Inference for Means (Chapters 19-22)

So far, we've discussed randomization, bootstrap, and mathematical models as methods to approximate/describe a sampling distribution and quantify variability, as well as how these methods can be used to answer research questions about categorical data. Now, we turn to how these three methods can be used to answer research questions for quantitative data, specifically a quantitative response. We'll consider scenarios where there is a single numerical variable measured (one mean, Chapter 19), scenarios where a categorical explanatory variable with two possible values is recorded/assigned and a numerical response variable is observed (two independent means, Chapter 20), and scenarios in which a categorical explanatory variable with more than two possible values is recorded/assigned and a numerical response variable is observed (many means, Chapter 22). We'll also encounter a new scenario: the difference between paired observations (Chapter 21). We'll also meet two new distributions!

In all of these scenarios the parameter(s) of interest is the mean (μ) of the population(s) under consideration. The natural estimator of the population mean is the sample mean, $\bar{X}$. In Chapter 19 (and, spoiler alert, Chapter 21) we'll have a single μ. In Chapter 20 we'll have two μ_is, and in Chapter 22 we'll have several μ_is.

Like we did with proportions, we'll rely on the Central Limit Theorem to model $\bar{X}$ using the normal distribution when we consider the mathematical model approaches. Also as with proportions, certain conditions must be met for this approach to be valid. We'll discuss those conditions in each of the data scenarios. We'll start with a single variable measured on each sample unit, where the observation results in a number.

Inference for a Single Mean (Chapter 19)

Bootstrap Confidence Intervals for a Mean

Consider the following scenario. We'd like to learn about the true average wait time at Starbucks for a particular drink. To learn about this, we go to 6 randomly selected Starbucks locations in the same city, all at 10:00 am on Monday. At each location we order the same drink and observe the waiting time in seconds until it is prepared. The parameter of interest is

$\mu =$

The sample statistic is

$\bar{X} =$

Suppose we observed wait times of: 110, 54, 76, 123, 91, and 101. Based on our sample of six locations, the sample average wait time is $\bar{x} = 92.5$ seconds with sample standard deviation $s = 24.76$ seconds.

Like we did with proportions, we can use the bootstrap method to approximate the variability we expect to see in sample means (calculated from 6 observations) from sample to sample:

Let's go to R! We'll start by setting up the data

```
waittime<-c(110, 54, 76, 123, 91, 101)
waittime
```

```
[1] 110  54  76 123  91 101
```

Now, we'll find the observed sample mean and standard deviation from our 6 observations:

```
mean(waittime)
```

```
[1] 92.5
```

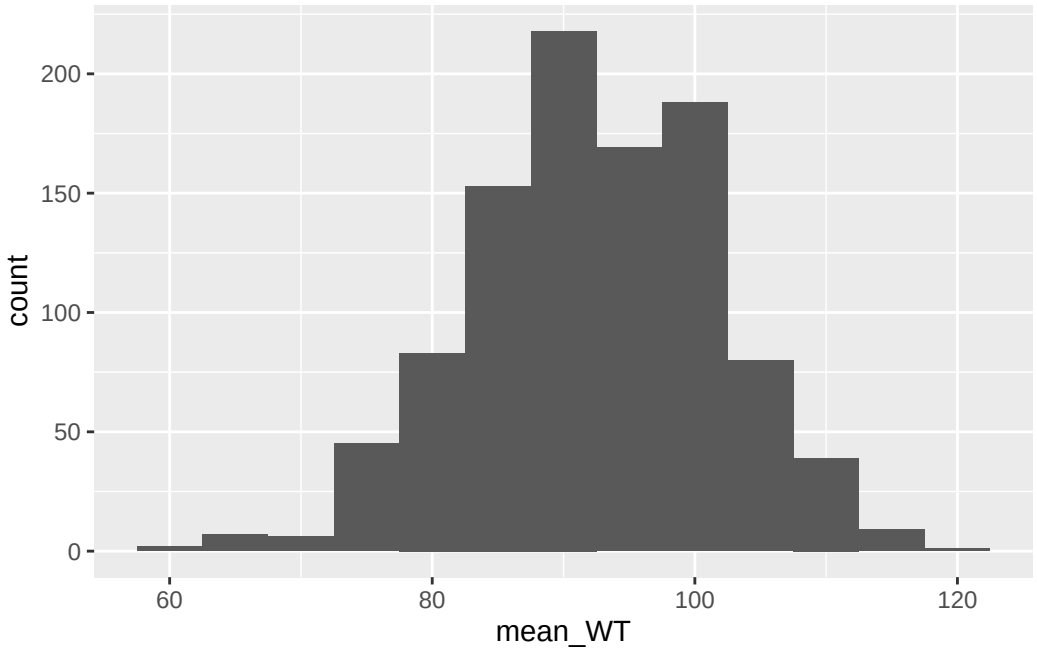
```
sd(waittime)
```

```
[1] 24.76086
```

We'll draw bootstrap samples just like we did with proportions, draw repeated samples of size 6 from our data.

```
library(ggplot2)

#Set up an empty data set with 2 columns: simulation number, bootstrap mean#

boot.samples<-data.frame(sim=1:1000,mean_WT=NA)
```

```r
#For each row in the data set, draw a bootstrap sample from the original data and find#
# mean_WT#

for(i in 1:1000){
  boot.samples$mean_WT[i]<-mean(sample(waittime,size=6,replace=TRUE))
}

#Histogram#
boot.hist<-ggplot(boot.samples, aes(mean_WT)) + geom_histogram(binwidth=5)

#See the plot#
boot.hist
```




```r
#To get the bootstrap percentile confidence interval, #
#start by ranking the bootstrap means from smallest to largest #
rankmean<-sort(boot.samples$mean_WT)

#Lower endpoint is the 2.5th percentile (95% confidence)#
lower<-rankmean[25]
lower
```

```
[1] 73.66667
```

```
#Upper endpoint is the 97.5th percentile (95% confidence)#
upper<-rankmean[975]
upper
```

```
[1] 110.1667
```

This will give us a **bootstrap percentile confidence interval**.

