SEIS 631

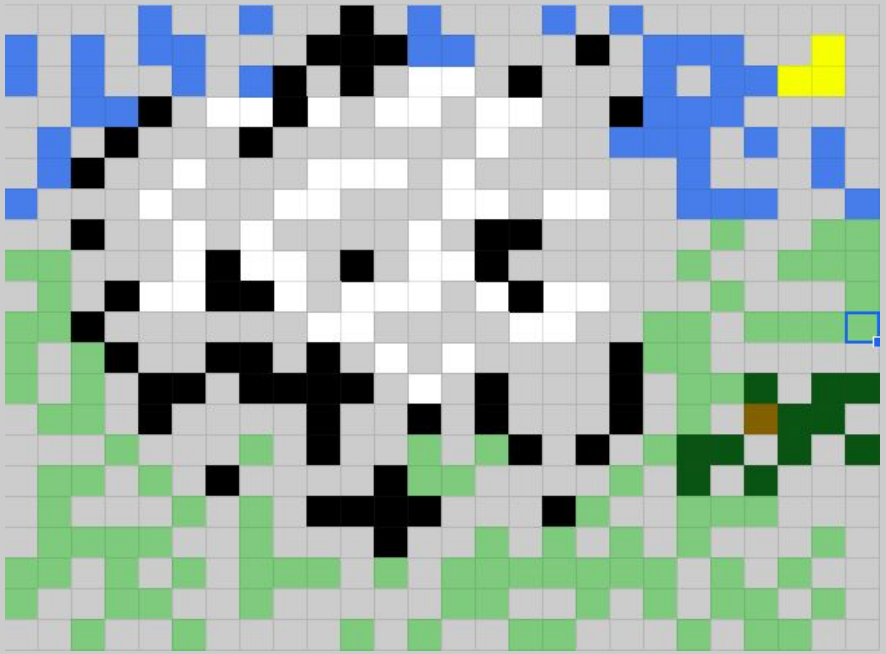
Populations vs. Samples vs. Sampling Distributions

Google Sheets

Sampling Distributions

Population

Sample



R Demo On Sampling Distributions

- R demo
- Shiny App https://aranglancy.shinyapps.io/SamplingDistributions/
- Other App http://onlinestatbook.com/stat_sim/sampling_dist/

Sampling Distribution for a Proportion

- Goal: Estimate the ***true proportion*** in a population, p
  - "What percent of the population support universal healthcare?"
- Method:
  - Gather a ***representative*** sample and calculate the ***sample proportion*** $\hat{p}$

- How Good is the Picture?
  - The ***sample proportion*** is an unbiased estimate of ***true proportion***
- Sampling Distribution
  - If we gathered hundreds of samples and calculated the **sample proportion** in each, the **distribution of sample proportions,** or **sampling distribution** is
    - Centered on the True Proportion
    - Normally Distributed
    - Spread with $SE = \sqrt{\frac{p(1-p)}{n}}$ or $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Sampling Distribution for a Mean

- Goal: Estimate the ***true mean*** in a population, μ
  - "What is the average cholesterol level of US adults?"
- Method:
  - Gather a ***representative*** sample and calculate the ***sample average*** $\bar{x}$

- How Good is the Picture?
  - The ***sample mean*** is an unbiased estimate of ***true mean***
- Sampling Distribution
  - If we gathered hundreds of samples and calculated the **sample mean** in each, the **distribution of sample means,** or **sampling distribution** is
    - Centered on the True Mean
    - Normally Distributed
    - Spread with $SE = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem

- When many uniform, independent events are combined (additively) the distribution of the results will tend toward a normal distribution as the sample size increase.

- Specifically: If samples are repeatedly drawn from a population and the mean or a proportion is calculated from those samples, the results will tend to be normally distributed.
    - The more trials, the closer the distribution will be to approximating N
    - The larger the sample size, the better the approximation

Standard Error: Application of the Central Limit Theorem

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right),$$

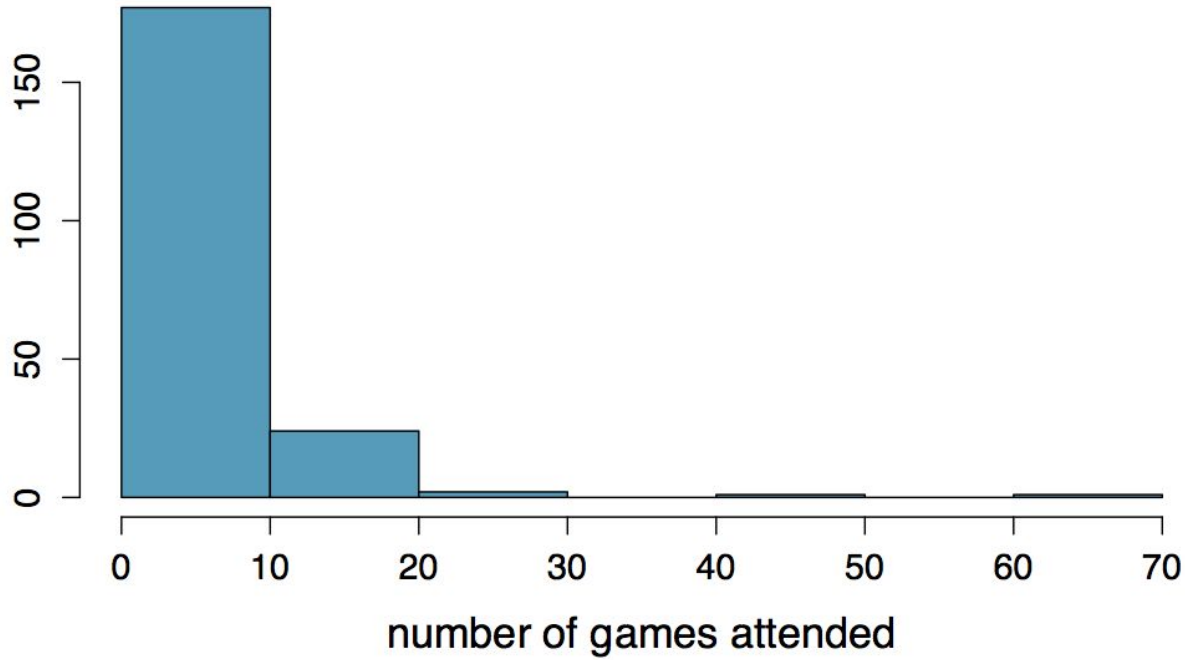
where *SE* is represents standard error, which is defined as the standard deviation of the sampling distribution. **If σ is unknown, use *s*.**

Conditions:

1. **Independence**: Sampled observations must be independent.
   - Random sample/assignment
   - If sampling without replacement, then n < 10% of the population
2. **Sample size and skew**: Either the population distribution is normal or if the population is skewed then sample size is large (usually n > 30).

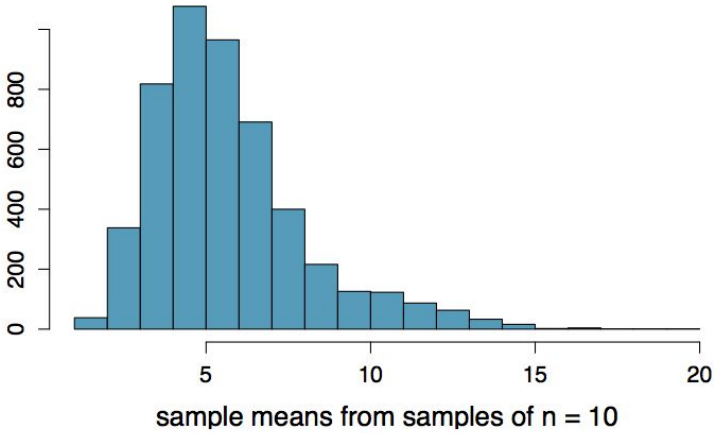
Average number of basketball games attended

Next let's look at the population data for the number of basketball games attended:



Average number of basketball games attended (cont.)

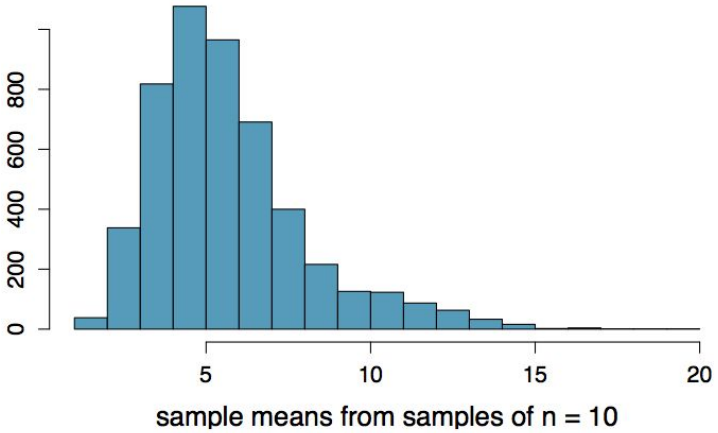
Sampling distribution, n = 10:



What does each observation in this distribution represent?

Average number of basketball games attended (cont.)

