
  2. Fail to reject H_0 when H_0 is true.
  3. Reject H_0 when H_0 is true (error).
  4. Fail to reject H_0 when H_0 is false (error).

Decision Errors

		Test Conclusion	
		Do not reject H_0	Reject H_0
Truth	H_0 true	Correct Decision	**Type I Error**
	H_0 false	**Type II Error**	Correct Decision

- A **Type 1 Error** is rejecting H_0 when it is actually true.
- A **Type 2 Error** is failing to reject H_0 when the H_A is actually true.

Example

Let's think about our criminal court example. Recall that the null hypothesis is innocence.

- A Type I error is when we decide that a person is guilty, even though they are innocent.
- A Type II error is when we decide that we do not have enough evidence to say that someone is guilty, but they are in fact guilty.

Example

How could we reduce the Type 1 Error rate in US criminal courts?

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.

Example

What influence might this have on the Type 2 Error rate?

Raising our standard for conviction would also make it more difficult to convict the people who are actually guilty, so we would make more Type 2 Errors.

Error Trade-Offs

- In general, reducing the Type I error rate increases the Type II error rate.
- Similarly, reducing the Type II error rate increases the Type I error rate.
- We see a lot of these trade-offs in statistics.

Decision Errors

- Hypothesis testing is built around rejecting or failing to reject the null hypothesis.
- But when do we have "strong enough" evidence?
- We usually build our tests around Type I error.
- If the null is actually true, we do not want to incorrectly reject any more than, say 5% of the time.
- This corresponds to a significance level of $\alpha = 0.05$

Significance Levels

We talked about significance level α in our discussion about confidence intervals. It comes into play again here!

- The significance level indicates how often the data will lead us to incorrectly reject H_0
- This is also how often we commit a Type I error!
- In fact, α is the probability of committing such an error

$$\alpha = P(\text{Type I error})$$

Significance Levels

If we use a 95% confidence interval for hypothesis testing and the null is true,

- The significance level is $\alpha = 0.05$.
- We 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

Hypothesis Testing Using Confidence Intervals

- Confidence intervals can be very useful in hypothesis testing.
- However, sometimes we are unable to construct a confidence interval.
- For example, what if we wanted to consider something like

$$H_0 : p_1 = p_2 = p_3 = p_4$$

- Therefore we want to develop a more general hypothesis testing framework.

Formal Testing Using P-Values and Test Statistics

- We want a way to consider the strength of the evidence against the null hypothesis and in favor of the alternative hypothesis.
- Instead of using confidence intervals, we use:
  - p-values.
  - test statistics.

P-Values

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.

Test Statistics

- A **test statistic** is a value based on the sample data.
- This is the z-score for the point estimate.
- The test statistic can be used to find a p-value (and vice versa).

In a hypothesis testing framework, using the test statistic and using the p-value are equivalent.

Critical Value

- We used critical values before when building confidence intervals:

$$z_{\alpha/2}$$

- Critical values in the hypothesis testing framework are the same idea.
- If the null hypothesis is true, the **critical value** corresponds to the maximum amount of Type I error allowed.

Example: Coal

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.

Example: Coal

- Let p be the true proportion who support coal.
- The uninteresting result is that there is no majority either way.
- In this case, half would support and half would oppose ($p_0 = 0.5$).
- Alternatively, there is a majority support or oppose.

$$H_0 : \quad p = 0.5$$
$$H_A : \quad p \neq 0.5$$

Hypothesis Testing

- We want to work with the normal distribution, so we need to check our success-failure condition.
- Whenever we use the Central Limit Theorem, we want to use the true parameter but typically don't have it.
- With hypothesis testing, p_0 is the *proposed* value for p.
- We will therefore use p_0 in place of p in our plug in method.

Hypothesis Testing

We use p_0 in place of p for good reason:

- We are interested in how unlikely our observed statistic is *under the condition* that the null hypothesis is true.
- If the null hypothesis is true, then $p = p_0$.

Example: Coal

What would the sampling distribution of $\hat{p}$ look like if the null hypothesis were true?

We assume that our poll is based on a random sample, so independence is satisfied. Using p_0 to check our success-failure condition,

$$np_0 = n(1 - p_0) \overset{H_0}{=} 1000 \times 0.5 = 500$$

So we are comfortable working with a normal distribution.

