SEIS 631

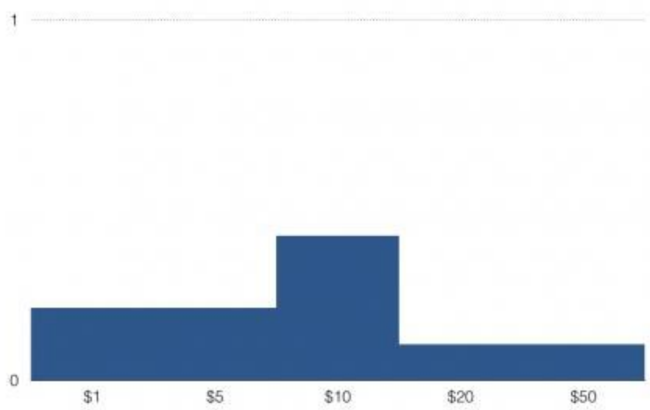
Money Duck

Your Price:

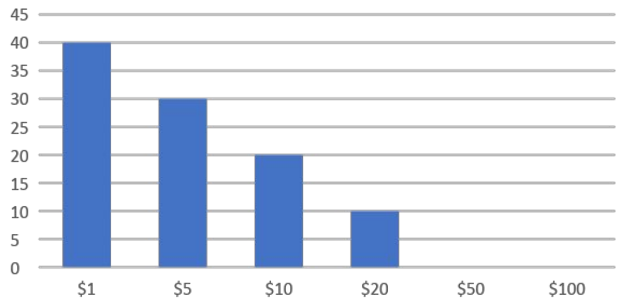
Buy

Money Duck

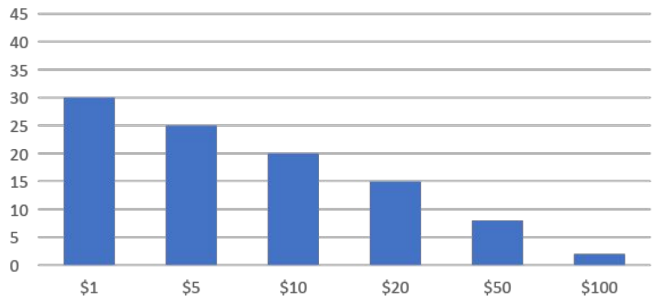
- How much would you pay?



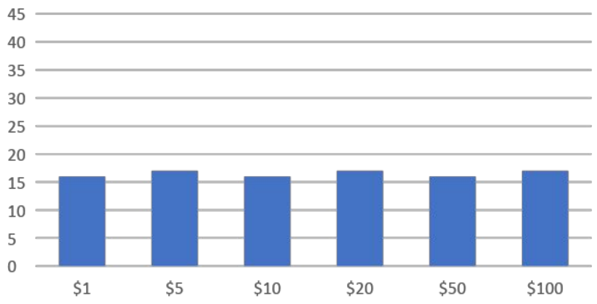
Which would you rather play?



A

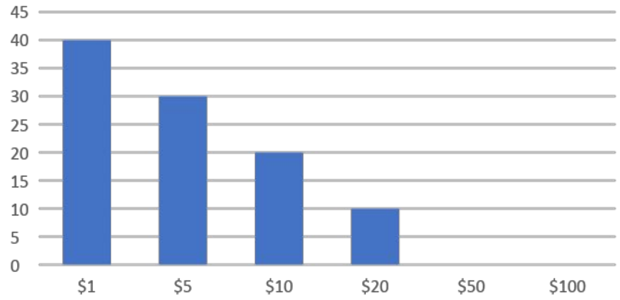


C



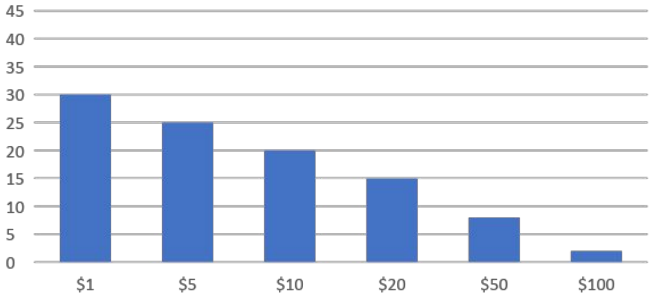
B

Which would you rather *pay* to play?



$6.00

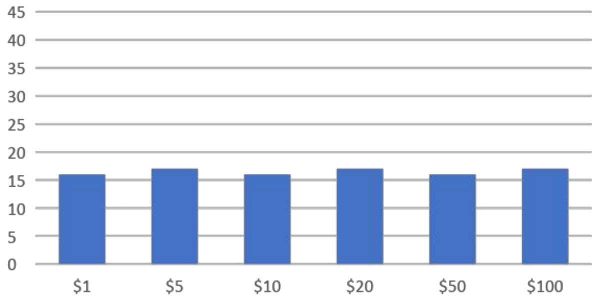
A



$13.00

C

$30.00



B

Expected Value

- The expected value of some event is the average of the values you get if you repeat the event many, many times.

- Expected Values are very similar to means (averages) in a lot of situations

Expected Value

Calculated slightly differently depending on the situation.

- Example: The average weight of an orange is 75g. If I randomly pick an orange, what's the expected value of it's weight?
  - E(weight of orange) = 75 g (the weight of an average orange)
- Example: Roll a die. What's the expected value of the roll?
  - E(Dice Roll) = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.5 (average value of the faces)
- Example: Money Duck
  - E(Money Duck) = 0.2 x \$1 + 0.2 x \$5 + 0.4 x \$10 + 0.1 x \$20 + 0.1 x \$50 = \$12.2

Expected value and its variability: Normal Distribution

In the normal distribution the *expected value* is the **mean**.

$$E(x) = \mu$$

The *variation* we expect in the value is the **standard deviation**

$$Var(x) = \sigma$$

Expected value: Binomial Distribution

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, 100 x 0.262 = 26.2.
- Or more formally, $\mu = np$ = 100 x 0.262 = 26.2.
- But this doesn't mean in every random sample of 100 people exactly 26.2 will be obese. In fact, that's not even possible. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Expected value and its variability: Binomial Distribution

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

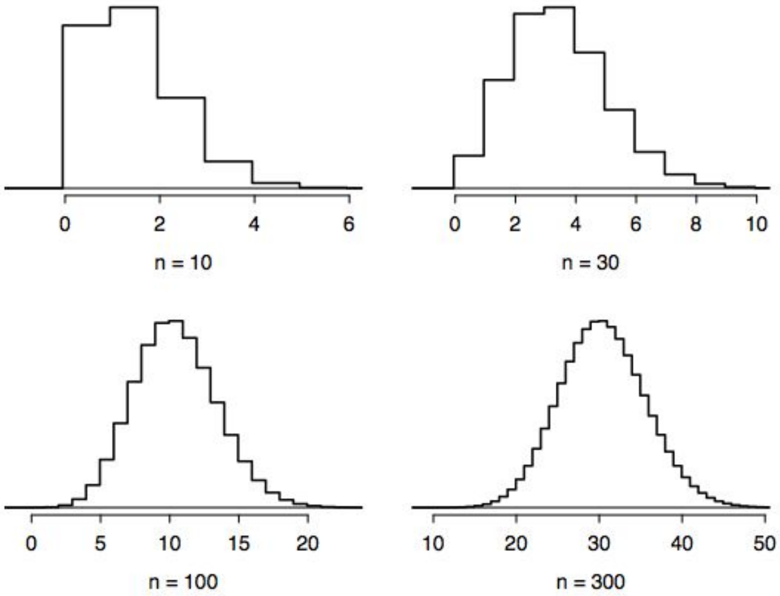
We would expect 26.2 out of 100 randomly sampled Americans to be obese, with a standard deviation of 4.4.

___

Note: Mean and standard deviation of a binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

Approximating Binomial with Normal

Distributions of number of successes



Hollow histograms of samples from the binomial model where p = 0.10 and n = 10, 30, 100, and 300. What happens as n increases?