Sampling distribution, n = 10:

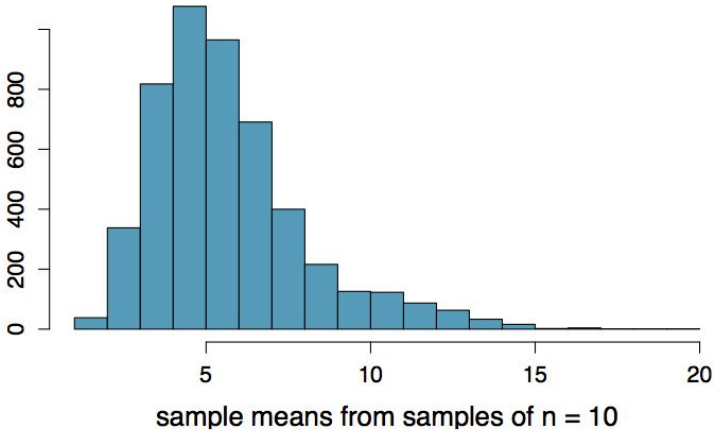


What does each observation in this distribution represent?

Sample mean ($\bar{x}$) of samples of size n = 10.

Average number of basketball games attended (cont.)

Sampling distribution, n = 10:



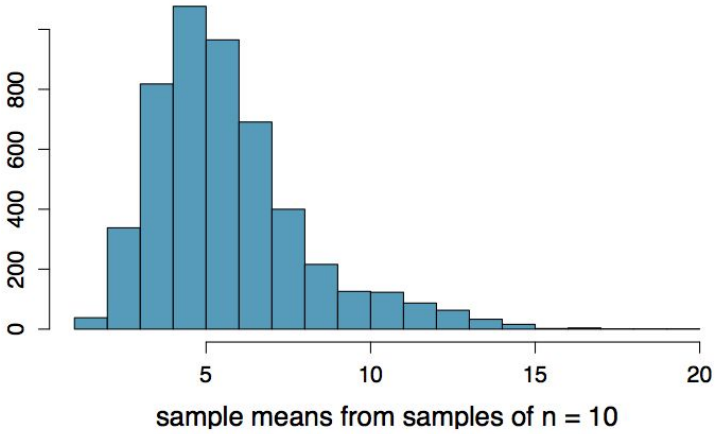
What does each observation in this distribution represent?

Sample mean ($\bar{x}$) of samples of size n = 10.

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Average number of basketball games attended (cont.)

Sampling distribution, n = 10:



What does each observation in this distribution represent?

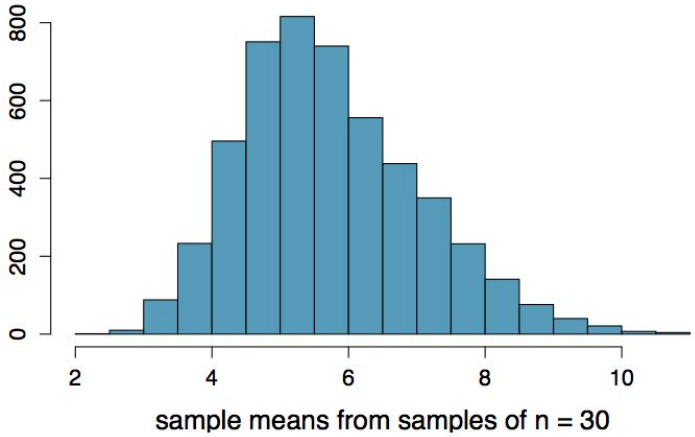
Sample mean ($\bar{x}$) of samples of size n = 10.

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Smaller, sample means will vary less than individual observations.

Average number of basketball games attended (cont.)

Sampling distribution, n = 30:

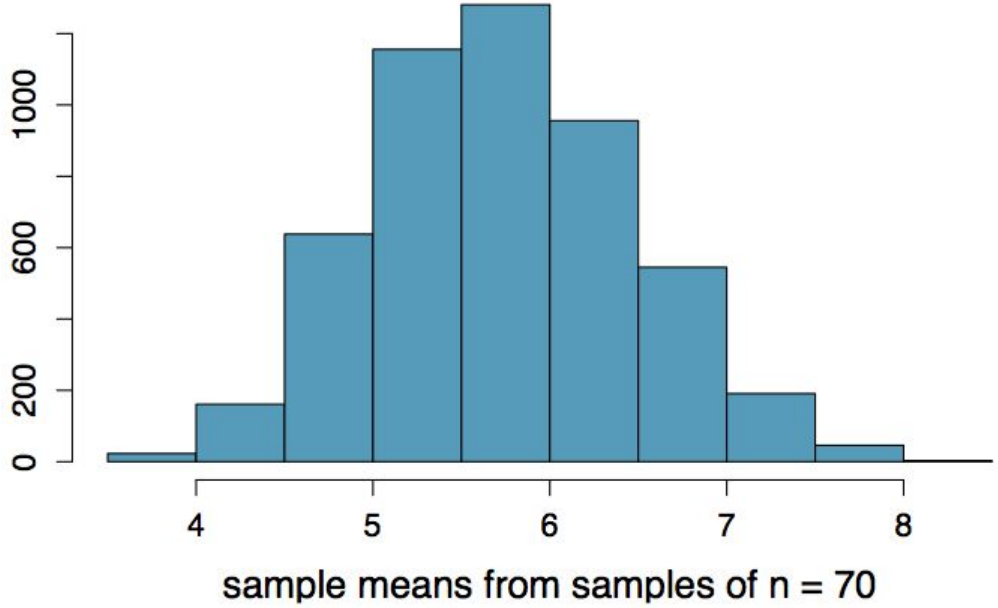


How did the shape, center, and spread of the sampling distribution change going from n = 10 to n = 30?

Shape is more symmetric, center is about the same, spread is smaller.

Average number of basketball games attended (cont.)

Sampling distribution, n = 70:

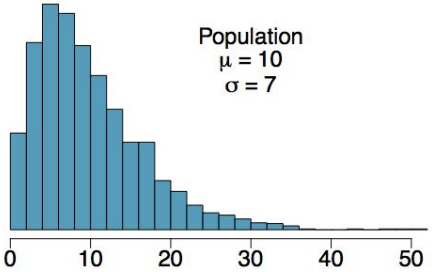


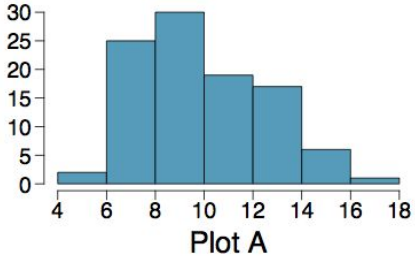
Question

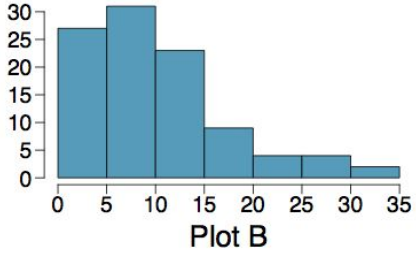
Four plots: Determine which plot (A, B, or C) is which.

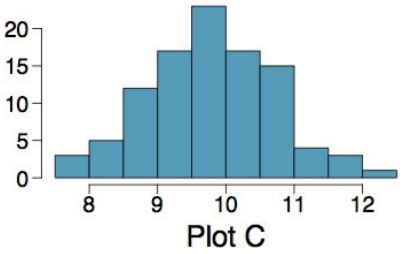
NOTE: First plot is distribution for a population ($\mu = 10, \sigma = 7$),

- a single random sample of 100 observations from this population,
- a distribution of 100 sample means from random samples with size 7, and
- a distribution of 100 sample means from random samples with size 49.









Confidence Intervals
Or
Compatibility Intervals

How many jelly beans?

What's your best guess?

What's the lowest number it could reasonably be?

What's the highest number it could reasonably be?

How confident are you that your guess is correct?

How confident are you that true value is between your low and high values?

How many jelly beans Part II?

Here's a new jar, but this time the rules are different.

1. Instead of a specific guess, give a range ("It's between ___ and ___")
2. If the actual number is in your range, you qualify for the finals.
3. The winner is the person in the finals with the smallest range.

Confidence Interval

- A range of values that you are “confident” captures the true value.
- The wider the range the more confident we can be.
- The narrower the range, the less confident.
- Confidence Intervals always refer to *Population Parameters* (not sample statistics).

Confidence intervals

Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

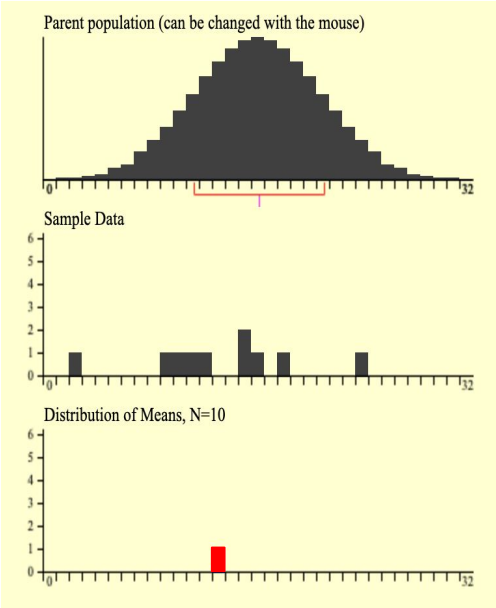
We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

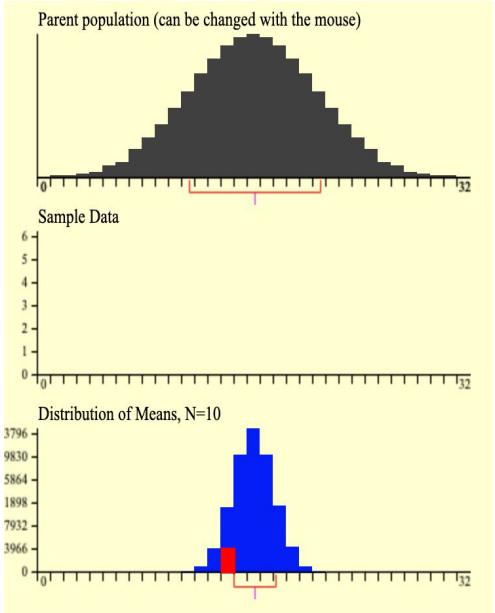
Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Confidence Intervals

