Review of Hypothesis Tests – Big Ideas

- We start with a review of hypothesis tests for one-sample. We can test a population proportion (when working with categorical data) or we can test a population mean (when working with quantitative data).

- First make a claim about a population parameter:

Proportions

$H_0: p = p_0$

$H_A: p \; < / > / \neq \; p_0$

Means

$H_0: \mu = \mu_0$

$H_A: \mu \; < / > / \neq \; \mu_0$

- Very important – which hypothesis is assumed true? The null, H_0

- Second, collect your data – **good data**. What conditions about the data do we check and why do we check them?

Always: ① Unbiased sample (random even better)

② Sample size less than 10% of Population size

Why? This helps ensure independence among data values

Proportions: At least 10 successes, failures

Means: $n \geq 30$ or population normal

Why? This ensures normal model works for our methodologies.

- Now we convert our data values into a test statistic. Usually there is a (somewhat) complicated formula to do this, so we like to rely on technology to handle the gritty details.

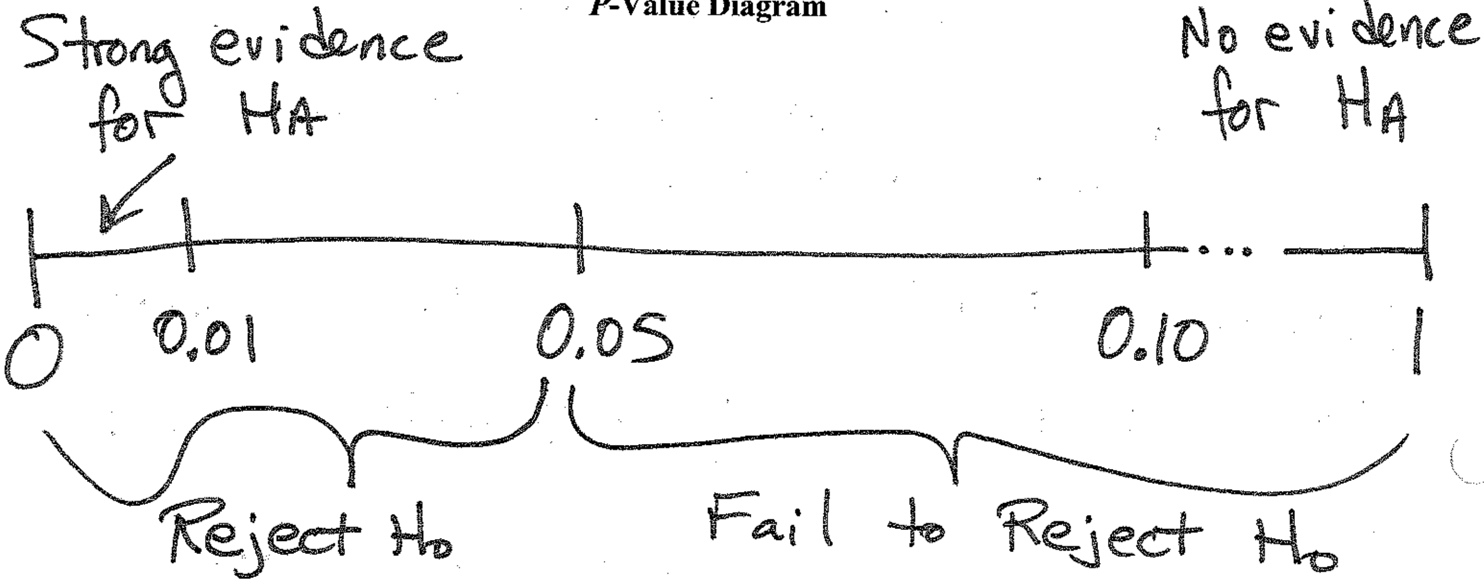
- What is a **test statistic**? Essentially, you convert your sample data values into the summary value you are testing, like a sample proportion or a sample mean. Then you convert that value into the number of standard errors it lies away from your hypothesized value.

- How many standard errors is unusual or rare? Generally speaking, more than ±2 is getting to be unusual and more than ±3 is getting to be rare. It depends somewhat on the exact kind of statistical test you are running. Therefore, it's easiest to now convert your test statistic into a probability – specifically, a P-Value.

- Define *P*-value: The probability of observing data like we did (or even more extreme data) if the null hypothesis is true.

- If the *P*-value is small, we Reject Ho. This is because it was unusual to collect the data that we did, if the null hypothesis is true.

P-Value Diagram



Example: Statewide, it is published online that 63% of all developmental math students pass all developmental math classes within four years. Here at Cecil, is there evidence that we are underperforming that benchmark? Run the one-proportion hypothesis test, showing all steps. Summary results obtained for a cohort of students who were followed for four years here at Cecil College:

	Passed Math 091, 092, 093	Did Not Pass	Total
Males	81	71	152
Females	161	87	248
Total	242	158	400

Hypotheses : $H_0 : p = 0.63$ (assumed true to start)
$H_A : p < 0.63$

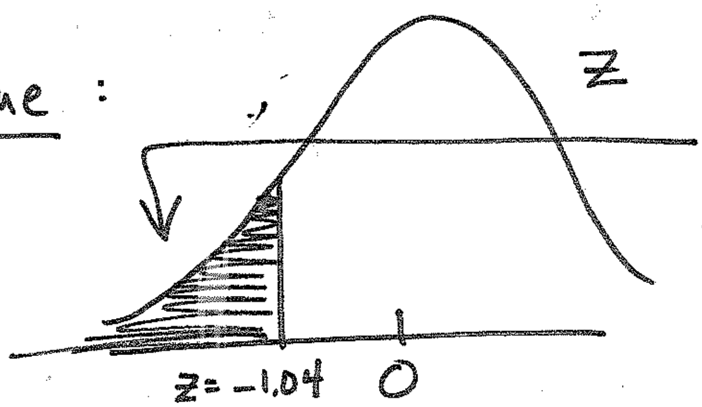
Conditions : $n = 400$ are unbiased dev. ed. students ✓
242 and 158 both exceed 10 ✓

Data : $\hat{p} = \frac{242}{400} = 0.605 = 60.5\%$ passed

Test Stat : $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.605 - 0.63}{\sqrt{\frac{0.63(1-0.63)}{400}}}$$

$= -1.04$ (our $\hat{p}$ is 1.03 SE below 0.63, not unusual.)

P-Value :



normalcdf(-E99, -1.04, 0, 1)
$= 0.150$

TI Calc: 1-Prop Z Test, demo in class

Decision: Since the P-Value = 0.150 exceeds 0.05, we fail to reject H_0.

Conclusion: There is no statistical evidence that, at Cecil, less than 63% of our dev. ed. students pass within 4 years.

The actual $\hat{p} = 60.5\%$ was likely just sample to sample variation.

P-Value Explanation:

If Cecil has 63% passing dev. ed. (i.e. if H_0 true), we'd get $\hat{p} = 60.5\%$ or worse about 15.0% of the time. This is not unusual enough to convince us H_0 isn't plausible.

(P-Value under 0.05, we'd be convinced).

Testing One Proportion – Reference Page

Step 1: Write down the correct set of hypotheses, based on the context of the problem:

$$H_0 : p = p_0 \qquad H_0 : p = p_0 \qquad H_0 : p = p_0$$
$$H_A : p < p_0 \qquad H_A : p > p_0 \qquad H_A : p \neq p_0$$

Step 2: Check the conditions:

a. The sample must be random or at least unbiased (know the difference between the two).
b. The sample size *n* is less than 10% of the population size *N*.
c. We can expect at least 10 successes and 10 failures. In other words, $np \geq 10$ and $nq \geq 10$.

Step 3: Convert the data $\hat{p} = \frac{x}{n} = \frac{\text{number of successes}}{\text{sample size}}$ into the test statistic:

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

Tests on proportions are *z*-tests. They use the standard normal model.

