SEIS 631

Calculating Confidence Intervals

For some point estimate, we can build a Confidence Interval (CI) for the population parameter if the sampling distribution of the point estimate is normally distributed with standard error SE.

If $Samp.\ Dist(Point\ Estimate) \sim N(Point\ Estimate, SE)$, then the CI is found with

$$\text{Point Estimate} \pm z^* \cdot SE$$

where $z*$ (the critical value) corresponds to the confidence level selected.

Margin of Error

In a confidence interval, $z^* \cdot SE$ is called the **Margin of Error**

Conditions: (When can we use this formula)

1. **Independence**: Sampled observations must be independent.
   - True if you have random sample/assignment with replacement
   - If sampling without replacement, then still true if n < 10% of the population
2. **Sample size and skew**:
   - If the population is (or can reasonably be assumed to be) normal, there is no restriction on *n*
   - If the population is heavily skewed, then sample size needs to be larger, n > 30 as a rule of thumb

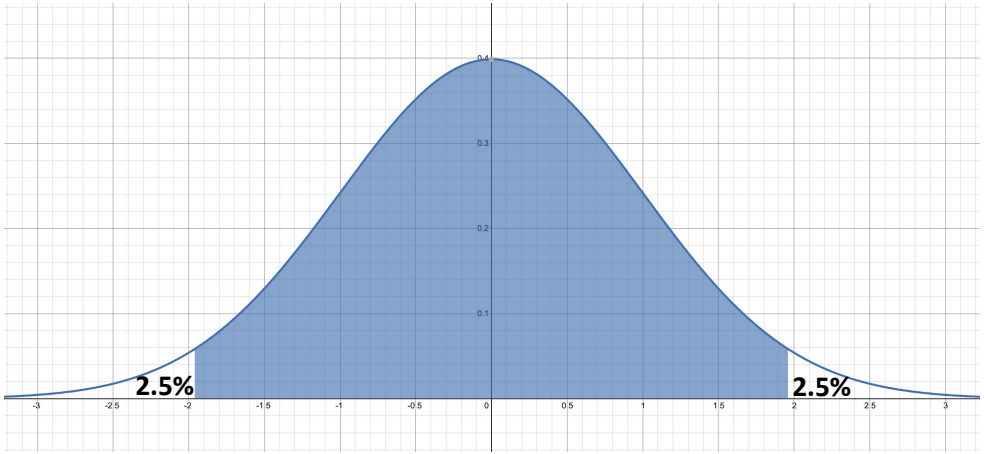
Finding the Critical Value

$$\text{point estimate} \pm z^{\star} SE$$

- For 95% CI, the sum of two tails is 5%
  - One tail is 2.5%

```
> qnorm(0.025) [1]
-1.959964
```

Z* = 1.96



Question

Which of the below Z scores is the appropriate z^* when calculating a 98% confidence interval?

(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = -2.33

(e) Z = -1.65

Question

Which of the below Z scores is the appropriate z^* when calculating a 98% confidence interval?

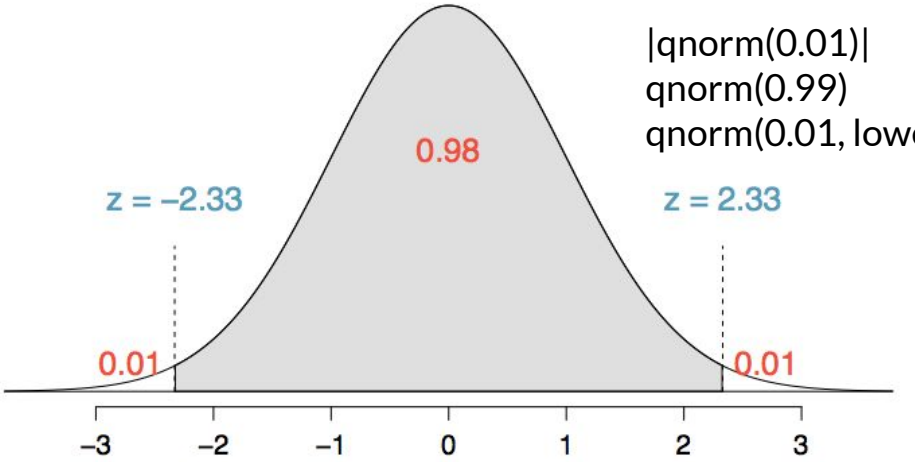
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = -2.33

(e) Z = -1.65



|qnorm(0.01)|
qnorm(0.99)
qnorm(0.01, lower.tail = FALSE

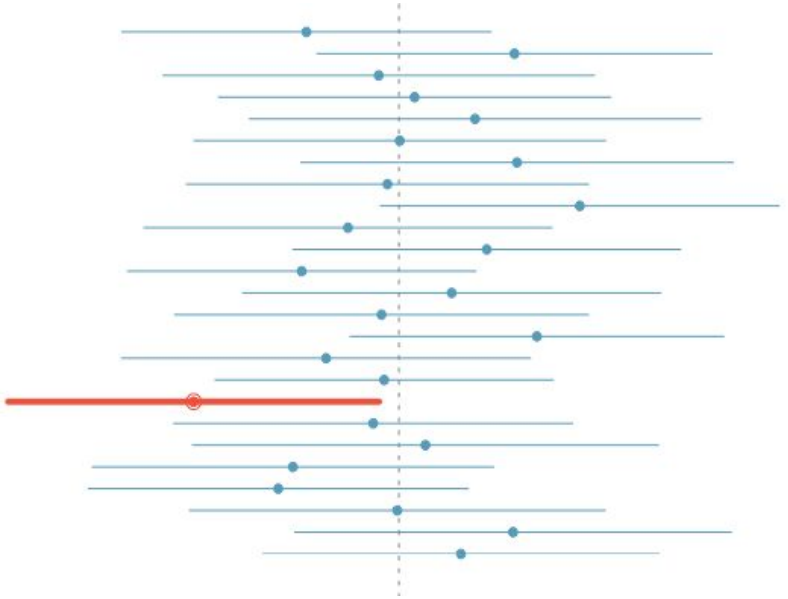
Accuracy vs Precision of CI

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate ± 2 SE*.

Then about 95% of those intervals would contain the true population mean (μ).

The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



Confidence Intervals and Growth Mindset

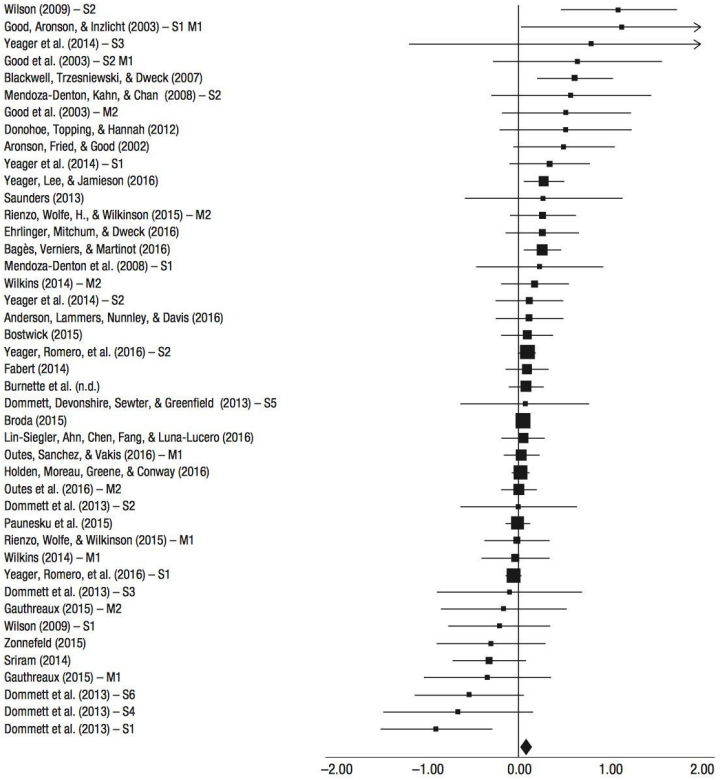


Fig. 4. Standardized mean differences (Cohen's *d*s) in academic achievement between students receiving a growth-mind-set intervention and students in the comparison group. Cohen's *d*s (squares) and 95% confidence intervals (error bars) are displayed for all effects entered

Compatibility Window vs. Confidence Interval

What does a 95% Compatibility Window mean?

- The parameters in the Window are compatible with the data at a 95% confidence level.
- It's reasonable to expect data like ours given values within the compatibility window.
- Values for the true parameter in this window are consistent with our data.
- The math for a Compatibility Window is identical to CI

How to Interpret Confidence Intervals

Correct: We are XX% confident that the true population parameter lies within (our interval)

Things to remember to include:

Our confidence level: ex. 99%, 95%, etc

What the population parameter we are estimating: ex. True average height of US men, true proportion of smokers in MN, etc.

Lower and upper limits of our interval: Ex. Lies between 5.4 and 6.6, lies betweend 0.4 and 0.9, etc.

Incorrect: Our confidence interval captures the true population parameter with a probability of 0.95

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see any drawbacks to using a wider interval?