Reporting Just A Point Estimate



Reporting a Confidence Interval

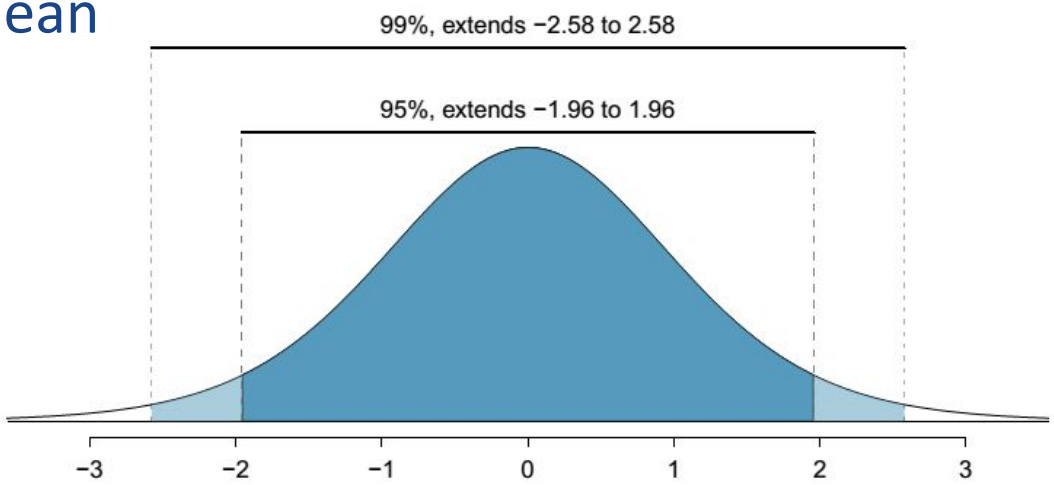


Confidence Intervals

CLT says that $\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$

Approximate 95% CI = x ± 2SE

95% of sample will have a mean in this range.



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

Question

One of the earliest examples of behavioral asymmetry is a preference in humans for turning the head to the right, rather than to the left, during the final weeks of gestation and for the first 6 months after birth. This is thought to influence subsequent development of perceptual and motor preferences. A study of 124 couples found that 64.5% turned their heads to the right when kissing. The standard error associated with this estimate is roughly 4%. Which of the below is *false*?

a) The 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 4%.
b) A higher sample size would yield a lower standard error.
c) The margin of error for a 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 8%.
d) The 99.7% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 12%.

Güntürkün, O (2003) Adult persistence of head-turning asymmetry. *Nature*. Vol 421.

Changing the confidence level

$$\textit{point estimate} \pm z^* \times SE$$

- In a confidence interval, $z^* \times SE$ is called the margin of error, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust z^* in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^* = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate z^* for any confidence level.

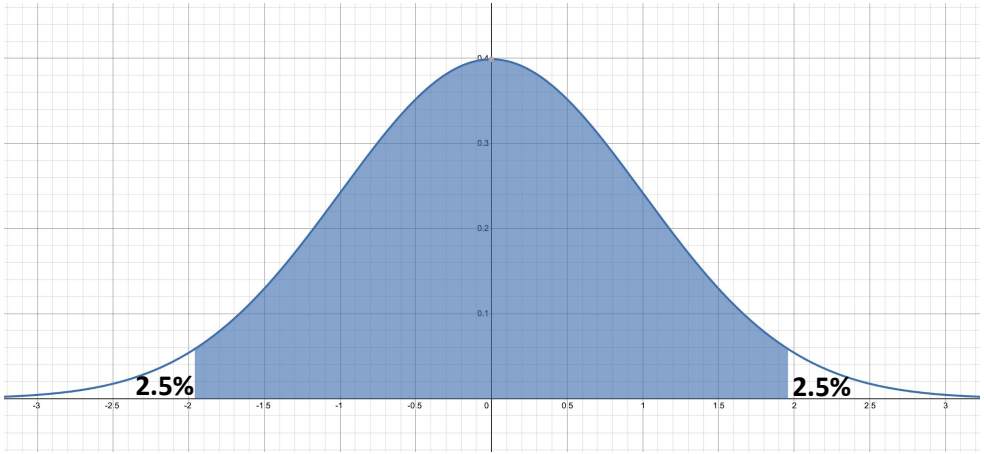
Finding the Critical Value

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

- 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?

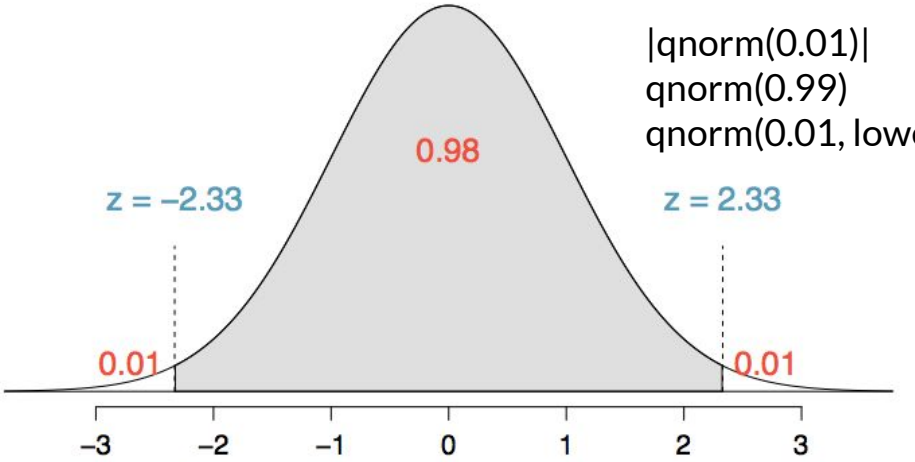
(a) Z = 2.05

(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

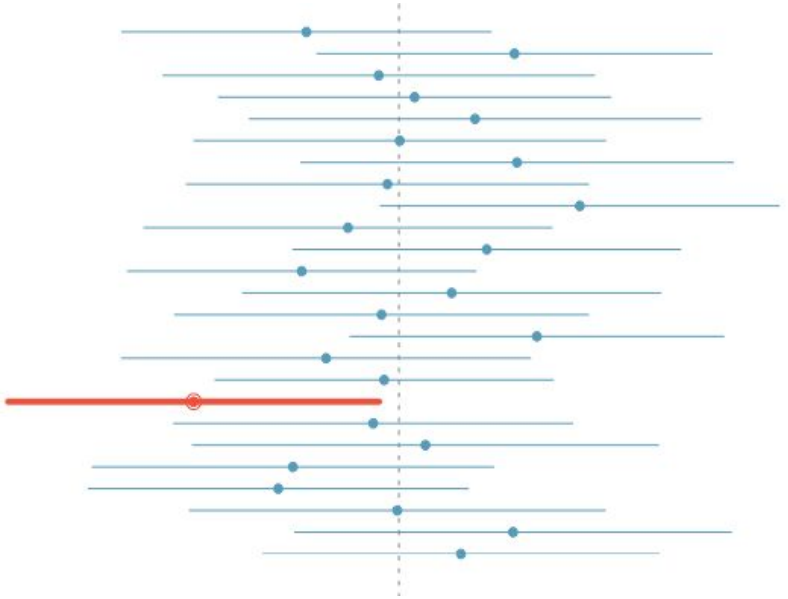
Accuracy vs Precision of CI

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

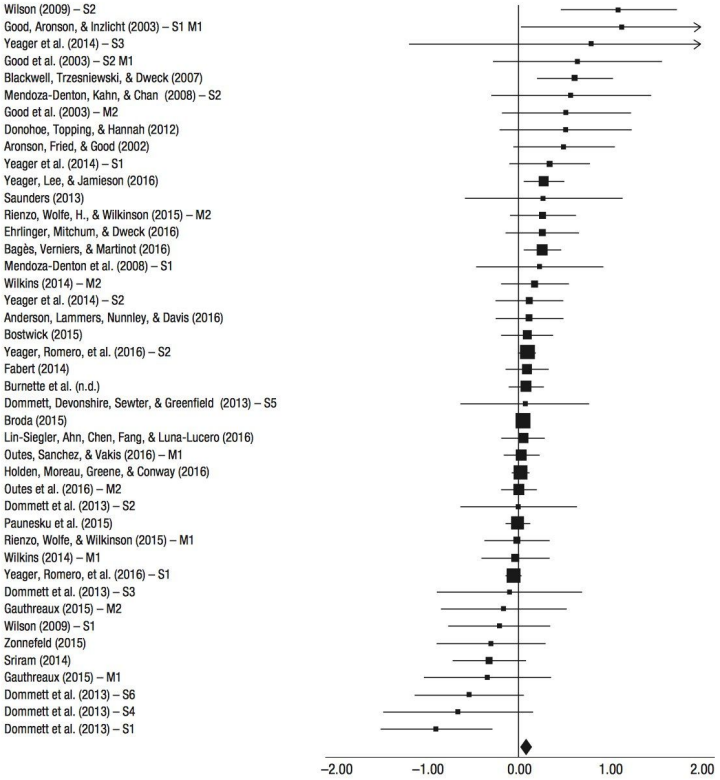


Fig. 4. Standardized mean differences (Cohen's *d*s) in academic achievement between students receiving a growth-mind-set intervention and students in the comparison group. Cohen's *d*s (squares) and 95% confidence intervals (error bars) are displayed for all effects entered

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.