Step 4: Determine the *P*-value by shading under the standard normal model in the H_A direction. Use normalcdf(left, right, 0, 1) on the TI.

NOTE: Steps 3 and 4 can be performed directly on the TI using **STAT → TESTS → 1-PropZTest** or on StatCrunch using **Stat → Proportions → One-sample.**

Step 5: If the *P*-value is low (rule of thumb, under 0.05, but even lower is better), reject the null hypothesis. This means we do have compelling evidence in favor of the alternative hypothesis.

Step 6: Write a concluding remark in terms of the alternative hypothesis.

If we rejected the null, phrase your conclusion that we **do** have evidence to conclude…

If we failed to reject the null, phrase your conclusion that we **do not** have evidence to conclude…

Example: A manufacturer of a sprinkler system used for fire protection in office buildings claims that the true average system-activation temperature is 130°. An engineer tests the system nine times and records the following activation temperatures:

131.2	**130.9**	**130.6**	**130.5**	**130.5**	**132.2**	**130.2**	**129.4**	**131.3**

Run a one-sample *t* test for the population mean to determine if the data collected by the engineer contradicts the manufacturer's claim. Show all steps.

Hypotheses: $H_0: \mu = 130°$ (Assumed true)

$H_A: \mu \neq 130°$ (Looking for evidence manufacturer's claim is false)

Conditions: $n = 9$ test values are unbiased

Since $n < 30$, normality checked ✓ (Histogram & QQ on StatCrunch)

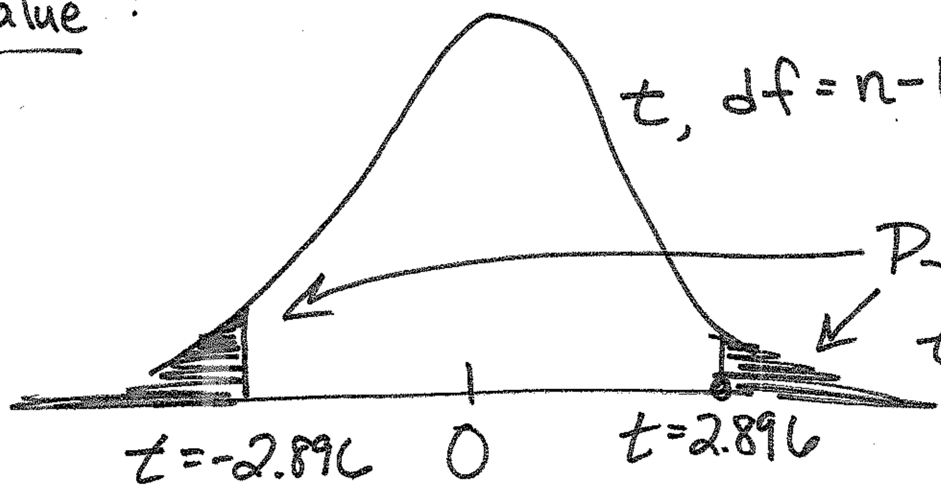
Data: $\bar{y} = 130.756$, $s = 0.783$

Test Stat: $t = \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}} = \dfrac{130.756 - 130}{0.783/\sqrt{9}}$

$= 2.896$ (Our $\bar{y}$ is 2.896 SE above 130)

P-Value:

t, $df = n - 1 = 9 - 1 = 8$



P-Value = 2 × tcdf(2.896, E99, df)

≈ 0.02

TI-Calc: T Test, demo in class

StatCrunch: Stat → T → one-sample

Decision: Since the P-Value is small (0.02 is under 0.05, close to 0.01)

Reject H_0.

Conclusion: There is statistical evidence that the mean activation temp. is different than 130° F. The engineers and manufacturer should look into the potential issue!

P-Value Explanation: If the true mean activation temperature really is 130° F, we'd get a sample mean of 130.756° F or one even farther from 130° F (in either direction) only 2% of the time.

Since 2% is quite unusual, the true mean is probably not 130° F as claimed.

Testing One Mean – Reference Page

Step 1: Write down the correct set of hypotheses, based on the context of the problem:

$$H_0 : \mu = \mu_0 \qquad H_0 : \mu = \mu_0 \qquad H_0 : \mu = \mu_0$$
$$H_A : \mu < \mu_0 \qquad H_A : \mu > \mu_0 \qquad H_A : \mu \neq \mu_0$$

Step 2: Check the conditions:

a. The sample must be random or at least unbiased.
b. The sample size n is less than 10% of the population size N.
c. The sample size n is at least 30 or the data come from a normal population (check using a histogram and QQ plot).

Step 3: Convert the data $\bar{y}$ = sample mean into the test statistic:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

Tests on means are t-tests. They use the Student's t model with degrees of freedom = $n - 1$.

Step 4: Determine the P-value by shading under the Student's t model in the H_A direction. Use tcdf(left, right, df) on the TI.

NOTE: Steps 3 and 4 can be performed directly on the TI using
STAT → TESTS → T-Test or on StatCrunch using **Stat → T statistics → One-sample.**

Step 5: If the P-value is low (rule of thumb, under 0.05, but even lower is better), reject the null hypothesis. This means we do have compelling evidence in favor of the alternative hypothesis.

Step 6: Write a concluding remark in terms of the alternative hypothesis.

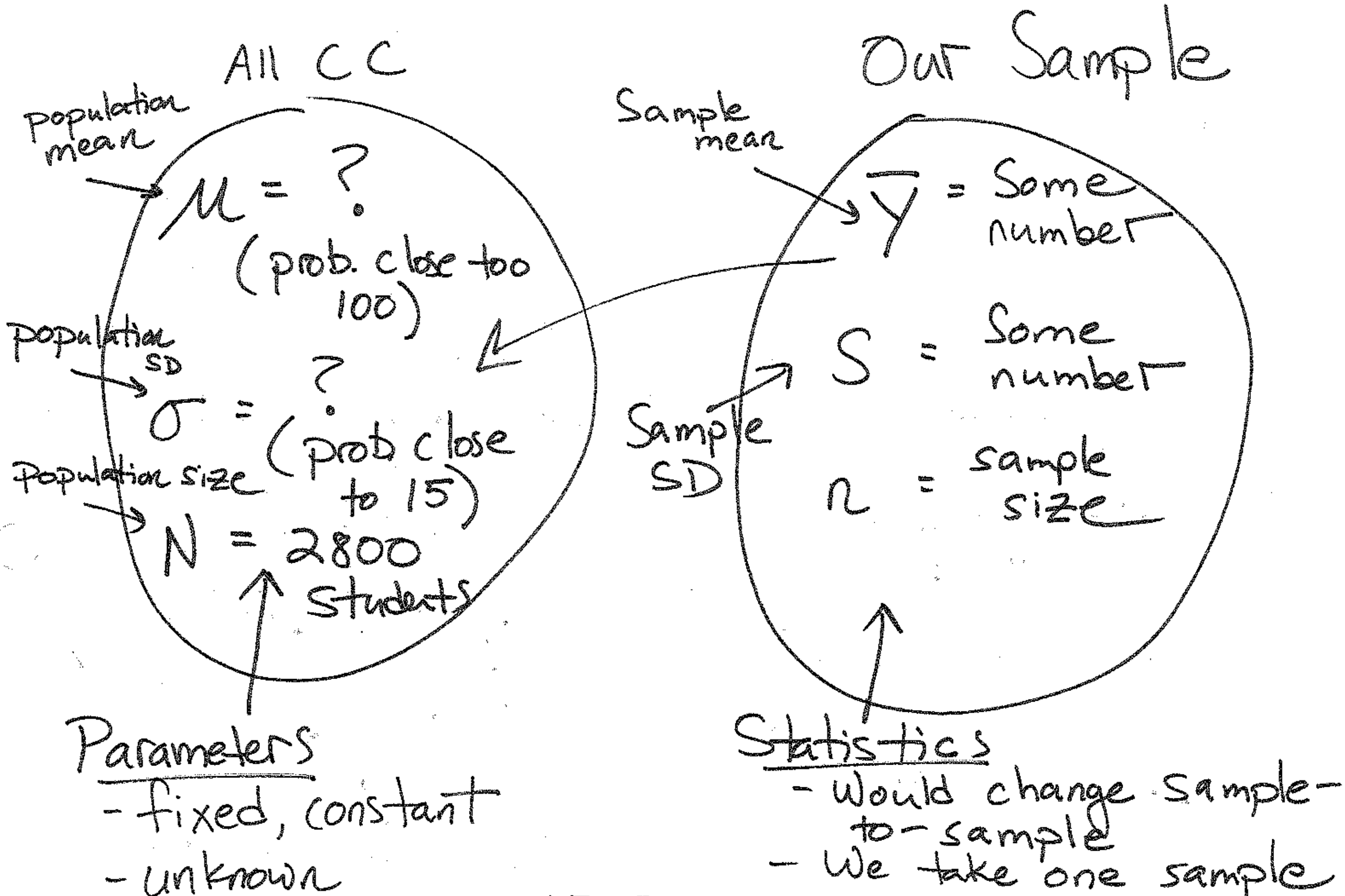
If we rejected the null, phrase your conclusion that we **<u>do</u>** have evidence to conclude…

