We want to know if the percent of families who eat together has changed or if more than half of our population support capital punishment.

General Education Statistics
Class Notes
Hypothesis Tests for Population Proportions (Section 10.2)

In 2001, 38% of adults with children under the age of 18 reported that their family ate dinner together seven nights a week. Has this percentage gone down since 2001?

Lipitor (a cholesterol lowering drug) caused 2.2% of patients in a clinical trial to develop flu-like symptoms. Other cholesterol lowering drugs report that 1.9% of their patients develop flu-like symptoms. Is there evidence that Lipitor causes this reaction more than the competition?

We want to test whether some proportion has changed over time or differs dramatically from another proportion. We will use hypothesis tests that were introduced in the previous section.

Recall: Definition: The **sample proportion,** denoted by $\hat{p}$, is given by $\hat{p} = x/n$ where x is the number of individuals in the sample (of size n) that share some characteristic. The mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$. The standard deviation of the sampling distribution of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.

The statistics we do here is based on the fact that the sample proportion is normally distributed. Otherwise, these hypothesis tests are *not* valid. Remember that requires that we verify the following. For a simple random sample of size n with a population proportion p,

1. the shape of the sampling distribution of $\hat{p}$ is approximately normal if $np(1-p) \geq 10$,

2. the sample is a simple random sample, and

3. the sampled values are independent of each other (or $n \leq .05N$).

Use the *assumed* proportion as *p.*

If we sample adults with children under the age of 18 today and we find that 34% report eating dinner together seven nights a week, is this evidence that the *population* itself has experienced a decrease in this percentage from 2001? Or maybe the sample we took just *happened* to give us a lower number? If we assume the population still eats dinner together at a rate of 38% (as was reported in 2001), but find the sample that got 34% to be **unusual**, we can conclude with some certainty that the population percentage actually has decreased.

We will investigate two methods of performing hypothesis tests. We will look at the **Classical approach** and the ***P*-value approach**.

Recall: Definition: Statistically significant: When observed results are *unlikely* under the assumption that the null hypothesis is true, we say the result is **statistically significant**. The sample result was so far from the assumed parameter's value that the difference is *significant* enough for us to say the assumption (about the population's parameter) was likely wrong. When results are found to be statistically significant, **we *reject* the null hypothesis**.

The Classical Approach:

We will say that the sample evidence is statistically significant (or sufficient) if the sample proportion is too many standard deviations away from the assumed population proportion. Again, this would indicate it was unlikely assuming the null hypothesis.

Summary of the Classical Approach:

Remember we first verify the following.

1. the sample is a simple random sample,
2. $np_0(1-p_0)\geq 10$ where n is the sample size, and
3. the sampled values are independent of each other (or $n \leq .05N$).

We will use the notation p_0 to mean the proportion assumed in the null hypothesis.

Step 1: Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

1. Equal versus *not* equal hypothesis (**two-tailed test**)
   $H_0: p = p_0$
   $H_1: p \neq p_0$

2. Equal versus less than (**left-tailed test**)
   $H_0: p = p_0$
   $H_1: p < p_0$

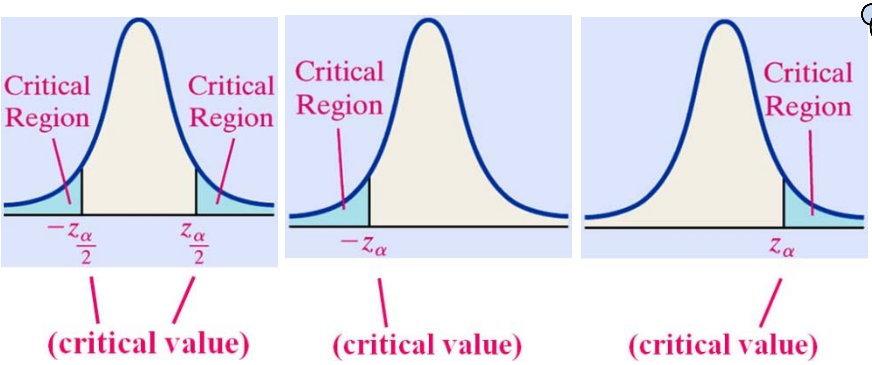
3. Equal versus greater than (**right-tailed test**)
   $H_0: p = p_0$
   $H_1: p > p_0$

Step 2: Select a level of significance, α, depending on the seriousness of making a Type I error.