Example: Coal

Under the null hypothesis, the normal distribution for this context has mean

$$\mu \overset{H_0}{=} p_0 = 0.5$$

and standard error

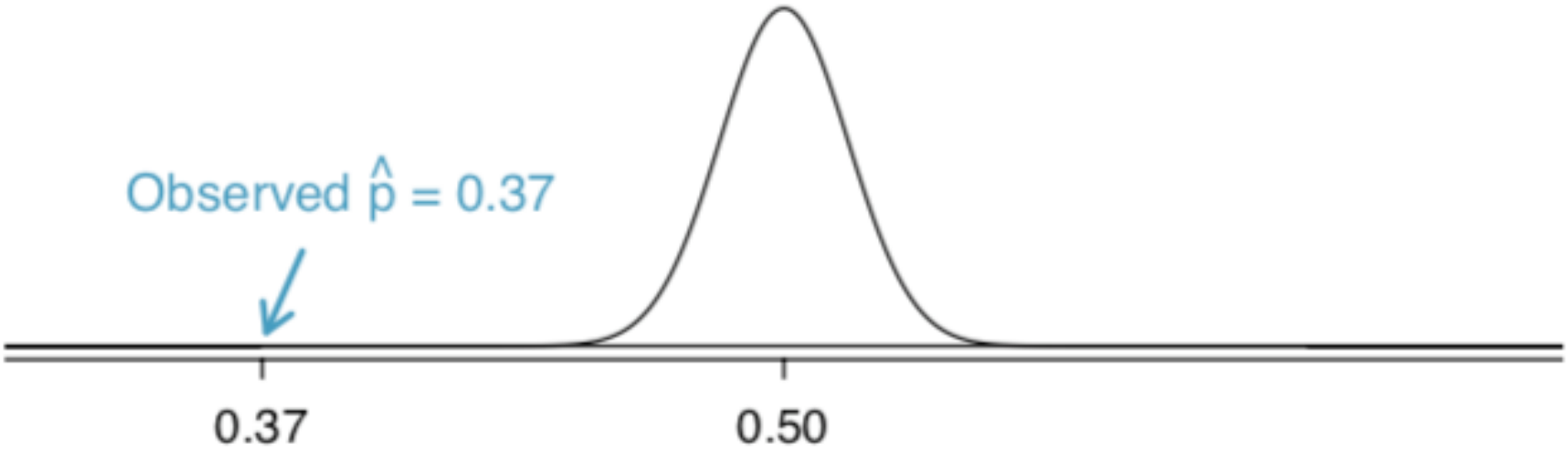
$$SE \overset{H_0}{=} \sqrt{\frac{p_0(1-p_0)}{n}} = 0.016$$

Under the null hypothesis, $\hat{p} \sim N(0.5, 0.016)$.

Example: Coal

- Pew Research's sample suggests that 37% of American adults support increased usage of coal.
- Does 37% represent a real difference from the null hypothesis of 50%?

Example: Coal



This is the sampling distribution under the null hypothesis. We call this the **null distribution**.

Example: Coal

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 $\hat{p} \sim N(0.5, 0.016)$.

Example: Coal

- This is a normal probability problem where $x = 0.37$.
- First, we draw a simple graph to represent the situation.
- We know that $\hat{p}$ is far in the tail, so the z-score should be far from 0.
- Equivalently, this tail area should be quite small.

Example: Coal

This Z-score is our test statistic.

$$\begin{aligned} ts = z &= \frac{\hat{p} - p_0}{SE} \\ &= \frac{0.37 - 0.5}{0.016} \\ &= -8.125 \end{aligned}$$

The observed proportion of 0.37 is over 8 standard deviations below the mean! If the null distribution were true, there would be almost no chance of seeing such an extreme observation.

Example: Coal

To find the p-value, we find the corresponding tail area.

- Using software, $P(Z < -8.125) = 2.2 \times 10^{-16}$.
- To account for values as least as extreme in the other tail area, we double this value.

$$2 \times P(Z < -8.125) = 4.4 \times 10^{-16}.$$

This means that there is essentially no chance that we would see a proportion of 0.34 in a sample size of 1000 if the null distribution were true!

Calculating a Test Statistic

In general, for proportions where the Central Limit Theorem holds, the test statistic is

$$ts = z = \frac{\hat{p} - p_0}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Calculating a P-Value

Once you've calculated the test statistic, the p-value is

$$2 \times P(|Z| > |ts|)$$

Hypothesis Testing Using Test Statistics

We compare the test statistic to the critical value to evaluate H_0.

When the test statistic is more extreme than the critical value,

$$|ts| > |z_{\alpha/2}|$$

we reject H_0. Otherwise, we do not reject H_0.

Hypothesis Testing Using P-Values

Equivalently, we may compare the p-value to α to evaluate H_0.

When the p-value is less than the significance level,

$$\text{p-value} < \alpha$$

we reject H_0. Otherwise, we do not reject H_0.

Hypothesis Testing

If either

$$|ts| > |z_{\alpha/2}|$$

or

$$\text{p-value} < \alpha$$

The data provide strong evidence supporting the alternative hypothesis.

Otherwise, we report that we do not have sufficient evidence to reject the null hypothesis. We will always describe the conclusion in the context of the data.

Example

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?

Example

Checking our conditions for normality,

- Independence: this is a simple random sample.
- Success-failure:

$$np_0 = n(1 - p_0) = 514 \geq 10$$

So we can model $\hat{p}$ using a normal distribution.

Example

Now we want to calculate the standard error:

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5 \times 0.5}{1028}} = 0.0156$$

Example: Test Statistic Approach

The test statistic can be computed in terms of our null model:

$$ts = z = \frac{\hat{p} - p_0}{SE} = \frac{0.56 - 0.5}{0.0156} = 3.75$$

The critical value for $\alpha = 0.05$ is $z_{0.05/2} = 1.64$. Since

$$|3.75| > |1.96|$$

we can reject H_0 at the $\alpha = 0.05$ level of significance and conclude that a majority of Americans support nuclear arms reduction.

Example: P-Value Approach

The p-value is the probability of being more extreme than the observed test statistic. We should draw a picture. Then using software:

$$2 \times P(Z > 3.75) = 0.0002$$

Since

$$\text{p-value} = 0.0002 < \alpha = 0.05$$

we can reject H_0 at the $\alpha = 0.05$ level of significance and conclude that a majority of Americans support nuclear arms reduction.

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:

1. Identify the parameter of interest, list hypotheses, identify the significance level, and identify $\hat{p}$ and n.
2. Verify that $\hat{p}$ is nearly normal under H_0. Use the null value in place of p.
3. If the conditions hold, compute the standard error under H_0, compute the Z-score, and (optionally) identify the p-value.
4. Evaluate by either comparing ts to $z_{\alpha/2}$ or p-value to α.

Make sure to provide your conclusion in the context of the problem!