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?

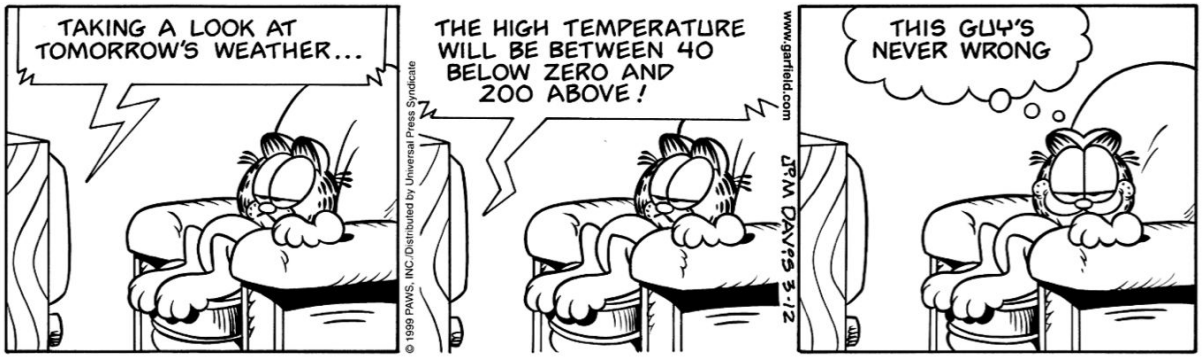
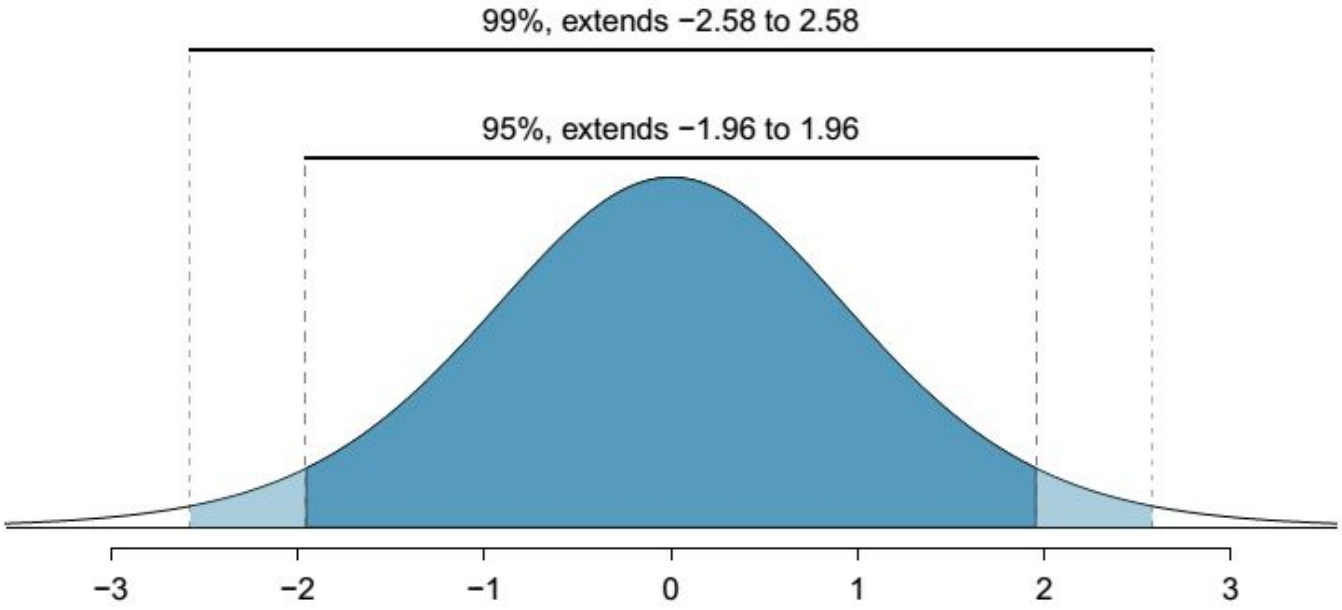
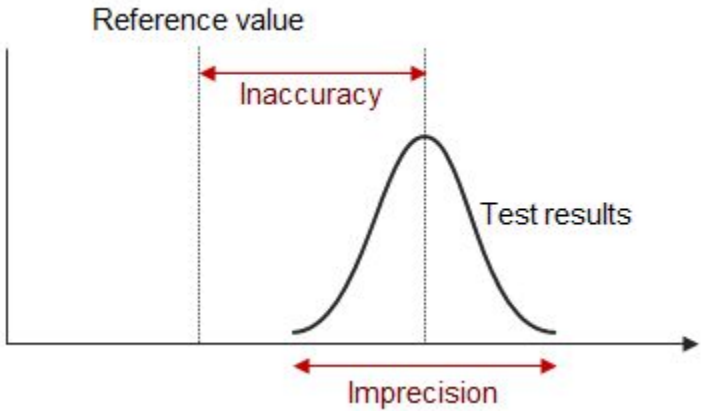


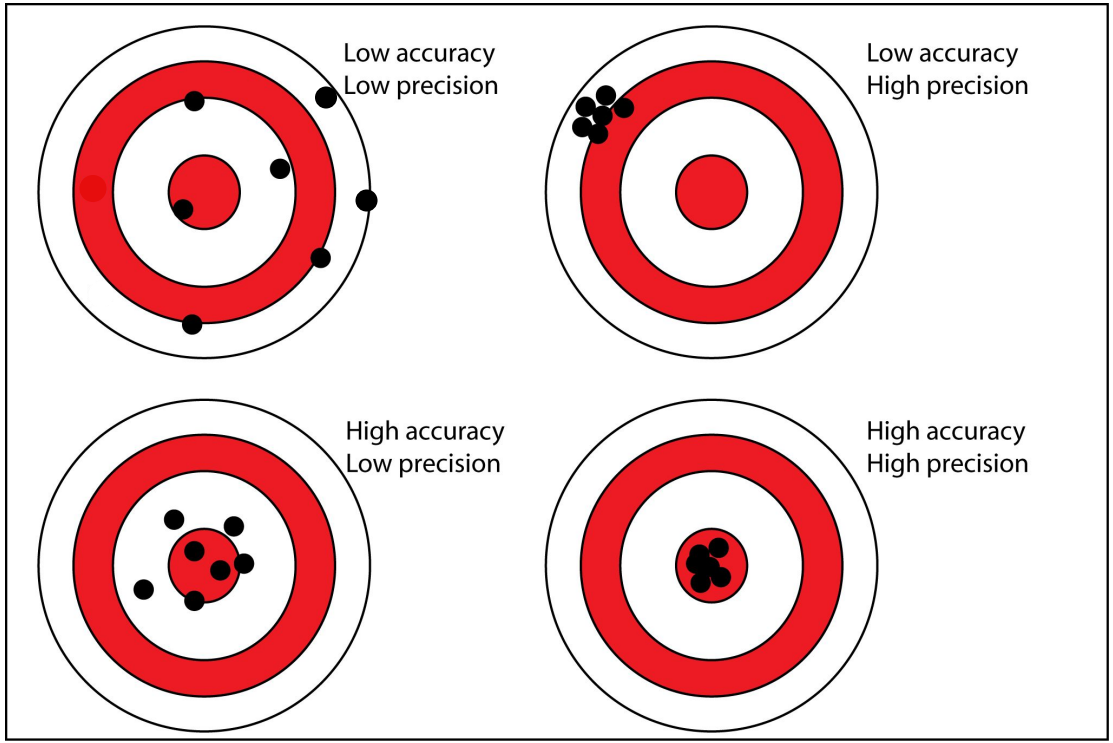
Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif



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

Slides adapted from work by Mine Çetinkaya-Rundel of OpenIntro
The slides may be copied, edited, and/or shared via the CC BY-SA license
Some images may be included under fair use guidelines (educational purposes)

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?

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

- 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

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

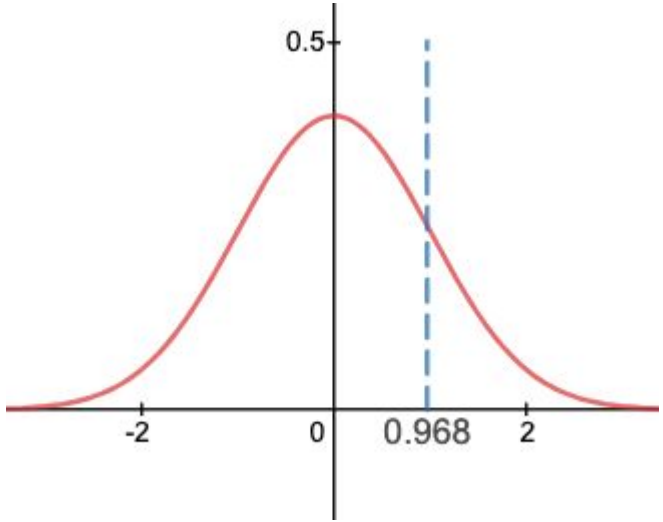
$$\bar{x} = 11.7$$

$$sd = 3.47$$

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

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?