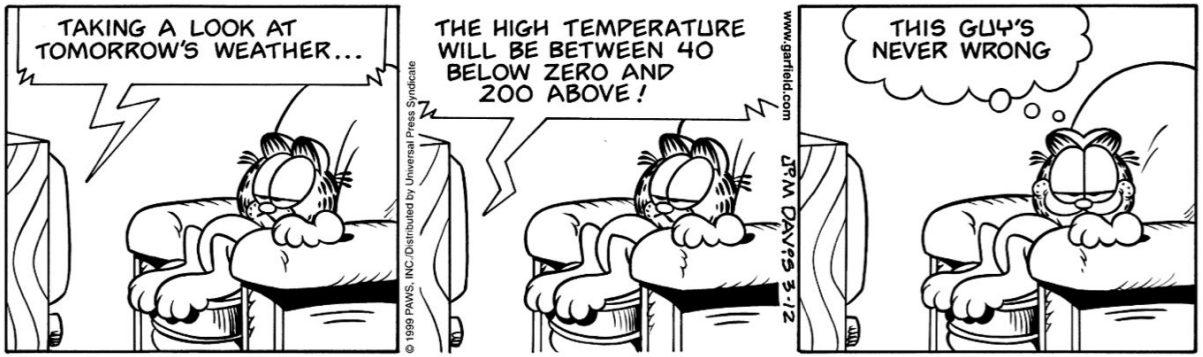
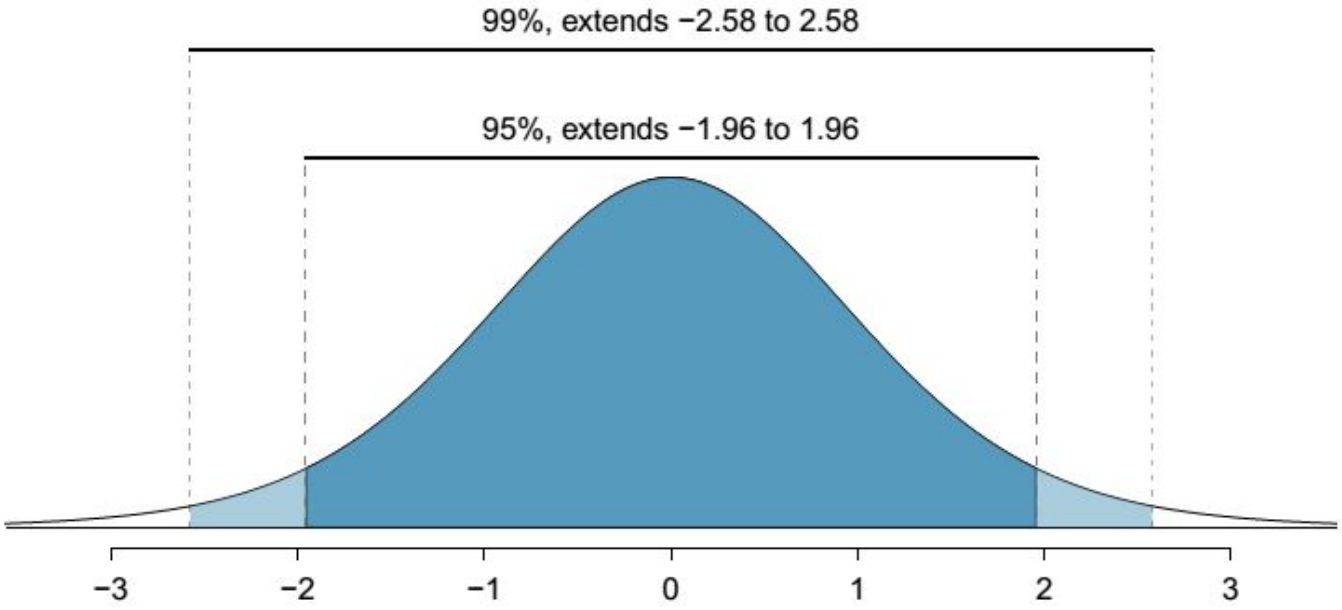
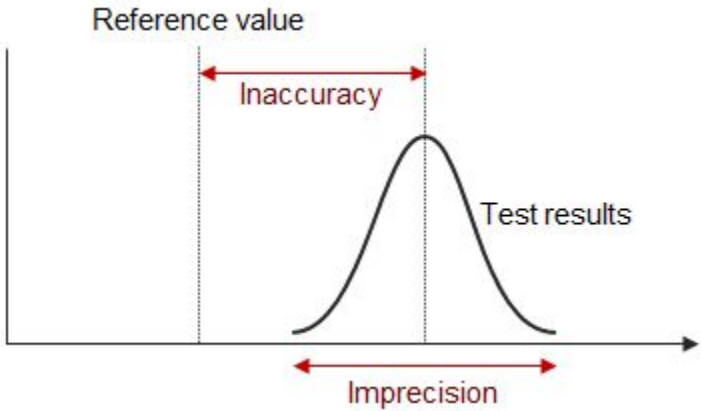


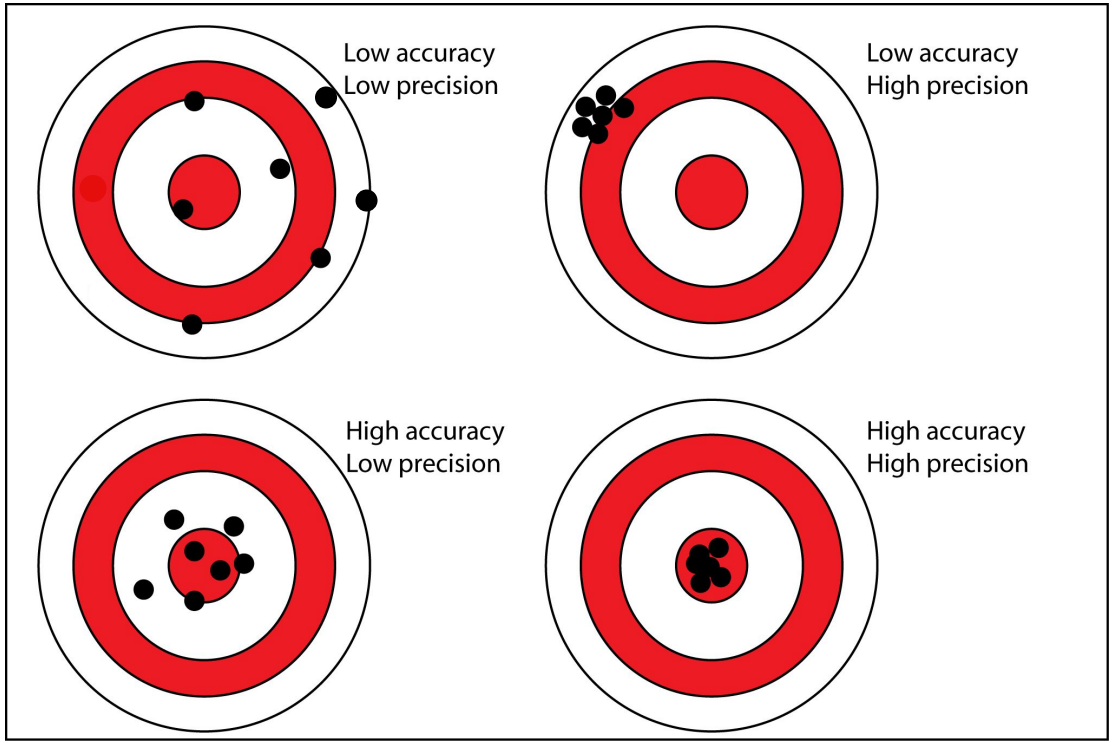
Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif



As Confidence Level goes up, the width goes up

- As Confidence Level goes up:
  - Width Increases
  - Accuracy Increases
  - Precision Decreases





Question

The mean of the sampling distribution is 5.75, and the standard deviation of the sampling distribution (also called the standard error) is 0.75. Which of the following is the most reasonable guess for the 95% confidence interval for the true average number of basketball games attended by students?

a) 5.75 ± 0.75

b) $5.75 \pm 2 \times 0.75$

c) $5.75 \pm 3 \times 0.75$

d) cannot tell from the information given

Sample Size

Finding a sample size for a certain margin of error

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

We know that the critical value associated with the 96% confidence level:

$z^* = 2.05$.

$$4 \geq 2.05 * 18/\sqrt{n} \rightarrow n \geq (2.05 * 18/4)^2 = 85.1$$

The minimum number of children required to attain the desired margin of error is 85.1. Since we can't sample 0.1 of a child, we must sample at least 86 children (round up, since rounding down to 85 would yield a slightly larger margin of error than desired).

Question

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

Confidence Interval Summary/Example

Question: How much time a day does a typical US resident spend in the car?

- How would you answer that question?
- 2016 American Driving Survey*
  - Average Time Spent in the Car (in 2016): 50.6 minutes
  - Standard Deviation: 65 minutes
  - Sample: 3,161 Drivers.
- Can we generalize this to the population at large?
- Calculate an 80% confidence interval

* Tefft, B. C. (2018, January). American Driving Survey: 2015-2016. (Research Brief). Washington, D.C.: AAA Foundation for Traffic Safety.

Confidence Interval Summary/Example

Question: How much time a day does a typical US resident spend in the car?

- What's the center of our confidence interval?

$$\bar{x} = 50.6$$

- What's the standard error (SE) for this sampling method?

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{65}{\sqrt{3161}} = 1.156$$

Confidence Interval Summary/Example

Question: How much time a day does a typical US resident spend in the car?

- What's the Margin of Error for an 80% CI?