An analysis of Facebook users

A recent study found that ``Facebook users get more than they give''. For example:

1. 40% of Facebook users in our sample made a friend request, but 63% received at least one request
2. Users in our sample pressed the like button next to friends' content an average of 14 times, but had their content ``liked'' an average of 20 times
3. Users sent 9 personal messages, but received 12
4. 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Source: Power users contribute much more content than the typical user.
http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

Example

This study also found that approximately 25% of Facebook users are considered ***power users***. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that **$n = 245$, $p = 0.25$**, and we are asked for the probability **$P(K \geq 70)$**. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$P(K \geq 70) = P(K = 70) + P(K = 71) + P(K = 72) + \ldots + P(K = 245)$$

Using R

We can calculate this directly using R

The probability of exactly 70:

```
> dbinom(70,size=245,p=0.25))
[1] 0.02509227
```

The probability of at least 70:

```
> sum(dbinom(70:245,size=245,p=0.25))
[1] 0.112763
> pbinom(69,245,.25,lower.tail=FALSE)
[1] 0.112763
```

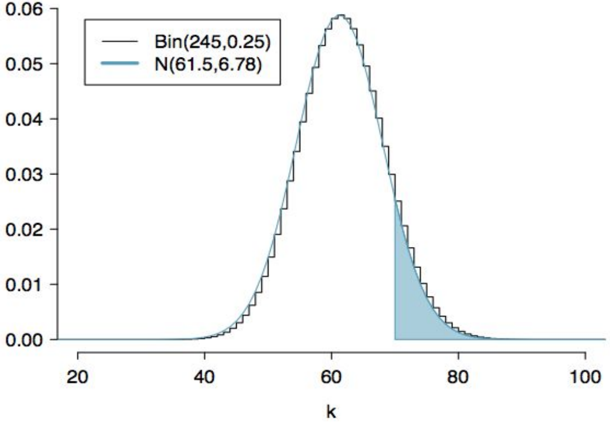
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

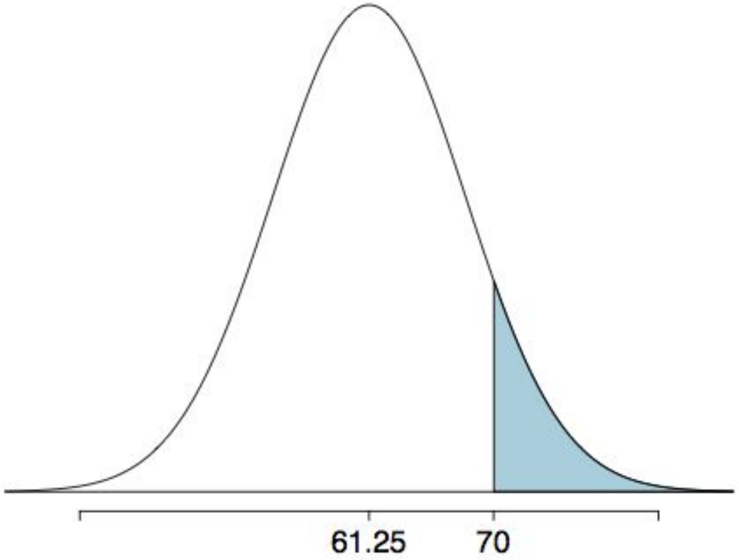
- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.



What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

Using R:

```
      pnorm(1.29, lower.tail = FALSE)
Or    1 - pnorm(1.29)
Or    pnorm(70, 61.25, 6.78, lower.tail =
FALSE)
Or    1 - pnorm(70, 61.25, 6.78)
```

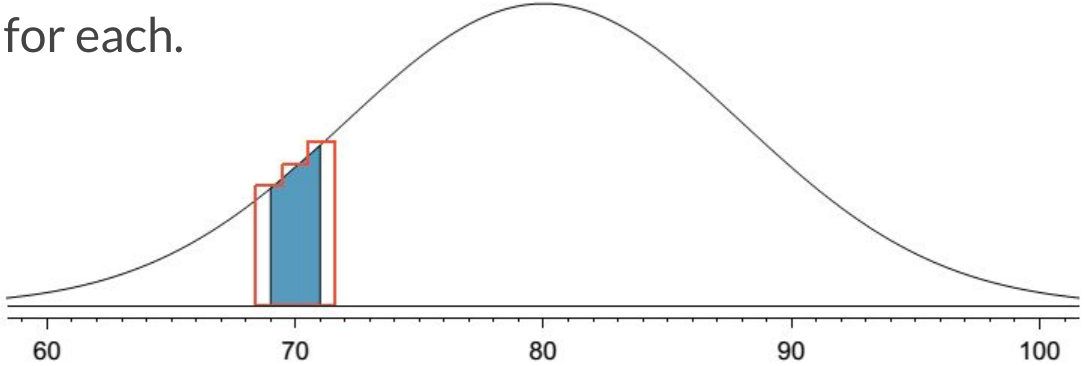
Each of them = 0.0985

$$P(X > 70) = P(Z > 1.29) = 0.0985$$

9.8% vs. 11% is not that great.

We can improve the accuracy by making an adjustment:

- The normal distribution is continuous but binomial trials are discrete.
- X ≥ 70 for a *Normal Distribution* includes everything even slightly larger that 70 (e.g. 70.1, 70.01, etc) but for the binomial distribution it only includes 70,71, 72, 73, etc.
- Compare X > 69 and X ≥ 70 for each.



Adjusting the Normal Approximation

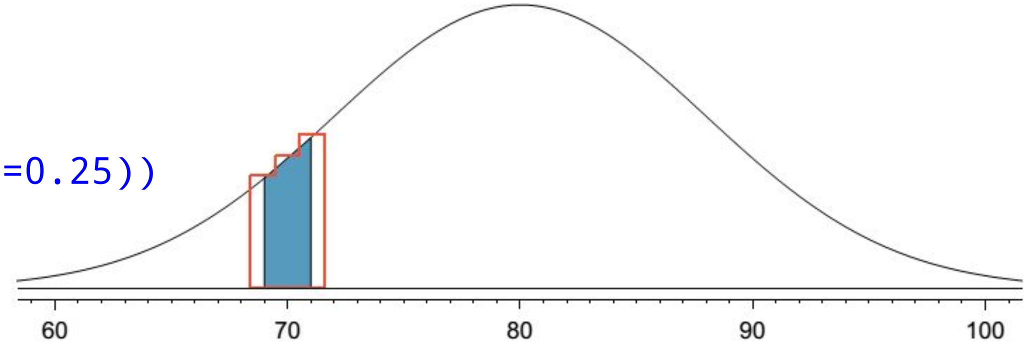
To adjust we lower the lower bound by 0.5 and raise the upper bound by 0.5.

Instead of `pnorm(70, 61.25, 6.78, lower.tail = FALSE) = 0.0985`

we use `pnorm(69.5, 61.25, 6.78, lower.tail = FALSE) = 0.1118`

Compare this to the value of

```
> sum(dbinom(70:245,size=245,p=0.25))
[1] 0.112763
```



Normal Approximation of Binomial - Condition

Normal approximation of the binomial distribution

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1-p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

i.e. $np \geq 10$ *and* $n(1-p) \geq 10$

Example

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

A. $n = 100, p = 0.95$

B. $n = 25, p = 0.45$

C. $n = 150, p = 0.05$

D. $n = 500, p = 0.015$

Example

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

A. $n = 100, p = 0.95$

B. $n = 25, p = 0.45 \rightarrow 25 \times 0.45 = 11.25, 25 \times 0.55 = 13.75$

C. $n = 150, p = 0.05$

D. $n = 500, p = 0.015$

Example

If the true proportion of smokers in a community is 20%, what is the probability of observing exactly 75 smokers in a sample of 400 people?

```
> dbinom(75,size=400,p=0.2)
[1] 0.04185711
```

Example

If the true proportion of smokers in a community is 20%, what is the probability of observing 70 or fewer smokers in a sample of 400 people?

```
> sum(dbinom(0:70,size=400,p=0.2))
[1] 0.1163917
```

Example

If the true proportion of smokers in a community is 20%, what is the probability of observing 70 or fewer smokers in a sample of 400 people?