(continued on next page)

Classical Approach (continued):

Step 3: Compute the **test statistic** $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$. Again, notice we use p_0 in the standard deviation on the bottom of this z-score. Use Table V to determine the critical value, based on α.



Here, we see the critical value in relation to α with two-tailed, left-tailed, and right-tailed tests.

In the above pictures, the shaded area is the level of significance α.

Step 4: Compare the critical value with the test statistic. The most common critical values are below. If z_0 is in the critical region, you *reject* the null hypothesis.

Hypothesis Testing Critical Values

Level of significance, α	**Left-tailed**	**Right-tailed**	**Two-tailed**
0.10 (10%)	–1.28	1.28	±1.645
0.05 (5%)	–1.645	1.645	±1.96
0.025 (2.5%)	–1.96	1.96	±2.24
0.01 (1%)	–2.33	2.33	±2.575

Step 5: State the conclusion. Either we say "There is sufficient evidence to conclude that the population parameter is as it's stated in the alternative hypothesis." (This is a *rejection* of the null hypothesis.) or "There is *not* sufficient evidence to conclude that the population parameter is as it's stated in the alternative hypothesis." (Here, we did *not* reject the null hypothesis.)

expl 1: Let’s take a closer look at the 2001 study of families’ eating habits. Assume the true population percentage is 38%. Does our (fictitious) sample of 300 adults that showed 34% mean that there is a decline in families eating together?

a.) First, determine the null and alternative hypotheses. Label them as H_0 and H_1.

b.) Assume that the sample is a simple random sample and that the sample size is less than 5% of the population size. Also, verify that $np(1-p) \geq 10$ by using $n = 300$ and $p = 0.38$.

c.) Calculate $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ by using $n = 300$ and $p = 0.38$. Also, we know that $\mu_{\hat{p}} = p$.

d.) Find the number of standard deviations below the mean that $\hat{p}$ = 34% is. In other words, find the z-score for $\hat{p}$ = 0.34.

$$z_0 = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

e.) We will use a level of significance of 0.025 (or 2.5%) which gives us a critical value of –1.96. Draw a normal curve with z_0 and this critical value marked. Shade the area under the curve to the left of the critical value, labeling this area as 0.025 (as in the pictures on the previous page.)

f.) We will say that the sample evidence is statistically significant if the sample proportion is 1.96 or more standard deviations away from the assumed population proportion. Look at your picture to see if it is. Complete the sentences by circling your choices.

We [reject *or* do *not* reject] the null hypothesis. We [do find *or* do *not* find] sufficient evidence that the true proportion of families who eat dinner together seven nights a week has decreased since 2001.

expl 2: If the sample from the previous example had 700 respondents, the standard deviation would have been $\sigma_{\hat{p}} = \sqrt{\frac{.38(1-.38)}{700}} \approx 0.018$. Answer the questions that follow.

a.) Recalculate the z-score under this condition. Redraw the normal curve with this value of z_0 and the critical value –1.96 marked. Shade the area under the curve to the left of the critical value, labeling it as 0.025.

If the standard deviation gets smaller, what happens to the z-score?

b.) We will say that the sample evidence is statistically significant if the sample proportion is 1.96 or more standard deviations away from the assumed population proportion. Complete the sentences by circling your choices.

We [reject *or* do *not* reject] the null hypothesis. We [do find *or* do *not* find] sufficient evidence that the true proportion of families who eat dinner together seven nights a week has decreased since 2001.

What does this mean?

Again, given sample information that is *so* far from the assumed population information, we have two possible conclusions.

One is that our sample (being random) just *happened* to differ from the true population that much.

Or, and this is the one we latch onto, the sample reveals that the assumed population information is *likely untrue*. Remember, we do *not* conclude this without a shadow of a doubt. That is why we say “there is sufficient evidence to …”

Because the area to the left of 1.96 standard deviations under the normal curve is 0.025, this procedure will cause us to make a Type I error 2.5% of the time.

What is a Type I error again?

The *P*-value Approach:

Here, we assume the statement in the null hypothesis is true and we find the probability of getting a sample statistic as extreme or more extreme than the one obtained. If this probability is small, **we reject the null hypothesis**.