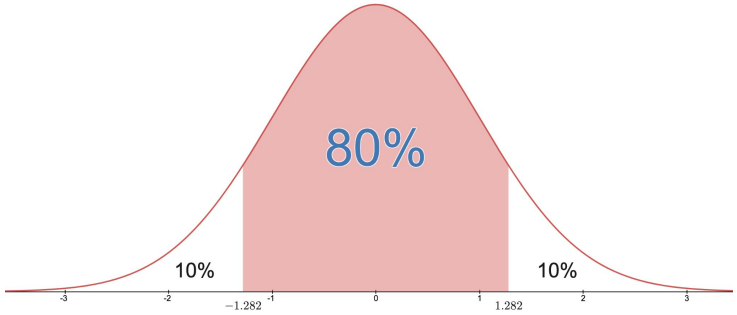
$$ME = z^* \cdot SE$$

Confidence Interval Summary/Example

Question: How much time a day does a typical US resident spend in the car?

- What's the Margin of Error for an 80% CI?

$$ME = z^* \cdot SE$$



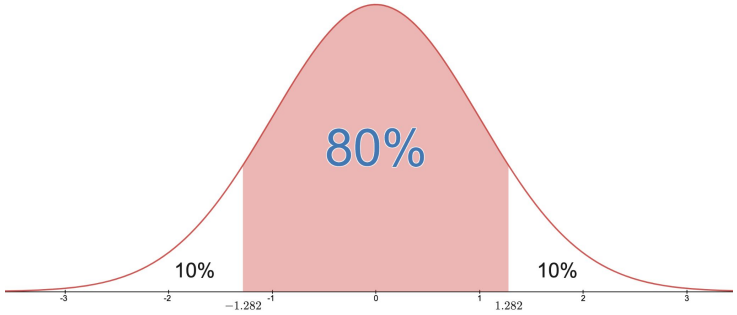
- What's z*?
  - Using R: **qnorm(0.1)**
    - Result: -1.28 → so z* = 1.28
    - Name two other ways (using R and the qnorm function) to get this.

Confidence Interval Summary/Example

Question: How much time a day does a typical US resident spend in the car?

- What's the Margin of Error for an 80% CI?

$$ME = z^* \cdot SE$$



- What's z*?
  - Using R: **qnorm(0.1)**
    - Result: -1.28 → so z* = 1.28
    - Name two other ways (using R and the qnorm function) to get this.
- Margin of Error:

$$ME = z^* \cdot SE = 1.28 \cdot 1.156 = 1.480$$

Confidence Interval Summary Example

Question: How much time a day does a typical US resident spend in the car?

- What's the 80% CI?

$$CI = \bar{x} \pm z^* \cdot SE$$
$$= \bar{x} \pm ME$$
$$= 50.6 \pm 1.48$$

(49.12, 52.08)

Confidence Interval Review

Question: How much time a day does a typical US resident spend in the car?

- What's the 80% CI? (49.12, 52.08)

"We are 80% confident that the true average time spent in a car for *all* US residents is between 49.12 minutes and 52.08 minutes."

"An average value for all US residents between 49.12 minutes and 52.08 minutes is compatible with our data at a confidence level of 80%"

Example

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

Conditions:

- Random Sample and 50 < 10% of all College Students.
- We can assume that the number of exclusive relationships one student in the sample has been in, is independent of another. So we have independent observations.
- Sample size is greater than 30, and the distribution of the sample is not so skewed. We can assume, that the sampling distribution of average number of exclusive relationships from samples of size 50 will be nearly normal.

$\bar{x} = 3.2 \qquad s = 1.74$

95% confidence interval is defined as

point estimate ± 1.96 SE

$SE = s / \sqrt{n} = 1.74 / \sqrt{50} \approx 0.246$

$\bar{x} \pm 1.96\ SE \rightarrow 3.2 \pm 1.96 \times 0.246$

$\rightarrow 3.2 \pm 0.48$

$\rightarrow (2.72, 3.68)$

We are 95% confident that college students on average have been in 2.72 to 3.68 exclusive relationships.

Confidence Intervals for Other Statistics

- In general confidence intervals are constructed by:

$$\text{point estimate} \pm ME$$
$$= \text{point estimate} \pm z^* \cdot SE$$

- SE (and thus ME) are calculated differently for different statistics. E.g.
  - Estimating a mean: $SE = \frac{\sigma}{\sqrt{n}}$
  - Estimating a proportion: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - Estimating a slope: $SE = \sqrt{\frac{\sum e^2}{(n-2)\sum(x-\bar{x})^2}}$
- In all cases it represents how spread out we expect the the value to be over different samples, i.e. if we repeated the analysis many times with different samples.

Sample Proportions

Sugary effervescent beverages: What do you call them?

Sample Proportions

Sugary effervescent beverages: What do you call them?

- We can estimate the proportion in the population who call it “pop” (for example) from our sample.

$$\hat{p} \rightarrow p$$

Sample Proportions

Sugary effervescent beverages: What do you call them?

- We can estimate the proportion in the population who call it “pop” (for example) from our sample.

$$\hat{p} \rightarrow p$$

- Standard Error (standard deviation of the sampling distribution)

$$SE = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Interval for the Proportion

- The CI is calculated just as before with the correct SE.

$$\hat{p} \pm ME$$

$$\hat{p} \pm z^* \cdot SE$$

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Hypothesis Testing

Hypothesis Testing Principles

1. Determine the hypothesis:
    a. Null Hypothesis: what if we are wrong, nothing interesting is happening, the status quo.
    b. Alternative Hypothesis: the opposite of the null.
2. Imagine (assume) the Null ***is true***. What data would you expect?
3. Collect data & compare it to what you expected assuming the Null?
4. Is it unlikely to see your data if you assume the null is true?
    a. You've got evidence against the null! And in favor of the alternative!!
    b. We "reject the null"
5. Is it likely that we would see data like ours if the null were true?
    a. Then we can't conclude the null is wrong. We "fail to reject the null."

Hypothesis Testing General Procedures

1. Identify the *Null* and *Alternative* hypotheses.
2. Calculate a sample statistic
3. Compare the sample statistic to the *Null Hypothesis* to calculate a ***Test Statistic***
4. Compare the ***Test Statistic*** to a theoretical distribution (like the Normal Distribution) to get a ***CI*** and ***p-value***

5. Use the CI and p-value to inform you decision about the Null Hypothesis

Example

- It is know that, historically, the leaves on a particular tree have an average length of 11 cm with a standard deviation of 2.8 cm.
- Here is a sample taken this fall. Has the average leaf changed in size?

Length (cm)		
13.6	16	12.4
12.6	13.8	14.6
11.9	6.5	11.5
15.3	13.1	4.9
12.6	11.7	5.1

Comparing a *sample statistic* $(\overline{x} = 11.7)$
To a "*known*" or "*historical*" value
(μ = 11 cm)

Guess the Population Applet

https://www.geogebra.org/m/JPMnJRjF

Hypothesis Testing Example

1-sample Z test

Scenario: Testing whether or not the mean of a certain group is equal to a hypothesized value

Test Statistic: The number of *standard errors* away from the null. Used to reject or fail to reject our null

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Comparing Leaves Again

- It is know that, historically, the leaves on a particular tree have an average length of 11 cm with a standard deviation of 2.8 cm.
- Here is a sample taken this fall. Has the average leaf changed in size?

Length (cm)		
13.6	16	12.4
12.6	13.8	14.6
11.9	6.5	11.5
15.3	13.1	4.9
12.6	11.7	5.1

Comparing Leaves Again

- It is know that, historically, the leaves on a particular tree have an average length of 11 cm with a standard deviation of 2.8 cm.
- Here is a sample taken this fall. Has the average leaf changed in size?

Length (cm)		
13.6	16	12.4
12.6	13.8	14.6
11.9	6.5	11.5
15.3	13.1	4.9
12.6	11.7	5.1

$$\bar{x} = 11.7$$

$$sd = 3.47$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Comparing Leaves Again

- It is know that, historically, the leaves on a particular tree have an average length of 11 cm with a standard deviation of 2.8 cm.
- Here is a sample taken this fall. Has the average leaf changed in size?

Length (cm)		
13.6	16	12.4
12.6	13.8	14.6
11.9	6.5	11.5
15.3	13.1	4.9
12.6	11.7	5.1

$$\bar{x} = 11.7$$

$$sd = 3.47$$

$$z = \frac{11.7 - 11}{2.8/\sqrt{15}}$$

Comparing Leaves Again

- It is know that, historically, the leaves on a particular tree have an average length of 11 cm with a standard deviation of 2.8 cm.
- Here is a sample taken this fall. Has the average leaf changed in size?

Length (cm)		
13.6	16	12.4
12.6	13.8	14.6
11.9	6.5	11.5
15.3	13.1	4.9
12.6	11.7	5.1

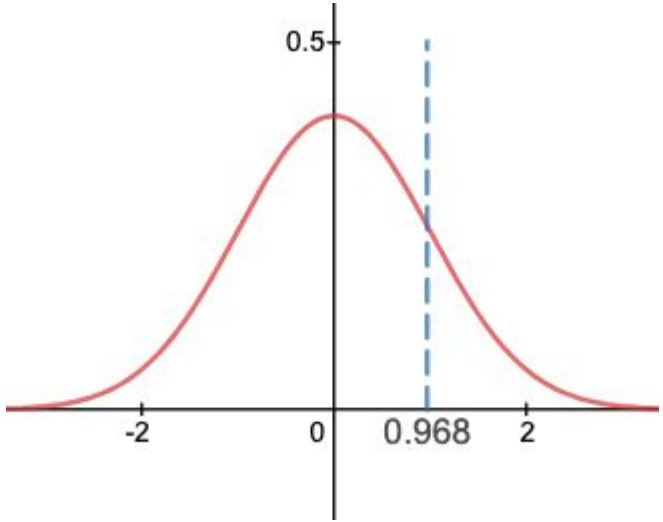
$$\bar{x} = 11.7$$

$$sd = 3.47$$

$$z = \frac{11.7-11}{2.8/\sqrt{15}} = 0.968$$

Calculating *p*-values

- What does the *z*-score tell us?
  - 11.7 is 0.968 ***standard errors*** above the mean.
- If there were ***truly*** no difference, how likely are we to get such a difference?