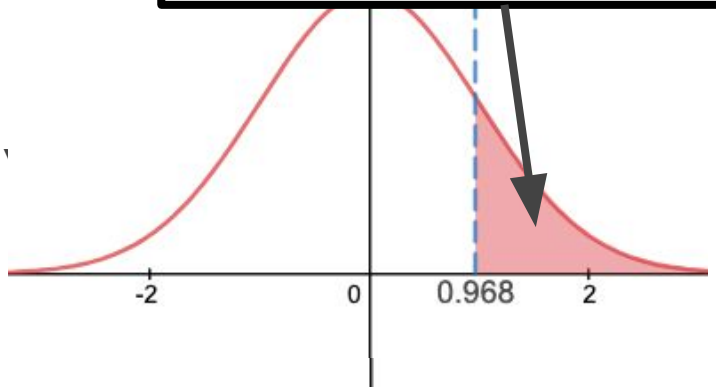
$$z = \frac{11.7-11}{2.8/\sqrt{15}} = 0.968$$

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

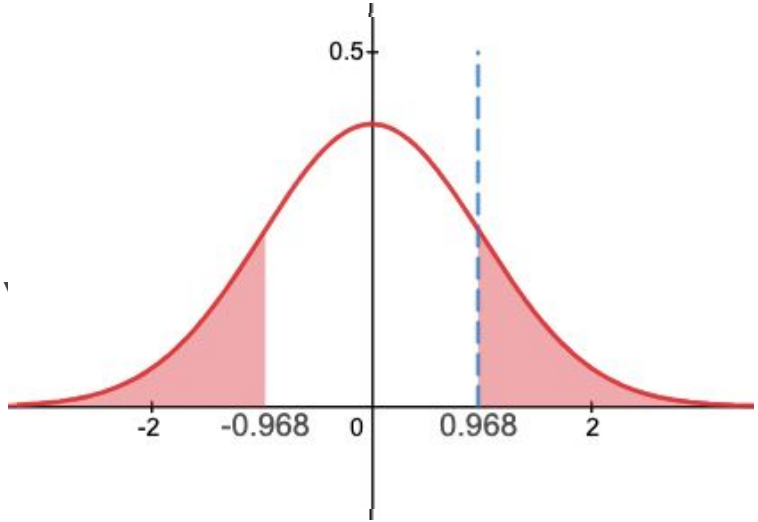
```
>pnorm(0.968, lower.tail = FALSE)
```



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

Calculating *p*-values

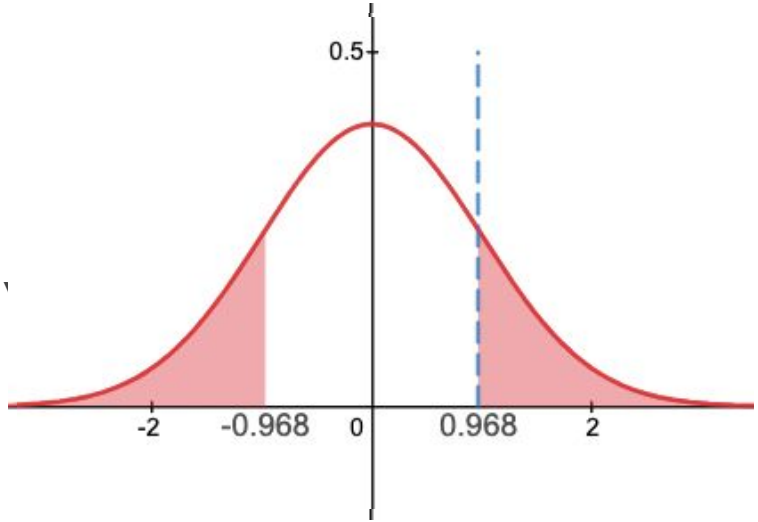
- What does the *z*-score tell us?
  - 11.7 is 0.968 ***standard errors*** above the mean.
- If there were ***truly*** no difference, how likel
  are we to get such a difference?



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

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 example (1 – sided)

A survey asks a sample of 206 Duke University students how many colleges they had applied to. The sample yielded an average of 9.7 applications with a standard deviation of 7. The College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all Duke students apply to is *higher* than recommended?

1 sample z test example (1 – sided)

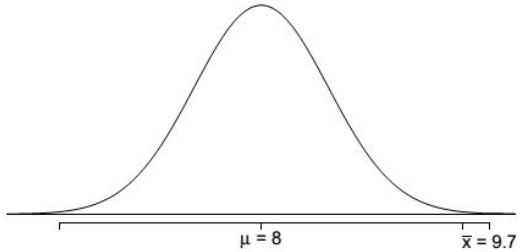
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

1. Set null and alternative:
   $H_o: \mu = 8$ vs. $H_A: \mu > 8$
2. Calculate test statistic

   $Z = \frac{9.7-8}{7/\sqrt{206}} = 3.4$

3. Determine p-value

```
> pnorm(3.4, lower.tail = FALSE)
[1] 0.0003369293
```



1 sample z test example (1 – sided)

4. Draw a conclusion about the null.

p = 0.0003, which is less small, therefore we can reject the null. We have strong evidence that Duke students apply to *more* colleges than the recommended amount.

1 sample z test example (2 – sided)

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?

1 sample z test example (2 – sided)

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

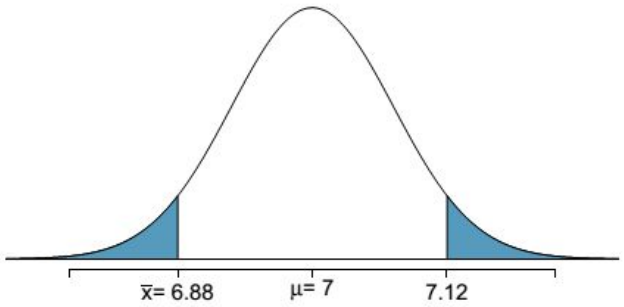
1. Set null and alternative:
   $H_o: \mu = 7$ vs. $H_A: \mu \neq 7$
2. Calculate test statistic
   $Z = \frac{6.88 - 7}{0.94 / \sqrt{169}} = -1.659$
3. Determine p-value

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

[1] 0.04849747

2 x 0.0485 = 0.097 – p value

- Hence the p-value would change as well:



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

1 sample z test example (2 – sided)

4. Draw a conclusion about the null.

P = 0.097, which is greater than 0.05, therefore we cannot reject the null and we have weak evidence that students get a different amount of sleep than 7 hours.

Note: 2 – sided tests are harder to reject because p-value is doubled.

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

- Examples

[3%, 17%] – effect is positive, *maybe* large

[9%,11%] – effect is positive *and* large

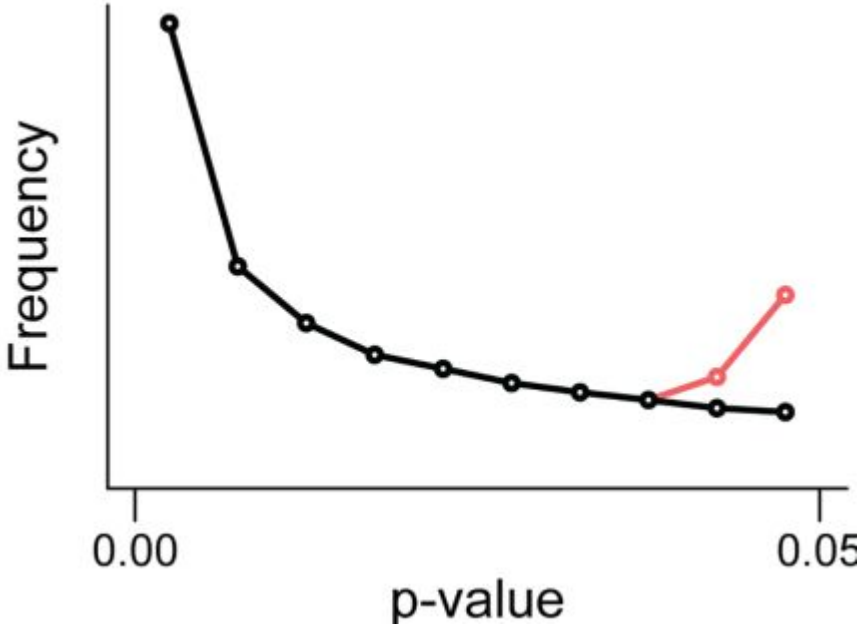
[-.5%,.5%] –maybe pos or neg, *and* small

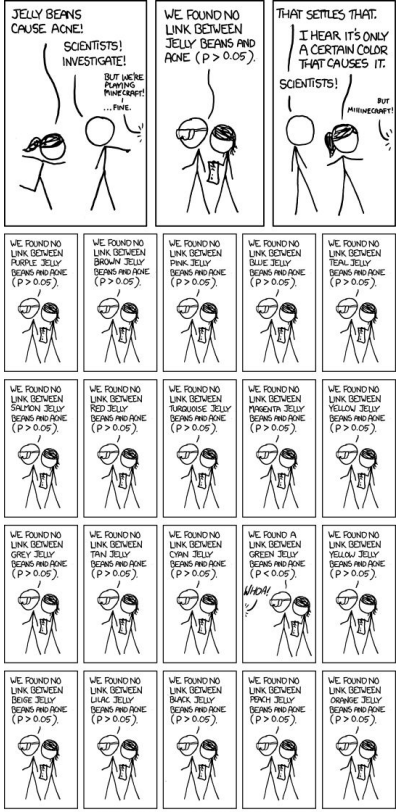
[-15%, 15%] – maybe pos or neg, *maybe* large?