Conditions: $np = 400 \times 0.20 = 80$ $\quad n(1-p) = 400 \times 0.8 = 320$

With these conditions checked, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

$$\mu = np = 80 \qquad \sigma = \sqrt{np(1-p)} = 8$$

Use the normal model $N(\mu = 80, \sigma = 8)$ to estimate the probability of observing 70 or fewer smokers.

Approximating a Binomial with a Normal Dist.

- Even with the adjustment it's still not a great approximation.
- Technology allows us to make the calculations directly.
- We can approximate a Binomial distribution with a Normal Distribution, but we shouldn't.

Hypothesis Testing

Swimming with Dolphins

- Does swimming with Dolphins help depression?
- How would you test this?
- What are the possible outcomes?
- Null hypothesis, H0, is the hypothesis that things aren't different, or haven't changed

H_0=Swimming with dolphins doesn't improve depression

H_a=Swimming with dolphins does improve depression

Swimming with Dolphins

- Antonioli & Reveley, British Medical Journal
  - 30 participants in total
  - 15 in a Dolphin Group (Treatment)
  - 15 in a Placebo Group (Control)
  - 13 (In *total*) Showed improvement
- What do we expect? How many improvers were in each group...
  - If the treatment worked?
  - If the treatment had no effect?
  - If the treatment had a negative effect?

Swimming with Dolphins

- 10 of 13 who improved were in the dolphin group.

- Is this convincing?
  - Even if the treatment is effective, could this have happened by chance?
  - How likely would we be to get these data if the treatment were not effective?
- Let's try to simulate using R:
  - How likely is it that we get these results by chance, assuming no impact from the treatment:
    - Assumptions: Assume that 13/30 ~ 43% of patients would improve on their own (What hypothesis does this follow?).

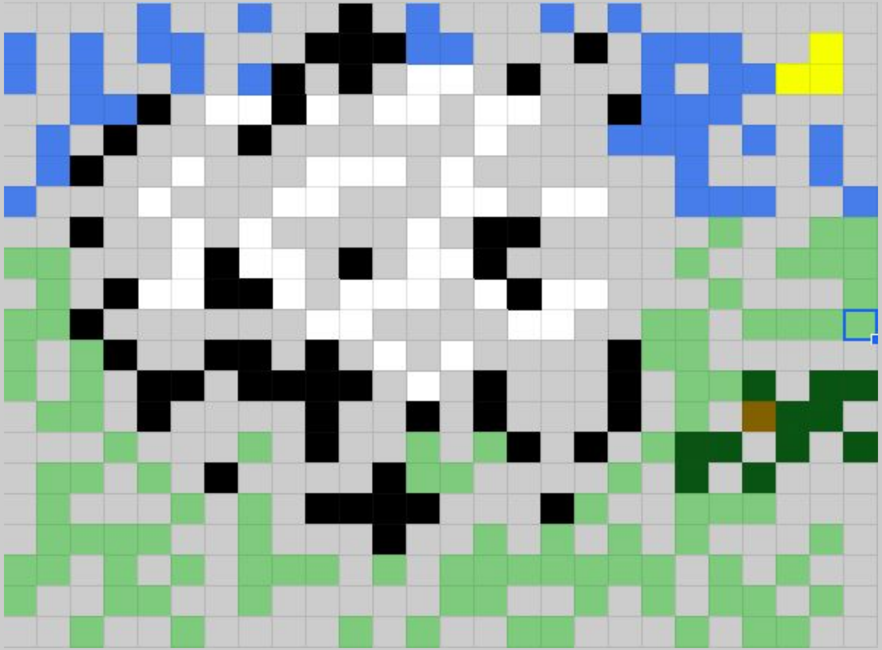
Populations vs. Samples vs. Sampling Distributions

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:

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

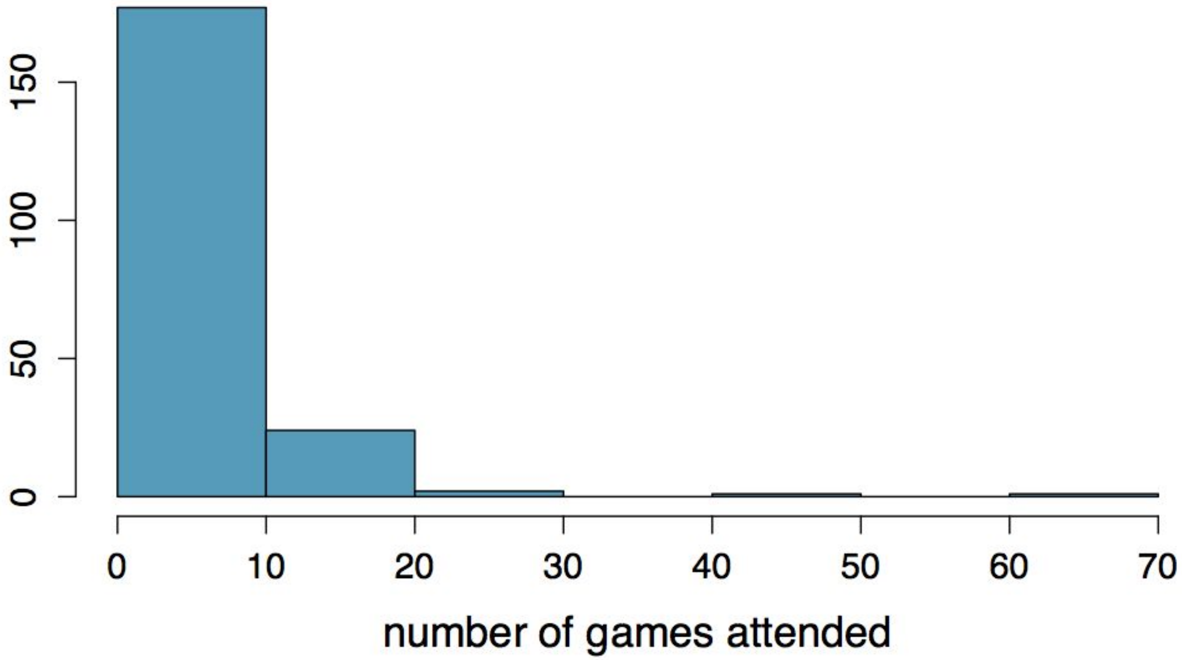
where *SE* is represents standard error, which is defined as the standard deviation of the sampling distribution. **If σ is unknown, use** *s*.

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).

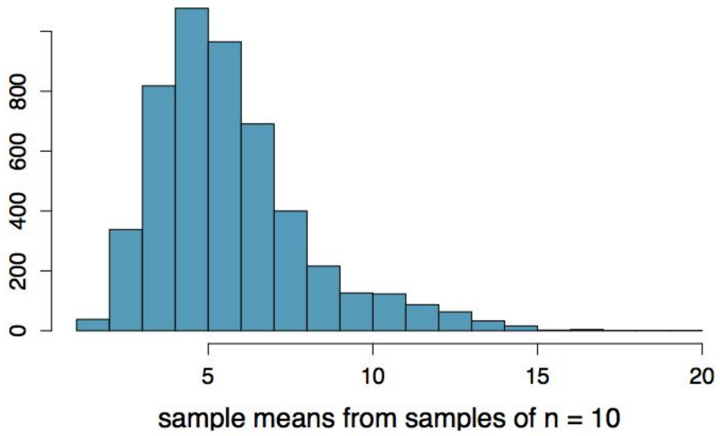
Average number of basketball games attended

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



Average number of basketball games attended (cont.)

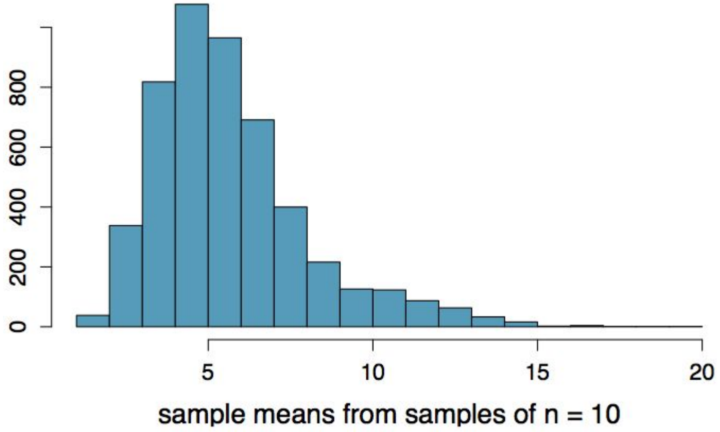
Sampling distribution, n = 10:



What does each observation in this distribution represent?