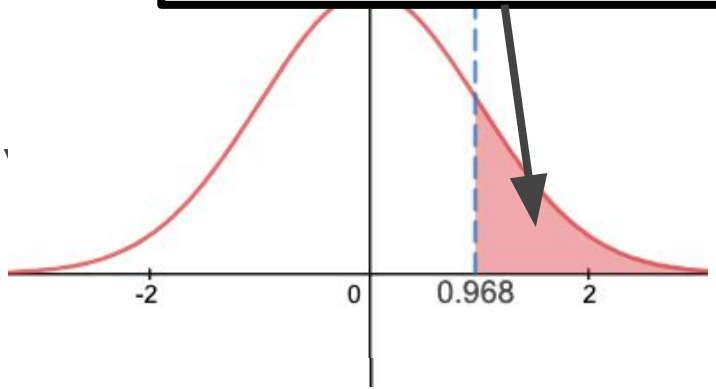
$$z = \frac{11.7-11}{2.8/\sqrt{15}} = 0.968$$

Calculating *p*-values

- What does the *z*-score tell us?
  - 11.7 is 0.968 ***standard errors*** above the mean.
- If there were ***truly*** no difference, how likel
  are we to get such a difference?

P(Z>0.968) = 0.167

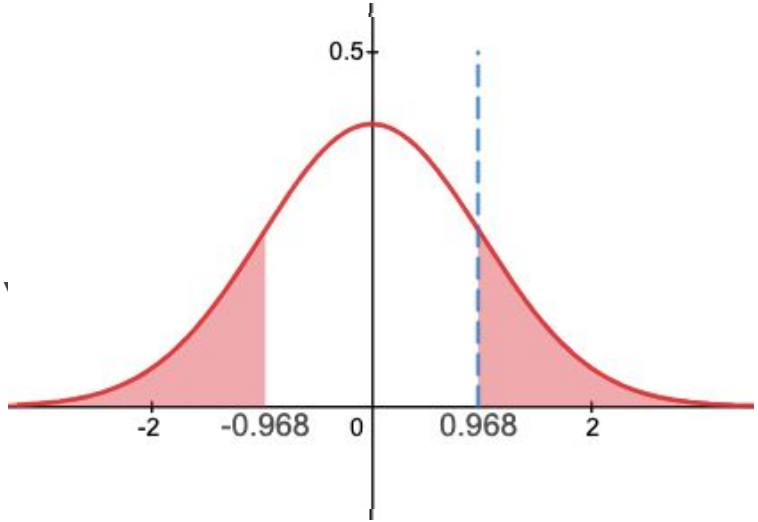
```
>pnorm(0.968, lower.tail = FALSE)
```



$$z = \frac{11.7-11}{2.8/\sqrt{15}} = 0.968$$

Calculating *p*-values

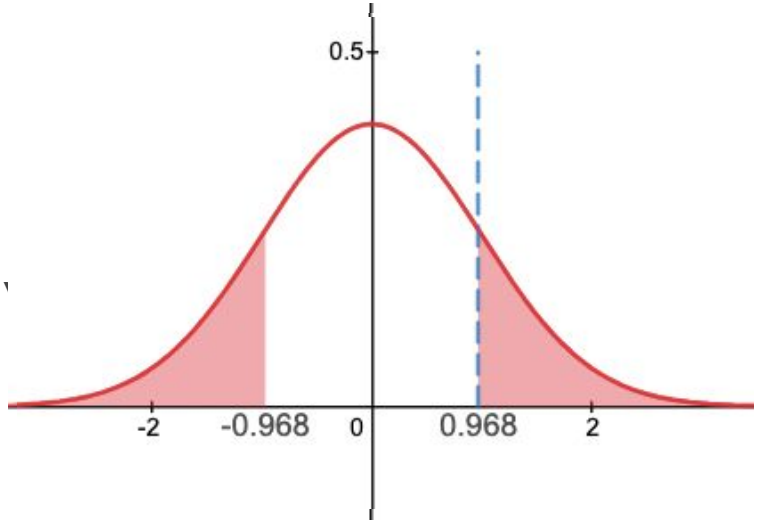
- What does the *z*-score tell us?
  - 11.7 is 0.968 ***standard errors*** above the mean.
- If there were ***truly*** no difference, how likel
  are we to get such a difference?



$$z = \frac{11.7 - 11}{2.8/\sqrt{15}} = 0.968$$

Calculating *p*-values

- What does the *z*-score tell us?
  - 11.7 is 0.968 ***standard errors*** above the mean.
- If there were ***truly*** no difference, how likel
  are we to see such a difference?



P(Z>0.968) = 0.167

P(Z> 0.968 *or* Z<-0.968) =
P(|Z|> 0.968) = 2 x P(Z > 0.968)
= 2 x 0.167
= 0.334

p-value: p = 0.334

$$z = \frac{11.7-11}{2.8/\sqrt{15}} = 0.968$$

Making Sense of p-values

- Mathematically, a *p*-value is the probability, assuming the ***null hypothesis*** is true, of seeing data that is *at least as extreme* as our data.
- Calculated based on z-scores and the standard normal distribution.
- If we include both sides of the distribution it is a **two-tailed** p-value.
- If we only include one tail it is a **one-tailed** p-value.

A More Typical Example of Hypothesis Testing

1-sample Z test

Scenario: Testing whether or not the mean of a certain group is equal to a hypothesized value

Test Statistic: The number of *standard errors* away from the null. Used to reject or fail to reject our null

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

σ vs. s

- If the population standard deviation (σ) is known, use this symbol and value for your calculations.

- If σ is not known AND we have a large enough sample (n>30 or so), we can use the *sample* standard deviation (s) in our equation.

  (Central Limit Theorem)

- We will learn a better way to do this if σ is not known in a little while.

Exploring z-scores

https://www.geogebra.org/m/JPMnJRjF

1 sample z test: One More Example

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A random sample of 169 college students yielded an average of 6.88, with a standard deviation of 0.94 hours. Does the data provide convincing evidence that the average amount of sleep college students get per night is *different* from the national average stated above?

Average Sleep Hypothesis Test

1. What are the null and alternative hypothesis?
2. Is this a two tailed or one tailed test? How do you know?
3. Calculate the test statistic (what is it called?).
4. Determine the p-value.
5. What does the p-value mean?
6. Calculate a 95% CI around the mean.
7. What conclusion do these data support?

Average Sleep Hypothesis Test

1. Set null and alternative:
   $H_o: \mu = 7$ vs. $H_A: \mu \neq 7$
2. Is it one sided or two? Two
3. Calculate test statistic (Z-test)

$$Z = \frac{6.88 - 7}{0.94/\sqrt{169}} = -1.659$$

4. Determine p-value

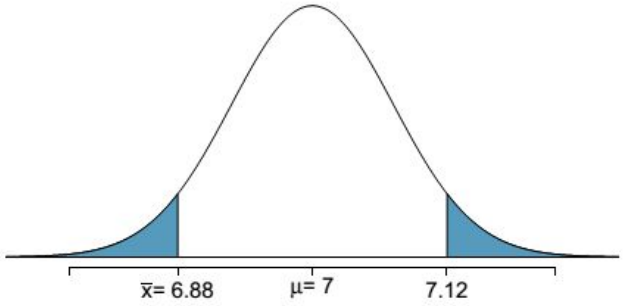
> pnorm(-1.6596, lower.tail = TRUE)

[1] 0.04849747

2 x 0.0485 = 0.097 (p value)

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- Hence the p-value would change as well:



p-value
$= 0.0485 \times 2$
$= 0.097$

Average Sleep Hypothesis Test

5. What does the p-value mean?

p = 0.097, which means that assuming the *null hypothesis*, that college students sleep 7 hours on average, we would expect samples of size 169 to have averages *at least as extreme* as our sample (e.g. 6.88, or ±1.659 standard errors) 9.7% of the time.

Average Sleep Hypothesis Test

6. Calculate a 95% CI around the mean.

$$
\begin{aligned}
CI &= \bar{x} \pm ME \\
&= \bar{x} \pm z^* \cdot SE \\
&= 6.88 \pm 1.96 \cdot \frac{0.94}{\sqrt{169}}
\end{aligned}
$$

$$CI = [6.74,\ 7.02]$$

Average Sleep Hypothesis Test

7. What conclusion do these data support?

According to the p-value, there is about a 1/10 chance of seeing data at least as extreme as this. Additionally, 7 (the mean under the *null hypothesis*) is ***compatible*** with our data at a 95% confidence level. Thus we **do not** have evidence against the null hypothesis. We fail to reject the null. Our data do not support the hypothesis that college students sleep an average different from 7 hours.

$$Z = -1.659$$
$$p = 0.097$$
$$CI = [6.74,\ 7.02]$$

1 sample z test vs. z score

1 sample Z test

- Deals with sample mean
- # of standard errors (σ/sqrt(n)) from the mean
- Both use standard normal dist.
- Calculate probabilities in same way

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Z Score

- Deals with single observation
- # of standard deviations (σ) from the mean
- Both use standard normal dist.
- Calculate probabilities in same way

$$z = \frac{x - \mu}{\sigma}$$

Critical Values

- How small is small for a p-value? How extreme is extreme enough for a z-score?
- It is common practice to set cut-offs prior to running your tests.
  - Z-score cut-off: z^*
    - Say $z^* = 1.96$. If $Z > z^*$, i.e. if $Z > 1.96$ it is considered extreme
  - P-value cut-off: α
    - Say $\alpha = 0.05$ (5%). If $p < \alpha$, i.e. if $p < 0.05$, it is considered small
  - "Statistically Significant"

- Current best practice: Don't set cut-offs, use a wholistic approach.