	critical values	p-values	confidence intervals
accept/reject	✔	✔	✔
degree of support	✘	✔	✔
estimate and uncertainty	✘	✘	✔

http://strata.uga.edu/6370/lecturenotes/pvalueconfidenceintervals.html

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

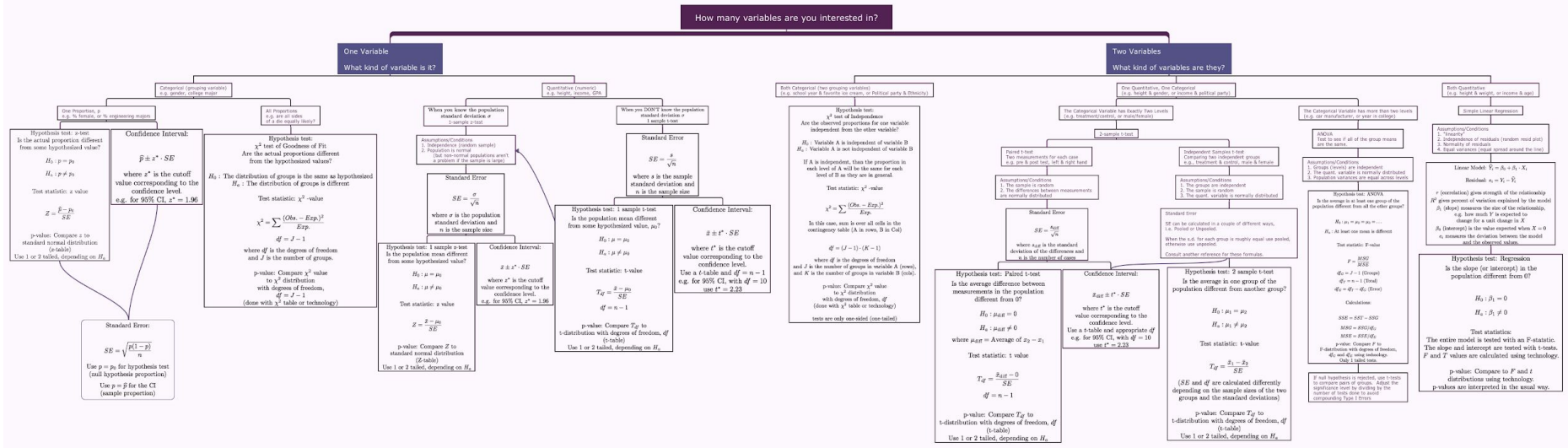
- Increase alpha
  - Set at the beginning of experiment
- Conduct one tailed test
  - Have to decide before experiment
- Decrease random error
  - Through more advanced sampling and experimental techniques
- Increase sample size
  - Depending on the situation, this is the most straightforward!

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.

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

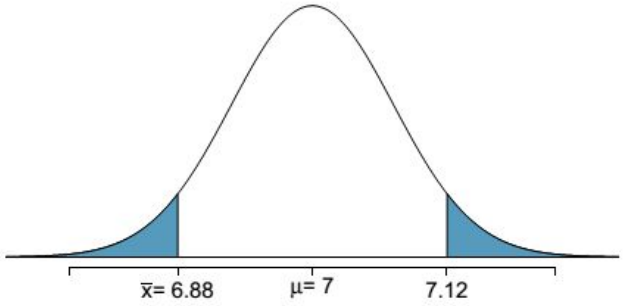
> pnorm(-1.6596, lower.tail = TRUE)

[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.

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

Hypothesis Testing with Numeric Data

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

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?

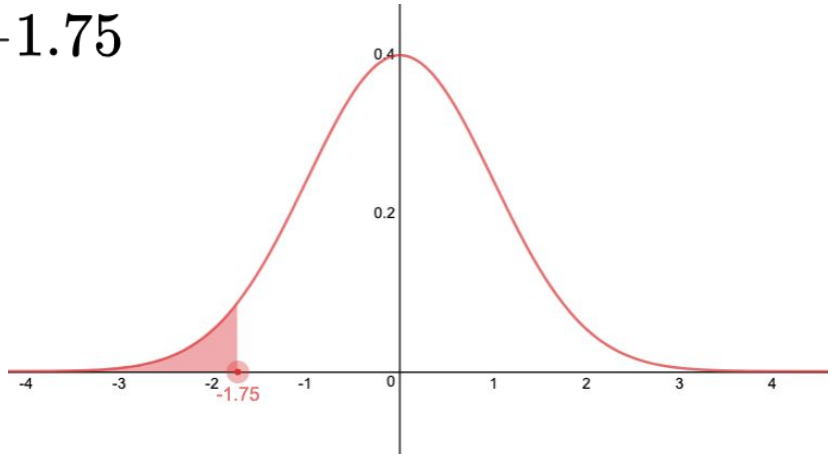
$$H_0 : \mu_{mc} = \mu_{population} \text{ or } \mu_{mc} = 85$$

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

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

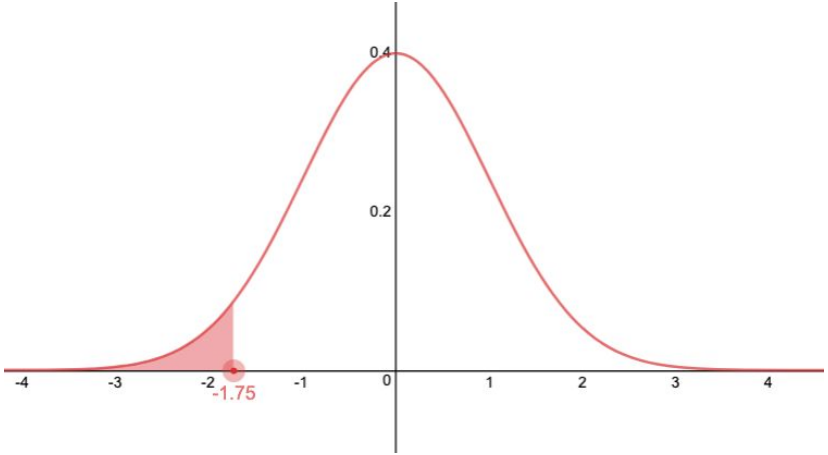
- Calculate the *p*-value (with R):

  > pnorm(-1.75)
  [1] 0.04005916

  or

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

  p = 0.04

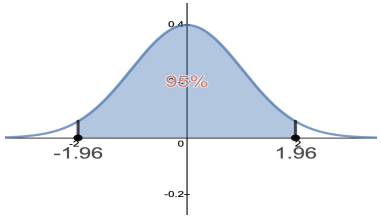


One sample z-test Example

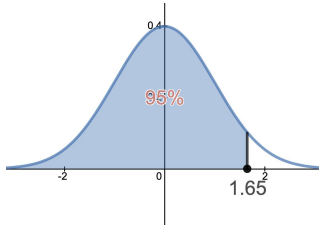
- Calculate the *95% CI*: This is a one sided test, so we can construct a 1-sided CI. For a 1-sided CI we put no limit on either the upper bound or lower bound.

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

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

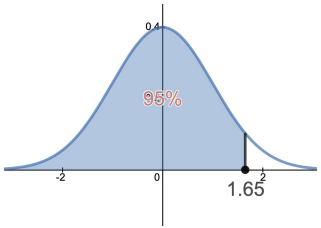


- In this case, we lump all the error on one side so we will get different z^* values.



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. How high could it reasonably be?
  - If it is actually lower than we estimated, we don't care (at least for this test) how much lower



- CI: $$CI = 80.94 + 1.65 \cdot 2.32 = 84.76$$
  - 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?

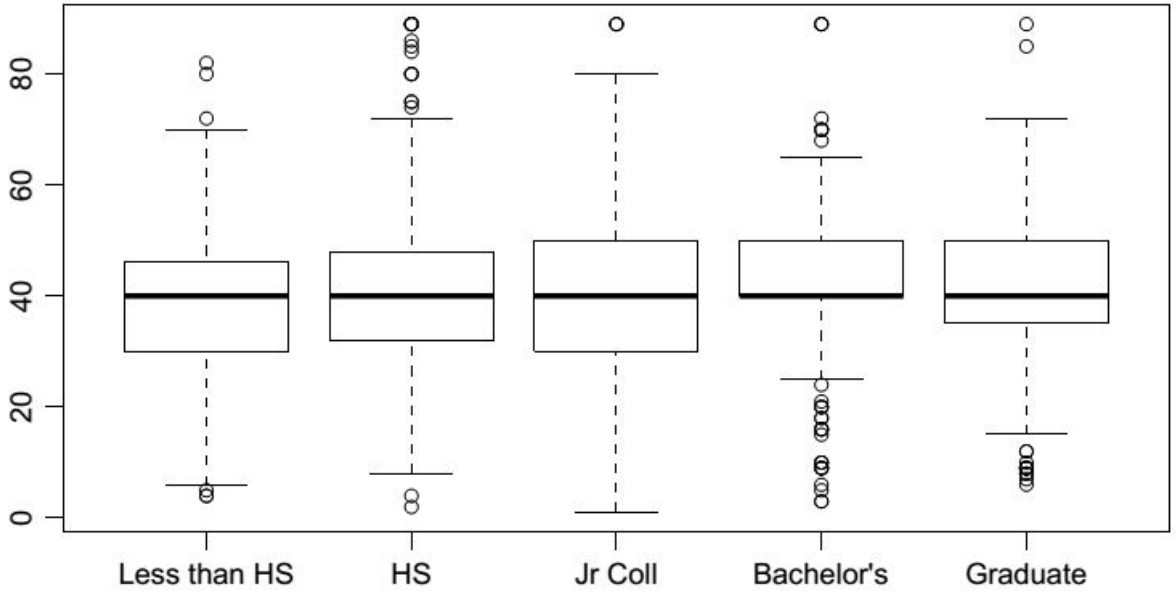
Difference of two means
- Confidence intervals for differences of means
- Hypothesis tests for differences of means

The General Social Survey (GSS) conducted by the Census Bureau contains a standard 'core' of demographic, behavioral, and attitudinal questions, plus topics of special interest. Many of the core questions have remained unchanged since 1972 to facilitate time-trend studies as well as replication of earlier findings. Below is an excerpt from the 2010 data set. The variables are number of hours worked per week and highest educational attainment.

	degree	hrs1
1	BACHELOR	55
2	BACHELOR	45
3	JUNIOR COLLEGE	45
⋮		
1172	HIGH SCHOOL	40

Exploratory analysis

What can you say about the relationship between educational attainment and hours worked per week?