The histogram of the bootstrapped sample means is relatively bell-shape, so we could also find a **bootstrap SE confidence interval**. For that, we'll need the bootstrap SE (the standard deviation of the bootstrapped sample means).

```
#Bootstrap standard error of the mean#
sd(rankmean)
```

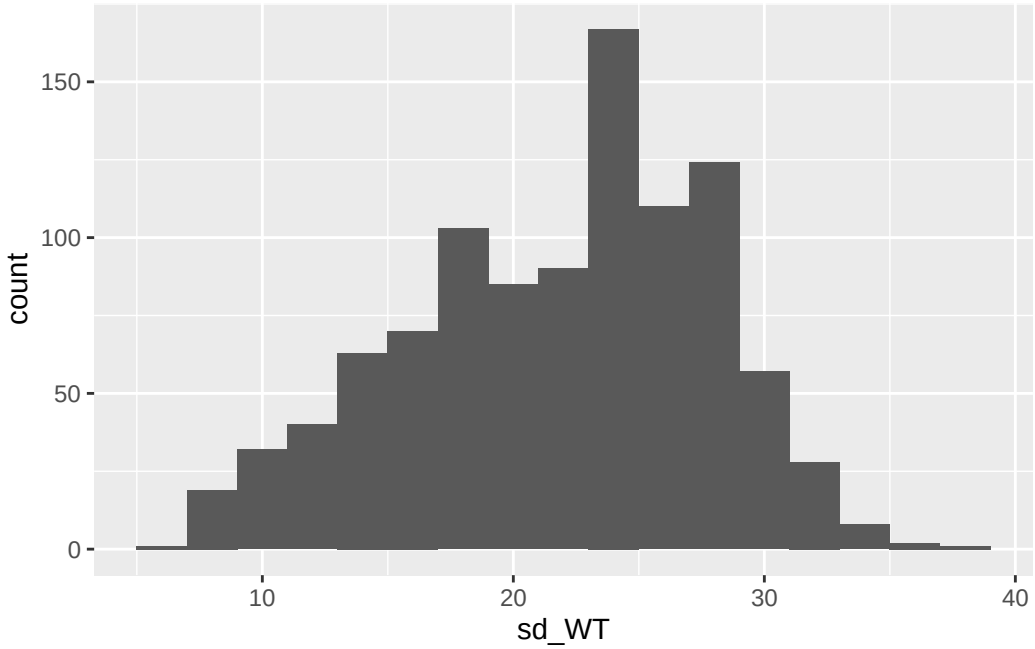
```
[1] 9.335909
```

The bootstrap method works for other statistics as well (even when the mathematical model does not)–like standard deviation, median, range, etc. With other stats we won't necessarily end up a bell-shaped distribution. That's okay–we can use the percentile method.

For example, we could use the bootstrap approach to get a confidence interval for σ, the true standard deviation of wait time.

```
#Set up an empty data set with 2 columns: simulation number, bootstrap SD#
boot.samples<-data.frame(sim=1:1000, sd_WT=NA)

#For each row in the data set, draw a bootstrap sample from the original data and find#
# sd_WT#
```

```r
for(i in 1:1000){
  boot.samples$sd_WT[i]<-sd(sample(waittime,size=6,replace=TRUE))
}

#Histogram#
boot.hist<-ggplot(boot.samples, aes(sd_WT)) + geom_histogram(binwidth=2)

#See the plot#
boot.hist
```




```r
#To get the bootstrap percentile confidence interval, #
#start by ranking the bootstrap sds from smallest to largest #
ranksd<-sort(boot.samples$sd_WT)

#Lower endpoint is the 5th percentile (90% confidence)#
lower<-ranksd[50]
lower
```

```
[1] 10.91788
```

```r
#Upper endpont is the 95th percentile (90% confidence)#
upper<-ranksd[950]
upper
```

```
[1] 30.49208
```

Example: Wildlife researchers trapped and measured six adult male collared lemmings. The data (in mm) are: 104, 99, 112, 115, 96, 109. Use bootstrap methods to find a 95% confidence interval for the true mean size of adult male collared lemmings. Use both the percentile approach and the bootstrap SE approach.

Mathematical Model Approach for a Mean

Like with proportions, we'll use the Central Limit Theorem here.

This presents a few complications:

The natural fix is to use s (the sample standard deviation) in place of σ, so $SE =$

But this leads to yet another complication: the normal distribution isn't quite right. We end up with a distribution that has heavier tails than the normal.

Instead, we use the t-distribution which, like the chi-squared, has the degrees of freedom parameter:

- degrees of freedom determines the shape of the t, with the distribution getting closer and closer to the normal as the df increase

  Demo: Compare t and Z

  As $df \to \infty$, the t goes to the standard normal.