1-sample z vs. Confidence Interval

Assume α =0.05

In a 2 sided z test, reject null if

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq -1.96 \quad \text{or} \quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq 1.96$$

Rearranging with some Algebra… We reject CI's if

$$\mu \geq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

Upper CI Bound

$$\mu \leq \bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$$

Lower CI Bound

Alternative to p-value: Report Confidence Intervals

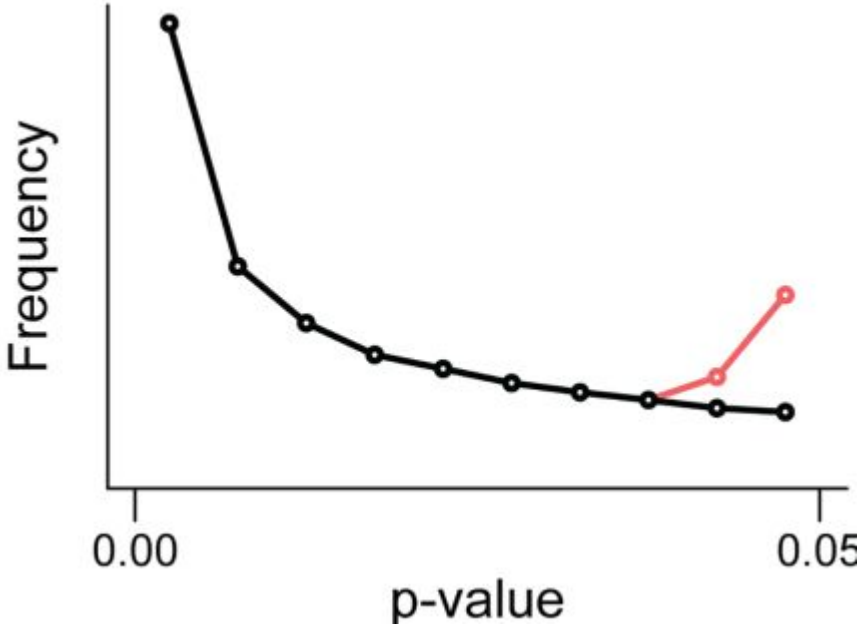
- Dance of the p-value video: https://www.youtube.com/watch?v=5OL1RqHrZQ8

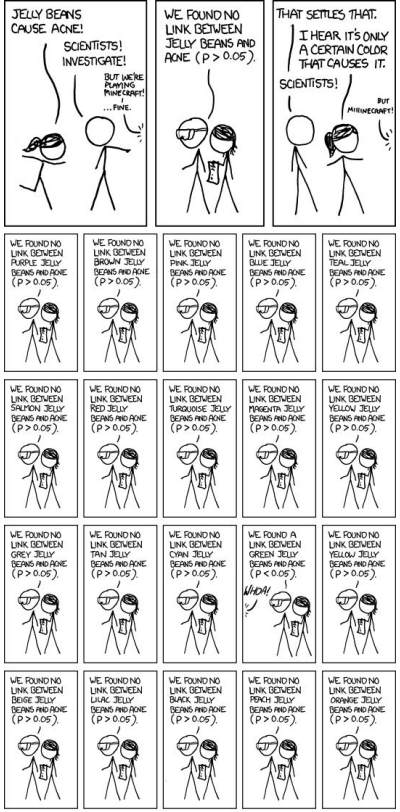
CI's might provide more meaningful information

A small p-value might mean large effect or large sample size, don't know!

CI's show effect size, and our uncertainty

Statistical Issues
P-value Hacking





Statistical Error

	Null Hypothesis (Truth)	
Decision	**False**	**True**
Reject	Correct Decision (prob = 1 – β)	Type I error (prob = α)
Fail to Reject	Type II Error (prob = β)	Correct Decision (prob = 1 – α)

Type I error: Rejecting the null when it is true in reality (with probability of α, typically set at 0.05)

Type II error: Failing to reject the null when it is false in reality (with probability of β)

Type I vs. Type II error

- The villagers in the boy who cried wolf:
  - “Wolf!!” They come running --- Type I Error
  - “Wolf!!” They come running --- Type I Error
  - “Wolf!!” They come running --- Type I Error
  - “No really, wolf!!” They DON’T come running --- Type II Error

Statistical Power

Power: The probability of rejecting the null hypothesis when it's false

Power = 1 – β = 1 – probability of Type II error

Significance level: (α) our threshold of whether or not to reject the null

	Null Hypothesis (Truth)	
Decision	**False**	**True**
Reject	Correct Decision (prob = 1 – β) *Power*	Type I error (prob = α) *Significance*
Fail to Reject	Type II Error (prob = β)	Correct Decision (prob = 1 – α)

Statistical Power - Intuition

Scientific test is like an instrument used to detect something (ex. Telescope)

- Powerful telescope will let you see the moons of Mars
- Cannot see them with binoculars (underpowered test)

The moons are still there, but our ability to detect them depends on the power of our test

Increasing statistical power

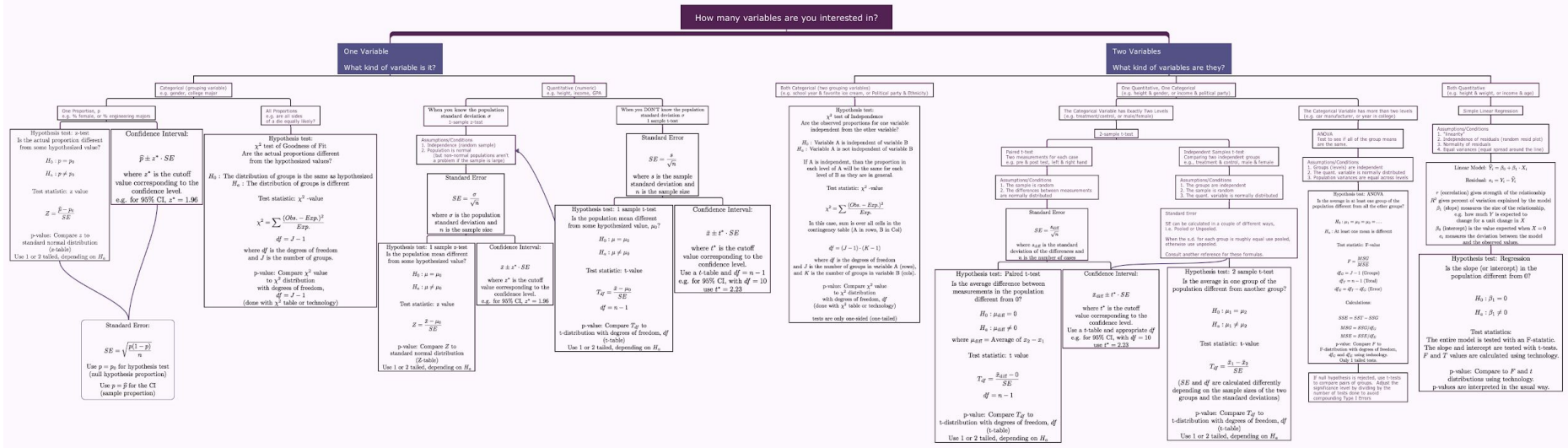
- Increase alpha
  - Set at the beginning of experiment
- Conduct one tailed test
  - Have to decide before experiment
- Decrease random error
  - Through more advanced sampling and experimental techniques
- Increase sample size
  - Depending on the situation, this is the most straightforward!

Tradeoff of Power

- UK 1995: Committee on Safety of Medications issued a warning that a certain birth control pill increased the risk of a dangerous embolism 100%
  - Risk went from about 1 in 7000 to 2 in 7000
  - Results were statistically significant, but not practically.
  - This warning was blamed for 13,000 unwanted pregnancies and may have saved 2 to 6 people per 10,000,000 users
- Powerful tests will pick up a “real” effect, however, the effect may not be meaningful, so we may be better off not being as sensitive.