Average number of basketball games attended (cont.)

Sampling distribution, n = 10:

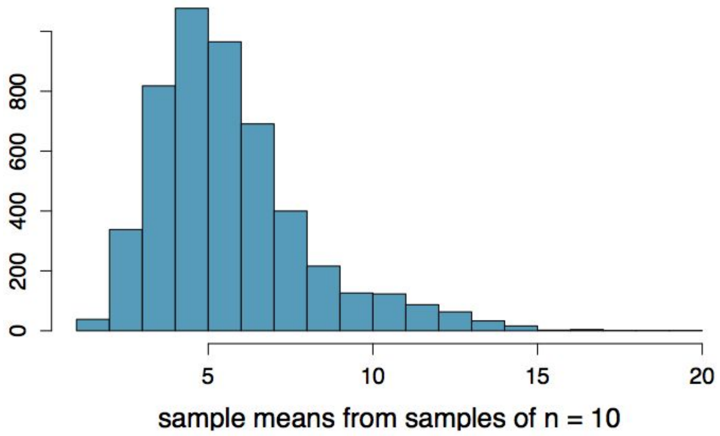


What does each observation in this distribution represent?

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

Average number of basketball games attended (cont.)

Sampling distribution, n = 10:



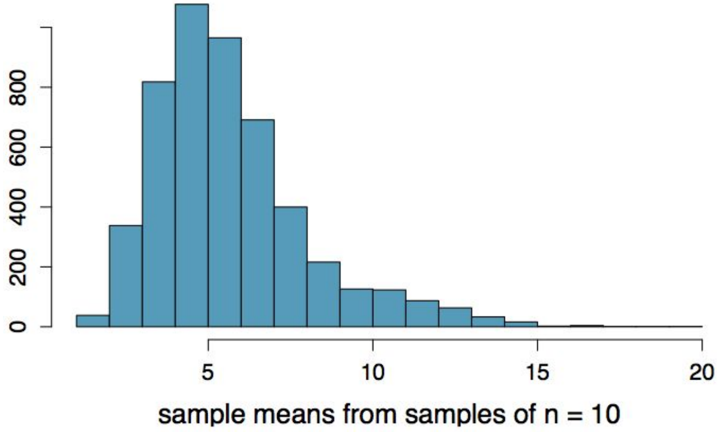
What does each observation in this distribution represent?

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?

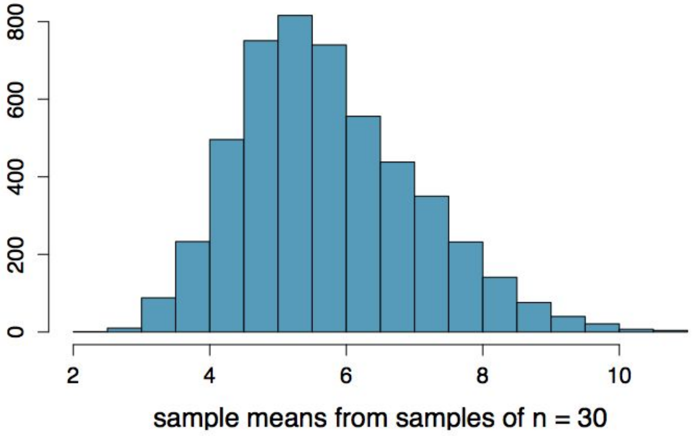
Sample mean ($\bar{x}$) of samples of size n = 10.

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:

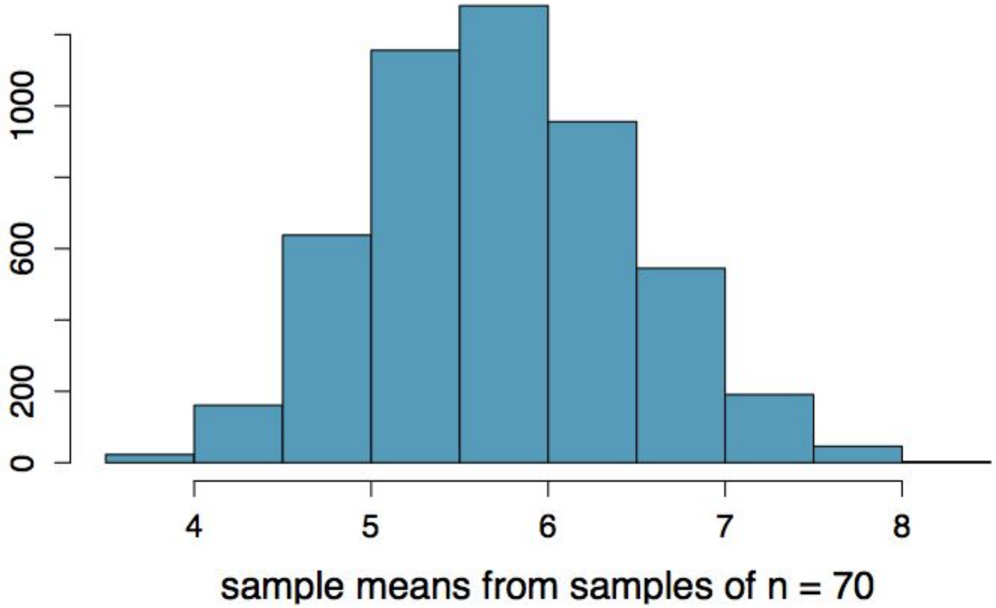


How did the shape, center, and spread of the sampling distribution change going from n = 10 to n = 30?

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

Average number of basketball games attended (cont.)

Sampling distribution, n = 70:

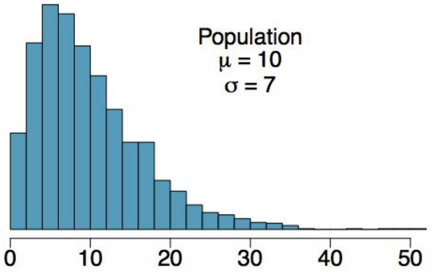


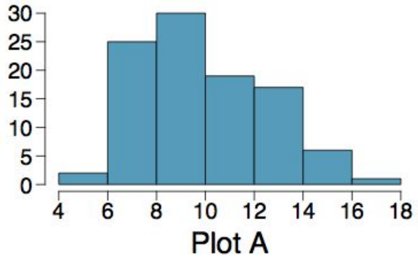
Question

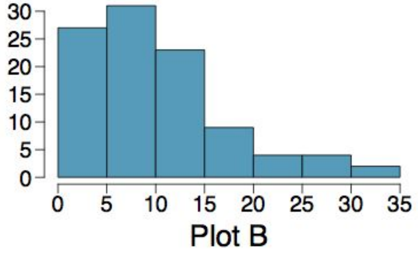
Four plots: Determine which plot (A, B, or C) is which.

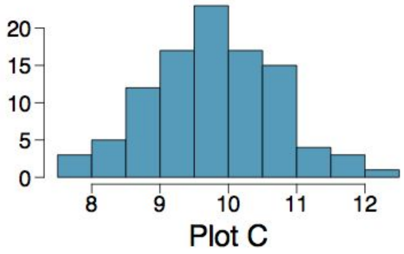
NOTE: First plot is distribution for a population ($\mu = 10, \sigma = 7$),

- 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.

