Stat 102 Notes

Table of contents

Course Goals for STAT 102

STAT 102 is an introduction to formal statistical inference. We will carry out inference using both simulation-based approaches and classical, theory-based methods. By the end of the course, you will:

- Read an example where the research question is explicitly stated, and then translate what's stated into a statistical statement involving parameters or other simple distributional characteristics.

- Identify whether the ideal data collection strategy would involve random assignment, random sampling, or both and explain why.

- Work with an example where the research question is explicitly stated, along with an existing data set, and propose and carry out an appropriate analysis to answer the research question.

- Explain the terms/components of a given statistical model, and connect those terms to the research question at hand.

- Check basic assumptions of various (simple) analysis methods and justify the use of the method.

- Apply existing functions and point-and-click software for implementing basic data analyses.

- Use tactile simulation to carry out a simple resampling procedure.

- Identify the steps and perform the calculations required for routine statistical procedures to address a given problem.

- Calculate simple analyses (t-test, chi-squared test for proportions) by hand, to verify the validity of the computational algorithm.

- Recognize when computational results do not make sense in the context of the problem.

1 Philosophy of Statistical Inference (Chapters 11-14)

In STAT 101, you focused on Exploratory Data Analysis. Exploratory data analysis aims to investigate the characteristics of a data set through visualizations and numerical summaries. Visualizations may include:

- box plots
- histograms
- bar charts
- pie charts
- scatterplots
- heat maps
-
-
-

Numerical summaries used to explore a data set may include:

- sample mean
- sample variance/standard deviation
- five number summary, and other order statistics
- sample proportions
- calculated regression slope and intercept
-
-
-

More often than not, the data were collected to answer a research question about a larger population for which the data collected are a (hopefully) representative sample. This notion of drawing conclusions beyond the data collected is at the heart of statistical inference.

Example: Bred in the Bone

- If each baby is really guessing/choosing blindly, what proportion would you expect to choose the good guy? Why?

- Based on this, what randomizing device could we use to model this experiment?

- Experiment! Add your results to the plot on the board.

- Applet

- What do you observe in the plot?

- Real experiment

Take away:

In exploratory data analysis, the visualizations and numerical summaries you choose are driven by the type of data at hand. This is true for statistical inference as well. The type of data will drive the appropriate inference techniques. However, the goal of the research study will also impact the selected method, as will the underlying assumptions of the technique (we'll talk **a lot** more about this). That said, there are some overarching approaches to quantifying variability, and thus drawing conclusions beyond the data set at hand.

Approaches to quantifying variability

- Randomization methods (Chapter 11)
- Bootstrap methods (Chapter 12)
- Mathematical models (Chapter 13)

We'll start the semester by talking about these three approaches fairly generally. For (most of) the rest of the semester, we'll see how these approaches fit with different types of data.

1.1 Randomization Tests (Chapter 11)

The goal of hypothesis tests is to use an **observed** data set to answer a yes/no question about a characteristic of a larger population from which the observed data set was drawn. For example, is swimming with dolphins therapeutic for patients with clinical depression? That is, we want to assess whether or not the explanatory variable causes changes in the response variable.

To answer this question, Antonioli and Reveley (2005) recruited 30 subjects with a clinical diagnosis of mild to moderate depression. The subjects were required to stop all other treatments (therapy and/or pharmaceuticals) 4 weeks prior the experiment, and the 30 subjects were all taken to an island off the coast of Honduras. The subjects were randomly assigned to one of two groups. Both groups spent one hour swimming and snorkeling each day, but one group did so in the presence of dolphins and the other group did not. At the end of two weeks, each subject's level of depressions was evaluated, and whether or not the subjects had a substantial improvement in their depression was recorded.

Explanatory variable:

Response variable:

Is this an observational study or an experiment? What does that imply about inference?

The question we will answer is whether the resulting data provide convincing evidence that subjects who swam with dolphins were more likely to see depression improvement than subjects who swam without dolphins.

If there really is no impact of swimming with dolphins, what does this imply about the explanatory and response variables?

If swimming with dolphins does improve depression, what does this imply about the explanatory and response variables?

This leads to two competing claims:

- **Null hypothesis**: H_0

- **Alternative hypothesis**: H_a

If the null hypothesis is true, how would this manifest in the observed data?

If the alternative hypothesis is true, how would this manifest in the observed data?

We will choose between the competing claims by assessing whether the data conflict so much with H_0 that the null hypothesis cannot be considered reasonable. If this happens, we'll reject the notion of H_0 and conclude that H_a must be true.

Up to now, we haven't seen the data! Here's a summary:

	Dolphin Therapy	Control Group	Total
Showed Improvement			
No Improvement			
Total			

We can see that

-

-

So,

The question remains...is this enough different from what we would expect under the null hypothesis to conclude that swimming with dolphins does make a difference in depression?

So far, nothing we've laid out is unique to a randomization test. Where does randomization come in?

Let's visualize these observations as a set of cards. Each card denotes a subject in the study. The color indicates the response: red for substantial improvement and black for no substantial improvement.

Any difference we see in the simulation is due to chance–the cards were randomly dealt into the dolphin/control groups.

It's not realistic to keep shuffling and dealing by hand...we need to turn to technology to do the randomization for us: Applet
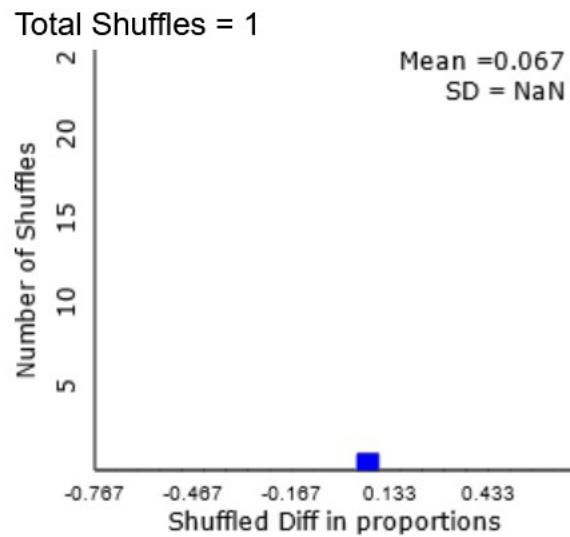
Figure 1.1: One Shuffle

We can do this over and over again to build up a **null distribution**. This distribution shows how we expect the variability to behave under the null hypothesis:
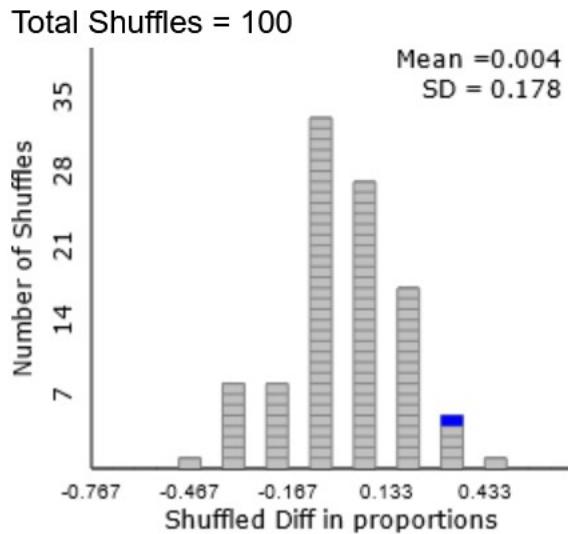
Figure 1.2: 100 Shuffles

What do you notice about this null distribution?

How rare is it to see our observed statistic **0.467** in this distribution? What does this imply?

So, we've just carried out a **statistical inference** technique! We might be wrong in our conclusions (more on this in Chapter 14), but we've made the best decision we could with the data available.

In summary:

Randomization Test Procedure:

- Frame the research question in terms of hypotheses
- Collect data from an observational study or experiment
- Model randomness that would occur if H_0 is true
- Analyze the data by comparing the observed data to the simulated distribution
- Conclusion

Now let's go to R!

1.2 Bootstrap Methods (Chapter 12)

Bootstrap methods are a relatively new statistical technique (proposed in 1979 by Efron), but they are based on a very simple idea. The goal is to characterize the variability of the statistic across many samples. One way we could do this is take lots and lots of samples from the population, and get a picture of how much variance there is among the samples. This is almost always impossible. So, rather than resample from the population, we could try resampling from the sample. This is the basic idea behind the bootstrap.

Bootstrapping is used in many different applications. For this general introduction to the approach, we're going to consider a confidence interval for a proportion.

A **confidence interval** is

Note the goal of the confidence interval is different from the goal of a hypothesis test!

However, like with hypothesis tests, we need to understand the variability inherent to the statistic. To figure out how wide the range of plausible values should be, we need to know how a statistic varies from sample to sample in the population.

For example, let's think back to the Baby scenario and suppose our goal is to estimate the population parameter

The researchers collected one sample of 16 babies, and found that 14 picked the good guy. This is our observed data. What do you think would happen if we took a sample of 16 different babies? And then a different sample of 16 babies?

Idea of the bootstrap:

Infinite populations are pretty tough to work with, though. However, we can produce an equivalent bootstrap distribution by

So, we'll repeatedly draw bootstrap samples of size 16 (why 16?) and calculate the proportion of successes in each bootstrap samples. After we do this many many times, we'll have an idea of a range of plausible values for the population parameter. We'll set the **confidence level** by opting for a wider or a narrower interval, based on how certain we need to be in the results.

Bootstrap Process

- Frame the research question in terms of a parameter to estimate
- Collect data using an observational study or an experiment
- Model the randomness by using the observed data as a proxy for the population
- Create the interval (in future chapters we'll see there are multiple ways to do this)
- Conclusion

Let's go to R!

1.3 Inference with Mathematical Models (Chapter 13)

So far, we've seen computational methods like randomization and bootstrapping to characterize the variability of a statistic. The use of computational methods is relatively recent, due to the increase in computing power. In pre-computing days, re-sampling and randomization was very difficult. As a result, mathematical approximations were used and are still pervasive. If you took AP Statistics or a different intro statistics course, you employed mathematical models. However, to be clear, all of the methods we'll talk about (randomization, bootstrap, mathematical models) are techniques to get a **sampling distribution**.

The sampling distributions we've seen so far have been (mostly):

This isn't coincidence...it's guaranteed by a very important theorem, the **Central Limit Theorem**.

> 💡 Central Limit Theorem

What are the requirements here?

- Independence:

- "Large enough":

Normal Distribution: Nothing follows it exactly–it's a mathematical construct. But, a lot of things follow it approximately, either:

- naturally:

- created to follow it:

The normal distribution depends on two parameters, $\mu =$ mean (where the distribution is centered) and $\sigma =$ standard deviation (how spread out it is). μ shifts the distribution up and down the number line, σ stretches and contracts the curve. The **standard normal** distribution has $\mu = 0$ and $\sigma = 1$ (this is the distribution tabulated in normal tables in textbooks).

The standard normal gives us a convenient way to compare observations, and any normal distribution can be transformed into a standard normal. The **Z-score** is

If the Z-score is positive

If the Z-score is negative

Z-scores can be used to

- gauge the unusualness of an observation

- find probabilities

Helpful R functions:

- `pnorm(x, mean=0, sd=1)`
- `normTail(m=0,s=1,L=x)` or `normTail(m=0,s=1,U=x)` will draw pretty pictures–need to use the OpenIntro library
- `qnorm(prob, mean=0, sd=1)` gives a Z-score with area to the left

Pictures are super-helpful!

Example: Full-term birth weights for single babies are normally distributed with a mean of 7.5 pounds and a standard deviation of 1.1 pounds.

1. A baby is born weighing 9.1 pounds. What is the weight percentile for this baby?

2. Babies that weigh less than 5.5 pounds are considered low birth weight. What proportion of babies are low birth weight?

3. What weight would make a baby at the 25th percentile?

4. What is the probability a randomly selected baby weighs between 7 and 8 pounds?

The **Empirical Rule** (aka the 68-95-99.7 Rule) presents a general rule for the probability of falling within one, two, and three standard deviations of the mean in a normal distribution.

This rule is useful in a wide range of settings when trying to make quick estimate (we'll use it with bootstraps too!).

Some more definitions we'll use throughout the semester:

- **Standard error**:

- **Margin of error**:

Example (13.11): In 2013, the Pew Research Foundation reported that "45% of US adults report that they live with one ore more chronic conditions." However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used. Create a 95% confidence interval for the proportion of US adults who live with one or more chronic conditions. Interpret the confidence interval in the context of the study.