- In this scenario of a single mean, $df =$
- R function: `pt(q,df)`

Mathematical model confidence intervals for a single mean

Let's work through confidence intervals for a single mean by way of example. Suppose we want to get a sense of the average number of goals scored per game in the NHL, and the average margin of victory. We record data on all 44 NHL games played over a Thursday-Monday in December. Let's first look at the number of goals. The data are in Canvas.

We'll start with visualizing the data in R

```
#Read in the NHL data#
hockey<-read.csv("NHLGames.csv",header=TRUE)
head(hockey)
```

```
  Goals MarginVictory
1     6             2
2     3             1
3     9             1
4     7             3
5     6             2
6     5             1
```

```
library(ggplot2)

#Summarize Number of Goals#
summary(hockey$Goals)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   5.000   6.000   6.114   7.000   9.000
```

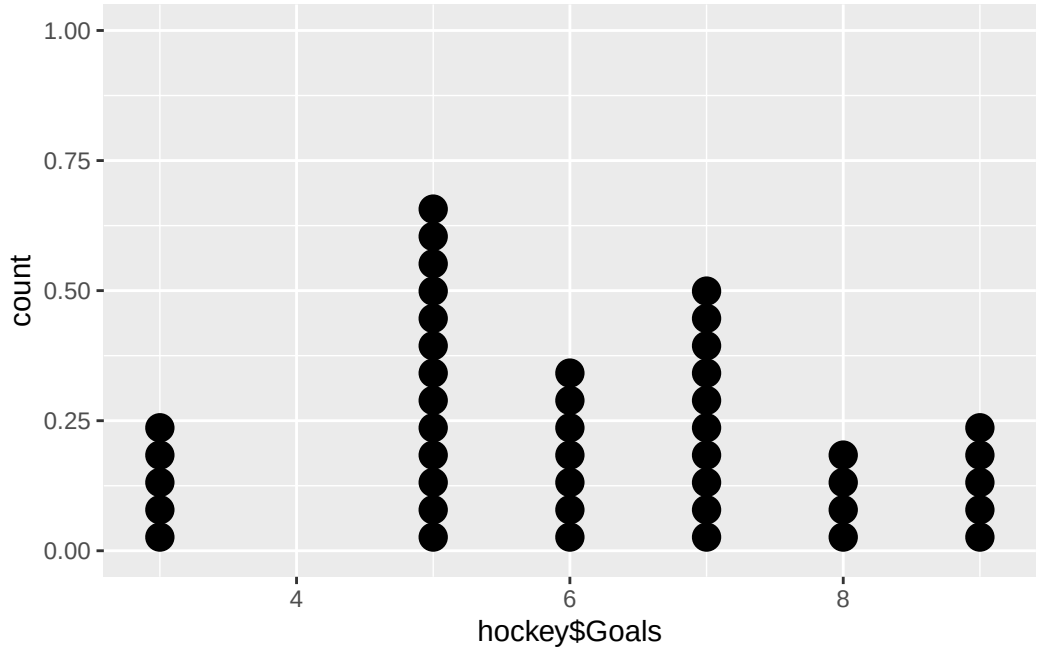
```
sGoals<-sd(hockey$Goals)
sGoals
```

```
[1] 1.728232
```

```
Goals.dot<-ggplot(hockey, aes(hockey$Goals)) + geom_dotplot()
Goals.dot
```

```
Warning: Use of `hockey$Goals` is discouraged.
i Use `Goals` instead.
```

```
Bin width defaults to 1/30 of the range of the data. Pick better value with
`binwidth`.
```



Are the conditions for the mathematical model met?

- Sample size
- Independence

The general form of the confidence interval hasn't changed:

$$\text{point estimate} \pm \text{multiplier} \times SE$$

In our data set set, there are $n = 44$ games.

```
#For a confidence interval, need the multiplier for a 95% confidence interval, puts 0.025 in
qt(0.05, df=43, lower.tail=FALSE)
```

```
[1] 1.681071
```

What about a 90% confidence interval? What would change?

What about a 95% confidence interval for margin of victory?