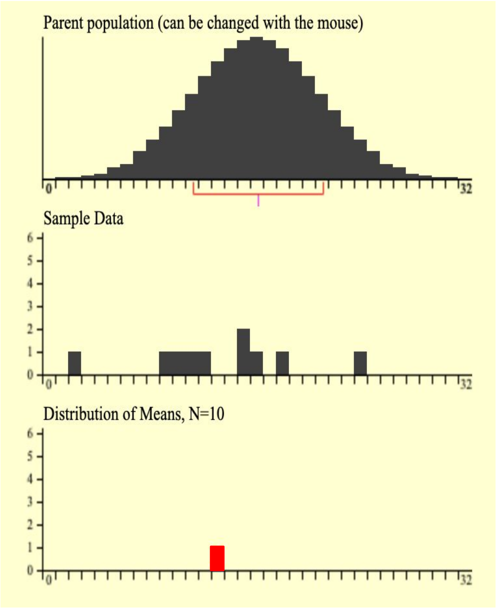
We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

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.

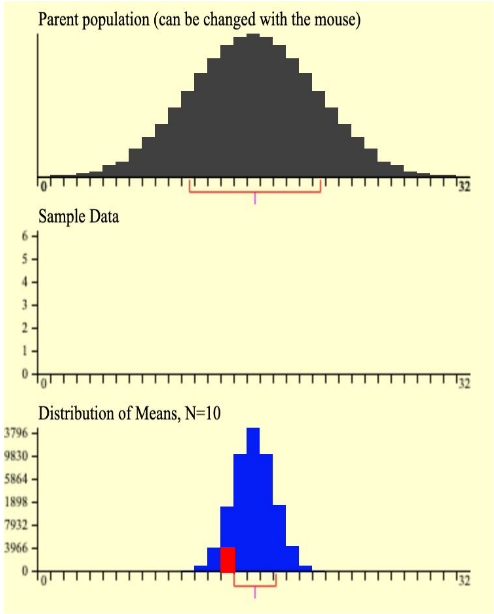
Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462)
and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Confidence Intervals

Reporting Just A Point Estimate



Reporting a Confidence Interval

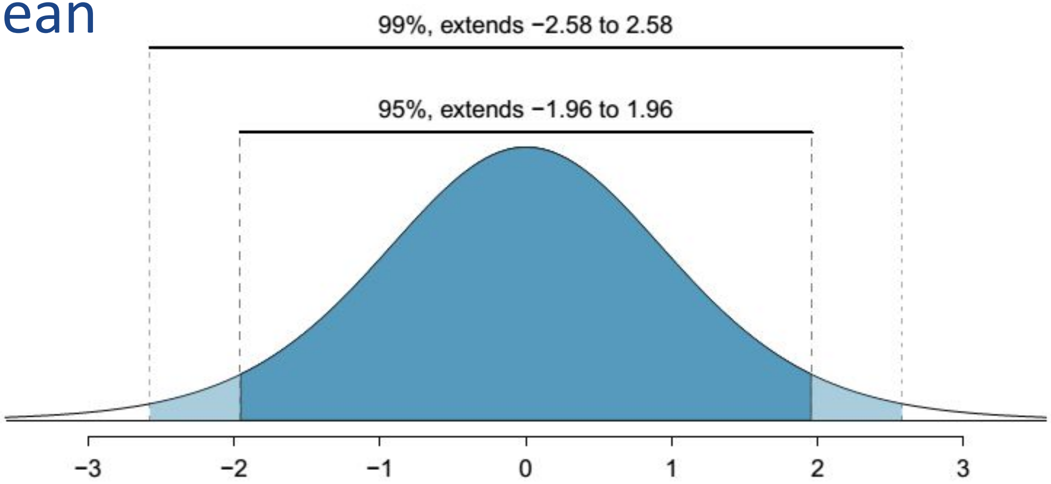


Confidence Intervals

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.

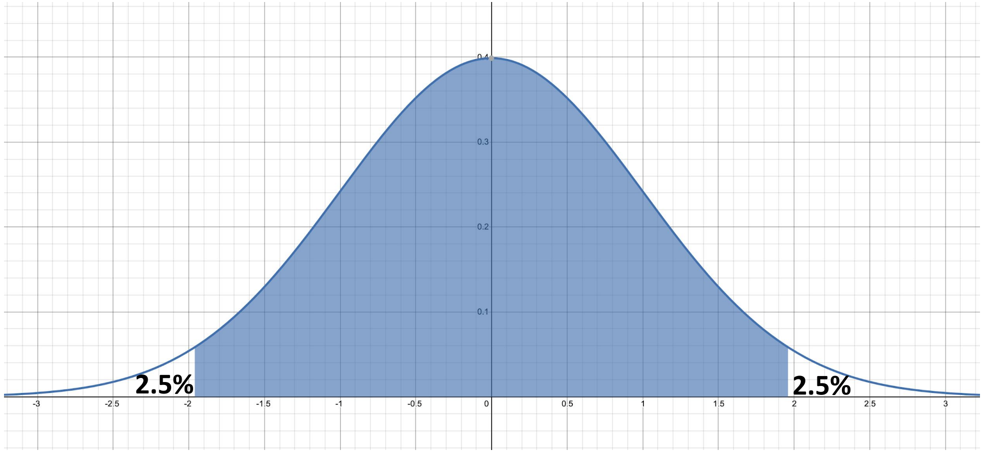
Finding the Critical Value

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

- 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?