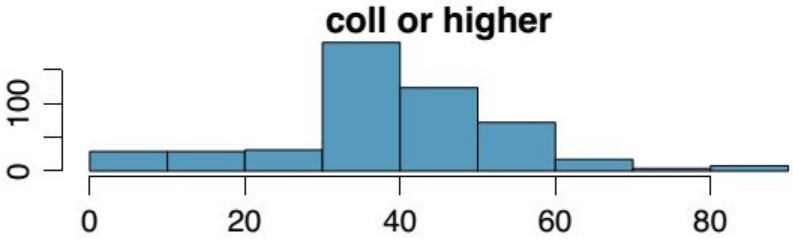


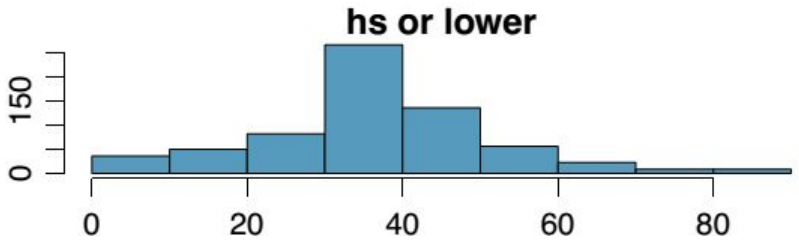
Collapsing levels into two

- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- Then we combine the levels of education into two:
  - `hs or lower` ← less than high school or high school
  - `coll or higher` ← junior college, bachelor's, and graduate

Exploratory analysis - another look

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667





Parameter and point estimate

We want to construct a 95% confidence interval for the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower. What are the parameter of interest and the point estimate?

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

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

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

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

Checking assumptions & conditions

1. *Independence within groups:*
   - Both the college graduates and those with HS degree or lower are sampled randomly.
   - 505 < 10% of all college graduates and 667 < 10% of all students with a high school degree or lower.

   We can assume that the number of hours worked per week by one college graduate in the sample is independent of another, and the number of hours worked per week by someone with a HS degree or lower in the sample is independent of another as well.
2. *Independence between groups:* ← *new!*

   Since the sample is random, the college graduates in the sample are independent of those with a HS degree or lower.
3. *Sample size / skew:*

   Both distributions look reasonably symmetric, and the sample sizes are at least 30, therefore we can assume that the sampling distribution of number of hours worked per week by college graduates and those with HS degree or lower are nearly normal. Hence the sampling distribution of the average difference will be nearly normal as well.

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- And all $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$
- So the only new concept is the standard error of the difference between two means...

Standard error of the difference between two sample means

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

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = \sqrt{\frac{s_{coll}^2}{n_{coll}} + \frac{s_{hs}^2}{n_{hs}}}$$

$$= \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}}$$

$$= 0.89$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$

$$\begin{aligned} (\bar{x}_{coll} - \bar{x}_{hs}) \pm z^\star \times SE_{(\bar{x}_{coll} - \bar{x}_{hs})} &= (41.8 - 39.4) \pm 1.96 \times 0.89 \\ &= 2.4 \pm 1.74 \\ &= (0.66, 4.14) \end{aligned}$$

Interpretation of a confidence interval for the difference

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

(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 on 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

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

Setting the hypotheses

What are the hypotheses for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

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

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

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

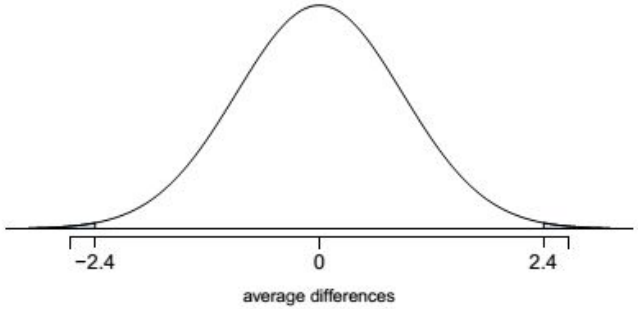
There is a difference in the average number of hours worked per week by college graduates and those with a HS degree or lower.

Calculating the test-statistic and the p-value

H_0: $\mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$

H_A: $\mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$

$\bar{x}_{coll} - \bar{x}_{hs} = 2.4$, $SE(\bar{x}_{coll} - \bar{x}_{hs}) = 0.89$



$$Z = \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} = \frac{2.4}{0.89} = 2.70$$

$$upper\ tail = 1 - 0.9965 = 0.0035$$

$$p - value = 2 \times 0.0035 = 0.007$$

Conclusion of the test

Which of the following is correct based on the results of the hypothesis test we just conducted?

(a) There is a 0.7% chance that there is no difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(b) Since the p-value is low, we reject H_0. The data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(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 between the average number of hours worked per week by college graduates and those with a HS degree or lower.

t – distribution

Are z-tests really practical?

- z-test for a mean:
  - We want to estimate a population mean, μ
  - We know that if population mean is μ, and population s.d. is σ then
    - Sample mean is approximately the population mean

$$\bar{x} \approx \mu$$

    - The standard error of the mean (i.e. the Standard Deviation of the sampling distribution)

$$S.E. = \frac{\sigma}{\sqrt{n}}$$

  - But how do we know the population s.d, σ?
  - Is s a good approximation for σ?

Estimating σ

- When we don't know σ, we estimate it with the sample standard deviation, s.

$$S.E. = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

- But this introduces more variability into our estimate…

William Gosset

- Guiness Brewery
- Quality Testing on Grain Samples
  - Quickly test a small sample
  - If bad, send off for more comprehensive testing (i.e. larger sample)
- With significance level of 5%, he expected 5% of batches to come back as actually ok…
- …But actually much more came back

Student's t-test

- Gosset took a sabbatical from Guinness to study this
- Found that if n is small, and σ is not known, then the sampling distribution isn't actually normal
- Guinness wouldn't let him publish these results under his own name
- Hence "Student's t-distrubtion" and "Student's t-test"

- When σ is not known, we modify the z-test slightly and use a t-test

t-Tests

- t-tests start the same as z-tests and are very similar.

- We'll about the t-test by starting the hypothesis test process then seeing where it deviates from a z-test.

Example:

- Is Friday the 13th unlucky?
- Do people change their behavior?

Friday the 13th

Between 1990 - 1992 researchers in the UK collected data on traffic flow, accidents, and hospital admissions on Friday 13th and the previous Friday, Friday 6th. Below is an excerpt from this data set on traffic flow. We can assume that traffic flow on given day at locations 1 and 2 are independent.

	type	date	6th	13th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

Scanlon, T.J., Luben, R.N., Scanlon, F.L., Singleton, N. (1993), "Is Friday the 13th Bad For Your Health?," BMJ, 307, 1584-1586.

Friday the 13th

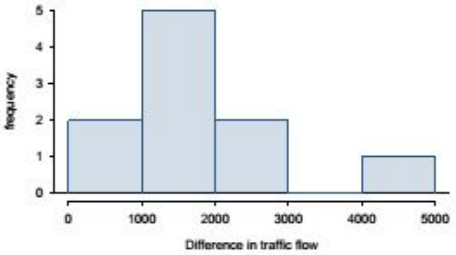
- We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.
- One approach is to compare the traffic flow on these two days.
- H_0 : Average traffic flow on Friday 6th and 13th are equal.
  H_A : Average traffic flow on Friday 6th and 13th are different.

Each case in the data set represents traffic flow recorded at the same location in the same month of the same year: one count from Friday 6th and the other Friday 13th. Are these two counts independent?

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6th and 13th?

Conditions

- *Independence:* We are told to assume that cases (rows) are independent.
- *Sample size / skew:*
  - The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not – probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.



  - $n < 30!$

So what do we do when the sample size is small?

Review: what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

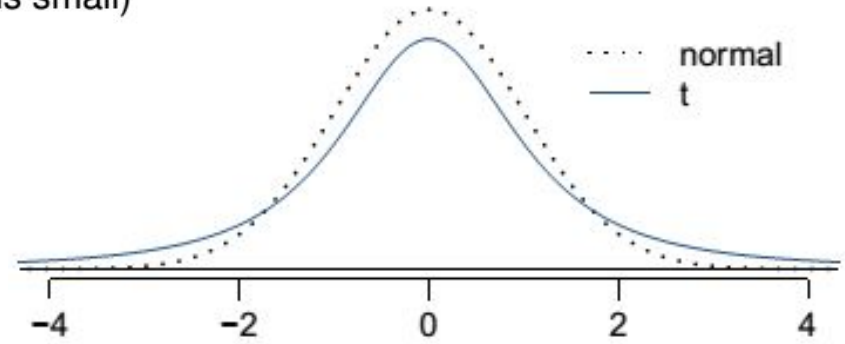
The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for *any* sample size as long as the population distribution is nearly normal.
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets.
- We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.
  - For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

The t distribution

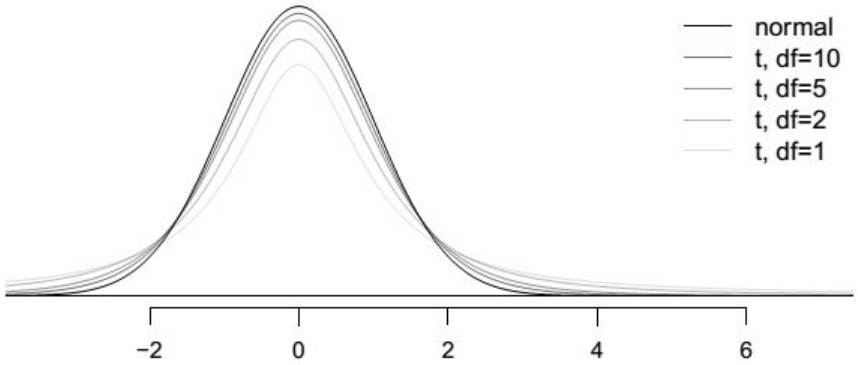
- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t distribution*.