1.4 Decision Errors (Chapter 14)

Anytime we're using sample data to make decisions about a larger population we can potentially make a mistake. We can make an incorrect decision in a hypothesis test or calculate a confidence interval that does not capture the true population parameter. In a hypothesis test, there are four possible outcomes:

Type I error:

Type II error:

Examples:

- Doping in the Olympics

- Criminal trial

- Diagnostic test for a serious disease

Errors require a balancing act. We want to reduce the chance of making a Type I error but this will necessarily increase the chance of making a Type II error. The best we can do is to set the probability of a Type I error. We can do through setting the **significance level**.

Significance level:

Another consideration that will impact the chance of making an error is the whether the test is one- or two-sided.

Two-sided hypotheses:

Example: Standard anticoagulant therapy to prevent blood clots requires frequent (expensive) lab monitoring. A new procedure called riva was tested because it did not require frequent monitoring. A randomized trial was conducted in 2012, with standard therapy randomly assigned to 2416 patients and riva randomly assigned to 2416 patients. A bad result was a recurrence of a blood clot in a vein. We want to know if the likelihood of a bad result is different between the two therapies.

Here are the results of the randomized trial

	Riva	Standard	Total
Clot	44	60	104
No Clot	2372	2356	4728
Total	2416	2416	4832

For two-sided tests, the p-value is the probability that we observe a result as least as favorable to the alternative hypothesis as the result we observe. That is, that we observe a result as extreme or more extreme in either direction.

When in doubt, use a two-sided test! Use a one-sided test only if you truly have interest in only one direction.

So, how can we control Type I error?

- Set up tests before seeing the data.
- Collect enough data that the test has sufficient **power**. We'll talk more about power later (and LOTS more in an experimental design course), but power is the probability of correctly rejecting a false null hypothesis. It's a function of how big the true difference is (which we don't know and can't control) and the sample size (which we can control).

2 Inference for Proportions (Chapters 16-18)

So far, we've discussed randomization, bootstrap, and mathematical models as methods to approximate/describe a sampling distribution and quantify variability. Now, we turn to how these three methods can be used to answer research questions for different kinds of data. The appropriate method will depend on both the type of data and the research question of interest.

During our class, we'll discuss two types of data: **categorical** and **quantitative**. Categorical data arise when the responses are categories. If you think about what is being measured on each unit in the sample, and could imagine checking a box to record the response, the data are categorical. For example:

We also have to consider the research question. The research question of interest will drive answers to the following:

1. Is a single variable being measured on each unit in the sample, or are two (or more) variables being recorded?

2. If two or more variables are being recorded for each unit in the sample, can one be considered the **response** variable and the other(s) be considered **explanatory**?

3. If a variable is categorical, does it have two possible outcomes (like yes/no) or more than two possible outcomes?

4. Is the research question focused around finding the answer to a yes/no question (like "Does a new teaching method improve student test scores?") or around estimating a value (like "By how much do student test scores change if a new teaching method is introduced?")?

5. Were the data collected obtained using random sampling, random assignment, both, or neither? (this is not typically driven by the research question, but can impact our analysis method and will **definitely** impact the conclusions we can draw.)

The answers to these questions will help us determine which method is most appropriate, as well as the specific analysis tool to implement that method. In this unit, we're going to focus on categorical variables. We're going to start with a single categorical variable measured on each unit in the sample, where that categorical variable has only two possible outcomes. From there we'll move to two categorical variables measured on each unit (one explanatory, one response), again with only two possible outcomes for each. Finally, we'll explore categorical variables with more than two possible outcomes.

2.1 Inference for a Single Proportion (Chapter 16)

Our first scenario involves a single categorical variable measured on each unit in the sample, with only two possible outcomes.

2.1.1 Bootstrap Tests for One Proportion

When we discussed bootstrapping earlier, we were sampling (with replacement) from our sample data, because we wanted to understand the variability inherent to our statistic, $\hat{p}$, assuming our sample is representative of all samples of the same size that could be drawn from the population. The goal of hypothesis testing is different: we want to understand the sampling distribution of $\hat{p}$ under the assumption

So, we need to repeatedly sample from a population with $p = p_0$. We can do this by simulating data sets of the same size as original sample, assuming that $p = p_0$. This is called a **parametric bootstrap**, because we are making an assumption about the value of the underlying parameter p and we are assuming a particular distribution to generate our simulated data sets. From each simulated data set, we could calculate the resulting $\hat{p}_{\text{sim}}$. Many simulated data sets will give us a good approximation for the distribution of $\hat{p}$ under our assumptions.

Example: Back to the babies picking the good guy or bad guy. We want to know if babies are more likely to pick the good guy puppet.

Under H_0, 50% of babies will pick the good guy. We'll assume this is true for all babies that could be tested. We'll simulate 16 babies undergoing this test to get a sample proportion from the null distribution.

Let's see how this works in R. We'll start by setting up the original data, so we can calculate the test statistic.

```
data_baby<-c(rep(1,14),rep(0,2))

obs_prop<-mean(data_baby)
obs_prop
```

```
[1] 0.875
```

Next, we'll set up the null model with 50% successes and 50% failures:

```
para_boot<-c(rep(1,8),rep(0,8))

para_boot
```

```
 [1] 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
```

Now, we'll repeatedly sample from the null distribution, and collect the resulting $\hat{p}_{pb}$ from each sample.

```
numsim<-100
boot.sample<-data.frame(sim=1:numsim,stat=NA)

head(boot.sample)
```

```
  sim stat
1   1   NA
2   2   NA
3   3   NA
4   4   NA
5   5   NA
6   6   NA
```

```
for(i in 1:numsim){
  boot.sample$stat[i]<-mean(sample(para_boot,size=16,replace=TRUE))
}

head(boot.sample)
```
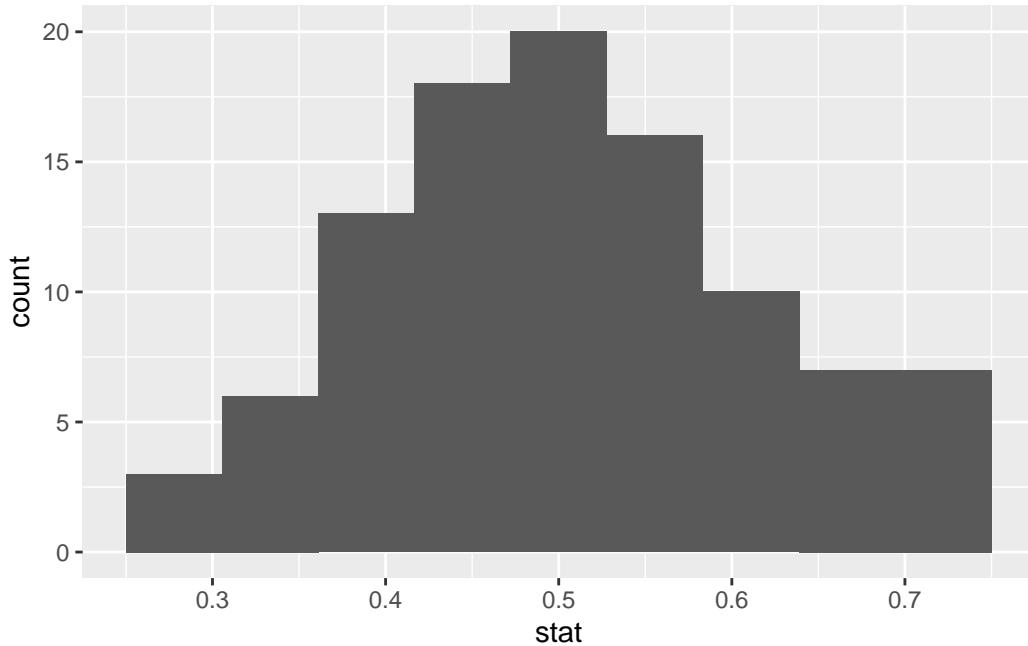
```
  sim   stat
1   1 0.3750
2   2 0.6250
3   3 0.6250
4   4 0.3750
5   5 0.5625
6   6 0.3750
```

We can plot these $\hat{p}_{pb}$, and see how unusual our observed test statistic is.

```
library(ggplot2)
boot.hist<-ggplot(boot.sample, aes(stat)) + geom_histogram(bins=10)

boot.hist
```

We've got a plot, but how do we know exactly how many/what proportion of $\hat{p}_{pb}$ were greater than our observed test statistic, $\hat{p} = 0.875$?

```
count<-(boot.sample$stat >= obs_prop)

sum(count)
```

```
[1] 0
```

```
sum(count)/numsim
```

```
[1] 0
```

So, we have an estimated p-value of

What if we want to change the number of simulated data sets? Let's try 1000. This gives an estimated p-value of

Why is this an estimated p-value?

Why does this histogram not look bell-shaped?

Try Problem 5 in Chapter 16, and see if you can modify the Babies R code to re-create the histogram provided in the book. It might help to try the applet first and see what has to change there.

Why Bootstrap?

- Works for any sample size!

- Intuitive way to explain what the p-value is actually measuring.

2.1.2 Bootstrap Confidence Intervals for One Proportion

We've already done these! Recall that with a confidence interval, we're not assuming the null hypothesis is true. Rather, our goal with the bootstrap is to characterize the variability of our statistic $\hat{p}$.

2.1.3 Mathematical Model for a Proportion

Sometimes, the sampling distribution of $\hat{p}$ can be well-approximated using a normal distribution. The conditions which must be met are:

-

-

If these conditions are met, then

[Note: This result is just another way of stating the Central Limit Theorem when dealing with proportions! The success/failure condition is playing the role of the "sample is large enough" requirement of the CLT.]

Let's think more about the standard error of $\hat{p}$. Remember the **standard error** is
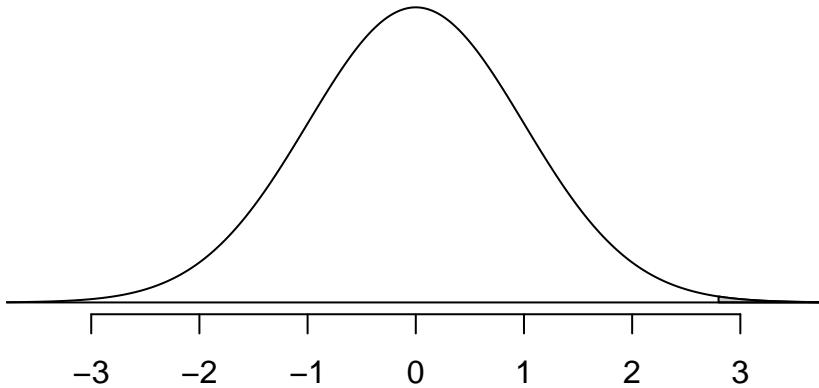
But this presents a problem. We don't know p (if we did, we wouldn't be doing tests or confidence intervals)!

How is this going to play out in hypothesis tests?

Example: Look at Problem 3 in Chapter 16. The journalist claims more than 1/5 adults living in Seattle support defunding the police. Is this true?

To find the p-value, we can use R.

```
normTail(m=0,s=1,U=2.79)
```



```
pvalue<-1-pnorm(2.79,mean=0,sd=1)
pvalue
```

```
[1] 0.002635402
```

Do you expect the p-value from the parametric bootstrap would be similar? Why or why not?

Let's see.

How is this going to play out in confidence intervals?

Example: Back to Chapter 16, problem 3. The journalist found that 159/650 Seattle residents support proposals to defund the police.

We can change the confidence level by changing z^*, and using the `qnorm` function.

- 90% confidence

```r
qnorm(0.05,mean=0,sd=1)
```

```
[1] -1.644854
```

Back to the bootstrap confidence interval...

So far, we've seen **bootstrap percentile confidence intervals**. We calculated these directly from the bootstrapped $\hat{p}_{\text{boot}}$. If we want a 90% confidence interval, we can find the 5th and 95th percentile values of the $\hat{p}_{\text{boot}}$ values.

We can also use the variability of the $\hat{p}_{\text{boot}}$ to calculate an estimate of the standard error of $\hat{p}$, and then calculate the interval using the mathematical model approach. This is a **bootstrap SE confidence interval**.

This is a rough approximation, using the 68-95-99.7 Rule which says that 95% of the observed differences should be no farther than 2 SE from the true parameter (p). To do this, the bootstrap histogram must be roughly symmetric and bell-shaped. So, it works for Chapter 16, problem 3; it doesn't work with the babies.

Why Z-Test/Z Confidence Intervals?

- In many cases, works as well as simulation methods
- Easy to calculate without technology/can do "back of the envelope" analyses
- Z-scores and "estimate $\pm$ margin of error" are easily interpretable
- Classical methods, used by scientists across disciplines

2.2 Inference for a Comparing Two Proportions (Chapter 17)

We now move on to consider situations in which two categorical variables are measured on each unit in the sample, and each variable has two possible values. In cases like these, typically one variable is considered the response and one variable is considered explanatory. The explanatory variable may be randomly assigned (like whether or not a subject swam with dolphins) or it may be merely observed (like smoking status). The two possible values of the explanatory variable lead to two groups, and we're interested in comparing the population proportions that arise from these two groups. We'll focus on the function of parameters $p_1 - p_2$. The natural estimate of this is $\hat{p}_1 - \hat{p}_2$: the difference in the sample proportions. We'll be constructing hypothesis tests to compare p_1 to p_2 and finding confidence intervals to estimate $p_1 - p_2$.