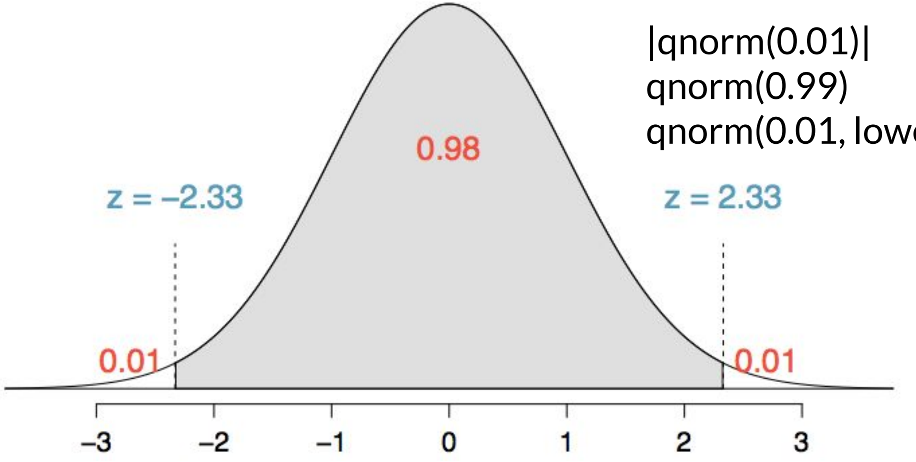
(a) Z = 2.05

(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

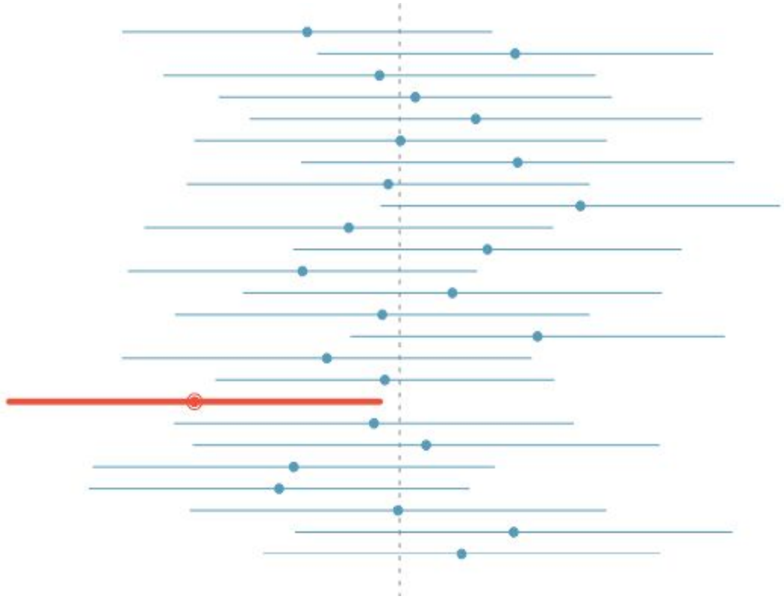
Accuracy vs Precision of CI

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

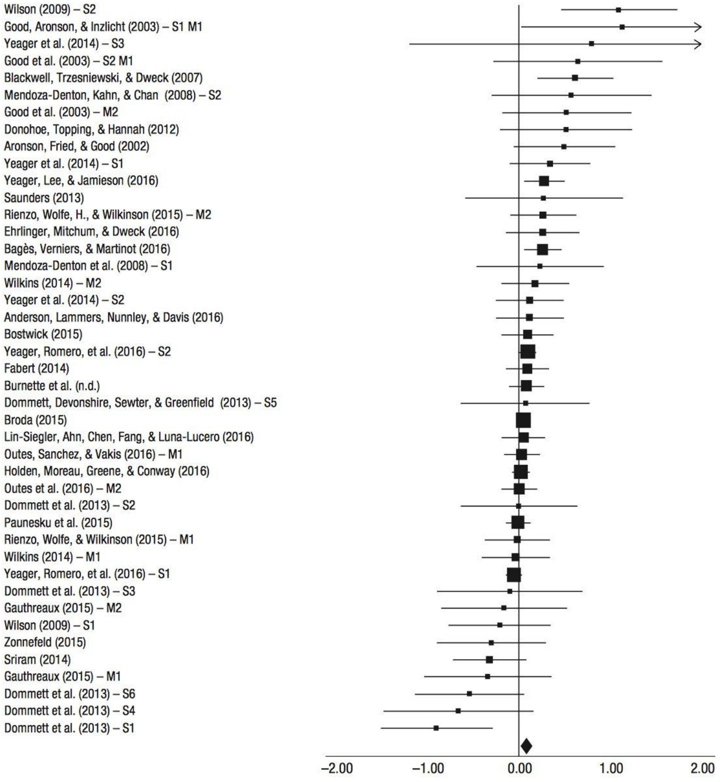


Fig. 4. Standardized mean differences (Cohen's *ds*) in academic achievement between students receiving a growth-mind-set intervention and students in the comparison group. Cohen's *ds* (squares) and 95% confidence intervals (error bars) are displayed for all effects entered

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

<u>**Correct:**</u> We are XX% confident that the true population parameter lies within (our interval)

Things to remember to include:

<u>Our confidence level</u>: ex. 99%, 95%, etc

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

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

<u>**Incorrect**</u>: 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?

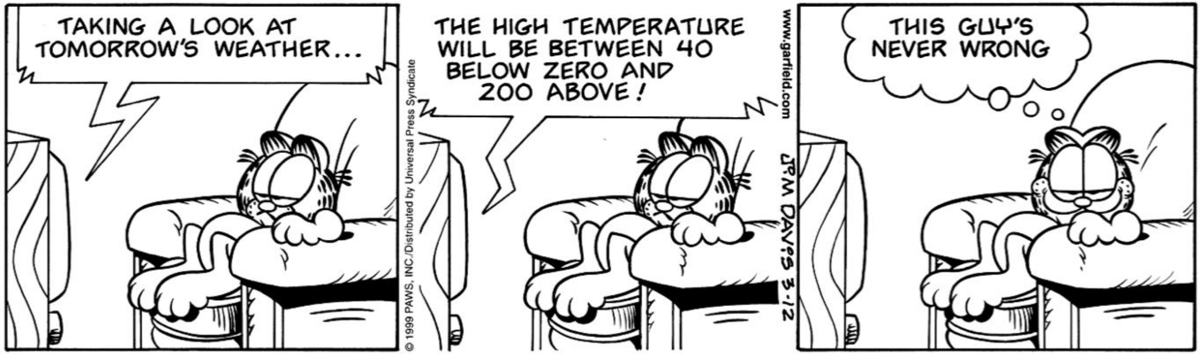
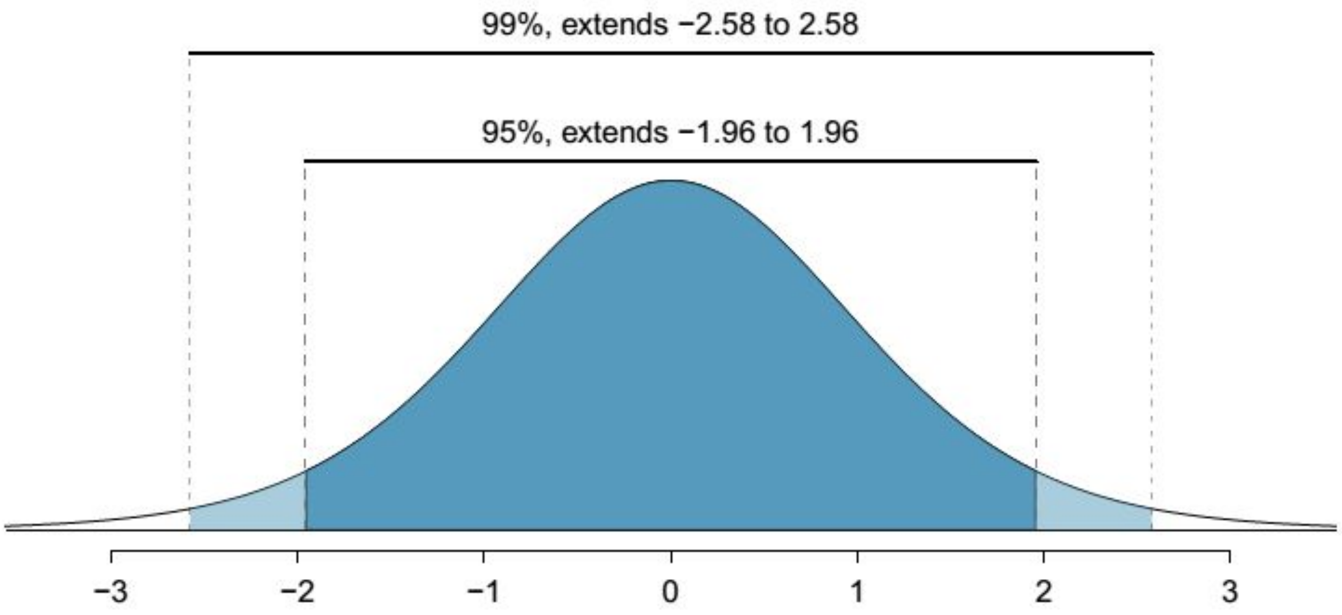
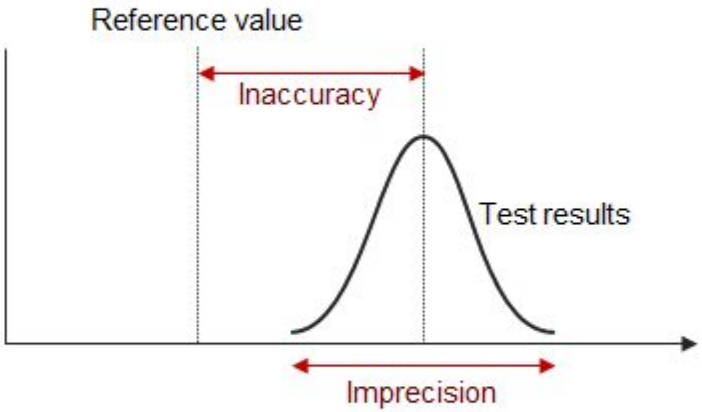