If we failed to reject the null, phrase your conclusion that we **<u>do not</u>** have evidence to conclude…

Hypothesis Testing – A Diagram

It is well-established that adults in general have an average IQ of 100 with a standard deviation of 15. We'd like to investigate IQ scores here at the college. Diagram the population and parameters, the sample and statistics.

* $\mu_{\text{ALL ADULTS}} = 100$, $\sigma_{\text{ALL ADULTS}} = 15$



A Few Important Notes

- If we could administer IQ tests to every Cecil student, there would be no need to run a hypothesis test. In other words, with Population data, we "have the answer".

- Samples vary, so sample statistics like Sample mean, $\bar{Y}$, Sample SD, S and Sample proportion $\hat{p}$ will change sample-to-sample. The hypothesis tests we run take this into consideration and evaluate the actual data using probabilities and sampling distributions.

Review of One-Sample Confidence Intervals

- When we have some idea about the value of a population parameter, and we'd like to evaluate whether or not our guess / claim is plausible, we run a hypothesis test.
- When we have little to no knowledge about the value of a population parameter, we can compute a confidence interval to give a range of plausible values for that parameter.

Example: What kind of gas mileage do Cecil students get out of their vehicles, on average? One way to answer this question would be to go online and look up the average miles per gallon for the whole country or some part of the country (one site claims 20.26 mpg). This would give us a ballpark guess, but as we know, here at Cecil, we are quite different than most other places!

We will collect data from our students and construct a 95% confidence interval for the true mean mile per gallon here at the college. Diagram the situation.

Population

N = all vehicles driven by

μ = mean mpg

σ = SD mpg

(We Take One) Sample

n = 37

$\bar{Y}$ = 18.202

S = 8.331

Calculated from our 37 mpg values

Using TI T-interval,

95% CI for μ is (15.424 mpg to 20.98 mpg)

One Sample Confidence Intervals – Reference Page

Step 1: Determine if you are working with means or proportions. With quantitative variables, we work with means. With categorical variables, we work with proportions.

Step 2: Check the conditions:

a. The sample must be random or at least unbiased.
b. The sample size n is less than 10% of the population size N.
c. **If working with means**, the sample size n should exceed 30 or the data should look normal.
If working with proportions, then we need at least 10 successes and 10 failures.

Step 3: Compute the confidence interval (formula or technology):

Means: **TI Calculator → STAT → TESTS → T-Interval**

StatCrunch → STAT → T Statistics → One-Sample

$$\bar{y} \pm t^*_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Use the Student's t model with degrees of freedom $= n - 1$.

Proportions: **TI Calculator → STAT → TESTS → 1-PropZInt**
StatCrunch → Proportions → One-Sample

$$\hat{p} \pm z^*_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Use the standard normal model to find z.

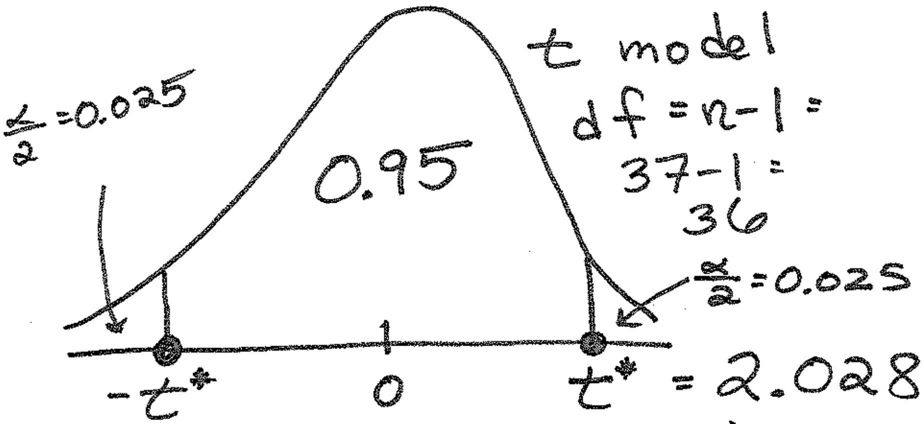
Step 4: Interpret your interval with a sentence in the context of the problem.

Example: For the Cecil student gas mileage data, demonstrate the formula and review the Student's *t* model. Interpret the interval in context.

Conditions: $n = 37$ (large), presume unbiased.

$\bar{y} = 18.202$, $s = 8.331$, $n = 37$

95% confidence so $\alpha = 0.05$



Formula: $\bar{y} \pm t^*_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$

$18.202 \pm 2.028\left(\frac{8.331}{\sqrt{37}}\right)$

18.202 ± 2.778 or $(15.424, 20.98)$

* We are 95% confidence this interval contains the true mean mpg ~~for~~ for all Cecil students.

ONE-SIDED CONFIDENCE INTERVALS

- So far, all confidence intervals have been two-sided. A 95% interval has $\alpha = 0.05$ split equally in the ~~tailed~~ tails of the standard normal model or the Student's *t* model.
- At times, we need one-sided intervals. We are looking for a lower bound or an upper bound for the value of the true parameter we are estimating with the interval.
- Rather than split the α in both tails, stick it all in one tail.
- Obtain the critical value z_α^* or t_α^* for usage in the confidence interval formula.
- Technology usually doesn't support this, so we must rely on formulas:

Lower Bound : $\hat{p} - z_\alpha^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\bar{y} - t_\alpha^* \left(s/\sqrt{n} \right)$

Upper Bound : Replace $-$ with $+$

- The intervals will **not** have two numbers (left bound, right bound). Instead, we get:

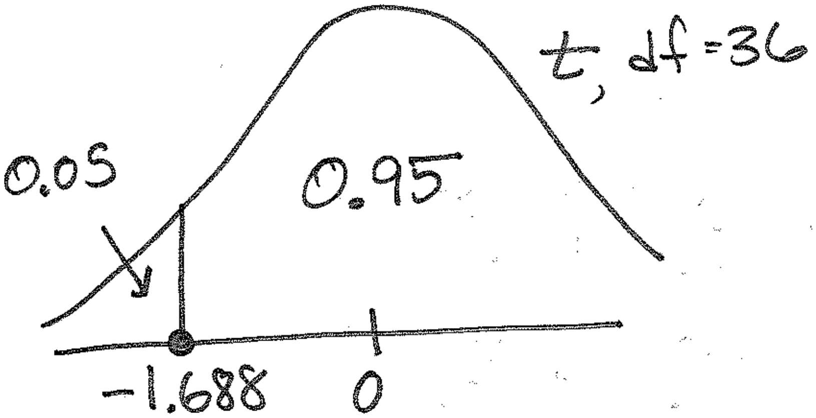
(lower bound, $+\infty$)

($-\infty$, upper bound)

Example: Suppose at Cecil College, instead we'd like to obtain a plausible lower bound for the minimum miles per gallon our students are achieving in their vehicles. Compute a one-sided 95% confidence interval for the mean miles per gallon for all Cecil College students. Also, suppose the county mandates that all cars should average at least 16 miles per gallon by the end of next year. Are Cecil students meeting that target?