Definition: A ***P*-value** is the probability of observing a sample statistic as extreme or more extreme than the one observed under the assumption that the statement in the null hypothesis is true.

In practice, we compare this ***P*-value** to the level of significance, α, discussed in the previous section, to see if it is "small".

Summary of the *P*-value Approach:

Remember we first verify the following.

This first bit and steps 1 and 2 are the same as the Classical Approach.

1. the sample is a simple random sample,
2. $np_0(1-p_0) \geq 10$ where n is the sample size, and
3. the sampled values are independent of each other (or $n \leq .05N$).

Step 1: Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

1. Equal versus *not* equal hypothesis (**two-tailed test**)
   $H_0: p = p_0$
   $H_1: p \neq p_0$

3. Equal versus less than (**left-tailed test**)
   $H_0: p = p_0$
   $H_1: p < p_0$

4. Equal versus greater than (**right-tailed test**)
   $H_0: p = p_0$
   $H_1: p > p_0$

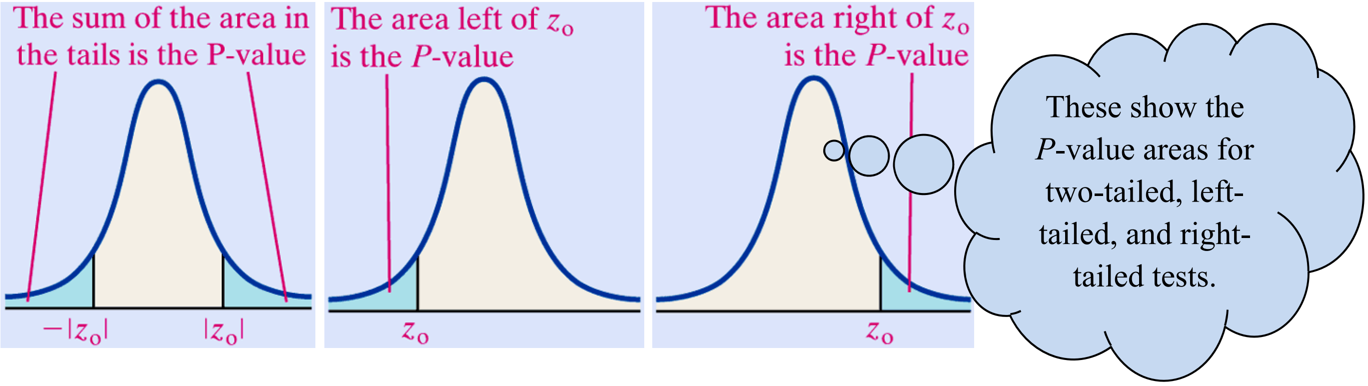
Step 2: Select a level of significance, α, depending on the seriousness of making a Type I error.

Steps 3 and 4 will differ (from the Classical Approach) when we use the *P*-value Approach. Step 5 will be the same…

(continued on next page)

The *P*-value Approach (continued):

Step 3: Compute the **test statistic** $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$. Again, notice we use p_0 in the standard deviation on the bottom of this z-score. Use Table V to determine the area to the left of z_0. The *P*-value is then figured, taking these pictures into account.



These show the *P*-value areas for two-tailed, left-tailed, and right-tailed tests.

You could use technology to find the *P*-value instead of looking it up on the table.

Step 4: If the *P*-value $<$ α, reject the null hypothesis.

Step 5: State the conclusion. Either we say "There is sufficient evidence to conclude that the population parameter is as its stated in the alternative hypothesis." (This is a *rejection* of the null hypothesis.) or "There is *not* sufficient evidence to conclude that the population parameter is as its stated in the alternative hypothesis." (Here, we did *not* reject the null hypothesis.)

More about that *P*-value:

Let's take a deeper dive into this *P*-value. Remember that it is the probability (assuming the null hypothesis is true) that a sample (of the same size) would result in a statistic as severe or more so as the one from our sample. We compare this to the level of significance α (the probability of rejecting the null hypothesis when it's true). I present this table for your enjoyment.