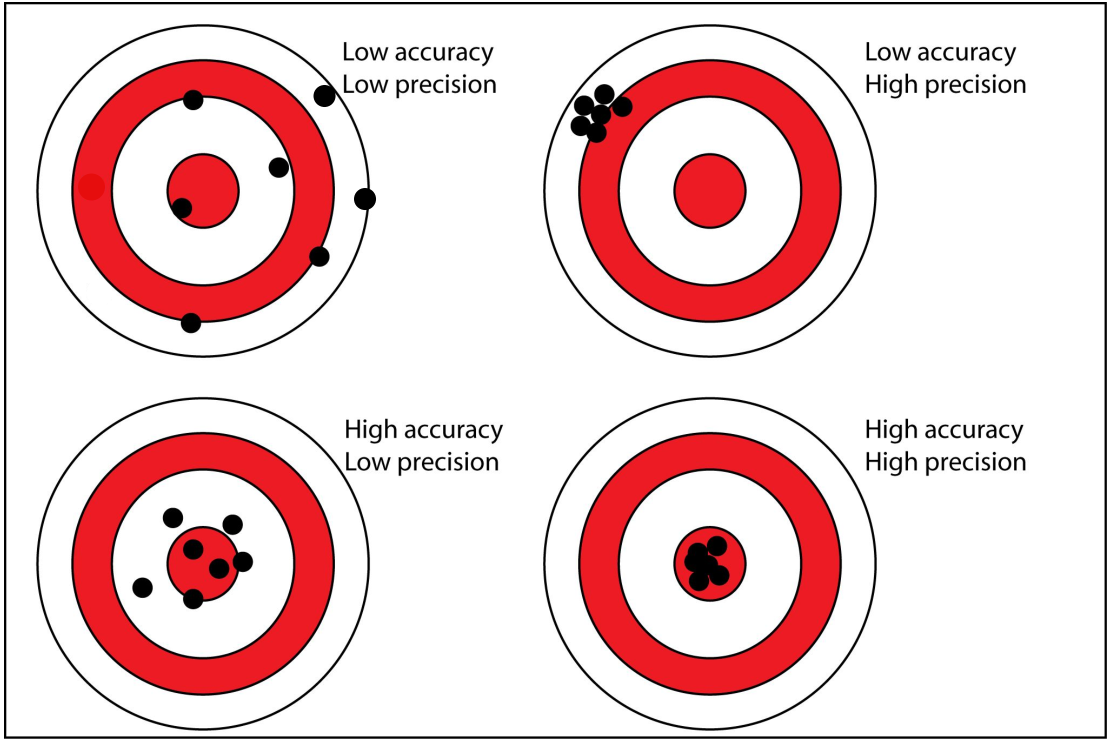
Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif



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?

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

- 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?

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

(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.

<u>Conditions:</u>

- 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

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

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

$\rightarrow \quad 3.2 \pm 0.48$

$\rightarrow \quad (2.72, 3.68)$

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:

$$\begin{aligned} &\text{point estimate } \pm ME \\ = \ &\text{point estimate } \pm z^* \cdot SE \end{aligned}$$

- 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

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

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

$$\bar{x} = 11.7$$

$$sd = 3.47$$

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

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

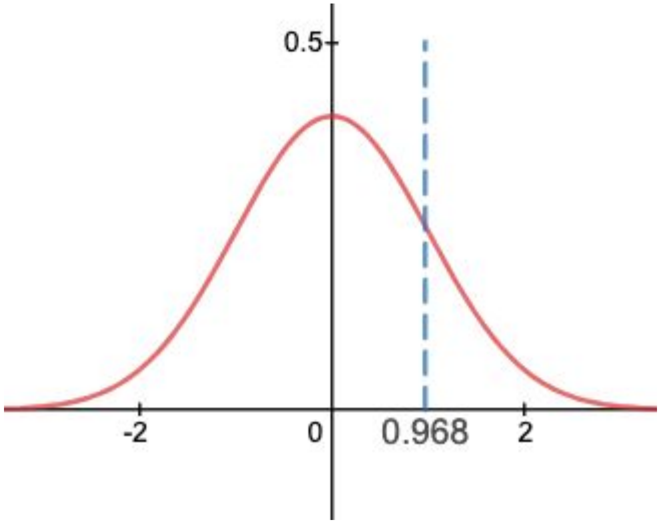
$$\bar{x} = 11.7$$

$$sd = 3.47$$

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

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?



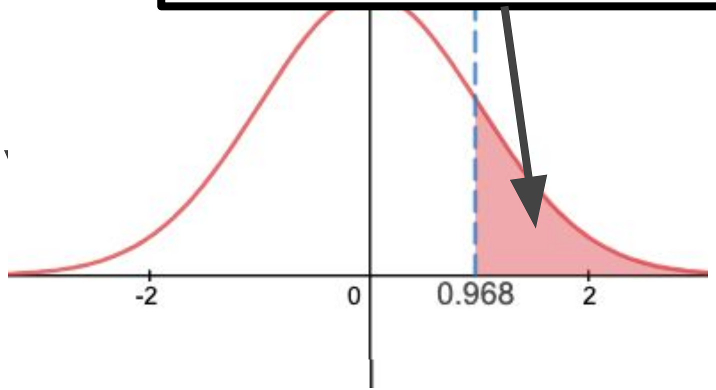
$$z = \frac{11.7 - 11}{2.8/\sqrt{15}} = 0.968$$

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

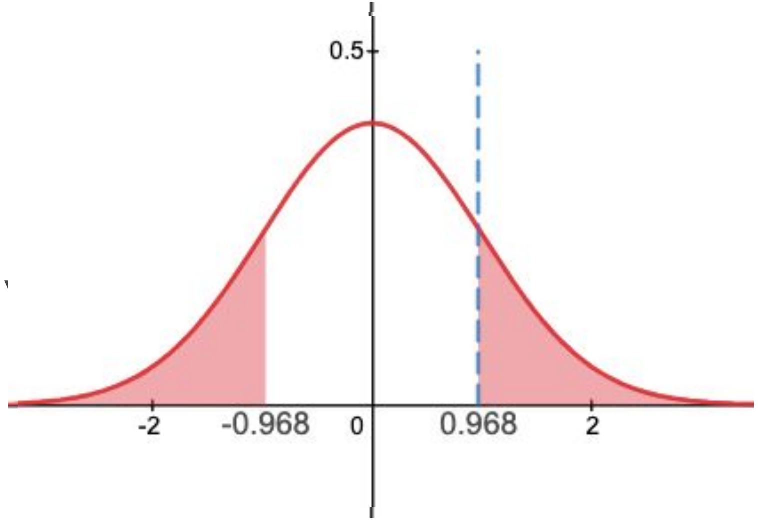
```
>pnorm(0.968, lower.tail = FALSE)
```



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

Calculating *p*-values

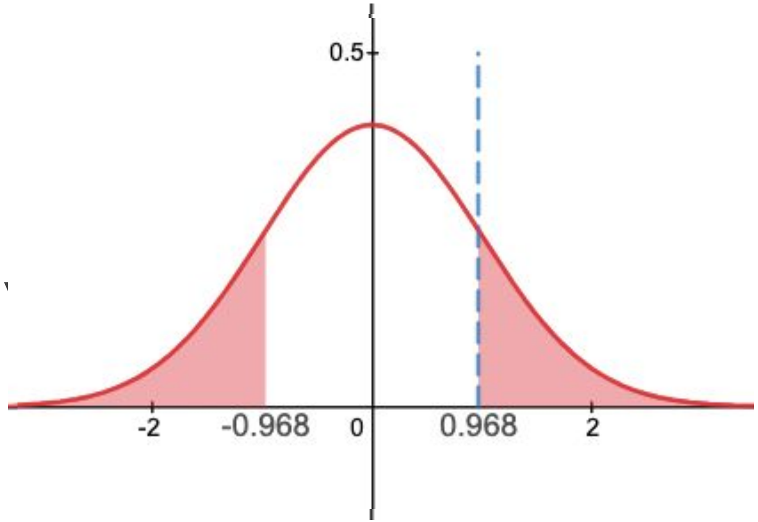
- What does the *z*-score tell us?
  - 11.7 is 0.968 ***standard errors*** above the mean.
- If there were ***truly*** no difference, how likel
  are we to get such a difference?