High power: Detect “truth”, but may be too “sensitive”

Low power: Fail to detect “truth”, but don’t “overreact”


How many variables are you interested in?
One Variable
What kind of variable is it?
Categorical (grouping variable)
e.g. gender, college major
One Proportion, p
e.g. % female, or % engineering majors
Hypothesis test: z-test
Is the actual proportion different from some hypothesized value?
$H_0 : p = p_0$
$H_a : p \neq p_0$
Test statistic: z value
$Z = \frac{\hat{p} - p_0}{SE}$
p-value: Compare z to standard normal distribution (z-table)
Use 1 or 2 tailed, depending on H_a
Confidence Interval:
$\hat{p} \pm z^* \cdot SE$
where z^* is the cutoff value corresponding to the confidence level.
e.g. for 95% CI, $z^* = 1.96$
Standard Error:
$SE = \sqrt{\frac{p(1-p)}{n}}$
Use $p = p_0$ for hypothesis test (null hypothesis proportion)
Use $p = \hat{p}$ for the CI (sample proportion)
All Proportions
e.g. are all sides of a die equally likely?
Hypothesis test:
χ^2 test of Goodness of Fit
Are the actual proportions different from the hypothesized values?
H_0 : The distribution of groups is the same as hypothesized
H_a : The distribution of groups is different
Test statistic: χ^2 -value
$\chi^2 = \sum \frac{(Obs. - Exp.)^2}{Exp.}$
$df = J - 1$
where df is the degrees of freedom and J is the number of groups.
p-value: Compare χ^2 value to χ^2 distribution with degrees of freedom, $df = J - 1$
(done with χ^2 table or technology)
Quantitative (numeric)
e.g. height, income, GPA
When you know the population standard deviation σ
1-sample z-test
Assumptions/Conditions
1. Independence (random sample)
2. Population is normal
(but non-normal populations aren't a problem if the sample is large)
Standard Error
$SE = \frac{\sigma}{\sqrt{n}}$
where σ is the population standard deviation and n is the sample size
Hypothesis test: 1 sample z-test
Is the population mean different from some hypothesized value?
$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$
Test statistic: z value
$Z = \frac{\bar{x} - \mu_0}{SE}$
p-value: Compare Z to standard normal distribution (Z-table)
Use 1 or 2 tailed, depending on H_a
Confidence Interval:
$\bar{x} \pm z^* \cdot SE$
where z^* is the cutoff value corresponding to the confidence level.
e.g. for 95% CI, $z^* = 1.96$
When you DON'T know the population standard deviation σ
1 sample t-test
Standard Error
$SE = \frac{s}{\sqrt{n}}$
where s is the sample standard deviation and n is the sample size
Hypothesis test: 1 sample t-test
Is the population mean different from some hypothesized value, μ_0?
$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$
Test statistic: t-value
$T_{df} = \frac{\bar{x} - \mu_0}{SE}$
$df = n - 1$
p-value: Compare T_{df} to t-distribution with degrees of freedom, df (t-table)
Use 1 or 2 tailed, depending on H_a
Confidence Interval:
$\bar{x} \pm t^* \cdot SE$
where t^* is the cutoff value corresponding to the confidence level.
Use a t-table and $df = n - 1$
e.g. for 95% CI, with $df = 10$ use $t^* = 2.23$
Both Categorical (two grouping variables)
(e.g. school year & favorite ice cream, or Political party & Ethnicity)
Hypothesis test:
χ^2 test of Independence
Are the observed proportions for one variable independent from the other variable?
H_0 : Variable A is independent of variable B
H_a : Variable A is not independent of variable B
If A is independent, than the proportion in each level of A will be the same for each level of B as they are in general.
Test statistic: χ^2 -value
$\chi^2 = \sum \frac{(Obs. - Exp.)^2}{Exp.}$
In this case, sum is over all cells in the contingency table (A in rows, B in Col)
$df = (J - 1) \cdot (K - 1)$
where df is the degrees of freedom and J is the number of groups in variable A (rows), and K is the number of groups in variable B (cols).
p-value: Compare χ^2 value to χ^2 distribution with degrees of freedom, df (done with χ^2 table or technology)
tests are only one-sided (one-tailed)
Two Variables
What kind of variables are they?
One Quantitative, One Categorical
(e.g. height & gender, or income & political party)
The Categorical Variable has Exactly Two Levels
(e.g. treatment/control, or male/female)
2-sample t-test
Paired t-test
Two measurements for each case
e.g. pre & post test, left & right hand
Assumptions/Conditions
1. The sample is random
2. The differences between measurements are normally distributed
Standard Error
$SE = \frac{s_{diff}}{\sqrt{n}}$
where s_{diff} is the standard deviation of the differences and n is the number of cases
Hypothesis test: Paired t-test
Is the average difference between measurements in the population different from 0?
$H_0 : \mu_{diff} = 0$
$H_a : \mu_{diff} \neq 0$
where μ_{diff} = Average of $x_2 - x_1$
Test statistic: t value
$T_{df} = \frac{\bar{x}_{diff} - 0}{SE}$
$df = n - 1$
p-value: Compare T_{df} to t-distribution with degrees of freedom, df (t-table)
Use 1 or 2 tailed, depending on H_a
Confidence Interval:
$\bar{x}_{diff} \pm t^* \cdot SE$
where t^* is the cutoff value corresponding to the confidence level.
Use a t-table and appropriate df
e.g. for 95% CI, with $df = 10$ use $t^* = 2.23$
Independent Samples t-test
Comparing two independent groups
e.g., treatment & control, male & female
Assumptions/Conditions
1. The groups are independent
2. The sample is random
3. The quant. variable is normally distributed
Standard Error
SE can be calculated in a couple of different ways, i.e. Pooled or Unpooled.
When the s.d. for each group is roughly equal use pooled, otherwise use unpooled.
Consult another reference for these formulas.
Hypothesis test: 2 sample t-test
Is the average in one group of the population different from another group?
$H_0 : \mu_1 = \mu_2$
$H_a : \mu_1 \neq \mu_2$
Test statistic: t-value
$T_{df} = \frac{\bar{x}_1 - \bar{x}_2}{SE}$
(SE and df are calculated differently depending on the sample sizes of the two groups and the standard deviations)
p-value: Compare T_{df} to t-distribution with degrees of freedom, df (t-table)
Use 1 or 2 tailed, depending on H_a
The Categorical Variable has more than two levels
(e.g. car manufacturer, or year in college)
ANOVA
Short for one If all of the group means are the same
Assumptions/Conditions
1. Groups (levels) are independent
2. The quant. variable is normally distributed
3. Population variances are equal across levels
Hypothesis test: ANOVA
Is the average in at least one group of the population different from all the other groups?
$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots$
H_a : At least one mean is different
Test statistic: F-value
$F = \frac{MSG}{MSE}$
$df_G = J - 1$ (Groups)
$df_T = n - 1$ (Total)
$df_E = df_T - df_G$ (Error)
Calculations:
$SSE = SST - SSG$
$MSG = SSG/df_G$
$MSE = SSE/df_E$
p-value: Compare F to F-distribution with degrees of freedom, df_G and df_E using technology.
Only 1 tailed tests.
If null hypothesis is rejected, use t-tests to compare pairs of groups. Adjust the significance level by dividing by the number of tests done to avoid compounding Type I Errors
Both Quantitative
(e.g. height & weight, or income & age)
Simple Linear Regression
Assumptions/Conditions
1. "Linearity"
2. Independence of residuals (random resid plot)
3. Equality of residuals
4. Equal variances (equal spread around the line)
Linear Model: $\hat{Y}_i = \beta_0 + \beta_1 \cdot X_i$
Residual: $e_i = Y_i - \hat{Y}_i$
r (correlation) gives strength of the relationship
R^2 gives percent of variation explained by the model
β_1 (slope) measures the size of the relationship, e.g. how much Y is expected to change for a unit change in X
β_0 (intercept) is the value expected when $X = 0$
e_i measures the deviation between the model and the observed values.
Hypothesis test: Regression
Is the slope (or intercept) in the population different from 0?
$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$
Test statistics:
The entire model is tested with an F-statistic. The slope and intercept are tested with t-tests. F and T values are calculated using technology.
p-value: Compare to F and t distributions using technology.
p-values are interpreted in the usual way.

Hypothesis Testing with Numeric Data

One Sample z-test Example

Boys of a certain age are known to have a mean weight of $\mu = 85$ pounds. A complaint is made that the boys living in a municipal children's home are underfed. As one bit of evidence, $n = 25$ boys (of the same age) are weighed and found to have a mean weight of $\bar{x} = 80.94$ pounds. It is known that the population standard deviation σ is 11.6 pounds (the unrealistic part of this example!). Based on the available data, what should be concluded concerning the complaint?

One sample z-test Example

- What question are we trying to answer?
  - Are the boys sufficiently underweight to convince us they are being underfed?

- What are the Null and Alternative Hypothesis?