$\bar{Y} = 18.202$, $S = 8.331$, $n = 37$

need t_{α}^{*} with $df = 36$



(T-Table)

Formula: $\bar{Y} - t_{\alpha}^{*}\left(S/\sqrt{n}\right)$

$18.202 - 1.688\left(\frac{8.331}{\sqrt{37}}\right)$

$18.202 - 2.312 = 15.89$

One-sided interval for μ is

$(15.89, +\infty)$

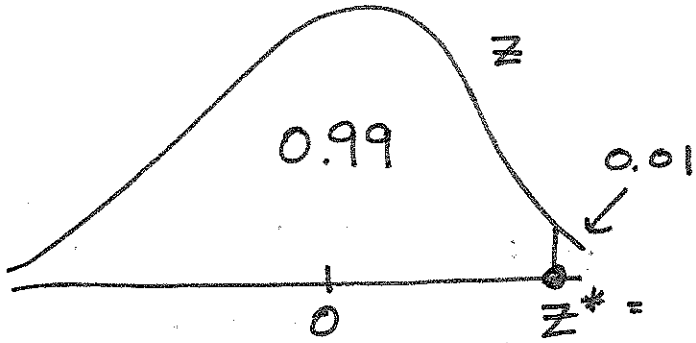
We're 95% confident Cecil students average at least 15.89 mpg.

Since the whole interval does not exceed 16 mpg, we do not have <u>statistical</u> evidence of meeting the goal.

Example: During Fall 2012, we asked Cecil students for their father's highest level of education. Twelve out of 138 respondents had Master's degrees or above. Give a 99% upper bound one-sided confidence interval for the proportion of all Cecil students whose father has at least a Master's degrees. Interpret the interval with a sentence.

Data: $\hat{p} = \frac{12}{138} = 0.087$

99% upper bound →



invNorm(0.99, 0, 1) = 2.33

Formula: $\hat{p} + z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} =$

$0.087 + 2.33 \sqrt{\frac{0.087(1-0.087)}{138}} =$

$0.087 + 0.056 = 0.143$

Interval: $(-\infty, +0.143)$

We are 99% confident that the percentage of students who have father's with a Master's + is at most 14.3%.

Hypothesis Test Errors and Power

Example: Lipitor, used to reduce cholesterol, has a number of side effects. From Pfizer's webpage, 4.7% of those taking the drug experienced dyspepsia. The maker of a new cholesterol drug, Drug Z, is running a clinical trial to compare their experimental drug to Lipitor, and among other things, wants to compare the percentage of patients experiencing dyspepsia. They are interested in advertising a lower percentage of those experiencing the side effect.

a. Write the appropriate hypotheses:

$H_0: p = 0.047$ (Assume 4.7% get dyspepsia)

$H_A: p < 0.047$ (Look for evidence that less than 4.7% get dyspepsia)

b. Explain what a Type I error would be:

Null is true, but we mistakenly reject it. "False positive" We conclude our drug has fewer patients experiencing dyspepsia, but Drug Z is no better than Lipitor.

Our trials produced nonrepresentative data.

c. Explain what a Type II error would be:

Alternative is true, but we mistakenly failed to reject the null. "False negative".

Drug Z is better than Lipitor (for side effect dyspepsia) but we conclude it is not better. "Unlucky" data.

d. The probability of making a Type I error is exactly α when we decide to run a hypothesis test using a significance level. Usually in this class, we reject the null when the P-value is low (under 0.05 or under 0.01) without setting a significance level, but sometimes we do specify this before running the test. IF we chose a significance level, we reject the null when the P-value is less than α.

If the makers of Drug Z must show evidence at the 5% level of significance, they are specifying the probability of making a Type I error beforehand.

e. What if Drug Z is better than Lipitor (for dyspepsia)? Then H_A is true. We can't possibly make a Type I error, but we could get data such that we fail to reject H_0. This mistake is a Type II error, and we use the letter β for the probability of making a Type II error.

β is harder to calculate because when H_A is true, we don't know the value of the parameter we are testing. Recall for this example, the hypotheses are:

$H_0: p = 0.047 \qquad H_A: p < 0.047$

And if H_A is true, p can be anything under 4.7%! It could be that using Drug Z, $p = 2.3\%$ experience dyspepsia (or $p = 3.1\%$ or $p = 4.65\%$, etc…)

We can compute β for any value of p, but the one we should choose depends on the situation.

f. You can reduce β by increasing the value of α. In other words, a larger significance level makes it easier to reject the null. With greater chances of rejecting the null, we reduce the chances of making a Type II error.

Therefore, α and β have an inverse relationship.

Many times, the consequences of making a certain kind of error (Type I versus Type II) are more severe, so we can run our test with this in mind.

The only way to simultaneously reduce the probability of making either error is to collect more data.

g. A test's ability to detect a false null hypothesis is key. For the Lipitor and Drug Z example, if Drug Z really does have a lower occurrence of dyspepsia:

① Less than 4.7% get dyspepsia

② We want our data to show it!

The probability of correctly rejecting a false null hypothesis is called the power of the test. We want this large

h. When the null hypothesis is actually false, our goal is to have a hypothesis test powerful enough to reject it. If an experiment or clinical trial fails to reject its null hypothesis, the test's power comes into question.

Did we collect enough data?

Was there too much variation in the data?

Was the effect size too small to detect?

i. **Definition:** Effect Size:

The distance between p_0 (the value in H_0, 4.7% here) and the true value p (% getting dyspepsia using Drug Z, whole population).

If using Drug Z, 1.9% of the whole population experienced dyspepsia, the effect size is:

Effect Size: $p_0 = 4.7\%$

$p = 1.9\%$

→ 2.8% difference (+ or − doesn't matter?)

j. If the effect size is large, it should be relatively easy to have our data collected reflect that. It should be relatively easy to reject the null.

If the effect size is small, it will be difficult to detect when we run our test. We will have lower power and we will commit more Type II errors

k. We also must ask ourselves, "What effect size will really matter?" For the Lipitor and Drug Z example:

1.9% getting dyspepsia would matter but if it were 4.65% or something closer to 4.7%, it probably wouldn't be important.

l. Step-by-step instructions to calculate power for a one-sample proportion test:

1. Begin by assuming the null hypothesis is true, so $p = p_0$. For the Drug Z example:

   Assume $p = 0.047$ (just like Lipitor)