2.2.1 Randomization tests for the difference in proportions

Example: Researchers are interested whether electrical brain stimulation will help with problem solving tasks. 40 volunteers were all trained to solve problems in a particular way. Half of the volunteers were randomly assigned to receive electrical stimulation and the other half received a sham stimulation (placebo). All volunteers were then presented with an unfamiliar problem and asked to solve it. The researchers are interested in testing whether the proportion able to solve the problem following electrical stimulation is greater than the proportion able to solve the problem without electrical stimulation.

There are a couple of different ways we could state the hypotheses of interest:

- H_0 :

- H_a :

Recall that hypothesis tests work by assessing how unusual our observed data are, if the null hypothesis is really true. A very unusual result implies that observed data are not likely to have occurred under the null hypothesis. Randomization tests allow us to assess that unusualness by estimating the null distribution–a simulated distribution of what we could expect the distribution of $\hat{p}_1 - \hat{p}_2$ to look like if H_0 is true. We assume H_0 is true by recreating the randomization that occurred in the experiment.

Here are the data:

	Solved	Not Solved	Total
Sham			20
Electrical			20
Total			40

To demonstrate what the randomization test is doing, we need 40 cards. Why 40?

Of these cards, how many should be red and how many should be black? What do these represent?

We'll shuffle, and deal into two stacks.

A randomization test is going through this shuffling/dealing over and over again, find the difference in proportions for each simulation.

Let's look at this in the applets.

What do you notice about the null distribution? How unusual is the observed $\hat{p}_1 - \hat{p}_2$?

Example: Try Problem 2 in Chapter 17. Set up the hypotheses and describe how a randomization test would work.

- How many cards are needed?
- How many red? How many black?
- How many should be dealt into each stack?
- What would you calculate from each shuffle/deal?

Let's do this in R!

First, we'll need to set up the data. This is honestly the hardest part.

```
VM<-data.frame(Treatment="Vaccine",Response="Malaria",obs=1:89)
head(VM)
```

```
  Treatment Response obs
1   Vaccine  Malaria   1
2   Vaccine  Malaria   2
3   Vaccine  Malaria   3
4   Vaccine  Malaria   4
5   Vaccine  Malaria   5
6   Vaccine  Malaria   6
```

```
VN<-data.frame(Treatment="Vaccine",Response="NoMal",obs=1:203)
CM<-data.frame(Treatment="Control",Response="Malaria",obs=1:106)
CN<-data.frame(Treatment="Control",Response="NoMal",obs=1:41)
malariadf<-rbind(VM,VN,CM,CN)
head(malariadf)
```

```
  Treatment Response obs
1   Vaccine  Malaria   1
2   Vaccine  Malaria   2
3   Vaccine  Malaria   3
4   Vaccine  Malaria   4
5   Vaccine  Malaria   5
6   Vaccine  Malaria   6
```

Next, we'll calculate the observed difference in the proportion of children who contracted malaria between those who received the vaccine and those who received the control. To use the `diffmean` function, we need to load the `mosaic` package.

```
observed<-diffmean(Response == "Malaria" ~ Treatment, data=malariadf)
observed
```

```
  diffmean
-0.4162939
```

Now, we need to shuffle the vaccine/control treatment labels many times and calculate the difference in proportions of malaria for each shuffle. To do this, we'll need the `mosaic` library but it's already been loaded.

```
malaria.null<-do(1000)*diffmean(Response == "Malaria" ~ shuffle(Treatment),data=malariadf)
head(malaria.null)
```
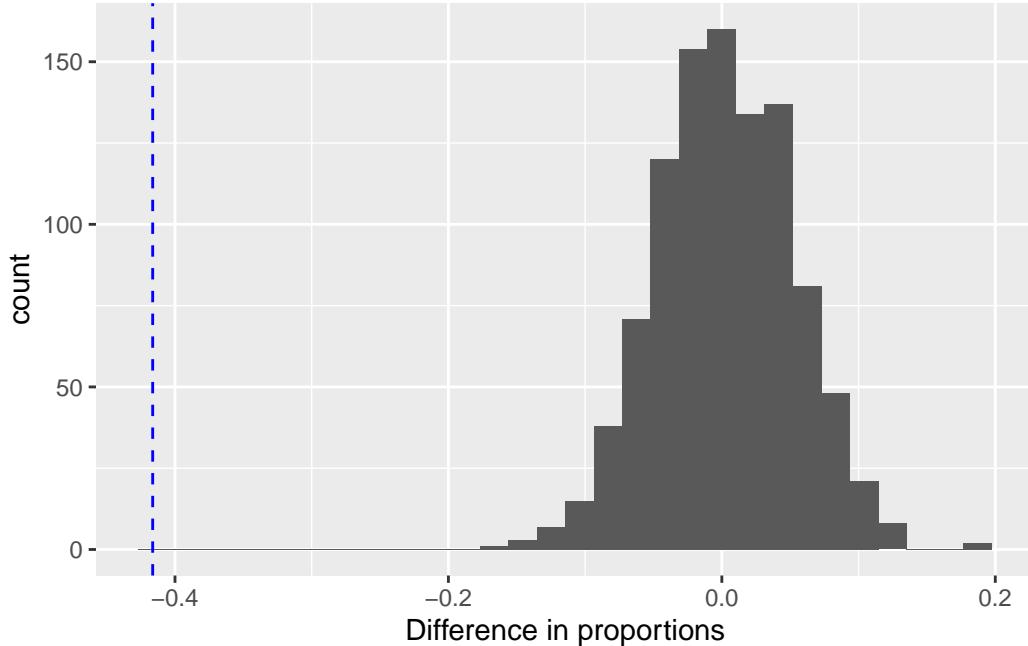
```
      diffmean
1 -0.07879042
2 -0.06856304
3 -0.07879042
4 -0.06856304
5 -0.04810828
6 -0.07879042
```

We'll plot the null distribution, and to do this we'll need the **ggplot2** library.

```
library(ggplot2)

ggplot(data=malaria.null) + geom_histogram(mapping=aes(x=diffmean)) +
  xlab("Difference in proportions") +
  geom_vline(xintercept = observed, linetype=2, color="blue")
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Finally, we can calculate the p-value by observing how unusual the observed difference in the proportions of malaria is under the null hypothesis.

```
prop(~diffmean <= observed, data=malaria.null)
```

```
prop_TRUE
        0
```

Why are we considering less than or equal to be more extreme in this example?

2.2.2 Bootstrap confidence interval for the difference in proportions

As we saw with a single proportion, bootstrapping will allow us to estimate the variability of $\hat{p}_1 - \hat{p}_2$ without assuming the null hypothesis is true. With a single proportion, we drew repeated samples (with replacement) from our sample data, and from each bootstrap sample calculated $\hat{p}_{\text{boot}}$. The distribution of the $\hat{p}_{\text{boot}}$ provided an estimation of the sampling distribution of $\hat{p}$.

Now, with two samples, our observed statistic of interest is

Let's go to R, and see how this works with the electrical stimulation example. We'll start by setting up the data and calculating the observed difference in proportion of solved problems between electrical stimulation and control.

```
#original Sample 1 data (Electrical), creating a data set with 10 S (1) and 10 F (0)#
electrical<-c(rep(1,10),rep(0,10))

#original Sample 2 data (Control), creating a data set with 6 S (1) and 14 F (0)#
```

```
control<-c(rep(1,6),rep(0,14))

#Calculate and print the observed statistic, p-hat_E - p-hat_C#
obs_stat<-mean(electrical)-mean(control)
obs_stat
```

[1] 0.2

Now, we'll repeatedly sample with replacement separately from each group. First we need to set up a place to store our summaries from each resample.

```
#Set up an empty data set with 4 columns: sim number, p_hat_boot_E, p_hat_boot_C, diff#

boot.samples<-data.frame(sim=1:1000,stat_E=NA,stat_C=NA, diff=NA)

head(boot.samples)
```

```
  sim stat_E stat_C diff
1   1     NA     NA   NA
2   2     NA     NA   NA
3   3     NA     NA   NA
4   4     NA     NA   NA
5   5     NA     NA   NA
6   6     NA     NA   NA
```

```
#For each row in the data set, draw a bootstrap sample from Sample 1 and Sample 2 and find#
# p_hat_boot_E and p_hat_boot_C #

for(i in 1:1000){
  boot.samples$stat_E[i]<-mean(sample(electrical,size=20,replace=TRUE))
  boot.samples$stat_C[i]<-mean(sample(control,size=20,replace=TRUE))
}

head(boot.samples)
```
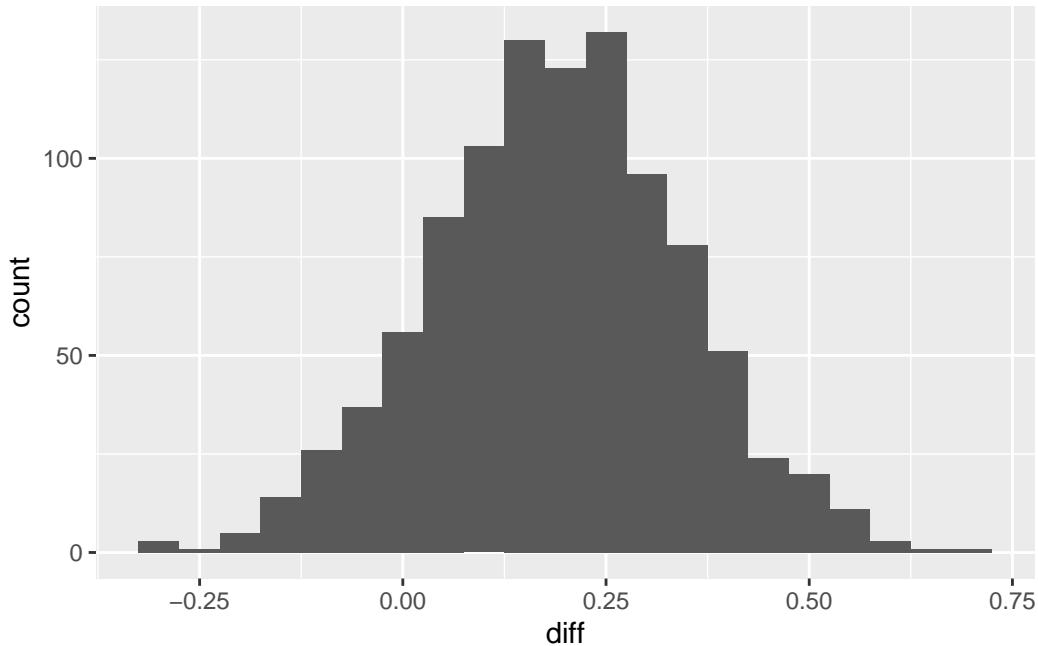
```
  sim stat_E stat_C diff
1   1   0.55   0.35   NA
2   2   0.55   0.30   NA
3   3   0.50   0.30   NA
4   4   0.30   0.15   NA
5   5   0.50   0.20   NA
6   6   0.50   0.30   NA
```

```
#Now find the differences#

boot.samples$diff<-boot.samples$stat_E - boot.samples$stat_C

head(boot.samples)
```

```
  sim stat_E stat_C diff
1   1   0.55   0.35 0.20
2   2   0.55   0.30 0.25
3   3   0.50   0.30 0.20
4   4   0.30   0.15 0.15
5   5   0.50   0.20 0.30
6   6   0.50   0.30 0.20
```

Now let's plot the bootstrap distribution.

```
boot.hist<-ggplot(boot.samples, aes(diff)) + geom_histogram(binwidth=0.05)

boot.hist
```


Notice where this distribution is centered!

We are re-using this example, since we've already carried out the randomization test in the applet. By doing so we are ignoring (maybe) the research question. But, we're doing both a randomization test and confidence interval so that we can compare the resulting sampling distributions of $\hat{p}_1 - \hat{p}_2$. What is different? What's the same?

Now that we have the bootstrap distribution, we can find the bootstrap percentile confidence interval. Let's do a 90% interval.

```
#start by ranking the bootstrap differences from smallest to largest #
rankdiff<-sort(boot.samples$diff)

#Print out just the first few#
head(rankdiff)
```

```
[1] -0.30 -0.30 -0.30 -0.25 -0.20 -0.20
```

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

```
[1] -0.05
```

```
#Upper endpoint is the 950th percentile (90% confidence)#
upper<-rankdiff[950]
upper
```

```
[1] 0.45
```

Because our bootstrap distribution is relatively bell-shaped, we could also a calculate a rough 95% boostrap SE confidence interval.