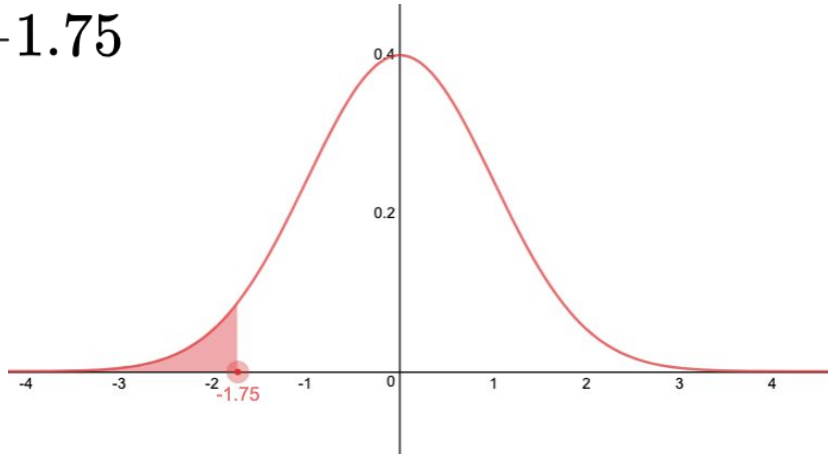
$$H_0 : \mu_{mc} = \mu_{population} \text{ or } \mu_{mc} = 85$$

$$H_a : \mu_{mc} < \mu_{population} \text{ or } \mu_{mc} < 85$$

One sample z-test Example

- What test?
  - We are comparing two means (μ_{mc} and $\mu_{population}$)
  - We know the population standard deviation
  - → Z-test
- Calculate the test statistic

$$Z = \frac{\bar{x}_{mc} - \mu_{population}}{SE} = \frac{80.94 - 85}{11.6/\sqrt{25}} = -1.75$$



One sample z-test Example

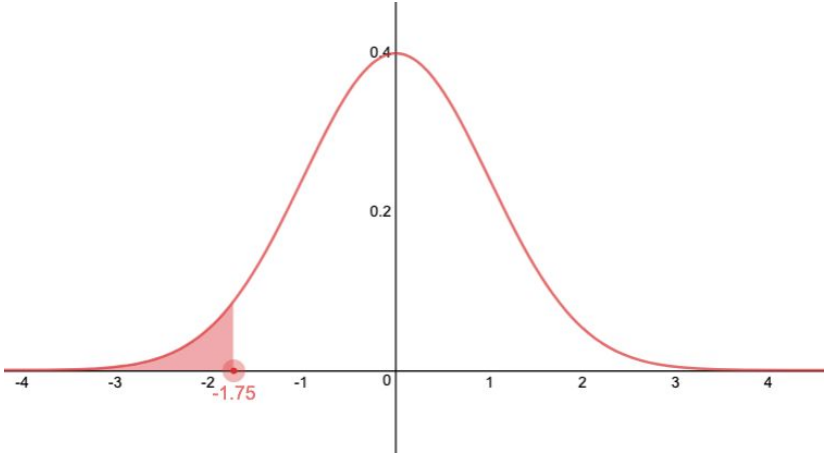
- Calculate the *p*-value (with R):

  > pnorm(-1.75)
  [1] 0.04005916

  or

  > pnorm(80.94,85,11.6/sqrt(25))
  [1] 0.04005916

  p = 0.04

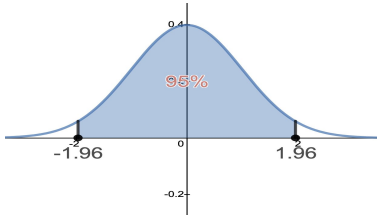


One sample z-test Example

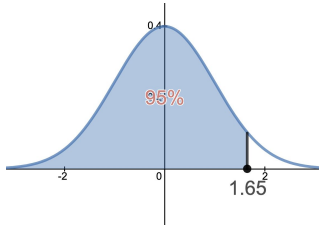
- Calculate the *95% CI*: This is a one sided test, so we can construct a 1-sided CI. For a 1-sided CI we put no limit on either the upper bound or lower bound.

  Instead of saying we are 95% confident μ is in [a, b]

  We say we are 95% confident μ < b (or μ > a).

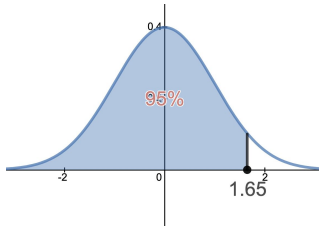


- In this case, we lump all the error on one side so we will get different z^* values.



One sample z-test Example

- For this example, we want all the error on the positive side.
  - We may have underestimated the weight, i.e. the actual weight is higher (which might change our conclusion). How high could it reasonably be?
  - If it is actually lower than we estimated, it wouldn't change our conclusion (at least for this test).



$$CI = 80.94 + 1.65 \cdot 2.32 = 84.76$$

- CI:
  - We are 95% confident that the true average is less than 84.76 lbs.

$$\mu < 84.76$$

One sample z-test Example: Conclusion

- Result of Test
  - P-value: p = 0.04
  - CI: $\mu < 84.76$
- Conclusion
  - We are 95% confident that true average weight for boys at the municipal center is less than 84.76 lbs. Values in this interval are reasonably compatible with our data.
  - The p value indicates that if we assume the average weight for boys at the municipal center is 85 lbs (and we have met all model assumptions) then the probability of collecting data like this or more extreme is 4% (1 in 25).
  - Are the boys being underfed?

Difference of two means
- Confidence intervals for differences of means
- Hypothesis tests for differences of means

Hours Worked & Education

- Do people with more education work more, less, or about the same as those with less education?
- How could we investigate this?

The General Social Survey (GSS)

- Large scale survey conducted by the Census Bureau each year
- First started in 1972
    - Many questions have remained the same

Hours Worked & Education

Here are some data from the 2010 GSS. The variables are

degree: Level of highest educational attainment

hrs1: Number of hours worked per week

Is there a relationship? How should we start?

A tibble: 1,172 x 2

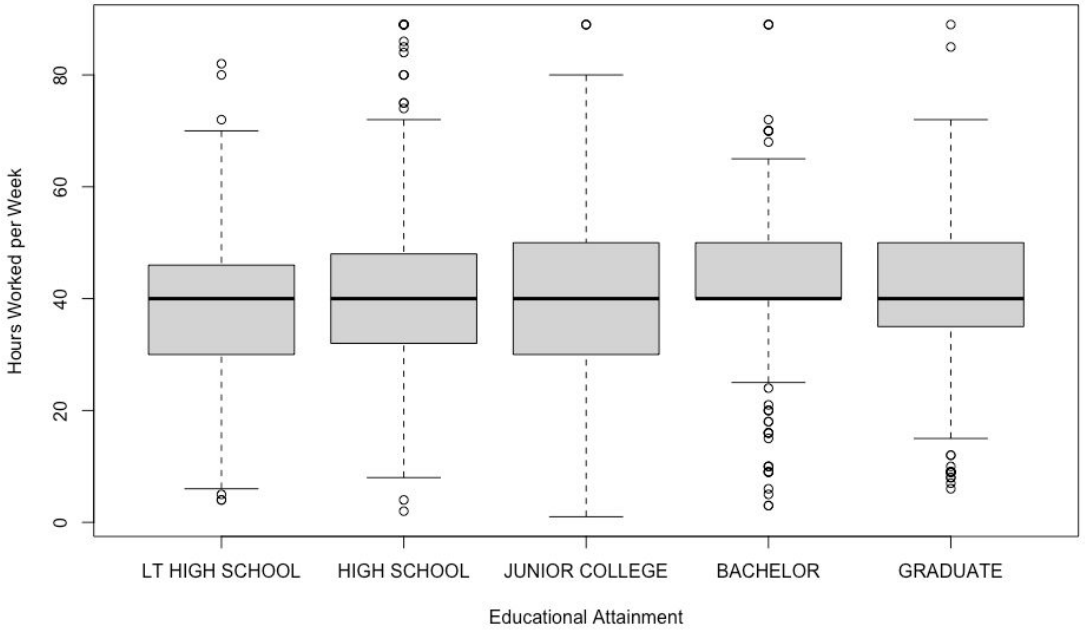
	degree	hrs1
	<fct>	<int>
1	BACHELOR	55
2	BACHELOR	45
3	JUNIOR COLLEGE	45
4	HIGH SCHOOL	40
5	HIGH SCHOOL	48
6	JUNIOR COLLEGE	26
7	HIGH SCHOOL	40
8	BACHELOR	50
9	HIGH SCHOOL	40
10	HIGH SCHOOL	25

… with 1,162 more rows

Hours Worked & Education

Side by Side Boxplots

What can you say about the relationship between the variables?



z-test

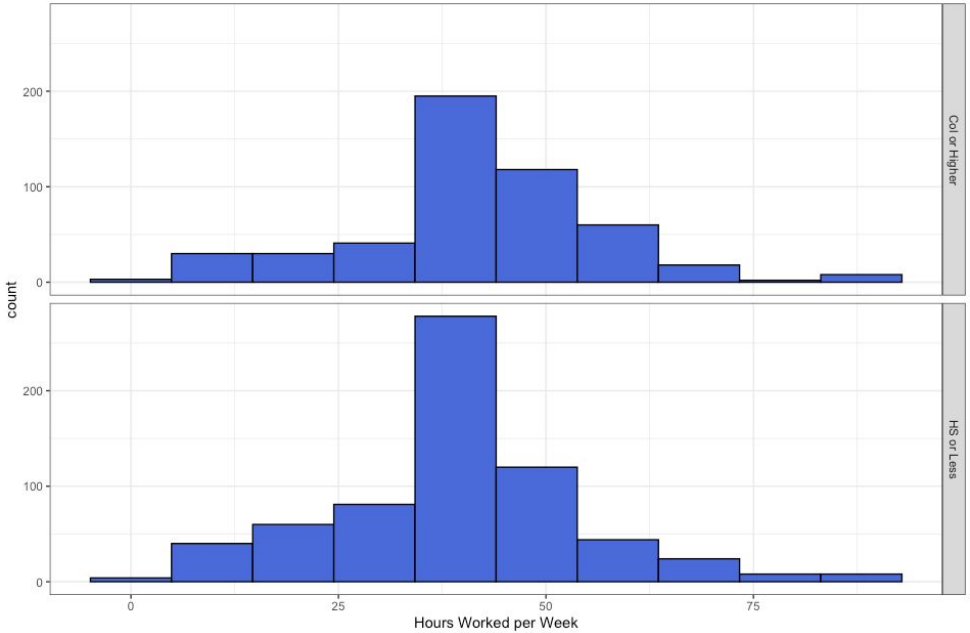
- Can we use a z-test to test our hypothesis?