2. Specify the level of significance. For Drug Z, we'll use:

   $\alpha = 0.05$

3. Determine the correct normal model for $\hat{p}$, and find it's mean and standard deviation, presuming that we will use a sample size of $n = 520$:

$$\mu_{\hat{p}} = p = 0.047$$

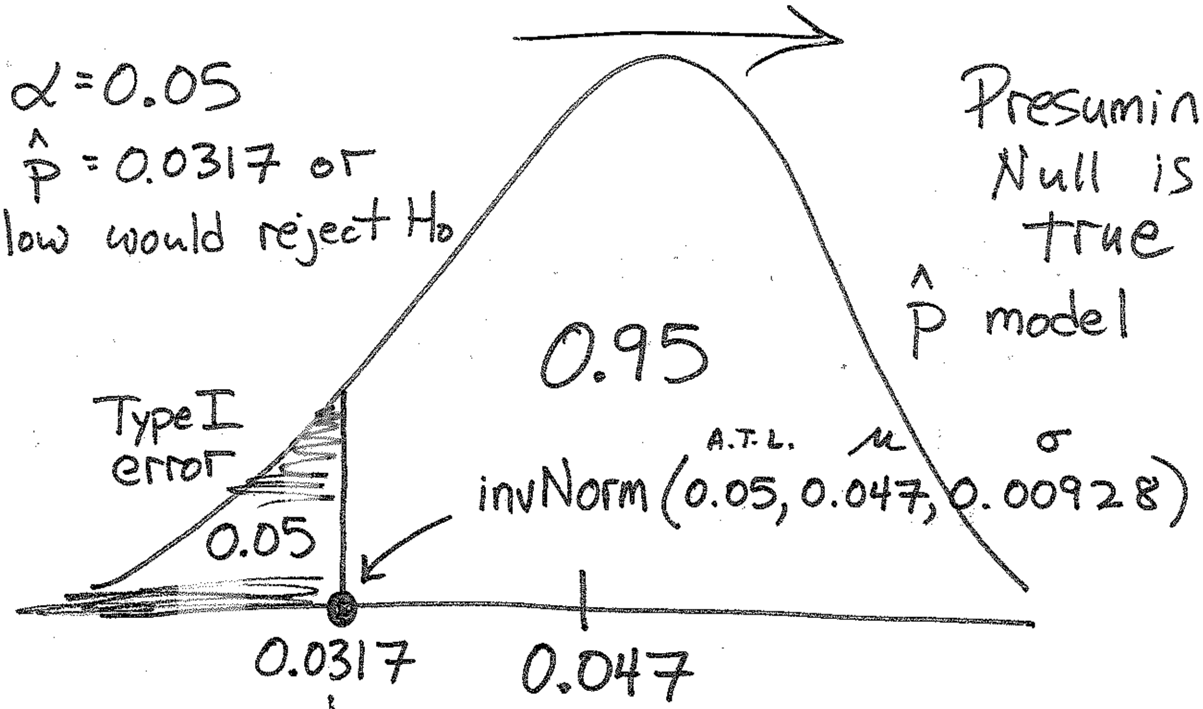
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.047(1-0.047)}{520}}$$

$$= 0.00928$$

4. Find the critical value of $\hat{p}$ that corresponds to the significance level of the test. This is the sample proportion that would lead to rejection of the null hypothesis. Use invNorm on the TI calculator to find this value. Draw the normal model, but leave space for another normal model below it.

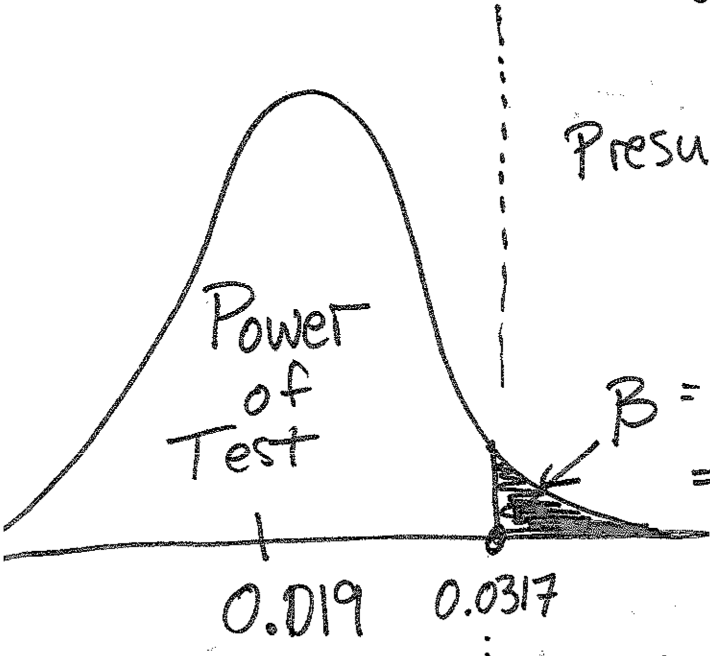
$\alpha = 0.05$

So $\hat{p} = 0.0317$ or below would reject H_0



Presuming Null is true $\hat{p}$ model

Presuming Null is False



β = ncdf(0.0317, E99, 0.019, 0.00599)
= 0.017

P(Type II) = 0.017, 1.7% chance

Power = 1 − β = 1 − 0.017
= 0.983

Reject H_0

Fail to Reject H_0

98.3% chance

5. Now suppose the null hypothesis is really false. In other words, $p \neq p_0$ like we assumed previously. For the Drug Z example, let's presume that in reality, 1.9% experience dyspepsia, so $p = 0.019$, and our sample size will be $n = 520$. Find the correct normal model, mean and standard deviation.

If $p = 0.019$, $\mu_{\hat{p}} = 0.019$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.019(1-0.019)}{520}} = 0.00599$$

6. Draw the second normal model below the first, lining up the critical value. The models will be offset, because they are centered at different values.

The area in the tail of the top model (assuming the null is true) is the probability of making a Type I error. We pre-specified this (usually 0.05 or 0.01).

The area in the opposite tail of the bottom model (assuming the null is false) is the probability of making a Type II error. We can determine this by using normalcdf on the TI.

The area in the bottom model in the other direction is the power of the test. We want this big.

Power depend on three things:

What we chose for α

The effect size / true value of p

Sample Size

Example: Shaquille O'Neal is a lifetime 52.7% foul shot shooter (horrible). He decides to make a comeback and works all summer long on his foul shooting. In reality, he improves to a 65% shooter, but the coach isn't convinced. For his comeback tryout, Shaq will shoot 100 free-throws and needs to make 60 of them to get picked up. Determine α, β, the power of the test and describe in context what each number represents.

Testing $H_0: p = 0.527$ vs. $H_A: p > 0.527$

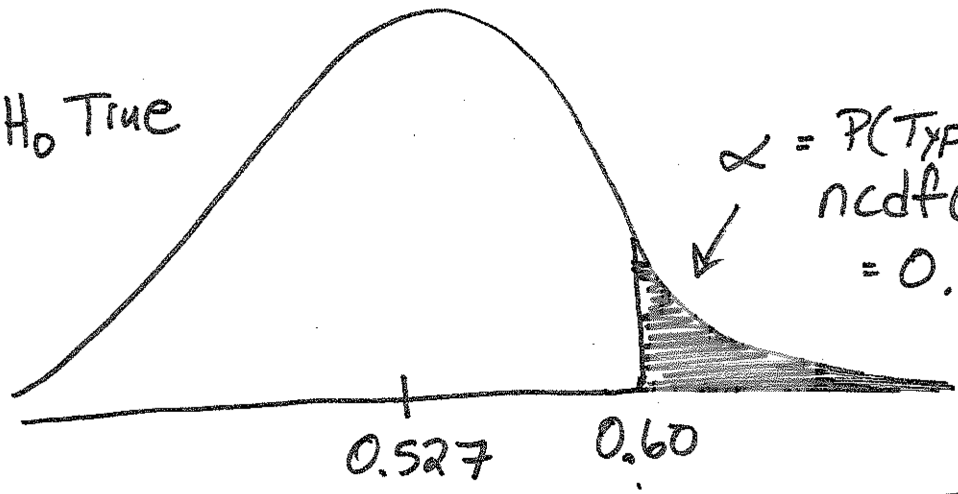
If H_0 true, $\hat{p}$ is normal, $\mu_{\hat{p}} = 0.527$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.527(1-0.527)}{100}} \approx 0.05$$