```
SE<-sd(rankdiff)
SE
```

```
[1] 0.1555809
```

2.2.3 Mathematical model for the difference in proportions

For a single proportion, we needed two conditions to be met to ensure the sampling distribution of $\hat{p}$ is approximately normal:

-

-

If these conditions are met, then

We must meet similar conditions to ensure the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal:

-

-

If these conditions are met, then
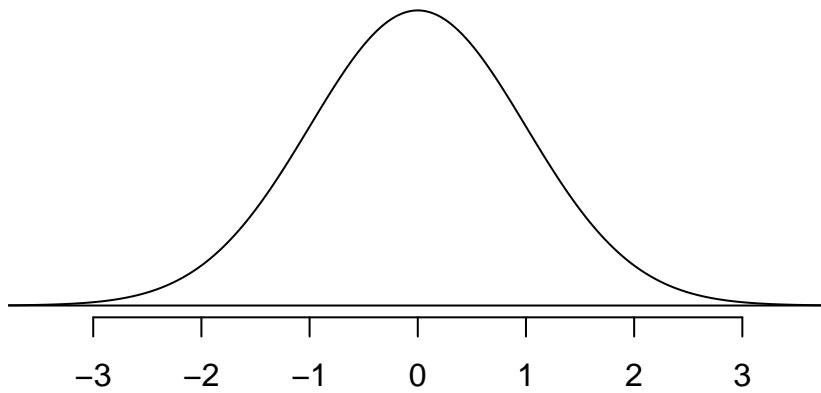
Like before we don't know p_1 and p_2, so we'll use our best guess. And, like before, our best guess will change depending on whether we're constructing a confidence interval or carrying out a hypothesis test.

How is this going to play out in a hypothesis test?

Example (17.5): (Aside: why can't we do the electrical stimuation example again?). A 2021 Gallup poll surveyed 3941 students pursuing a bachelor's degree and 2064 students pursuing an associate's degree. The survey found that 51% of the bachelor's students (2010) and 44% of the associate's students (908) said that COVID-19 will negatively impact their ability to complete the degree. We want to decide whether the proportion of bachelor's students who believe the pandemic will negatively impact degree completion is different from the proportion of associate's students who believe they will be negatively affected. Let p_B be the proportion of bachelor's students who believe they'll be negatively affected and let p_A be the proportion of associate's students who believe they'll be negatively affected.

To visualize the p-value:

```
normTail(m=0,s=1,L=-5.15, U=5.15)
```

And to get the p-value:

```
pnorm(-5.15,mean=0,sd=1)+(1-pnorm(5.15,mean=0,sd=1))
```

```
[1] 2.604865e-07
```

How is this going to play out in a confidence interval?

Example(17.9): A Kaiser Family Foundation poll for US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic "National Health Plan." There were 347 Democrats, 298 Republicans, and 617 Independents surveyed (Foundation, 2019). We want to estimate the difference between the proportion of Democrats and Independents who support a National Health Plan.

What impacts the width of a confidence interval?

There are three main things that impact the width of a confidence interval:

- Confidence level

- Sample size

- Standard Error

No matter what method we use to calculate the confidence interval, the **confidence level** is a statement about the long run percentage of confidence intervals that would succeed in capturing the true value of the parameter. What does this mean? Applet

Hypothesis tests vs. Confidence Intervals

While we should be matching analysis method to the research question, there is a nice relationship between hypothesis tests and confidence intervals. Recall that confidence intervals give a set of plausible values for the unknown parameter.

2.3 Inference for Two-Way Tables (Chapter 18)

So far, we've considered categorical variables with only two possible outcomes: success and failure. Many categorical variables have more than two possible outcomes, so we can't easily define the proportion of "successes." Instead, we'll summarize categorical data with more than two levels using two-way tables. In this class, we're still going to restrict ourselves to only two variables (often explanatory and response, but not necessarily), both with two or more levels. However, there are certainly statistical methods for more complicated situations.

Typically, research questions focus on how the proportions of the possible outcomes in the response variable change (or don't) across the levels of the explanatory variable. However, we can also consider questions about a single variable with more than two outcomes (are the possible outcomes all equally likely? do the possible outcomes follow a particular pattern?) or just whether the two categorical variables are independent or dependent without assigning an explanatory/response relationship. Due to the structure of the variable(s), there really isn't a population parameter of interest. We can't (usually) make a function of proportion of successes that makes sense to estimate, like we can with $p_1 - p_2$. That means we'll be considering only tests, not confidence intervals. We'll focus on the randomization test and the mathematical model approach. Both methods start with the same set-up.

Example: When surveys are administered, we hope that the respondents give accurate answers. Does the mode of survey delivery affect this? Schober et al (2015) investigated this question. They had 147 people who agreed to be interviewed on an iPhone, and they were randomly assigned to one of three interview modes: human voice, automated voice, text. One question asked was whether they exercise less than once per week during a typical week (a yes is mostly likely considered socially undesirable). The explanatory variable here is survey mode and the response is whether or not the respondent said yes. Here are the data:

	Text	Human Voice	Automated Voice	Total
Exercise Yes	34	21	20	75
Exercise No	124	139	139	402
Total	158	160	159	477

Based on these data, it looks like the answer to the question does change depending on survey mode, with respondents more likely to say yes via text. However, we don't know if this result could have happened by chance.

2.3.1 Expected Counts

We don't expect the proportion of 'yes' to be exactly the same across all survey modes, but we want to know if these vary enough to convince us that survey mode and answer are not independent. To do this, we need to find **expected counts** for each cell in the table.

Again, here are the observed data

	Text	Human Voice	Automated Voice	Total
Exercise Yes	34	21	20	75
Exercise No	124	139	139	402
Total	158	160	159	477

	Text	Human Voice	Auto Voice	Total
Exercise Yes	34 (_____)	21 (_____)	20 (_____)	75
Exercise No	124 (_____)	139 (_____)	139 (_____)	402
Total	158	160	159	477

So now the key question...are the observed and expected cell counts different enough?

- Cell(1,1) obs - exp = 34 -
- Cell(1,2) obs - exp = 21 -
- Cell(1,3) obs - exp = 20 -
- Cell(2,1) obs - exp = 124 -
- Cell(2,2) obs - exp = 139 -
- Cell(2,3) obs - exp = 139 -

New test statistic!

In our example:

- Cell(1,1) (obs - exp)^2/exp = $(34 - 24.84)^2/(24.84) = 9.16^2/24.84 = 3.3778$
- Cell(1,2) (obs - exp)^2/exp = $(21 - 25.16)^2/(25.16) = (-4.16)^2/25.16 = 0.6878$
- Cell(1,3) (obs - exp)^2/exp = $(20 - 25)^2/(25) = (-5)^2/25 = 1$
- Cell(2,1) (obs - exp)^2/exp = $(124 - 133.16)^2/(133.16) = (-9.16)^2/133.16 = 0.6301$
- Cell(2,2) (obs - exp)^2/exp = $(139 - 134.84)^2/(134.84) = 4.16^2/134.84 = 0.1283$
- Cell(2,3) (obs - exp)^2/exp = $(139 - 134)^2/(134) = 5^2/134 = 0.3731$

To see if this is 'big' we need the sampling distribution of our new test statistic. We can estimate that sampling distribution using either a randomization test or the mathematical model approach.

2.3.2 Randomization Test

The randomization test for a two-way table works just like it does with two samples. We'll randomize by shuffling and dealing/assigning the 75 yes answers and 402 no answers to the three survey modes at random.

- How many colors of cards?

- How many stacks to deal them into? How many in each stack?

- What do we find for each deal/shuffle?

Applet

Conclusion:

2.3.3 Mathematical Model

Based on what we just observed in the applet, the normal distribution is not going to be a good approximation to sampling distribution. It turns out this test statistic follows a different mathematical distribution, the **chi-squared distribution** (proof: see STAT 462). The normal distribution has two parameters that determine its shape: the mean (μ) and standard deviation (σ). The shape of the chi-square distribution is determined by a parameter called the **degrees of freedom (df)**. Figure 18.2 on page 307 shows how the shape of the distribution changes depending on the df.

So how can we use this?

Again, we have conditions that need to be met for the mathematical model to be a good approximation:

-

-

Example: Let's go back and do the survey mode example using the mathematical model approach. First, we'll need to check the conditions are met:

-

-

To find the p-value, we can use the R function `pchisq()`. Like `pnorm` it gives area to the left. So,

```
pchisq(6.1971,df=2,lower.tail=FALSE)
```

```
[1] 0.04511457
```

We can also do this directly in R.

```
surveymode<-read.csv("surveymode.csv",header=TRUE)

#make a table#
mode<-table(surveymode$Response,surveymode$Mode)

#see the table, note alphabetical order#
mode
```

```
     Avoice Hvoice Text
  No    139    139  124
  Yes    20     21   34
```

```
#Chi-square mathematical model test#
chisq.test(mode)
```

```
    Pearson's Chi-squared test

data:  mode
X-squared = 6.0069, df = 2, p-value = 0.04962
```

```
#Chi-square randomization test#
chisq.test(mode,simulate.p.value=TRUE, B=1000)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 1000
    replicates)

data:  mode
X-squared = 6.0069, df = NA, p-value = 0.05894
```

3 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.

3.1 Inference for a Single Mean (Chapter 19)

3.1.1 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)

#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
```
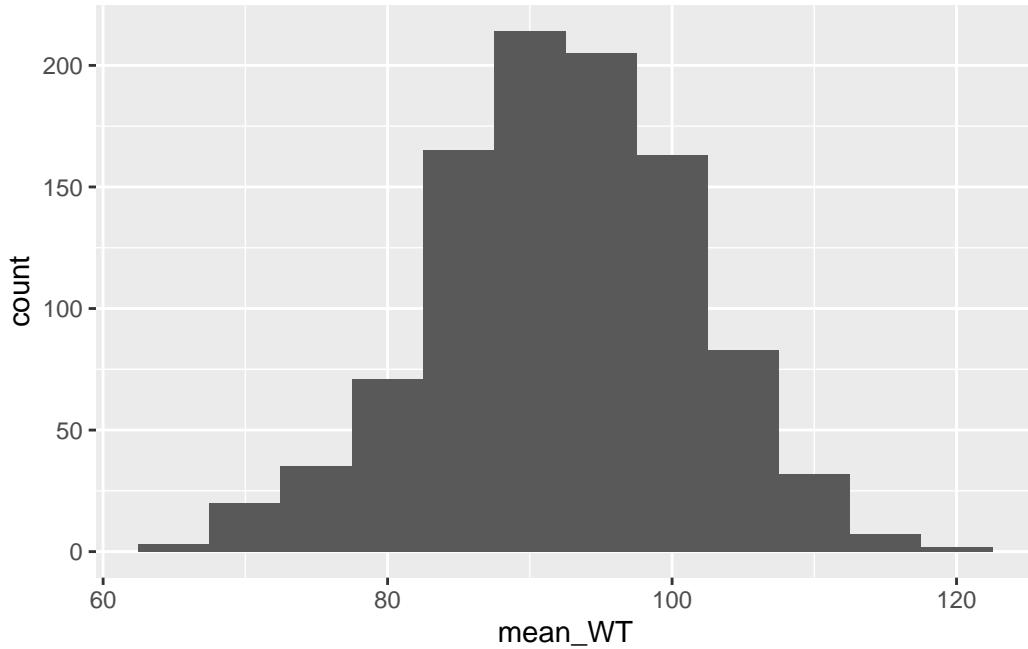


```
#To get the bootstrap percentile confidence interval, #
#start by ranking the bootstrap means from smallest to largest #
rankmean<-sort(boot.samples$mean_WT)

#Lower endpoint is the 2.5th percentile (95% confidence)#
```

```
lower<-rankmean[25]
lower
```

[1] 72.83333

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

[1] 109.1667

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] 9.057314

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#

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
```
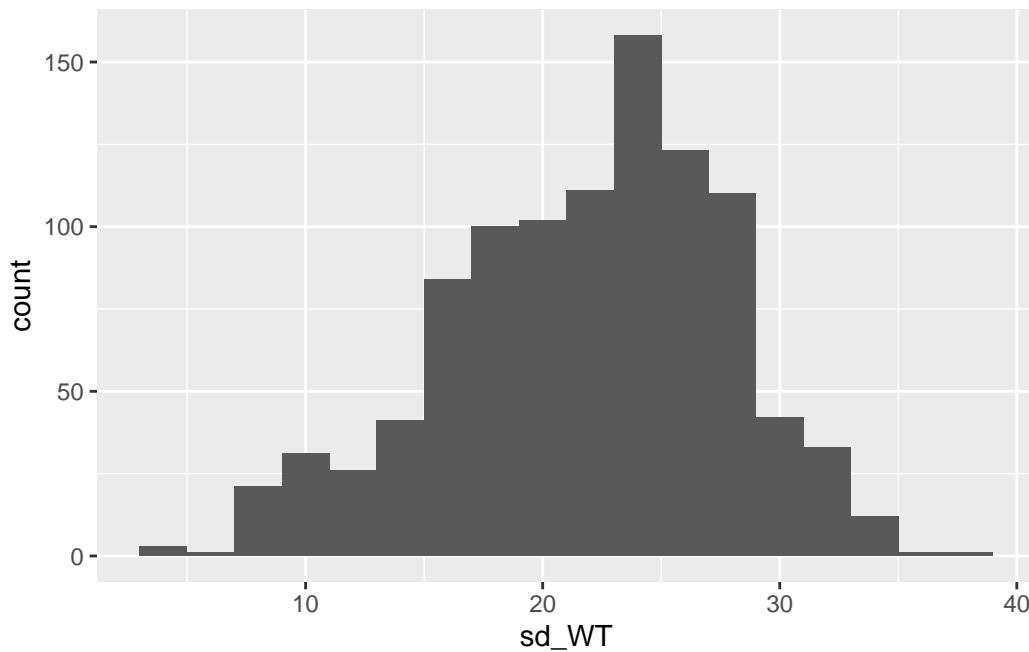


```
#To get the bootstrap percentile confidence interval, #
#start by ranking the bootstrap sds from smallest to largest #
ranksd<-sort(boot.samples$sd_WT)

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

```
[1] 10.80123
```

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

```
[1] 30.82477
```

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.