- No
  - z-tests can only compare two means
- Can we manipulate our data so that a z-test will apply?

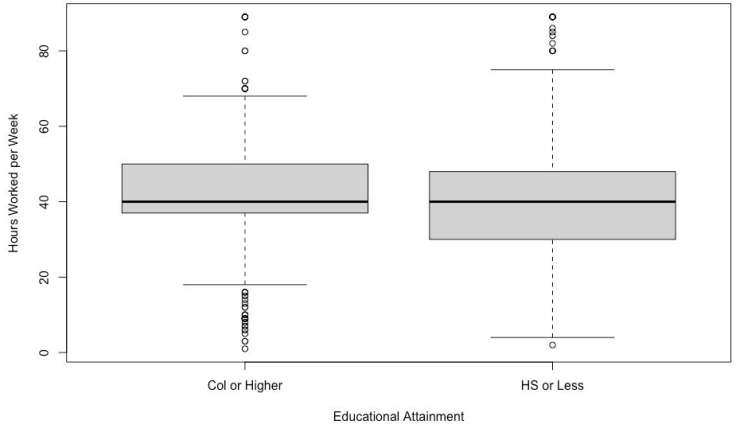
Collapsing Data

- Sometimes it makes sense to collapse similar or related groups together. In this case we might think of HS as a dividing line:
  - HS or Lower: Less than HS and HS from the original data set
  - College or Higher: Junior College, Bachelors, and Graduate from the original data set
- Now we have two groups that we can compare

College or More vs. HS or Less



Education	mean	sd	n
* <chr>	<dbl>	<dbl>	<int>
1 Col or Higher	41.8	15.1	505
2 HS or Less	39.4	15.1	667



Do you think there is a true difference?

Difference Between Two Means

The difference between the sample means is

$$\bar{x}_{coll} - \bar{x}_{hs} = 41.8 - 39.4 = 2.4 \text{ hours}$$

Education	mean	sd	n
* <chr>	<dbl>	<dbl>	<int>
1 Col or Higher	41.8	15.1	505
2 HS or Less	39.4	15.1	667

Can we use this sample difference to estimate the true difference in the population?

$$\mu_{coll} - \mu_{hs}$$

Difference Between Two Means

What is the difference in the average number of hours worked per work by Americans with a college degree or higher and those with a high school diploma or lower?

- Parameter of interest: Difference between the average number of hours worked per week by *all* Americans with a college degree and the average number of hours worked per week by *all* Americans with high school diploma or lower.

$$\mu_{coll} - \mu_{hs}$$

- Point estimate: Difference between the average number of hours worked per week by *sampled* Americans with a college degree and the average number of hours worked per week by *sampled* Americans with high school diploma or lower.

$$\bar{x}_{coll} - \bar{x}_{hs}$$

Applying the Central Limit Theorem

We can apply what we know about the sampling distribution from the Central Limit Theorem as long as the model assumptions or conditions are met.

- Independence within groups: Are individuals independent of each other?
  - ✓ Both groups are sampled randomly (GSS methodology)
  - ✓ Both groups are much smaller than the populations so sampling without replacement is fine
- Independence between groups: Are the two groups independent?
  - ✓ Again, since all individuals are sampled randomly this condition is satisfied.
- Sample size / skew: Will the sampling distribution be sufficiently Normal?
  - ✓ Both groups have fairly large sample sizes (505 and 667)
  - ✓ Both histograms and boxplots, although not perfectly symmetric, don't show any major skewness, so we can be confident that the sampling distribution will be nearly Normal.

Confidence Interval

The difference between samples was 2.4 hours, which leads to an estimate of the true difference as 2.4 hours as well. Can we build a confidence interval around that difference?

- Confidence Interval: *point estimate ± ME* (as always)
- Margin of Error: *ME = critical value x SE of the point estimate*

But what's the SE for the difference between two means?

Standard Error

Standard Error for a single mean:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

Standard Error for the difference between two means:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Standard Error

Calculate the SE for the difference in the mean number of hours worked for college and above vs. high school and below.

Education	mean	sd	n
* <chr>	<dbl>	<dbl>	<int>
1 Col or Higher	41.8	15.1	505
2 HS or Less	39.4	15.1	667

$$SE_{\bar{x}_1 - \bar{x}_2} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= \sqrt{\frac{15.1^2}{505} + \frac{15.1^2}{667}}$$

$$= 0.8925$$

Confidence Interval

Build a 95% Confidence Interval for the difference between the two means.

$$(\bar{x}_{coll} - \bar{x}_{hs}) \pm z^{\star} \times SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = (41.8 - 39.4) \pm 1.96 \times 0.89$$

$$= 2.4 \pm 1.74$$

$$= (0.66, 4.14)$$

Education	mean	sd	n
* <chr>	<dbl>	<dbl>	<int>
1 Col or Higher	41.8	15.1	505
2 HS or Less	39.4	15.1	667

$SE_{x1-x2} = 0.893$

Question

Which of the following is the ***best*** interpretation of the confidence interval we just calculated?

We are 95% confident that...

a) The difference between the average number of hours worked per week by college grads and those with a HS degree or lower is between 0.66 and 4.14 hours.
b) College grads work an average of 0.66 to 4.14 hours more per week than those with a HS degree or lower.
c) College grads work on average 0.66 hours less to 4.14 hours more per week than those with a HS degree or lower.
d) College grads work on average 0.66 to 4.14 hours less per week than those with a HS degree or lower.

Careful with Differences

There is a subtle difference between looking at the average of the differences and the difference of the averages…

$$\frac{1}{n}[(x_1 - y_1) + (x_2 - y_2) + \cdots + (x_n - y_n)]$$

Vs

$$\bar{x} - \bar{y}$$

Hypothesis Test

Is there a difference between the average number of hours worked per week by people with college education and the average hours worked per week by those with a HS diploma or lower?

What are the hypothesis?

$$H_0 : \quad \mu_{coll} = \mu_{hs}$$

There is no difference in the average number of hours worked per week by those with a college education and those with a HS diploma or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

$$H_A : \quad \mu_{coll} \neq \mu_{hs}$$

There is a difference in the average number of hours worked per week by those with a college education and those with a HS diploma or lower.

Test Statistic

TEST: z-test

TAILS: two-tailed

TEST STATISTIC FOR THE DIFFERENCE OF MEANS:

$$Z = \frac{\bar{x}_{coll} - \bar{x}_{hs} - 0}{SE_{\bar{x}_{coll} - \bar{x}_{hs}}} \quad \text{or} \quad Z = \frac{\bar{x}_{coll} - \bar{x}_{hs}}{SE_{\bar{x}_{coll} - \bar{x}_{hs}}}$$

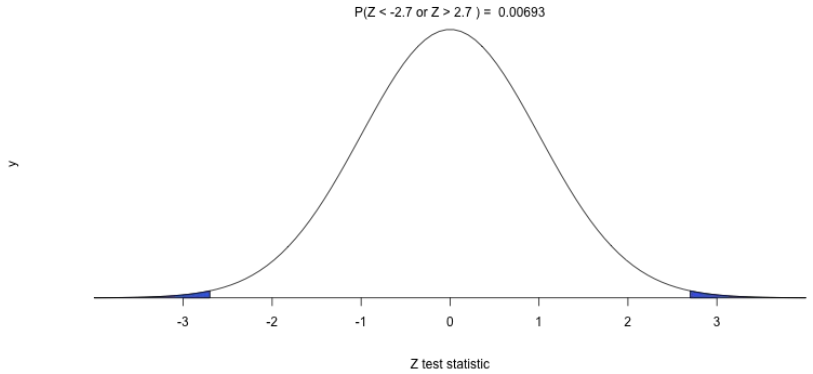
Calculating the Test Statistic & p-value

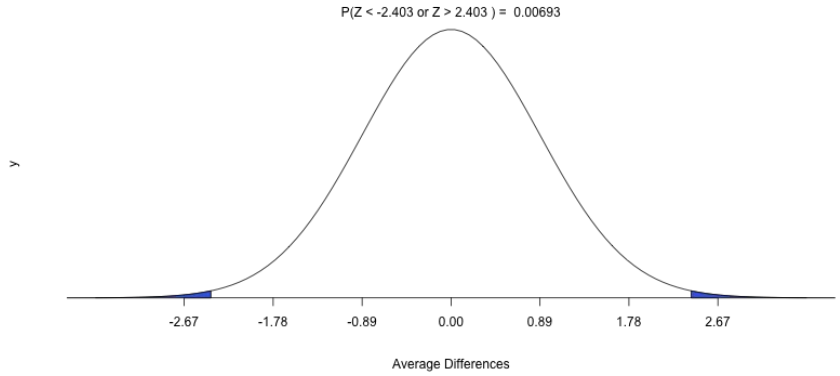
$$Z = \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}}$$

$$= \frac{2.4}{0.89} = 2.70$$

Upper or Lower tail: 0.00347

p-value = 0.00693





What's the conclusion of the test?

Which of the following is best supported by the results of the hypothesis test we just conducted?

a. There is 0.7% chance that there is no difference in the average number of hours worked per week between the two groups.

b. Since the p-value is low, we reject H_0. The data provide convincing evidence of a difference in the average number of hours worked per week between the two groups.

c. Since we rejected H_0, we may have made a Type 2 error.

d. Since the p-value is low, we fail to reject H_0. The data do not provide convincing evidence of a difference in the average number of hours worked per week between the two groups.