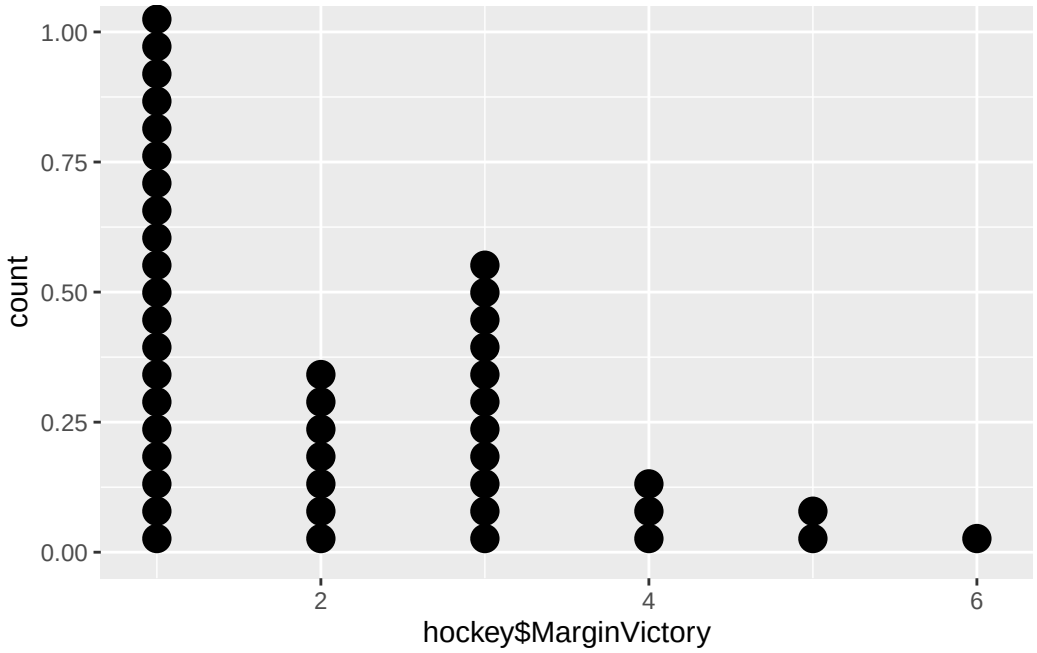
```
#Summarize Margin of Victory#
summary(hockey$MarginVictory)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.159   3.000   6.000
```

```
Margin.dot<-ggplot(hockey, aes(hockey$MarginVictory)) + geom_dotplot()
Margin.dot
```

```
Warning: Use of `hockey$MarginVictory` is discouraged.
i Use `MarginVictory` instead.
```

```
Bin width defaults to 1/30 of the range of the data. Pick better value with
`binwidth`.
```




```
sMV<-sd(hockey$MarginVictory)
sMV
```

```
[1] 1.328457
```

Mathematical model hypothesis tests for a single mean

Just as with confidence intervals, the form of the test statistic doesn't (typically) change as we move from data type to data type:

$$\text{test statistic} = \frac{\text{observed value - null value}}{SE}$$

So now,

If the null hypothesis is true and the conditions are met, then our test statistic follows a $t-$distribution with $df = n - 1$. The conditions are the same: independent observations and a large enough sample size with no extreme outliers. We can use the R function `pt(T,df=)` to get p-values.

Example: The Lincoln Marathon is the 51st largest marathon in the US, and is a qualifier for the Boston Marathon. From 2003 to 2019, the average finish time was 253.25 minutes (4 hours and 13 minutes, 15 seconds). The race was not run in 2020. We want to see if the break changed the average finish time. We took a random sample of 50 finishers from the 2021 race. For this random sample of 50, $\bar{x} = 261.38$ minutes and $s = 51.87$ minutes.

Example: Consider a manufacturing process for hypodermic needles used for blood donation. The needles need to have a diameter of 1.65 mm. If the needles are too big, they hurt the donor. Too small, and they'll rupture the red blood cells, making the donated blood useless. During every shift, quality control staff take a random sample of several needles and measure their diameter. If there's a problem, they shut down the manufacturing process to correct it. Suppose the most recent sample of 35 needles had an average diameter of 1.64 mm and a standard

deviation of 0.07 mm. Suppose the diameters of needles have a bell-shaped distribution. Based on these data, should the process be shut down?

Example: The General Social Survey (GSS) is a survey of a representative sample of U.S. adults who are not institutionalized. A 2018 General Social Survey asked a random sample of 1,118 adults how often they contacted their closest friend by either phone, internet, other communication device, or face-to-face. Of the 1,118 responses, the average number of times per week the respondents contacted their closest friend was 2.87, with a standard deviation of 2.46. The sample data are not strongly skewed. We want to estimate the mean number of closest friend contacts per week.

Aside: How do we know from the sample mean and standard deviation that the distribution of contact times cannot be bell-shaped? Why is it still okay to use the mathematical model?

We can also do these in R using `t.test`, but we'll need the full data set, not just the summary statistics.

For the NHL data,

```
t.test(hockey$Goals, mu=5, alternative="greater")
```

```
    One Sample t-test

data:  hockey$Goals
t = 4.2743, df = 43, p-value = 5.221e-05
alternative hypothesis: true mean is greater than 5
95 percent confidence interval:
 5.675649      Inf
sample estimates:
mean of x
 6.113636
```

Inference for a Two Independent Means (Chapter 20)

Now, we'll extend the methods for a single mean to differences in population means that come from two groups. So, we'll now focus on constructing hypothesis tests about and estimating the function of parameters $\mu_1 - \mu_2$, where μ_1 is the mean of Group 1 an μ_2 is the mean of Group 2. A reasonable point estimate is $\bar{x}_1 - \bar{x}_2$, the difference in sample means.

As we did with two proportions, we'll look at analysis three different ways: randomization test; bootstrap to find an interval estimate; mathematical framework for tests and confidence intervals (assuming the conditions are met to use a normal approximation. One note: one of the conditions for these techniques (no matter which) is the groups are independent. What happens in Group 1 has no bearing on Group 2. If there is any dependence among the groups (twin studies, before-and-after studies, for example) these are not appropriate. This was not really a concern with proportions, but can occur quite naturally with means. We'll consider dependence between the groups in a future section.

Randomization test for the difference in means

When we were working with proportions, we carried out a randomization test using two colors of cards. One color represented success, and the other color represented failure. We shuffled the cards, and dealt them into two stacks, representing our two groups. We then found the proportion of successes in each stack, and took the difference in proportions. We then did this

shuffling/dealing many times. We'll see through an example how our process changes when dealing with quantitative data.