3.1.2 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)`

3.1.2.1 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
```
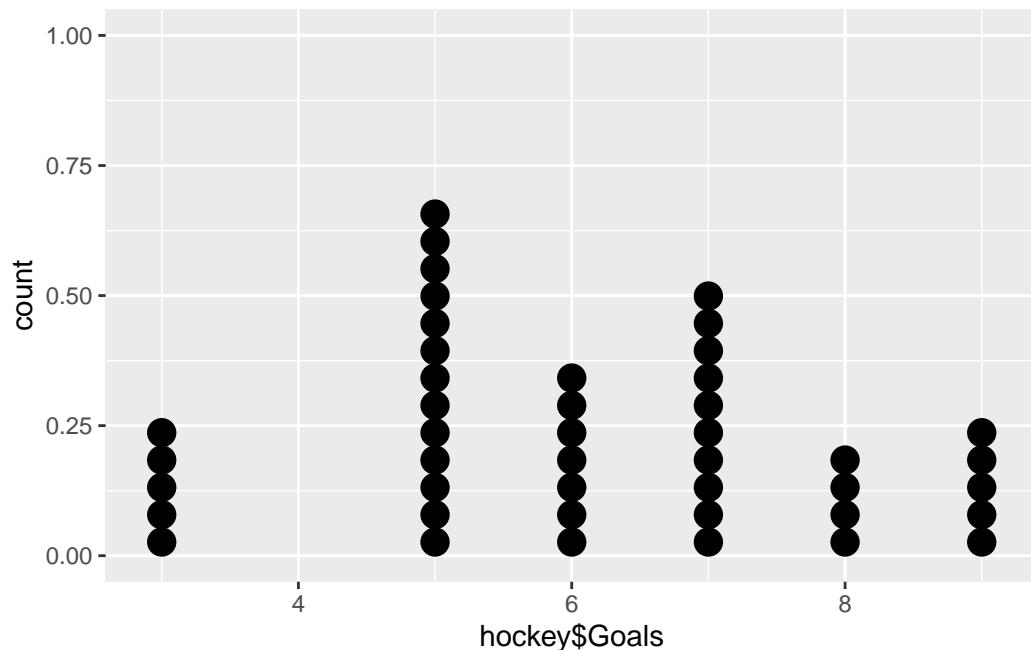
```
sGoals<-sd(hockey$Goals)
sGoals
```

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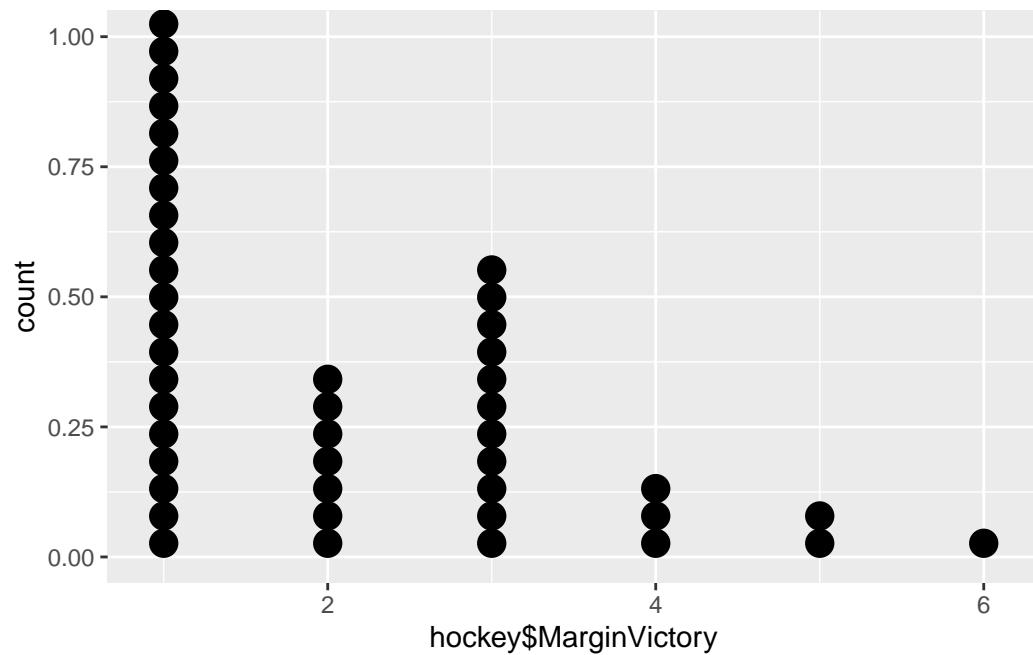
```
#Summarize Margin of Victory#
summary(hockey$MarginVictory)
```

```
   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
```

3.1.2.2 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
```

3.2 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.

3.2.1 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.

66

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

We'll start by loading the necessary packages and reading in the data.

```
library(ggplot2)
library(mosaic)
```

```
chili<-read.csv("chili.csv",header=TRUE)
```

```
head(chili)
```

```
  VideoGame ChiliSauce
1   violent         42
2   violent          4
3   violent         27
4   violent          2
5   violent         10
6   violent          5
```
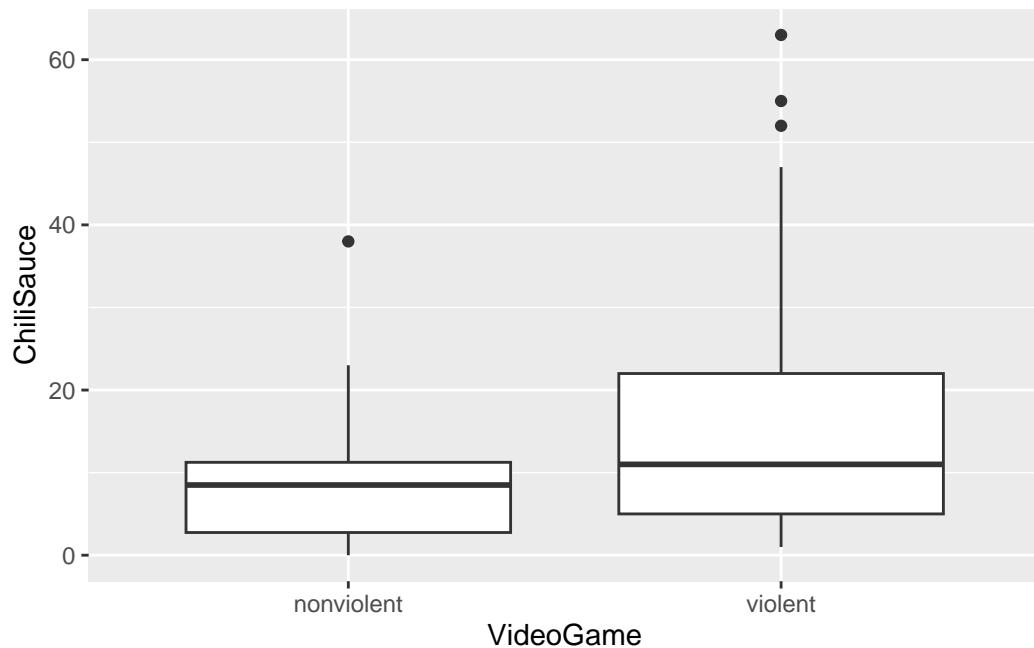
```
#Get the sample mean for each group#
mean(ChiliSauce~VideoGame, data=chili)
```

```
nonviolent     violent
  9.057692  16.122449
```

```
#Find the difference in sample means#
obs_diff <- diff(mean(ChiliSauce~VideoGame, data=chili))
obs_diff
```

```
 violent
7.064757
```

```
#Construct box plots for each group#

gf_boxplot(ChiliSauce~VideoGame,data=chili)
```



```
#Scramble the treatment groups with respect to outcome many times to get the null distributio

null_dist <- do(1000)*diff(mean(ChiliSauce~shuffle(VideoGame), data=chili))

head(null_dist)
```
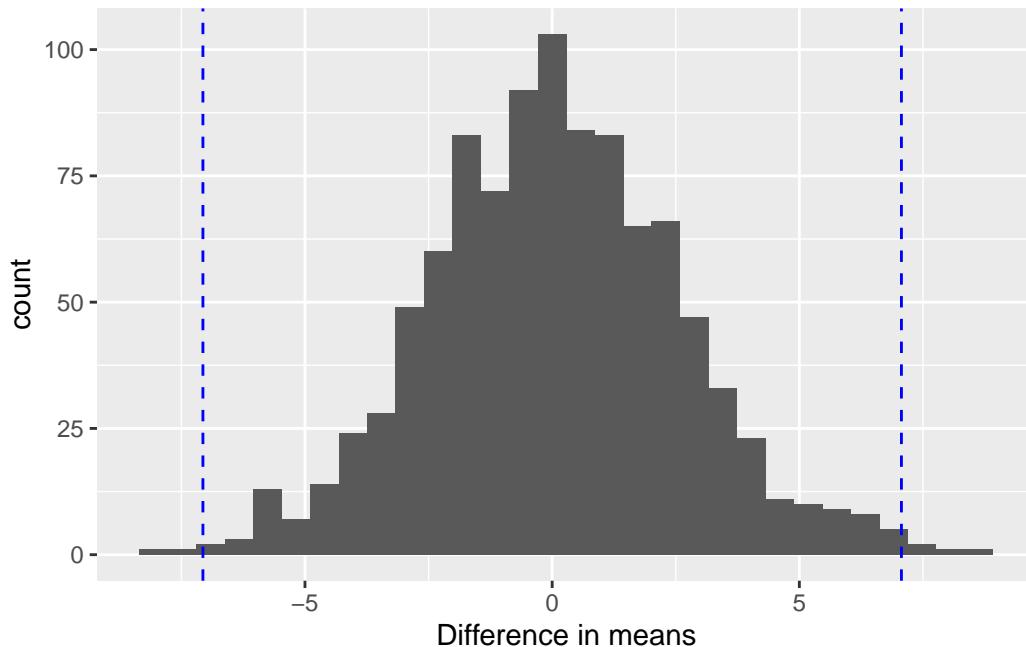
```
       violent
1 -2.349489796
2 -1.854003140
```

```
3  0.009026688
4  3.259419152
5 -2.052197802
6  0.009026688
```

```
#Histogram of the null distribution#
ggplot(data=null_dist) + geom_histogram(mapping=aes(x=violent))+
  xlab("Difference in means") +
  geom_vline(xintercept = obs_diff, linetype=2, color="blue") +
  geom_vline(xintercept = -obs_diff, linetype=2, color="blue")
```
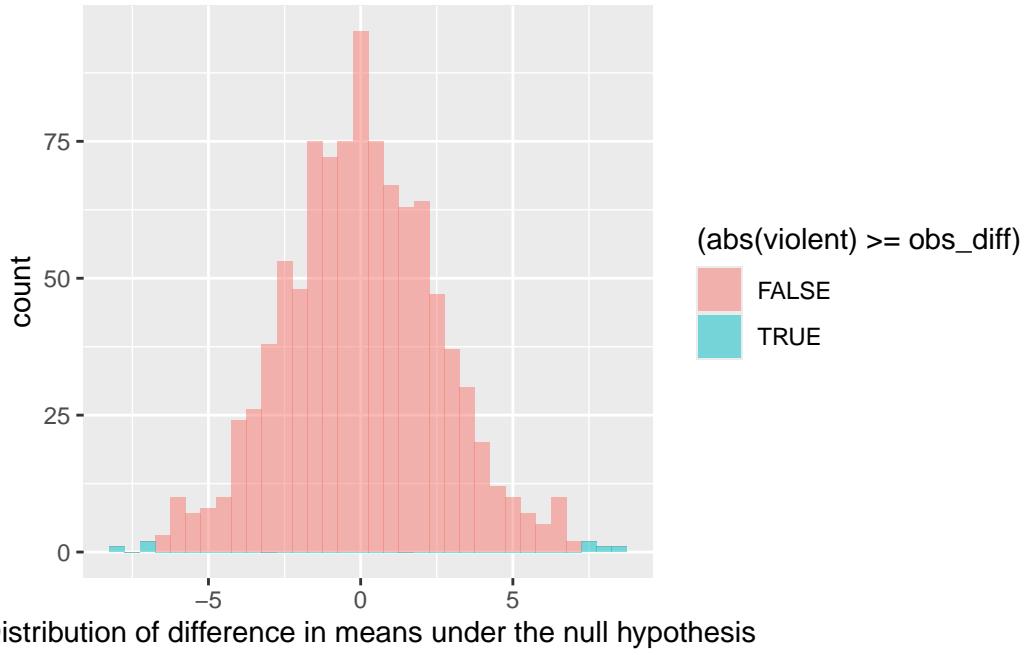
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.


```
#Calculate the proportion of simulated differences in mean as or more extreme than obs#
prop(~violent >= obs_diff, data=null_dist)+prop(~violent <= -obs_diff, data=null_dist)
```

```
prop_TRUE
    0.007
```

```
#Another way to visualize!#
gf_histogram(~violent, fill = ~(abs(violent) >= obs_diff), data=null_dist,
             binwidth=0.5,xlab="Distribution of difference in means under the null hypothesis
```