If $p = 0.65$, $\hat{p}$ is normal, $\mu_{\hat{p}} = 0.65$

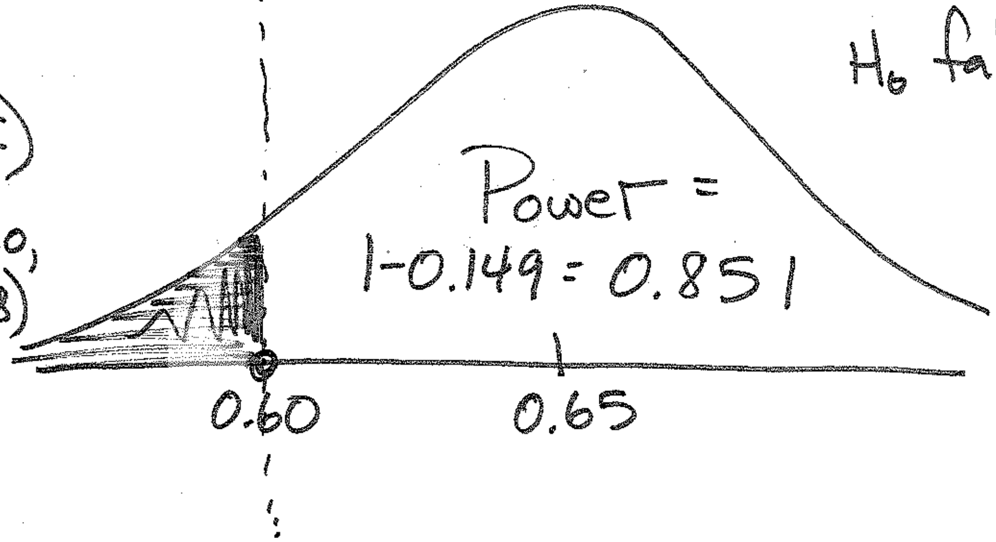
$$\sigma_{\hat{p}} = \sqrt{\frac{0.65(0.35)}{100}} \approx 0.048$$

H_0 True



$\alpha = P(\text{Type I}) =$
ncdf(0.6, E99, 0.527, 0.05)
$= 0.072$

H_0 false



$\beta = P(\text{Type II})$
= ncdf(−E99, 0.60, 0.65, 0.048)
$= 0.149$

Power = $1 - 0.149 = 0.851$

$\alpha = 0.072 = P(\text{Type I error})$

* There is a 7.2% chance Shaq makes 60 + Free Throws, if he did <u>not</u> improve (i.e. he still shoots 52.7% in).
(We know he <u>did</u> improve).

$\beta = 0.149 = P(\text{Type II error})$

* There is a 14.9% chance that newly-improved shaq does <u>not</u> hit 60 + Free Throws. i.e. 14.9% chance he incorrectly does <u>not</u> get picked up.

Power $= 1 - \beta = 100\% - 14.9\% = 85.1\%$
chance that newly-improved Shaq does get picked up (correct decision)

<u>Thought</u>: If coach gives Shaq 1000 shots and he needs to make 600 in $\rightarrow$ Power goes <u>up</u>.

Two Proportions (Intervals and Tests) – Reference Page

Step 1: Check the conditions:

a. Both samples must be random or at least unbiased.
b. Both sample sizes n_1 and n_2 are less than 10% of the population sizes N_1 and N_2.
c. We need at least 10 successes and 10 failures from both groups.

Step 2a: Compute the confidence interval (formula or technology):

TI Calculator → STAT → TESTS → 2-PropZInt (Recommended)

Formula:

$$(\hat{p}_1 - \hat{p}_2) \pm z^*_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Use the standard normal model to find z.
Write a concluding remark as per usual.

Step 2b : Run the test:

$$H_0 : p_1 - p_2 = 0 \qquad H_0 : p_1 - p_2 = 0 \qquad H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 < 0 \qquad H_A : p_1 - p_2 > 0 \qquad H_A : p_1 - p_2 \neq 0$$

TI Calculator → STAT → TESTS → 2-PropZTest (Recommended)

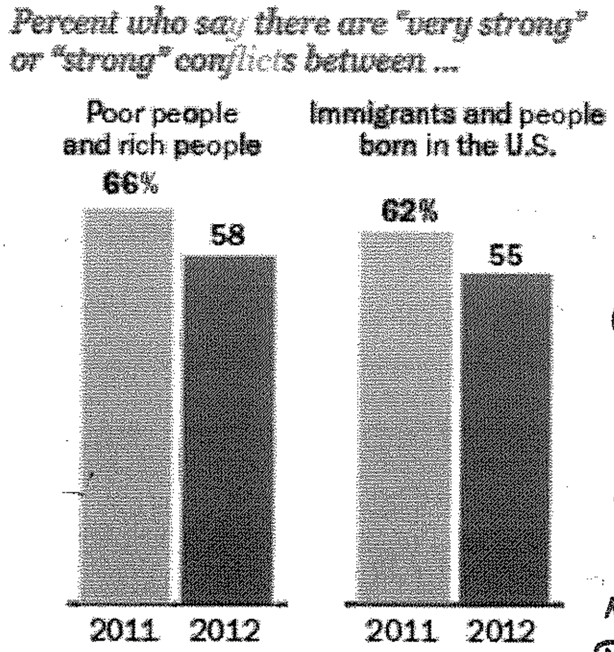
Test Statistic Formula:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (\text{Usually Zero})}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}}, \quad \text{with } \hat{p}_{\text{pooled}} = \frac{\text{success}_1 + \text{success}_2}{n_1 + n_2}$$

Determine the P-value by shading under the standard normal model in the H_A direction. Use normalcdf(left, right, 0, 1) on the TI.

Make a decision and write a concluding remark as per usual.

Example: In 2011 and 2012, Pew Research polled 1500 and 1300 Americans respectively.



Run a two-sample hypothesis test to determine if in 2012, the proportion of Americans who think there is a "very strong" or "strong" conflict between poor and rich has declined.

A mouthful. In other words, are there fewer people now who think there is a big conflict?

If the difference is statistically significant, support your conclusion with a 95% confidence interval for the difference in proportions.

Data

$\hat{p}_{2011} = 0.66 = \frac{990}{1500}$

$\hat{p}_{2012} = 0.58 = \frac{754}{1300}$

Conditions ✓✓

$H_0: p_{2011} - p_{2012} = 0$

$H_A: p_{2011} - p_{2012} > 0$

TI: 2PropZTest

Test Stat: $z = 4.36$

P-Value = 0.0000066

Reject H_0. There is decisive evidence that the proportion of Americans who think there's a big conflict between rich and poor has declined, 2011 to 2012.

TI: 2PropZInt : (4.40% to 11.60%)

We're 95% confident that between 4.40% and 11.60% fewer Americans think there is a big problem.

Usually, when we test for a difference in proportions, we test for "a" difference, or "any" difference. Sometimes, though, we test for a certain difference, as noted in this example.

Example: Using the Pew data from the previous page, run a two-sample hypothesis test to determine if in 2012, the proportion of Americans who think there is a "very strong" or "strong" conflict between immigrants and US-born people has declined by 5%. We will need to use the formula for this one.

$H_0: P_{2011} - P_{2012} = 0.05$ (Use decimals for calcs.)