Example: The research question we are interested in investigating is whether playing violent video games lead people to more or less aggressive behavior. Hollingdale and Greitemeyer (2014) approached the question in this way. They randomly assigned 49 students from a UK university to play Call of Duty: Modern Warfare (violent) and 52 students to play LittleBigPlanet (not violent). After 30 minutes playing the video games, the subjects were asked to complete a marketing survey investigating a new hot chili sauce recipe. They were told to prepare some chili sauce for a taste tester and that the taste tester "couldn't stand hot chili sauce but were taking part due to good payment.'' They were then presented with that appeared to be a very hot chili sauce and asked to spoon what they thought would be an appropriate amount into a bowl for a new recipe. The amount of chili sauce was weighed in grams after the participant left the experiment. The amount of sauce was used as a measure of aggression: the more chili sauce, the greater the subject's aggression.

- Is this an experiment or an observational study? How do you know?

- How do we know this is involves quantitative data?

- Parameters:

- Hypotheses:

The resulting data are:

Group	n	Mean	SD	Min	Max
Violent	49	16.12	15.30	1	63
Nonviolent	52	9.06	7.65	0	38

So our observed statistic is:

Our goal is the same as it was with proportions: to determine whether the observed difference in sample means is likely to have occurred by chance if the null hypothesis is really true.

Just like we shuffled cards in the two-proportions case, we're going to have cards again. Like before, the shuffling implements the null hypothesis model–there is no effect of the violent video game. The amount of chili sauce selected doesn't depend on whether or not the participant just played a violent game. We'd sometimes expect participants to use slightly more chili sauce if they'd just played a violent game ($\bar{x}_{\text{violent}} > \bar{x}_{\text{nonviolent}}$ and sometimes expect participants to use slightly less chili sauce if they'd just played a violent video game ($\bar{x}_{\text{violent}} < \bar{x}_{\text{nonviolent}}$ just due to natural variability.

Before, we looked at red and black cards, shuffled, and dealt into two stacks representing our two groups. Now, color isn't enough. Instead, we'll still have $n_{total} = 49 + 52$ total cards, but we'll write on the cards the observed amount of chili sauce. Then shuffle, and deal into stacks. One stack will get 49 cards (representing the 49 violent players) and the other will get 52 cards (representing the 52 nonviolent players). We'll find the difference in sample means after the shuffling/dealing. We'll repeat this process many times, and look to see how unusual our observed $\bar{x}_{\text{violent}} - \bar{x}_{\text{nonviolent}} = 7.065$ is.

- In the applet

- In R

Bootstrap confidence interval for the difference in means

When we used bootstrapping to find confidence intervals for the difference in two proportions, we took a bootstrap sample separately from each group and calculated the difference in the resulting proportions. We're going to do the same thing here–take a bootstrap sample from each group, find the two sample bootstrap means, and then find the difference in the bootstrap means. Doing this over and over again will allow us to explore the variability/sampling distribution of the difference in sample means.

Example: A random sample of college baseball players and a random sample of (male) college soccer players were obtained independently and weighed. The table below shows the weights (in pounds) (also a .csv file in Canvas).

Baseball	Soccer	Baseball	Soccer
190	165	186	156
200	190	210	168
187	185	198	173
182	187	180	158
192	183	182	150
205	189	193	172
185	170	200	180
177	182	195	184

We are interested in estimating the differences in mean weight between baseball players and soccer players.

Practice: (from book, page 382-383) Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? The data are in the library `openintro`. Try to find a 95% confidence interval for $\mu_{ESC} - \mu_{Control}$ using the bootstrap percentile confidence interval and the bootstrap SE confidence interval.

Mathematical model for the difference in means

Just like with mathematical model methods for single means, we need to check conditions to determine whether we can the t-distribution to construct tests and form confidence intervals for the difference in means.

- Independence–both between and within groups
- Check normality of each group separately (basically checking for extreme outliers)
- If these are both met, then the standard error of $\bar{x}_1 - \bar{x}_2$ is $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ with $df =$ really complicated (you'll see we get non-integers in R–it's doing the complicated calculation). We'll use $\min(n_1 - 1, n_2 - 1)$ if we're not using R. We won't know σ_1^2 and σ_2^2, so we'll approximate the standard error using $SE \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

As with tests for a single mean (and one proportion, and two proportions), our test statistic will have the usual form:

$$\text{test statistic} = \frac{\text{observed value - hypothesized value}}{SE}$$

In the case of two means, this is

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When the null hypothesis is true and the conditions are met, T has a t-distribution with $df = \min(n_1 - 1, n_2 - 1)$.

Confidence intervals will also have the same form:

$$\text{observed statistic} \pm \text{multiplier} \times SE$$

For this specific situation of comparing two independent means, this is

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and we'll again use $df = \min(n_1 - 1, n_2 - 1)$ (or let R calculate it for us).

With two proportions, our SE depending on whether we were doing a hypothesis test or calculating a confidence interval. Here, it doesn't. Any guesses why?