Distribution of difference in means under the null hypothesis

3.2.2 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.

Let's look at R, and read in the data.

```
athletes<-read.csv("Athletes.csv",header=TRUE)

head(athletes)
```

```
  Baseball Soccer
1      190    165
2      200    190
3      187    185
4      182    187
5      192    183
6      205    189
```

```
#observed mean and sd#
basemean<-mean(athletes$Baseball)
basemean
```

```
[1] 191.375
```

```
basesd<-sd(athletes$Baseball)
basesd
```

```
[1] 9.464847
```

```
soccermean<-mean(athletes$Soccer)
soccermean
```

```
[1] 174.5
```

```
soccersd<-sd(athletes$Soccer)
soccersd
```

```
[1] 12.49533
```

```
#Set up an empty data set with 4 columns: simulation number,
# bootstrap mean for baseball, bootstrap mean for soccer, difference#

boot.samples<-data.frame(sim=1:1000,mean_base=NA,mean_soccer=NA,diff=NA)

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

```
for(i in 1:1000){
  boot.samples$mean_base[i]<-mean(sample(athletes$Baseball,size=16,replace=TRUE))
  boot.samples$mean_soccer[i]<-mean(sample(athletes$Soccer,size=16,replace=TRUE))
  boot.samples$diff[i]<-boot.samples$mean_base[i]-boot.samples$mean_soccer[i]
}

head(boot.samples)
```
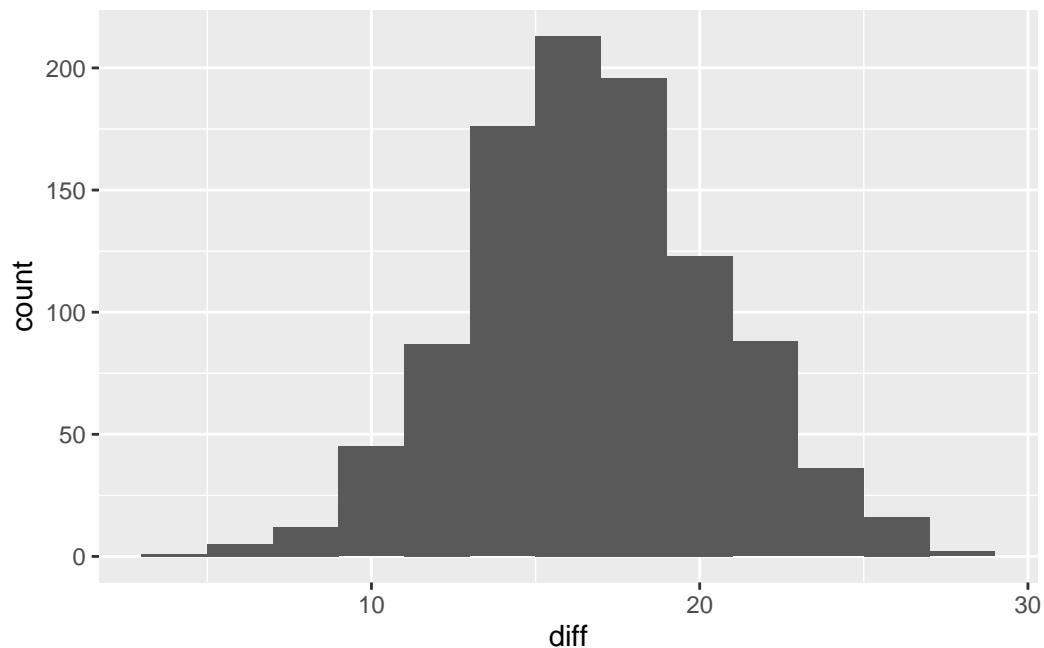
```
  sim mean_base mean_soccer     diff
1   1  194.5000    176.3750 18.1250
2   2  189.8750    173.3750 16.5000
3   3  191.1250    173.5625 17.5625
4   4  189.1250    173.1875 15.9375
5   5  190.3125    175.5625 14.7500
6   6  193.9375    171.6250 22.3125
```

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

#See the plot#
boot.hist
```

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

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

```
[1] 9.6875
```

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

```
[1] 24.375
```

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

```
[1] 3.790677
```

So we can get a bootstrap percentile interval or a bootstrap SE interval.

Practice: (from book, page 346-348) 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.

More Practice: The Canvas file 'shoes.csv' contains data on many styles of athletic shoes. Find a confidence interval for the difference in mean price for road versus trail shoes.

3.2.3 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?

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

```
sleep<-read.csv("SleepStudy.csv",header=TRUE)

#Those with normal stress#
summary(sleep$GPA[sleep$Stress=="normal"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.250   3.209   3.500   4.000
```

```
sum(with(sleep,Stress=="normal"))
```

```
[1] 197
```

```
sd(sleep$GPA[sleep$Stress=="normal"])
```

```
[1] 0.412318
```

```
#Those with high stress#
summary(sleep$GPA[sleep$Stress=="high"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.500   3.065   3.350   3.366   3.603   4.000
```

```
sum(with(sleep,Stress=="high"))
```
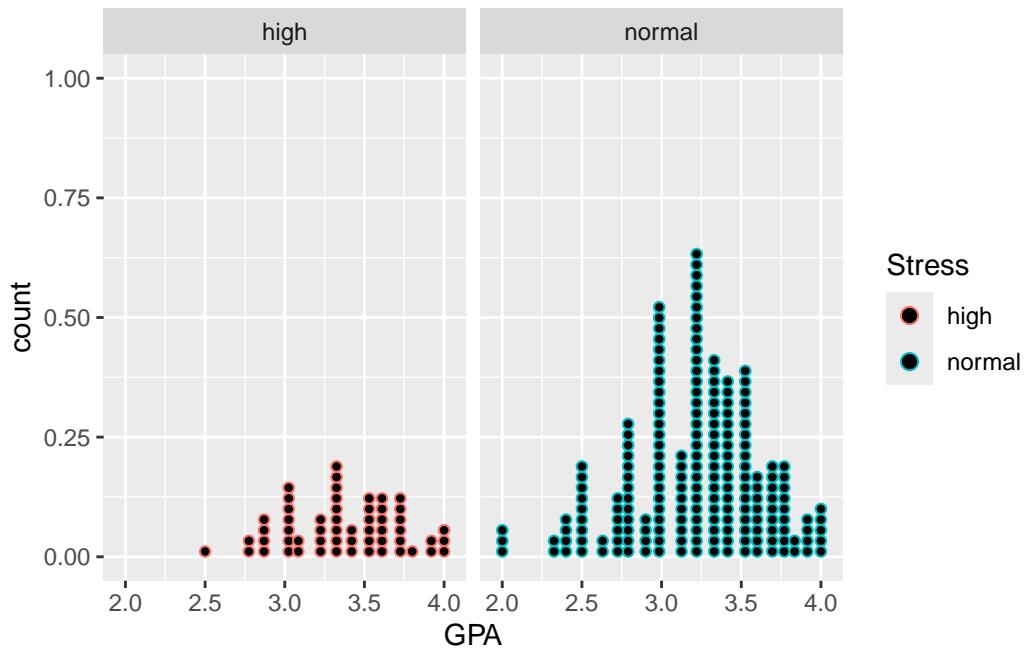
```
[1] 56
```

```
sd(sleep$GPA[sleep$Stress=="high"])
```

```
[1] 0.3513277
```

```
dotplot<-ggplot(sleep, aes(GPA)) + geom_dotplot(aes(color=Stress)) + facet_grid(~Stress)

dotplot
```

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

And we can get the p-value:

```
2*pt(-2.835,df=55)
```

```
[1] 0.00639864
```

In R, we can carry out this test very easily:

```
t.test(GPA~Stress,data=sleep)
```

```
    Welch Two Sample t-test

data:  GPA by Stress
t = 2.8397, df = 102.11, p-value = 0.005451
alternative hypothesis: true difference in means between group high and group normal is not e
95 percent confidence interval:
 0.04741781 0.26711265
sample estimates:
  mean in group high mean in group normal
           3.366250             3.208985
```

We can also use the `subset` function 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:

```
sleep2<-subset(sleep,LarkOwl=='Owl')
```

```
head(sleep2)
```

	Gender	ClassYear	LarkOwl	NumEarlyClass	EarlyClass	GPA	ClassesMissed
3	0	4	Owl	0	0	2.97	12
5	0	4	Owl	0	0	3.20	4
20	1	3	Owl	3	1	2.80	20
26	1	2	Owl	0	0	3.07	20
27	1	2	Owl	4	1	3.00	10
46	1	3	Owl	0	0	3.50	2

	CognitionZscore	PoorSleepQuality	DepressionScore	AnxietyScore	StressScore
3	0.38	18	18	18	9
5	1.22	9	7	25	14
20	-0.57	15	6	5	10
26	-0.62	9	23	5	16
27	0.16	5	14	13	14

```
46               0.47              6          9          5         19
   DepressionStatus AnxietyStatus Stress DASScore Happiness AlcoholUse Drinks
3         moderate        severe normal       45        17      Light      3
5           normal        severe normal       46        15   Moderate      4
20          normal        normal normal       21        19   Moderate      7
26          severe        normal   high       44        24    Abstain      0
27        moderate      moderate normal       41        28   Moderate      8
46          normal        normal   high       33        24   Moderate      9
   WeekdayBed WeekdayRise WeekdaySleep WeekendBed WeekendRise WeekendSleep
3      27.44        6.55         3.00      28.00       12.59        10.09
5      25.90        8.67         6.09      23.75        9.50         7.00
20     25.25        8.12         7.20      25.67       10.50         8.00
26     24.70       11.02        10.32      27.50       12.25         8.75
27     24.40        7.95         8.45      24.00       11.75         9.50
46     24.95        8.86         7.97      27.00       11.63         9.13
   AverageSleep AllNighter
3          5.02          0
5          6.35          0
20         7.43          0
26         9.87          0
27         8.75          0
46         8.30          0
```

```
t.test(WeekdaySleep~EarlyClass, data=sleep2)
```

```
    Welch Two Sample t-test

data:  WeekdaySleep by EarlyClass
t = 0.79051, df = 30.009, p-value = 0.4354
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -0.5768755  1.3055071
sample estimates:
mean in group 0 mean in group 1
      7.986316        7.622000
```

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 do need to check the conditions to see if we can use the mathematical model. We can use dotplot to check for extreme outliers. However, this data set is structured differently than we have seen before, and we'll need to restructure. There are two ways to do this, but one will require a couple of new packages: `tidyr` and `tidyverse`.

```
library(tidyverse)
library(tidyr)
```

One way to restructure:

```
athletes2 <- athletes %>% pivot_longer(cols=c('Baseball','Soccer'), names_to='Sport',values_t

athletes2
```

```
# A tibble: 32 x 2
   Sport     Weight
   <chr>      <int>
 1 Baseball     190
 2 Soccer       165
 3 Baseball     200
 4 Soccer       190
 5 Baseball     187
 6 Soccer       185
 7 Baseball     182
 8 Soccer       187
 9 Baseball     192
10 Soccer       183
# i 22 more rows
```