$H_A: P_{2011} - P_{2012} > 0.05$

<u>Data</u>: $\hat{P}_{2011} = 0.62 = \frac{930}{1500}$

$\hat{P}_{2012} = 0.55 = \frac{715}{1300}$

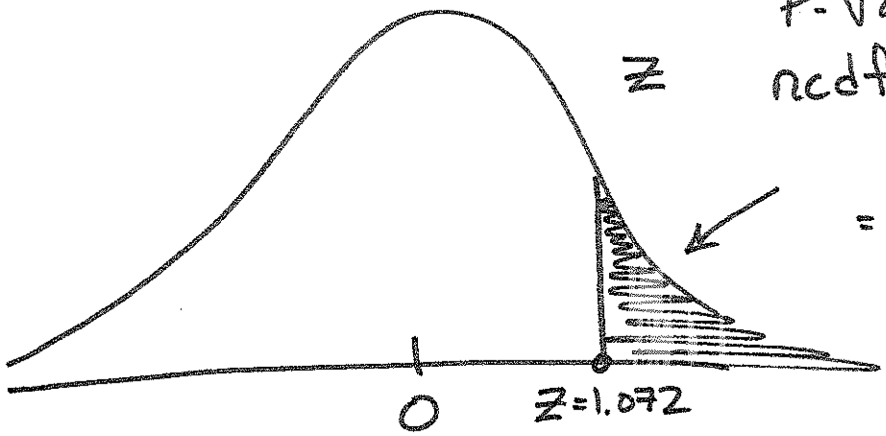
$\hat{P}_{POOLED} = \frac{930 + 715}{1500 + 1300} = 0.5875$

<u>Test Stat</u>:

$$Z = \frac{(0.62 - 0.55) - 0.05}{\sqrt{\frac{0.5875(1-0.5875)}{1500} + \frac{0.5875(1-0.5875)}{1300}}}$$

$$= \frac{0.02}{0.01865} = 1.072$$

Get P-Value:



P-Value = ncdf(1.072, E99, 0, 1) = 0.142

Decision: Since P-Value = 0.142 exceeds any reasonable α level, fail to reject Ho.

Conclusion: There is no evidence that the proportion of Americans who think there's a big conflict between immigrants and US-born citizens has declined by 5% since 2011.

Two Means, Independent Samples (Intervals and Tests) – Reference Page

Step 1: Check the conditions:

a. Both samples must be random or at least unbiased.
b. Both sample sizes n_1 and n_2 are less than 10% of the population sizes N_1 and N_2.
c. We need at both samples to look normal, or sample sizes exceed 30, or some combo.

Step 2a: Compute the confidence interval (~~formula~~ or technology):

TI Calculator → STAT → TESTS → 2-SampTInt (Recommended)

StatCrunch → STAT → T-Statistics → 2-Sample (Recommended)

Formula:

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Use the Student's t model to find t. The degrees of freedom formula is messy (p. 582, De Veaux)
Write a concluding remark as per usual.

Step 2b : Run the test:

$$H_0: \mu_1 - \mu_2 = 0 \qquad H_0: \mu_1 - \mu_2 = 0 \qquad H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 < 0 \qquad H_A: \mu_1 - \mu_2 > 0 \qquad H_A: \mu_1 - \mu_2 \neq 0$$

TI Calculator → STAT → TESTS → 2-SampTTest (Recommended)

StatCrunch → STAT → T-Statistics → 2-Sample (Recommended)

Test Statistic Formula:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\text{Usually Zero})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Determine the P-value by shading under the Student's t model in the H_A direction. Use tcdf(left, right, ~~0, 1~~) df on the TI. The degrees of freedom formula is messy (p. 582, De Veaux)

Make a decision and write a concluding remark as per usual.

Example: The article "***Effects of Internal Gas Pressure on the Compression Strength of Beverage Cans and Plastic Bottles***" includes the accompanying data on compression strength (lb) for a sample of 12-oz aluminum cans filled with strawberry drink and another sample filled with cola. Does the data suggest that the extra carbonation of cola results in a higher average compression strength? Base your answer on a *P*-value. What assumptions are necessary for the analysis?

Beverage	Sample Size	Sample Mean	Sample SD
Strawberry Drink	15	540	21
Cola	15	554	15

Conditions: Each sample unbiased
Data look normal since $n_1 = n_2 = 15$

$H_0 : \mu_{COLA} - \mu_{STRAW} = 0$

$H_A : \mu_{COLA} - \mu_{STRAW} > 0$

Use TI : 2 Sample T Test

Test Stat : $t = 2.101$
P-Value = 0.023 (Between 0.01 and 0.05)

There is some evidence that the mean compression strength in cola cans exceeds that of strawberry drink cans.

Support with 95% CI for diff. in means.
(0.286 lbs. to 27.714 lbs.)

Example: The following summary data give proportional stress limits for specimens constructed using two different types of wood.

Type of Wood	Sample Size	Sample Mean	Sample SD
Red Oak	14	8.48	0.79
Douglas Fir	10	6.65	1.28

Assume conditions met.

Assuming both samples came from Normal distributions, carry out a test of the hypotheses to decide whether the true average proportional stress limit for red oak joints exceeds that for Douglas fir joints by more than 1 MPa. The degrees of freedom are 13.854 (messy formula, De Veaux p. 582).

$H_0: \mu_{RED} - \mu_{DOUG} = 1$ MPa

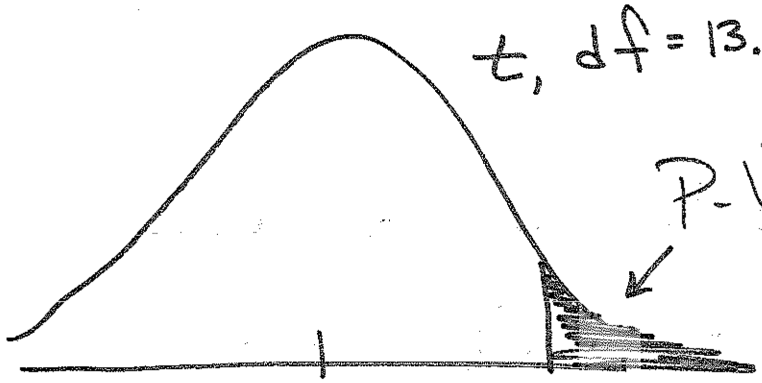
$H_A: \mu_{RED} - \mu_{DOUG} > 1$ MPa

Test Stat:

$$t = \frac{(8.48 - 6.65) - 1}{\sqrt{\frac{0.79^2}{14} + \frac{1.28^2}{10}}} = 1.818$$

Get P-Value:

t, $df = 13.854$



0 $t = 1.818$

P-Value = tcdf(1.818, E99, 13.854)

= 0.0454

Fail to Reject H_0

At $\alpha = 0.01$, we do not have evidence that the mean MPa for Red Oak Exceeds that for Douglas Firs by 1 MPa.

Two Means, <u>Dependent</u> Samples (Intervals and Tests) – Reference Page

Step 1: Check the conditions:

a. The data are matched pairs. We have two observations for each data point, linked somehow.
b. The number of differences in our sample are less than 10% of the population size.
c. We need the differences to look normal, or the number of differences to exceed 30.

Step 2: The "data" we work with are the differences.

Example: Do wives have higher IQs than their husbands? Take an unbiased sample of 50 married couples and give IQ tests to both spouses. Our data points are the differences:

Wife IQ – Husband IQ = Difference

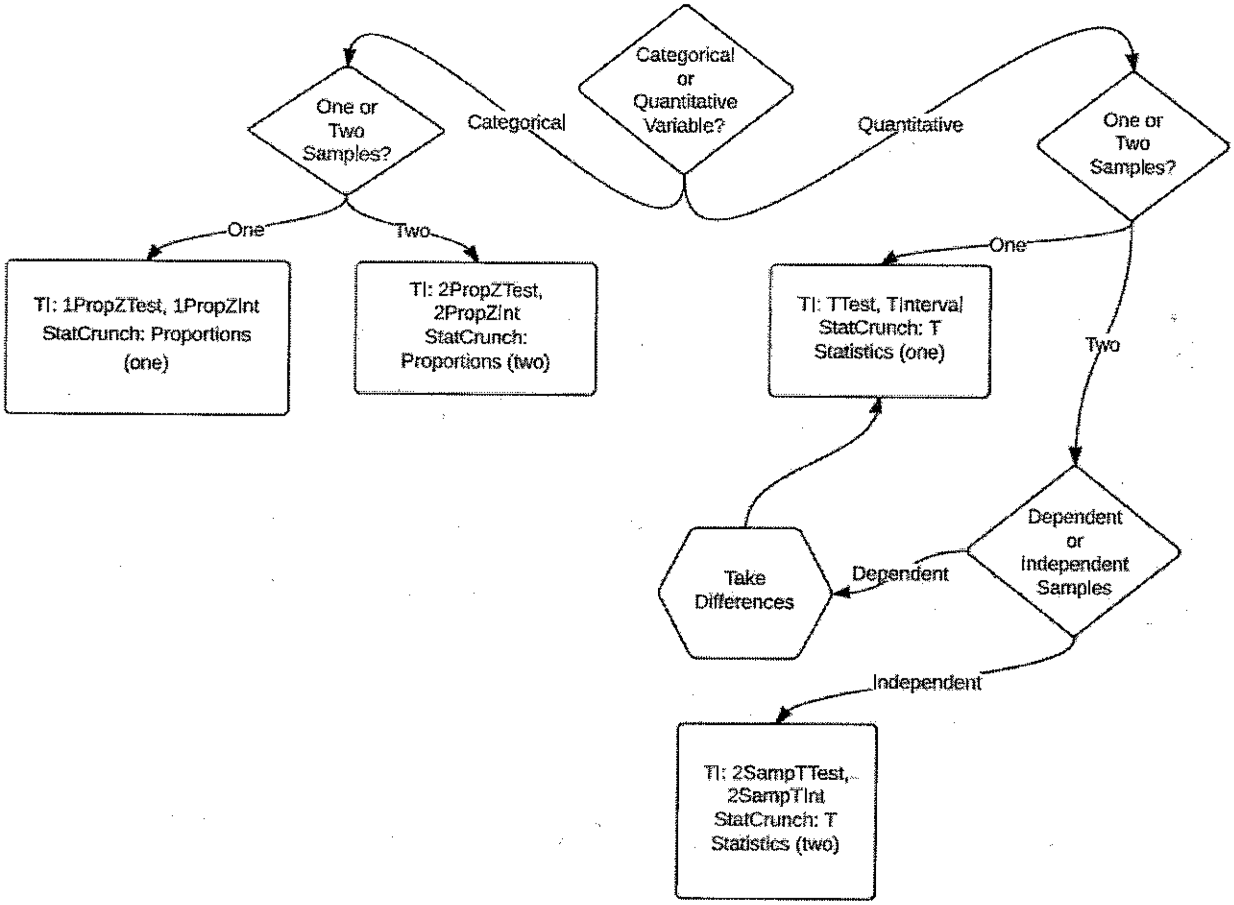
We can compute a confidence interval on the difference of means and run a hypothesis test on the difference of means.

Resort to one-sample *t* intervals and one-sample *t* tests, discussed previously on pages 8 and 11.

Example: On average, is our dominant hand stronger than our other hand? In class, we will measure the strength of each hand and run the appropriate hypothesis test at the 5% level of significance. If there is a significant difference, support your answer with the appropriate 95% one-sided confidence bound.

Data From Class:

Flowchart for Statistical Inference



StatCrunch Problems

Example: Are supermarkets more expensive than Wal-Mart? If so, by how much on average per item? Use the "***Wal-Mart Supermarket***" dataset on StatCrunch.

100 Cecil students visited Wal-Mart and a supermarket and recorded the price of the same food item at both locations. Run the appropriate test to determine if supermarkets are indeed more expensive than Wal-Mart, on average. Then give a 95% lower bound for the mean minimum amount per grocery item.

This is a matched-pairs (dependent) mean problem.

d = Supermarket − Walmart (Data → Compute Expression)

Data: $\bar{d} = 0.5154$, $S_d = 0.827$, $n = 100$

$H_0: \mu_d = 0$ (Assume prices equal)

$H_A: \mu_d > 0$ (Supermarket more expensive)

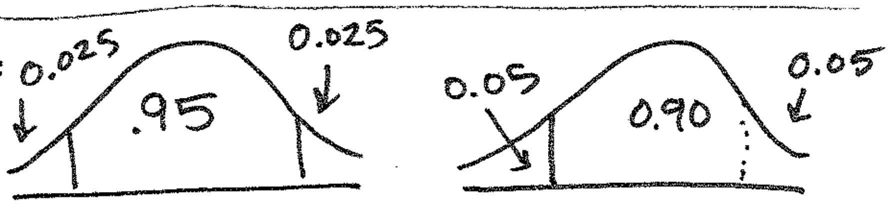
StatCrunch: T → ONE → DATA

Test Stat: $t = 6.23$

P-Value < 0.0001 Reject H_0

Conclude Supermarket is more expensive (on average).

Trick StatCrunch :



Do 90% CI and take lower bound

($0.38, +∞) ← 95% lower bound.

* We're 95% confident Walmart is at least ~$0.38 cheaper on average per item.

Example: In the Math 127 classes for Spring 2013, we asked all the students what grade they expected to get in the course. Open the "***Kupresanin Quiz 1 Data***" dataset on StatCrunch and run the appropriate hypothesis test to determine if the proportions who think they'll get an "A" differ, men versus women.

This is a two-sample proportion problem.

Conditions: Unbiased sample of Math 127 Students

10, 10 S/F Both Samples ✓

Data: $\hat{p}_M = \frac{17}{37} = 0.4595$

$\hat{p}_W = \frac{44}{85} = 0.5176$

$H_0: p_M - p_W = 0$

$H_A: p_M - p_W \neq 0$

STAT CRUNCH → PROPS → 2 → DATA

TEST STAT: $z = -0.591$

P-Value = 0.5546

Fail to Reject H_0

There is no evidence that men and women differ in the proportion of Math 127 students who think they'll earn an "A".

Example: Professor Kupe "knows" that his 2nd bowling game each week is his best and his 3rd game each week is his worst. Use the "***Kupe Bowling***" dataset on StatCrunch. Treat the games as independent each week (sometimes he excels in game 2 and for no good reason, tanks game 3). Run a test to determine if on average, his 2nd game is better than his 3rd game. Use a 10% level of significance. Also, give a 95% confidence interval for the difference in mean scores.

This is a two-sample (independent) means problem.

$H_0: \mu_{2nd} - \mu_{3rd} = 0$

$H_A: \mu_{2nd} - \mu_{3rd} > 0$

Conditions: $n_1 = n_2 = 18$, normality checked - barely plausible

STATCRUNCH: $\bar{Y}_{2nd} = 184.28$ $\bar{Y}_{3rd} = 171.28$

$S_{2nd} = 18.74$ $S_{3rd} = 23.28$

STAT → T → 2 → DATA

Test Stat: $t = 1.85$

P-Value = 0.037

At $\alpha = 0.10$, Reject H_0 and conclude on average, game 2 is better than game 3.

95% CI for difference in means:

$(-1.34, +27.34)$

(α mismatch.)

Example: In 2011, 6.3% of the county residents checked "Black or African American" on the most recent U.S. Census. Here at the college, do we have any evidence that we differ? Use the "***Quiz 3 / 4 Data Fall 2012***" dataset on StatCrunch. Run the appropriate test, using StatCrunch functions.

This is a one-sample proportion. problem.

$H_0: p = 0.063$ vs. $H_A: p \neq 0.063$

Data: $\hat{p} = \frac{14}{137} = 0.1022$

Conditions → Unbiased, 10/10 S/F.

stat → Prop → one →

Test Stat: $z = 1.89$

P-Value = 0.059

There is some evidence at Cecil, the proportion of Black/African American students differs from 6.3%.

Example: On average, are Cecil students "full time" students? Use the "***Fall 2012 Survey 1 Student Data***" dataset on StatCrunch. The variable is "Credits", and the data was collected from 117 students that semester. What must we assume about this sample?

This is a one-sample mean problem.

Conditions: Stat students mimic the general pop at Cecil. ? ✓

$n = 117 > 30$ ✓

Full Time = 12 credits

$H_0: \mu = 12$ credits

$H_A: \mu < 12$ credits

Data: $\bar{y} = 11.59$, $s = 3.72$

Stat → T → one → data

Test Stat: $t = -1.15$

Pvalue = 0.1253

Based on the data, we're not convinced our students average credit hours is under 12.