***P*-value**	**Conclusion**
P-value ≥ 0.10	We do *not* reject the null hypothesis (at the 10% level). The sample evidence is consistent with the null hypothesis.
$0.05 \leq P-value < 0.10$	There is some evidence *against* the statement in the null hypothesis. We would reject it at the 10% but *not* the 5% level.
$0.01 \leq P-value < 0.05$	There is moderate evidence *against* the statement in the null hypothesis. We would reject it at the 5% but *not* the 1% level.
$0.001 \leq P-value < 0.01$	There is strong evidence *against* the statement in the null hypothesis. We would reject it at the 1% but *not* the 0.1% level.
P-value < 0.001	There is very strong evidence *against* the statement in the null hypothesis. We would reject it at the 0.1% level.

Technology Instructions (Use only if you know x and n from sample):

Instructions for TI Calculators:
1. Press **STAT** > **TESTS** > **5:1-PropZTest**.

2. You'll enter the assumed population proportion from the null hypothesis (***p*₀**) along with the values of x and n of the sample. If you do *not* know the sample size, do *not* use the calculator.

3. Select the appropriate alternative hypothesis and highlight **Calculate**, pressing **ENTER**. The **Draw** option will draw the z-curve with the P-value shaded.

4. The (Calculate) output includes the test statistic (**z**) and the P-value (**p**) along with the basics of the test you performed.

Instructions for StatCrunch:
1. Enter raw data in column 1 and label it if available. (Probably *not* doing this.)

2. Select **Stat** > **Proportion Stats** > **One Sample** > **With Summary**. Here, you enter the values of x and n of the sample followed by the assumed population proportion from the null hypothesis (***p***). You will have selected the **Hypothesis test for p** radio button under **Perform**. Select the appropriate alternative hypothesis from the drop-down menu. (If you happened to enter raw data, you choose **With Data** here. You'll tell it which column that data is in and what is to be thought of as a **Success**. Then select the appropriate hypothesis test as discussed.) Click **Compute!**.

3. The output includes the test statistic (**Z-Stat**) and the P-value along with the basics of the test you performed.

expl 3: Perform this two-tailed test using the P-value Approach with technology. Follow the steps outlined on the previous pages. Also, draw a normal curve for the standard normal distribution with z_0 (and its opposite) labeled. That picture shows us the area (of the tails) that is equal to the P-value. Be sure to state whether you reject the null hypothesis or not.

$H_0: p = 0.9$
$H_1: p \neq 0.9$
$n = 500;\ x = 440;\ \alpha = 0.05$

expl 4: Suppose we are testing the hypothesis $H_0: p = 0.57$ versus $H_1: p < 0.57$ and we find the P-value to be 0.023. What is your conclusion if…

a.) …it is important that we *not* make a Type I error so the level of significance is set at $\alpha = 0.01$? Explain.

b.) …the level of significance is set at $\alpha = 0.05$? Explain.

Some Concerns:
We must be careful when accepting the results of a hypothesis test. Does it come from a reputable source? We have seen (in example 2) how a statistically significant result can be massaged from the data only by increasing the sample size. You can simply change the level of significance if you do not get a significant result the first time, can't you? Let's investigate another statistical shame-on-you tactic called p-hacking.

Definition: p-hacking (aka **data dredging** or **data snooping**)**:** This is the misuse of data analysis to find patterns in data that can be presented as statistically significant. This is done by performing many statistical tests on data but only reporting on those that are significant.

In 2006, Peter Austin and associates (https://pubmed.ncbi.nlm.nih.gov/16895820) found some statistically significant associations between your zodiac sign and disease. A Leo (born between July 23 and August 22) apparently has an increased probability of gastrointestinal hemorrhage (P-value = 0.0447) while Sagittarians (born between November 22 and December 21) have a higher probability of humerus (upper arm bone) fractures (P-value = 0.0123) compared with all other zodiac signs combined.

They "discovered" 72 similar zodiac-related maladies by searching through medical data for 10,674,945 residents of Ontario. Luckily, these statisticians were *not* just a bunch of dumbasses or shysters. They were studying the testing of multiple, non-prespecified hypotheses and how they may lead to detecting implausible associations. They performed 14,000 hypothesis tests but cherry-picked only 72 to report.

Researchers, so eager to publish (and getting published often depends on if you are reporting statistically significant results), can often convince themselves that such results are valid. Having been alerted to the problem, journals are starting to require researchers to declare what they will test and at what significance level *before* the study is completed.