Another way:

```
athletes3<-cbind(stack(athletes[1:2]))

athletes3
```

```
     values        ind
1       190 Baseball
2       200 Baseball
3       187 Baseball
4       182 Baseball
5       192 Baseball
6       205 Baseball
7       185 Baseball
8       177 Baseball
9       186 Baseball
10      210 Baseball
11      198 Baseball
12      180 Baseball
13      182 Baseball
14      193 Baseball
15      200 Baseball
16      195 Baseball
17      165   Soccer
18      190   Soccer
19      185   Soccer
20      187   Soccer
21      183   Soccer
22      189   Soccer
23      170   Soccer
24      182   Soccer
25      156   Soccer
26      168   Soccer
27      173   Soccer
28      158   Soccer
29      150   Soccer
30      172   Soccer
31      180   Soccer
32      184   Soccer
```

```r
#rename columns#
#Do I need to do this? No. Does it help me keep track of which variable is which? Yes#
colnames(athletes3)[1]="weight"
colnames(athletes3)[2]="sport"

athletes3
```
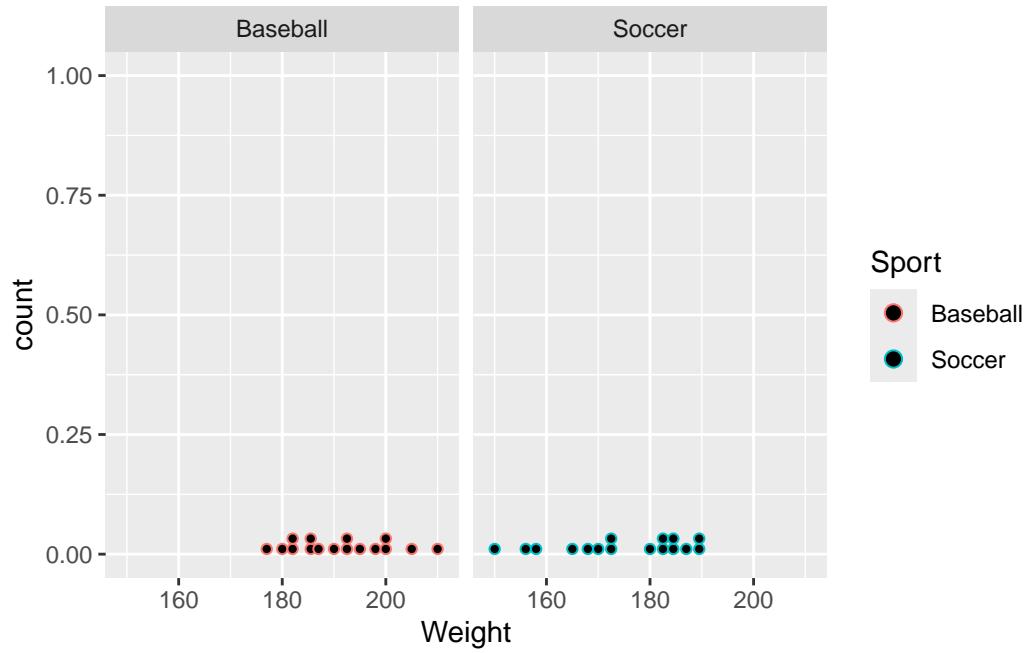
```
     weight    sport
1       190 Baseball
2       200 Baseball
```

```
3      187 Baseball
4      182 Baseball
5      192 Baseball
6      205 Baseball
7      185 Baseball
8      177 Baseball
9      186 Baseball
10     210 Baseball
11     198 Baseball
12     180 Baseball
13     182 Baseball
14     193 Baseball
15     200 Baseball
16     195 Baseball
17     165    Soccer
18     190    Soccer
19     185    Soccer
20     187    Soccer
21     183    Soccer
22     189    Soccer
23     170    Soccer
24     182    Soccer
25     156    Soccer
26     168    Soccer
27     173    Soccer
28     158    Soccer
29     150    Soccer
30     172    Soccer
31     180    Soccer
32     184    Soccer
```

Now we can get the dotplots.

```
dotplot<-ggplot(athletes2, aes(Weight)) + geom_dotplot(aes(color=Sport)) + facet_grid(~Sport)
dotplot
```

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

We can also get the confidence interval in R:

```
t.test(Weight~Sport,data=athletes2)
```

```
    Welch Two Sample t-test

data:  Weight by Sport
t = 4.3061, df = 27.95, p-value = 0.0001846
alternative hypothesis: true difference in means between group Baseball and group Soccer is r
95 percent confidence interval:
  8.846974 24.903026
sample estimates:
mean in group Baseball    mean in group Soccer
              191.375                 174.500
```

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

```
t.test(Weight~Sport,data=athletes2,conf.level=0.90)
```

```
    Welch Two Sample t-test

data:  Weight by Sport
t = 4.3061, df = 27.95, p-value = 0.0001846
alternative hypothesis: true difference in means between group Baseball and group Soccer is n
90 percent confidence interval:
 10.20813 23.54187
sample estimates:
mean in group Baseball    mean in group Soccer
             191.375                 174.500
```

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

3.3 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!
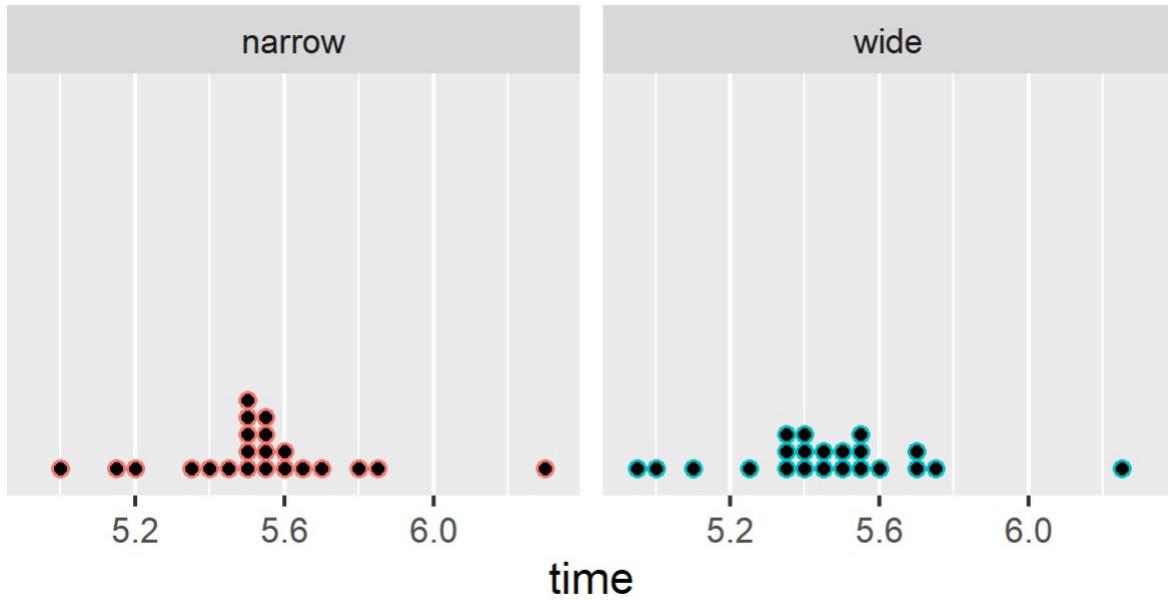
3.3.1 Randomization test for mean difference (matched pairs)

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 a `wide 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.)

Let's look at the data:

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. \end{itemize}

3.3.2 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**
- Let's look at the R code

```
bases<-read.csv("bases.csv",header=TRUE)
bases
```

```
   id narrow wide
1   1   5.50 5.55
2   2   5.70 5.75
3   3   5.60 5.50
4   4   5.50 5.40
5   5   5.85 5.70
6   6   5.55 5.60
7   7   5.40 5.35
8   8   5.50 5.35
```

```
9    9    5.15 5.00
10  10    5.80 5.70
11  11    5.20 5.10
12  12    5.55 5.45
13  13    5.35 5.45
14  14    5.00 4.95
15  15    5.50 5.40
16  16    5.55 5.50
17  17    5.55 5.35
18  18    5.50 5.55
19  19    5.45 5.25
20  20    5.60 5.40
21  21    5.65 5.55
22  22    6.30 6.25
```

```
bases$diff<-bases$narrow-bases$wide

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

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

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

for(i in 1:1000){
  boot.samples$mean_diff[i]<-mean(sample(bases$diff,size=22,replace=TRUE))
}

#Histogram#
boot.hist<-ggplot(boot.samples, aes(mean_diff)) + geom_histogram(binwidth=0.005)

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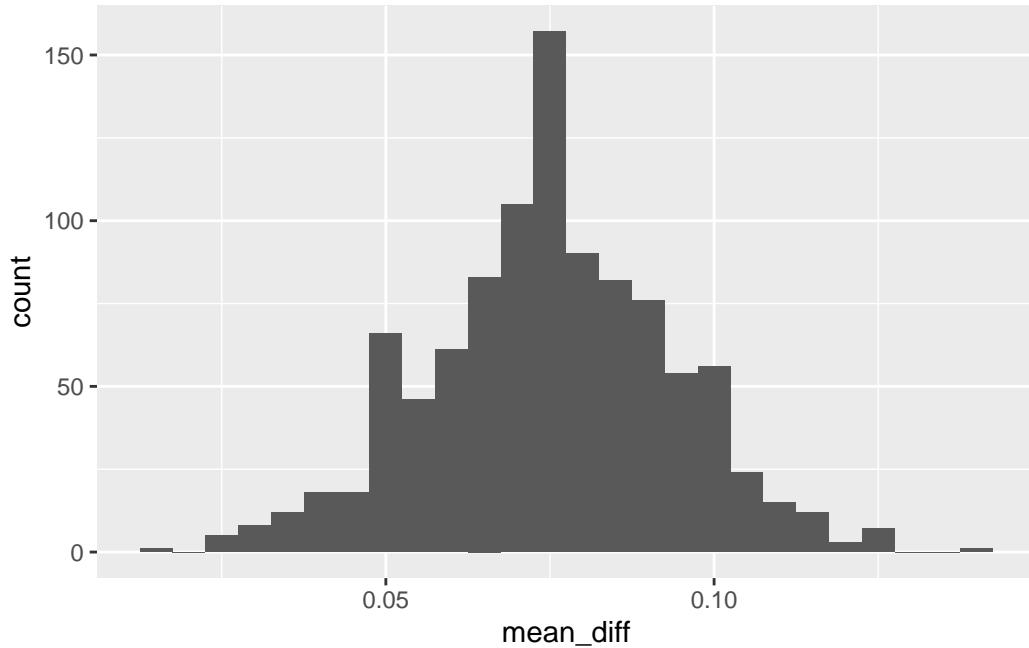
```
#To get the bootstrap percentile confidence interval, #
#start by ranking the bootstrap means from smallest to largest #
rankmean<-sort(boot.samples$mean_diff)

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

[1] 0.03636364

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

[1] 0.1113636

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

[1] 0.01858387

- 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.

```
##INCORRECT ANALYSIS##
#Set up an empty data set with 4 columns: simulation number, bootstrap mean for narrow, boots

boot.samples<-data.frame(sim=1:1000,mean_narrow=NA,mean_wide=NA,diff=NA)

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

for(i in 1:1000){
  boot.samples$mean_narrow[i]<-mean(sample(bases$narrow,size=22,replace=TRUE))
  boot.samples$mean_wide[i]<-mean(sample(bases$wide,size=22,replace=TRUE))
  boot.samples$diff[i]<-boot.samples$mean_narrow[i]-boot.samples$mean_wide[i]
}

#Histogram#
boot.hist<-ggplot(boot.samples, aes(diff)) + geom_histogram(binwidth=0.05)

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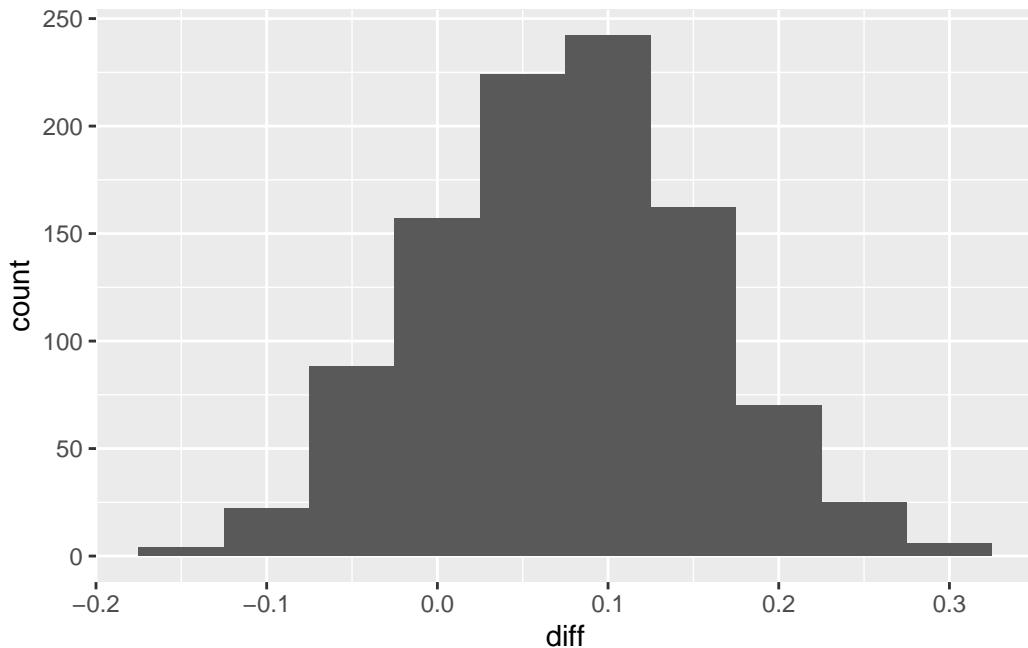
```
#To get the bootstrap percentile confidence interval, #
#start by ranking the bootstrap means from smallest to largest #
rankmean<-sort(boot.samples$diff)