- This distribution also has a bell shape, but its tails are *thicker* than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)



The t distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: *degrees of freedom* (df).



What happens to shape of the t distribution as df increases?

Back to Friday the 13th

	type	date	6th	13th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

$\downarrow$

$\bar{x}_{diff} = 1836$

$s_{diff} = 1176$

$n = 10$

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

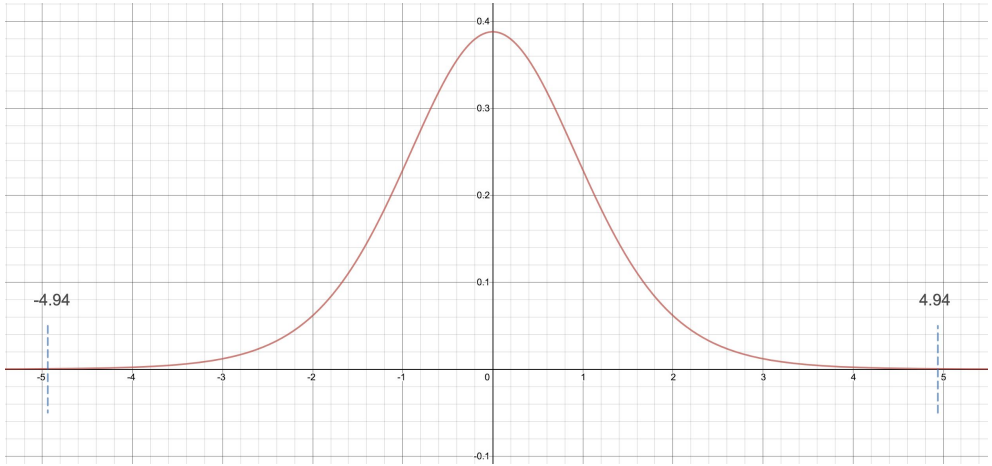
$$\begin{aligned} \textit{point estimate} &= \bar{x}_{diff} = 1836 \\ SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \\ T &= \frac{1836 - 0}{372} = 4.94 \\ df &= 10 - 1 = 9 \end{aligned}$$

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)

[1] 0.0008022394
```



What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6^{th} and 13^{th}.
- But it would be more interesting to find out what exactly this difference is.
- We can use a confidence interval to estimate this difference.

Confidence interval for a small sample mean

- Confidence intervals are always of the form

$$\text{point estimate} \pm ME$$

- ME is always calculated as the product of a critical value and SE.
- Since small sample means follow a t distribution (and not a z distribution), the critical value is a $t^\star$ (as opposed to a $z^\star$).

$$\text{point estimate} \pm t^\star \times SE$$

Finding t^*

- As usual, for a 95% CI, we split the 5% uncertainty between the two tails: 2.5% in each.
- Use the `qt(p, df)` to find the cuttoff.

```
qt(0.025,9)

[1] -2.262157

> qt(0.975,9)

[1] 2.262157

> qt(0.025,9,lower.tail = FALSE)
```

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the difference between the traffic flow between Friday 6th and 13th?

$$\bar{x}_{diff} = 1836 \qquad s_{diff} = 1176 \qquad n = 10 \qquad SE = 372$$

(a) $1836 \pm 1.96 \times 372$

(b) $1836 \pm 2.26 \times 372$ → *(995, 2677)*

(c) $1836 \pm -2.26 \times 372$

(d) $1836 \pm 2.26 \times 1176$

Interpreting the CI

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

$$\mu_{diff:6th-13th} = (995, 2677)$$

We are 95% confident that ...

(a) the difference between the average number of cars on the road on Friday 6th and 13th is between 995 and 2,677.

(b) on Friday 6th there are 995 to 2,677 fewer cars on the road than on the Friday 13th, on average.

(c) on Friday 6th there are 995 fewer to 2,677 more cars on the road than on the Friday 13th, on average.

(d) on Friday 13th there are 995 to 2,677 fewer cars on the road than on the Friday 6th, on average.

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - $n < 30$ and no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

- Confidence interval:

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$

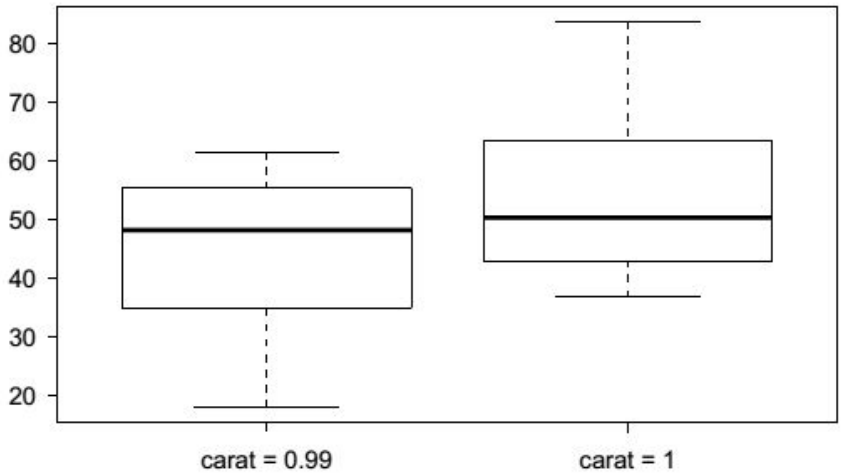
Note: The example we used was for paired means (difference between dependent groups). We took the difference between the observations and used only these differences (one sample) in our analysis, therefore the mechanics are the same as when we are working with just one sample.

t distribution for the difference of two means

Diamonds

- Weights of diamonds are measured in carats.
- 1 carat = 100 points, 0.99 carats = 99 points, etc.
- The difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but does the price of a 1 carat diamond tend to be higher than the price of a 0.99 diamond?
- We are going to test to see if there is a difference between the average prices of 0.99 and 1 carat diamonds.
- In order to be able to compare equivalent units, we divide the prices of 0.99 carat diamonds by 99 and 1 carat diamonds by 100, and compare the average point prices.

Data



	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

Parameter and point estimate

- *Parameter of interest:* Average difference between the point prices of *all* 0.99 carat and 1 carat diamonds.

$$\mu_{pt99} - \mu_{pt100}$$

- *Point estimate:* Average difference between the point prices of *sampled* 0.99 carat and 1 carat diamonds.

$$\bar{x}_{pt99} - \bar{x}_{pt100}$$

Hypotheses

Which of the following is the correct set of hypotheses for testing if the average point price of 1 carat diamonds ($_{pt100}$) is higher than the average point price of 0.99 carat diamonds ($_{pt99}$)?

(a) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} \neq \mu_{pt100}$

(b) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} > \mu_{pt100}$

(c) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} < \mu_{pt100}$

(d) $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$
$H_A : \bar{x}_{pt99} < \bar{x}_{pt100}$

Test statistic

Test statistic for inference on the difference of two small sample means

The test statistic for inference on the difference of two small sample means ($n_1 < 30$ and/or $n_2 < 30$) mean is the T statistic.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

where

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{and} \qquad df = n_1 + n_2 - 2$$

Assuming the variances of the samples are about the same.

Note: The calculation of the df is actually much more complicated. For simplicity we'll use the above formula to estimate the true df. If you are using R, the df will be calculated for you. Trust that number.

Test statistic (cont.)

	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}}$$

$$= \frac{-8.93}{3.56}$$

$$= -2.508$$

p-value

With T=-2.508, and df = 51 how do we find the p-value?

Synthesis

What is the conclusion of the hypothesis test?

- *p-value is small so reject H_0. The data provide convincing evidence to suggest that the point price of 0.99 carat diamonds is lower than the point price of 1 carat diamonds.*

Building a 90% CI

What is the appropriate $t^\star$ for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

(a) 1.32

(b) 1.72

(c) 2.07

(d) 2.82

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79

Confidence interval

Calculate the interval, and interpret it in context.

$$\textit{point estimate} \pm ME$$

$$\begin{aligned}(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^{\star}_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\ &= -8.93 \pm 6.12 \\ &= (-15.05, -2.81)\end{aligned}$$

We are 90% confident that the average point price of a 0.99 carat diamond is $15.05 to $2.81 lower than the average point price of a 1 carat diamond.

Recap: Inference using difference of two small sample means

- If $n_1 < 30$ and/or $n_2 < 30$, difference between the sample means follow a t distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - independence between groups
  - $n_1 < 30$ and/or $n_2 < 30$ and no extreme skew in either group
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = min(n_1 - 1, n_2 - 1)$$

- Confidence interval:

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$