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

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

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

σ VS. S

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

- If σ is not known <u>AND</u> 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:

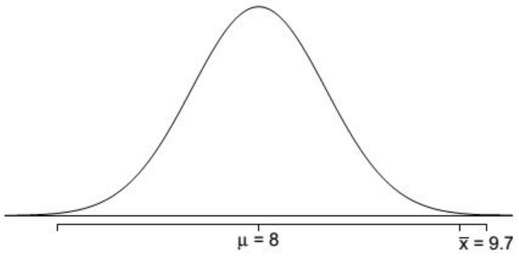
   $H_o: \mu = 8$ vs. $H_A: \mu > 8$

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

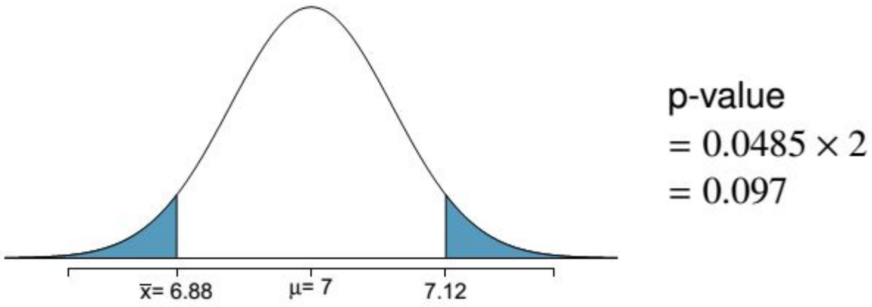
   $Z = \dfrac{6.88 - 7}{0.94 / \sqrt{169}} = -1.659$

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:



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.

<u>Note</u>: 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

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

Upper CI Bound

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

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

http://strata.uga.edu/6370/lecturenotes/pvaluecor
eintervals.html

- Examples

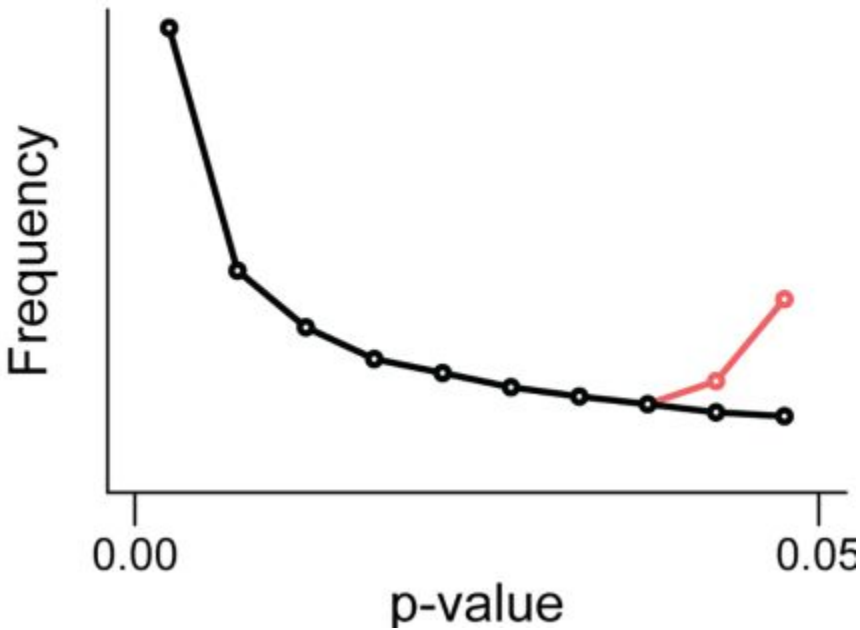
[3%, 17%] – effect is positive, *maybe* large

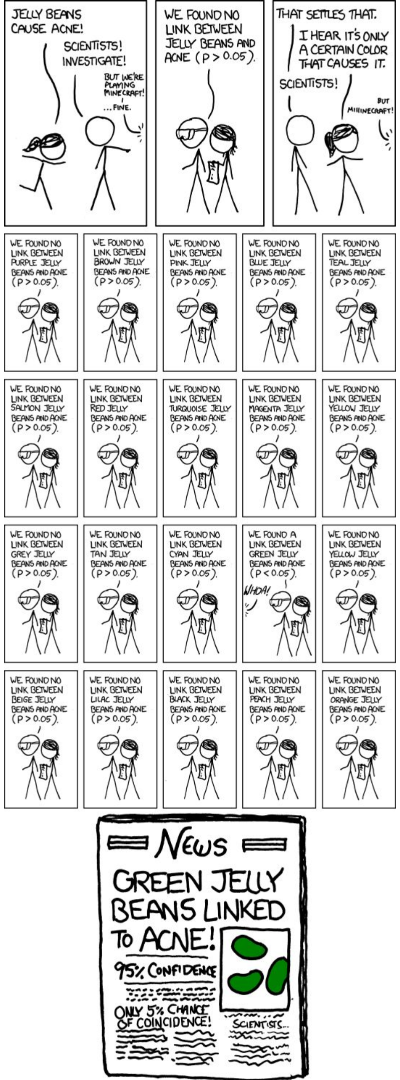
[9%,11%] – effect is positive *and* large

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

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

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

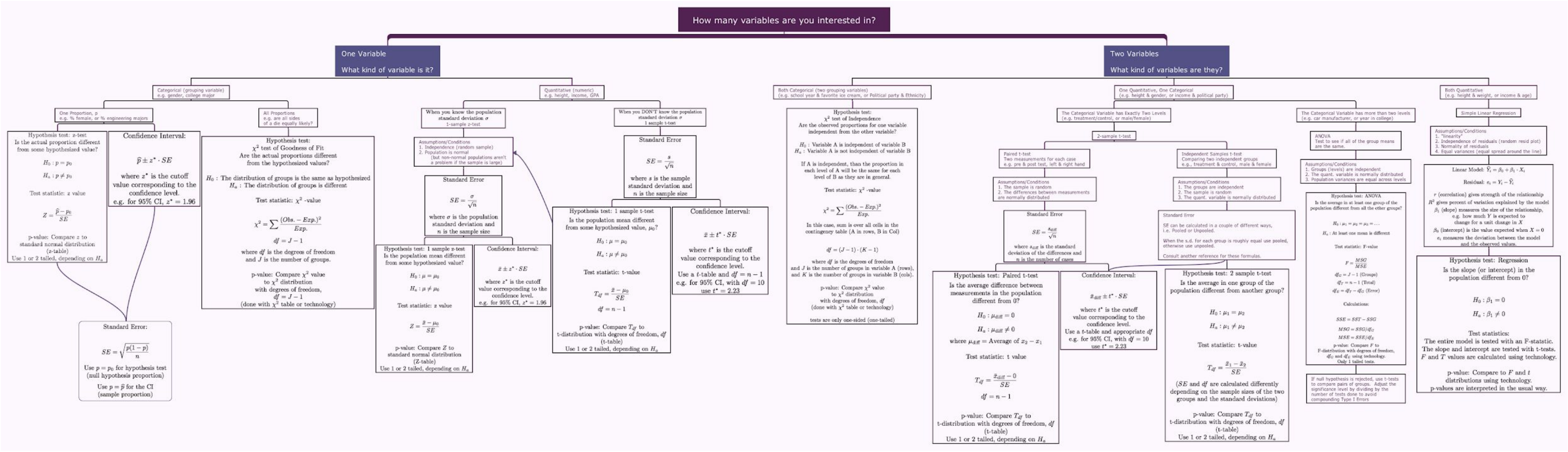
- Increase alpha
  - Set at the beginning of experiment
- Conduct one tailed test
  - Have to decide before experiment
- Decrease random error
  - Through more advanced sampling and experimental techniques
- Increase sample size
  - Depending on the situation, this is the most straightforward!

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.

<u>High power</u>: Detect "truth", but may be too "sensitive"

<u>Low power</u>: 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
$Z = \frac{\widehat{p} - p_0}{SE}$
p-value: Compare z to standard normal distribution (z-table)
Use 1 or 2 tailed, depending on H_a
Confidence Interval:
$\widehat{p} \pm z^* \cdot SE$
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 $p = \widehat{p}$ 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 $t^* = 2.23$
Two Variables
What kind of variables are they?
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)
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
$SE = \frac{s_{\text{diff}}}{\sqrt{n}}$
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?
$H_0 : \mu_{\text{diff}} = 0$
$H_a : \mu_{\text{diff}} \neq 0$
where $\mu_{\text{diff}} =$ Average of $x_2 - x_1$
Test statistic: t value
$T_{df} = \frac{\bar{x}_{\text{diff}} - 0}{SE}$
$df = n - 1$
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:
$\bar{x}_{\text{diff}} \pm t^* \cdot SE$
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
Test to see 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. Normality 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-statstic.
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.