Example: The data set SleepStudy contains data on 253 students who did skills tests to measure cognitive function, completed a survey about attitude and habits, and kept a sleep diary. The data were reported in Onyper, et al. (2012). There are lots of different potential variables to consider. The data set includes:

Gender	1=male, 0=female
ClassYear	Year in school, 1=first year, …, 4=senior
LarkOwl	Early riser or night owl? Lark, Neither, or Owl
NumEarlyClass	Number of classes per week before 9 am
EarlyClass	Indicator for any early classes
GPA	Grade point average (0-4 scale)
ClassesMissed	Number of classes missed in a semester
CognitionZScore	Z-score on a test of cognitive skills
PoorSleepQuality	Measure of sleep quality (higher values are poorer sleep)
DepressionScore	Measure of degree of depression
AnxietyScore	Measure of amount of anxiety
StressScore	Measure of amount of stress
DepressionStatus	Coded depression score: normal, moderate, or severe
AnxietyStatus	Coded anxiety score: normal, moderate, or severe
Stress	Coded stress score: normal or high
DASScore	Combined score for depression, anxiety, and stress
Happiness	Measure of degree of happiness
AlcoholUse	Self-reported: Abstain, light, moderate, or heavy
Drinks	Number of alcoholic drinks per week
WeekdayBed	Average weekday bedtime (24.0 = midnight)
WeekdayRise	Average weekday rise time (8.0 = 8 am)
WeekdaySleep	Average hours of sleep on weekdays
WeekendBed	Average weekend bedtime (24.0 = midnight)
WeekendRise	Average weekend rise time (8.0 = 8 am)
WeekendSleep	Average hours of sleep on weekends
AverageSleep	Average hours of sleep for all days
AllNighter	Had an all-nighter this semester? 1=yes, 0=no

Let's consider the variables GPA and stress (coded normal or high). We'd like to know if there is convincing evidence that those with high stress levels have a different mean GPA than those with normal stress levels.

- Hypotheses:

- Check conditions:
  - Independent observations?
  - Large enough sample sizes?

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} =$$

In R, we can carry out this test very easily: `t.test(GPA ~ Stress, data=sleep)`

We can also use the `subset` option to restrict our attention to a specific value of another variable. For example, suppose we want to determine whether average hours of weekday sleep (WeekdaySleep) differs between those who have at least one early class and those who do not (EarlyClass; 0=no; 1=yes). However, we want to restrict our attention only to those who consider themselves Night Owls (LarkOwl=Owl). We could do this with:

```
t.test(WeekdaySleep ~ EarlyClass, data=sleep, subset=LarkOwl=="Owl")
```

For confidence intervals, let's go back to the baseball and soccer players. The summary statistics are:

Group	n	Mean	SD
Baseball	16	191.375	9.465
Soccer	16	174.5	12.495

We can also get the confidence interval in R: `t.test(weight ~ sport, data=athletes2)`

By default, R calculates 95% confidence intervals. We can change this with the `conf.level=` statement

For practice, find a research question you can answer using variables in the sleep study data.

Inference for Comparing Paired Means (Chapter 21)

Everything we've done so far has assumed independence among observations. If we only had one group, it was just independence among observations. If we had two or more groups, it was independence between and within groups. Now, we'll turn our attention to a common situation: dependence between groups. Specifically, a particular dependency–pairing. This occurs in before/after studies, other studies in which subjects are matched. For example, considering the price of a item purchased from two different retailers.

In these situations, we generally take the difference between the two values, and consider the difference as our observation. So, for example, if we want to compare cost of textbooks between the campus bookstore and Amazon, we'd randomly select a set of book titles, and find their price at both the bookstore and Amazon. We'd find the difference in price, and use those differences as our observations.

Note that we're distinguishing between **difference in means** (Chapter 20) and **mean difference** (Chapter 21).

- Parameters:
- Observed Statistics:

Good news: we've already seen how to construct mathematical model tests and confidence intervals here! We just use the same techniques we used for a single mean (Chapter 19), but on the differences.

However, randomization tests didn't really work with one mean in Chapter 19 because there was nothing to randomize. They will work here!

Randomization test for mean difference (matched pairs)

Example: Suppose you are playing baseball and hit a hard line drive. You want to turn a single into a double. Does the path you take to round first base make a difference? A masters thesis way back in 1970 considered the difference between a `narrow angle'' and` awide angle'' around first base. Suppose we have 22 baseball players who have volunteered to participate. There are a couple ways we could design an experiment to see if there is a difference.

- Randomly assign 11 players to run a wide angle and 11 players to run a narrow angle. Problems: some players may be faster than others. Ideally, randomization will equally distribute the speedy runners between the two groups, but there is no guarantee. Speed could be a confounding variable.
- Have each of the 22 runners run both angles, with the angle run first randomized using a coin. This allows each player to serve as their own control.

The second option is what the thesis writer did–he randomly determined the angle the player would take first. He then used a stopwatch the time the run from going from a spot 35 feet past home to a spot 15 feet before 2nd base. After a rest period, the runner then ran the second angle. This controls for runner-to-runner variability. It's important to randomize the order of the treatments, where possible! (This isn't possible in before-and-after type studies.)

Parameter of interest:

Hypotheses of interest:

Observed statistic:

Like before, we're trying to determine if it's surprising to see such a large difference as $\bar{x}_d = 0.075$ just by chance, if running strategy has no effect on running time.

Here's how the randomization test works: if running strategy really doesn't make a difference, then the two times for each runner were going to be the same two times regardless of which strategy was used. Any difference was just by chance, perhaps which one they ran first. That is, it really doesn't matter which value we call wide angle time and which value we call narrow angle time–the two times are completely interchangeable or swappable. This idea of swapping is how we'll do the randomization.

In the two sample randomization test, the explanatory variable was randomly assigned to the response. We shuffled all the cards, and randomly dealt them into the two stacks. Here, randomization occurs within an observational unit (in our example, a baseball player). So, the two times will stay assigned to the same player, but we'll randomly decide which time is narrow and which is wide using a coin flip. If the coin comes up, we swap the times. If the coin comes up tails, we don't swap.

We're going to do these in the applet, because I think it's easiest to see the swapping. Let's try it:

- Go to applet, do one randomization. In our randomization, how many players had their times swapped? The mean difference from this first randomization is shown in the applet. Like other randomization tests, we'll need to do this over and over. We've built up an estimate of the sampling distribution for the mean difference.

- Now, just like before, we'll see how unusual our observed $\bar{x}_d = 0.075$ is. Remember this is a two-sided test. None of our randomizations resulted in a mean difference more extreme than $\bar{x}_d = 0.075$. So, our p-value is approximately 0, and it looks like we do have evidence that base-running strategy has an impact on running time. We can reject the null hypothesis, and conclude that there is a difference in the strategies.

Bootstrap confidence intervals for mean difference (matched pairs)

The bootstrap approach to finding a confidence interval for μ_d is almost identical to the method for finding a bootstrap confidence interval for a single mean. The difference is in the interpretation.

- Take bootstrap samples from the observed differences
- Go to R code–all that's really changed is sampling from the differences
- Bootstrap percentile confidence interval:

- Bootstrap SE confidence interval:

What would happen if we (incorrectly) ignored the pairing? Let's find a 95% confidence interval, assuming the two samples are independent.

The hardest part is determining whether we are dealing with independent samples or matched pairs. Let's talk through 21.2, 21.3, 21.4, and 21.5

Mathematical model approach for mean difference (matched pairs)

The mathematical model approach to matched pairs is the same as the one sample analysis, but carried out on differences. The changes come in the form of the hypotheses and interpretation of the confidence interval.

We still need to check conditions!

- Independence: among observations (we know the observations within an observation are not independent)
- Large enough sample size: no extreme outliers or strong skew

Example: A study carried out by Cai et al. (2019) aimed to determine whether laugh tracks make dad jokes seem funnier. The researchers had a professional comedian record 40 dad jokes. They had people listen to the jokes and rate how funny they were on a 7 point scale, with 1 being not funny at all and 7 being extremely funny. Other people listened to the same 40 jokes, but this time the researchers added a laugh track to the recording. The volunteers were randomized to either no laugh track or laugh track.

- Why is this a paired scenario?

- What is the parameter?

- What are the hypotheses?

Here are the observed statistics:

Laugh track?	n	Sample mean	Sample SD
With laugh track	40	3.010	0.490
No laugh track	40	2.715	0.507
Difference = with - without	40	$\bar{x}_d = 0.295$	$s_d = 0.427$

Hypothesis test:

90% confidence interval:

We can also do this in R, adding `paired=TRUE` to the `t.test` code. Let's try it with the base running data.

Inference for Comparing Many Means (Chapter 22)

We're going to start this section by considering an example. The data are in the file on Canvas called 'mice.csv'.

Example: These data come from an experiment to determine if exercise confers some resilience to stress. Mice were randomly assigned to either an enriched environment (exercise wheel) or standard environment, and spent three weeks there. After that time, they were exposed for five minutes per day for two weeks to a "mouse bully' '–a mouse very strong, aggressive, and territorial. After those two weeks, anxiety in the mice was measured, as amount of time hiding in dark compartment. Mice that are more anxious spend more time in darkness. We want to determine if there is a difference in time spent in darkness for the two groups of mice.

- We know how to answer this question using a t-test.

- Plot the data, and add sample means to the plot
- We're testing $\mathrm{H}_0 : \mu_1 = \mu_2$, and we assume this is true to construct the test. Let's assume it is true, and the common mean is μ. How could we estimate μ?

- It looks like there is a difference between the two groups, and the test statistic is $T = -9.65$.
- This difference between the two groups will also manifest itself in the variances. There will be variation between the group means and the overall mean, as well as variation between the data points and their group means.
- Remember how variance is calculated:

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

  We're exploring how far, on average, observations are from the mean (squared). So, variance has to be positive.

- If there is a difference between the group means, the first kind of variation (between the group means and the overall mean) will be much greater than the second kind of variance (between the data points and their group mean). We can test whether the first variance is bigger than the second using an F statistic.

$$F = \frac{\text{variance between group means and overall mean}}{\text{variance between the data points and their group mean}}$$

- If the variances are about equal, there's no evidence of a difference between the group means–they vary as much from the overall mean as data points vary from their group mean. This will result in an F statistic of about 1. If there is a difference between the group means, the first kind of variation (between the group means and the overall mean) will be much greater than the second kind of variance (between the data points and their group mean). This will result in an F statistic greater than 1.
- For the mice data:

- Is $F = 93.1$ enough bigger than 1 to convince us there's a difference between the groups?
- **Notice:**

We made some assumptions to carry out the t-test:

- approximate normality (no extreme outliers, no strong skew)
- independence
- constant variance (not mentioned at the time)

We can summarize these assumptions very succinctly, and to do so we're going to introduce some new notation.

Consider a random sample of observations from a normal distribution with mean μ and variance σ^2. If we let $Y_1, Y_2, ..., Y_n$ represent our data points we can summarize this as:

Or another way:

If we have two samples:

If we have more than two samples:

Let's start with some summary statistics

$$
\begin{aligned}
Y_{i\cdot} &= \sum_{j=1}^{n_i} Y_{ij} = i^{th} \text{ sample total} \\
\bar{Y}_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = i^{th} \text{ sample mean} \\
Y_{\cdot\cdot} &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij} = \text{ grand total} \\
\bar{Y}_{\cdot\cdot} &= \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij} = \text{ grand mean } (N = \sum_{i=1}^{n_i} n_i)
\end{aligned}
$$

Remember how to calculate the sample variance, $S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$. We're going to look at three difference variances. Let's assume for simplicity that $n_i = n$ (all groups have equal sample size, this is not really necessary, it's just to make it easier to look at notation):

1. Total Variance. Another name for the numerator is total sum of squares.

2. Error (Within-Group) Variance. Another name for the numerator is the error sum of squares.

   This is a combined estimate of the common σ^2, essentially an average of the S^2 values from each group.

3. Model (Between-Group) Variance. Another name for the numerator is the treatment (model) sum of squares.

   To see what this is measuring, first consider the 'inside' sum:

   This is still an estimate of variance, but it's an estimate of σ^2/n, because these are means. In order to be able to compare fairly to the error variance we must multiply by n (only works with equal sample sizes) or, equivalently, take the sum from $i = 1$ to n:

We can't lose sight of what we're interested in here: testing $H_0 : \mu_1 = \mu_2$. If H_0 is true, $\bar{y}_1$ and $\bar{y}_2$ should not be different from $\bar{y}_{..}$. This means that error variance should be about equal to model variance (both would estimate σ^2). If H_0 is not true, model variance will be larger because of the deviations of the group averages from the grand average. If it's much larger, this gives us evidence against H_0.

Why do we worry about three variances when we only use two (error and model) to get the F stat? It turn out that:

$$\text{Total SS} = \text{Model SS} + \text{Error SS}$$

For the mice data: {

$$\begin{aligned}
\text{Total SS} &= (259-328.07)^2 + \cdots + (231-328.07)^2 + (394-328.07)^2 + \cdots + (454-328.07)^2 = 193459 \\
\text{Error SS} &= (259-217.43)^2 + \cdots + (231-217.43)^2 + (394-438.71)^2 + \cdots + (454-438.71)^2 = 22073 \\
\text{Model SS} &= 6(217.43-328.07)^2 + 6(438.71-328.07)^2 = 171386
\end{aligned}$$

}

To convert these sums of squares into variances (which we call mean squares), they must be divided by denominators noted above. These are degrees of freedom, and have the same relationship as the sums of squares do:

$$\text{Total } df = \text{ Model } df + \text{ Error } df$$

In our mice example, we have

Total df	=	Model df	+	Error df
	=		+	
	=		+	
	=		+	

We often summarize our calculations in a table (assuming equal sample sizes):

Source	df	SS	MS
Model	$t-1$	SSModel	MSModel
Error	$t(n-1)$	SSError	MSError
Total	$nt-1$	SSTotal	

The MSError (usually called MSE) is our estimate of σ^2. In our mice example, we get the table:

Source	df	SS	MS
Model	1	171386	171386
Error	12	22073	1839
Total	13	193459	

To test $H_0 : \mu_1 = \mu_2$ we use the F stat:

$$F = \frac{\text{MSModel}}{\text{MSError}} = \frac{171386}{1839} = 93.2$$

and we add this to the table:

Source	df	SS	MS	F
Model	1	171386	171386	93.2
Error	12	22073	1839	
Total	13	193459		

What we've just done is called an **Analysis of Variance (ANOVA)**, and the resulting table is called an ANOVA table. It's a single hypothesis test to check whether the means across many groups are equal. Specifically, it's testing:

- H_0 :
- H_a :

We still have assumptions:

- Independence
- Responses are approximately normal
- Variability across groups is about equal

We still don't know if 93.2 is enough greater than 1 to determine there's a difference! We have two options:

- Mathematical model: F Test

  Assuming H_0 is true and the assumptions are met, F follows and F-distribution with $df_1 = t - 1$ and $df_2 = N - t$ (N is the total number of observations). We can use `1 - pf()` in R to find p-values

  In R, `lm()` and `anova(lm())`.

- Randomization test: Like for two means, write all responses on cards. Shuffle, and deal into as many stacks as there are groups with stack size corresponding to group size. Find F for the shuffle. Repeat many times, and see how unusual our observed F statistic is. We can do this in the applet or in R.

Example: The data set GPA gives a random sample of student GPAs, along with where they chose to sit in a classroom. We want to see if mean GPA differs based on where a student sits. Let's build up the ANOVA table.Then run in R.

Example: Baseball run time. The data gives run time in seconds for 50 yards for players at three different positions (OF, IF, C). Let's build up the ANOVA table. Then run in R.

Example: A group of college students wanted to see whether there was an association between students' major and the time (in seconds) to complete a small paper-and-pencil puzzle. They grouped majors into four categories: applied science (as), natural science (ns), social science (ss), and arts/humanities (ah). Let's build up the ANOVA table. Then run in R.