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

[1] -0.07727273

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

[1] 0.2295455

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

3.3.3 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	40	3.010	0.490
Without	40	2.715	0.507
Diff=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.

```r
t.test(bases$narrow,bases$wide,paired=TRUE)
```

```
    Paired t-test

data:  bases$narrow and bases$wide
t = 3.9837, df = 21, p-value = 0.0006754
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.03584814 0.11415186
sample estimates:
mean difference
          0.075
```

3.4 Inference for Comparing Many Means (Chapter 22)

We're going to start this section by considering an example. The data are in the file 'mice.csv.'
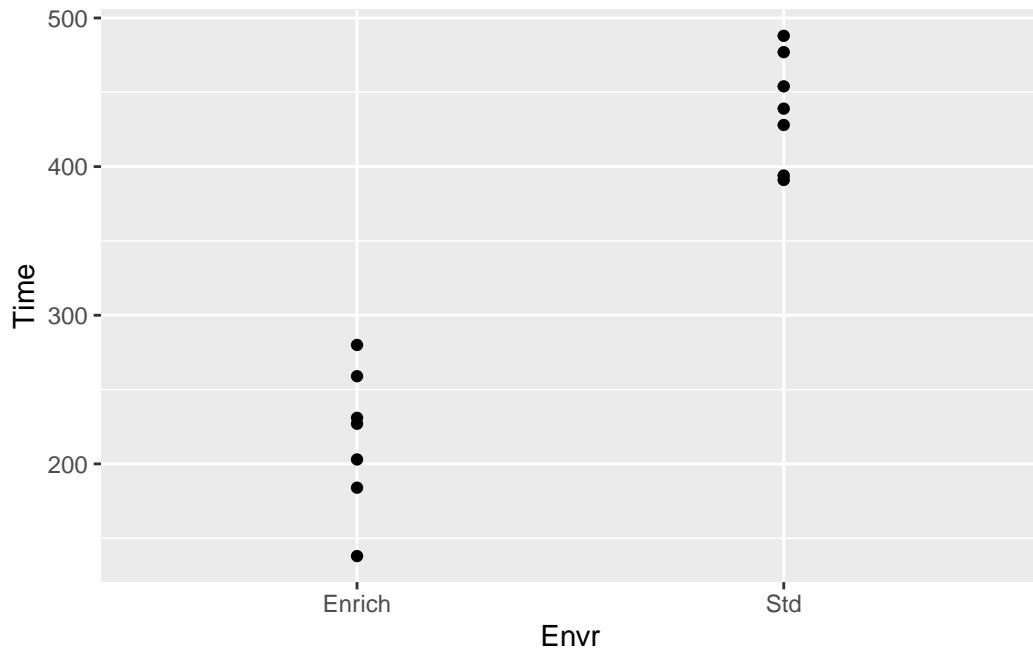
Example: These data come from an experiment to determine if exercise confers some resilience to stress. Mice were randomly assigned to either an enriched environment (exercise wheel) or standard environment, and spent three weeks there. After that time, they were exposed for five minutes per day for 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 a 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.

```
mice<-read.csv("mice.csv",header=TRUE)
head(mice)
```

```
    Envr Time
1 Enrich  259
2 Enrich  280
3 Enrich  138
4 Enrich  227
5 Enrich  203
6 Enrich  184
```

We already know how to answer this research question!

Let's first plot the data
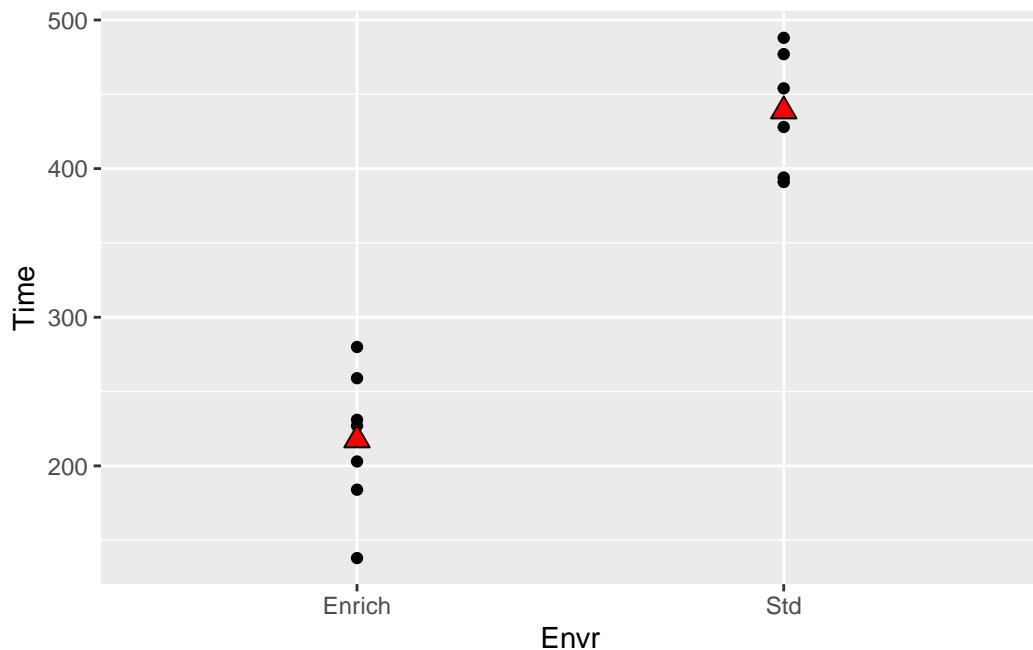
It definitely looks like there's a difference between the groups! We can find the group means and standard deviations. We'll also add the sample means to the plot.

```
aggregate(mice$Time, by=list(mice$Envr), FUN=mean)
```

```
  Group.1        x
1  Enrich 217.4286
2     Std 438.7143
```

```
aggregate(mice$Time, by=list(mice$Envr), FUN=sd)
```

```
  Group.1        x
1  Enrich 47.52844
2     Std 37.68162
```
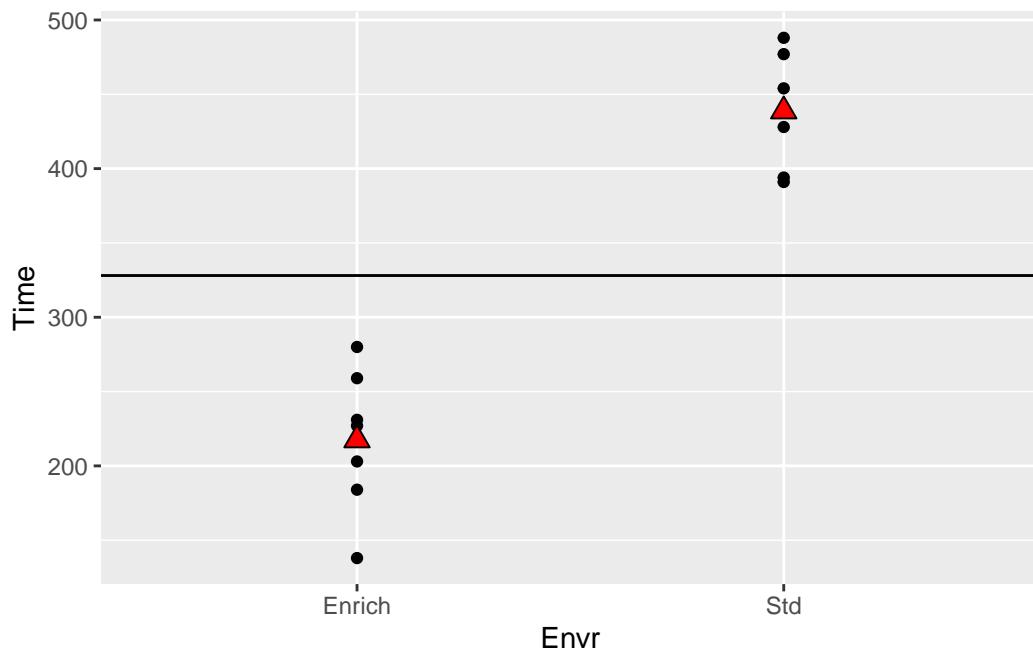
We're testing $H_0 : \mu_1 = \mu_2$, and assume this is true to construct the test. The overall common sample mean is $\bar{x} = 328.07$.

```
t.test(Time~Envr,data=mice)
```

```
	Welch Two Sample t-test

data:  Time by Envr
t = -9.6526, df = 11.407, p-value = 7.885e-07
alternative hypothesis: true difference in means between group Enrich and group Std is not eq
95 percent confidence interval:
 -271.5245 -171.0470
sample estimates:
mean in group Enrich    mean in group Std
          217.4286               438.7143
```

It turns out the 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 sample variance is calculated:

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, just like we did in the last section when we were comparing two variances:

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

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:

Notice!

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

- approximate normality (no extreme outliers, no strong skew)
- independence between groups and between observations
- constant variance (we didn't make a big deal of this one, but mentioned it)

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 $Y_1, Y_2, \dots, Y_n$ represent our data points we can summarize this as:

Or another way:

This is a **statistical model** with 2 parameters: μ and σ^2.
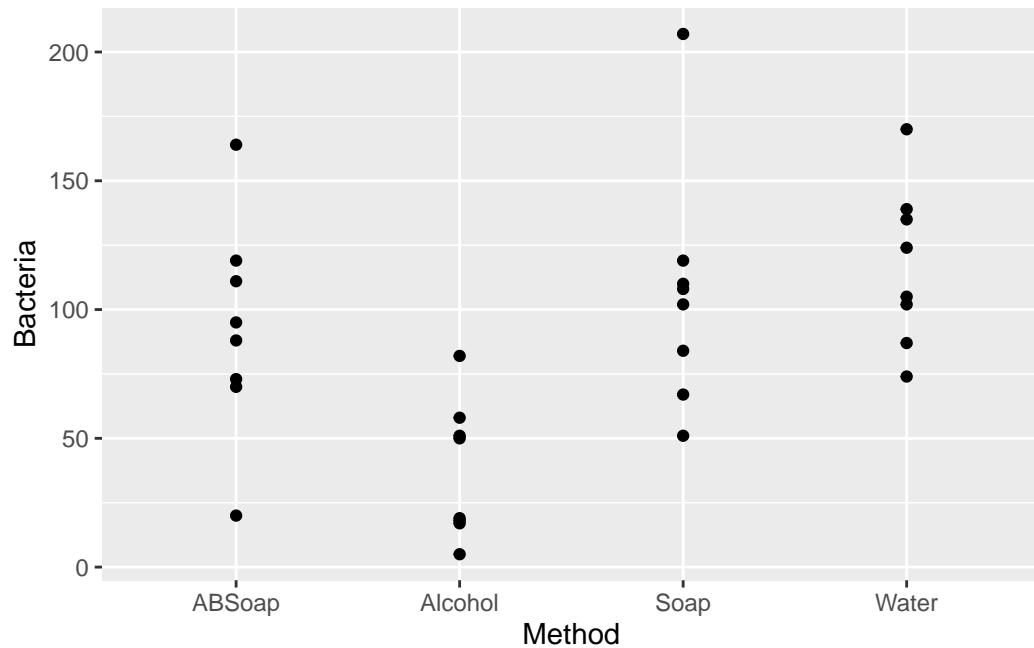
If we have two samples:

If we have more than two samples:

Let's start with some summary statistics
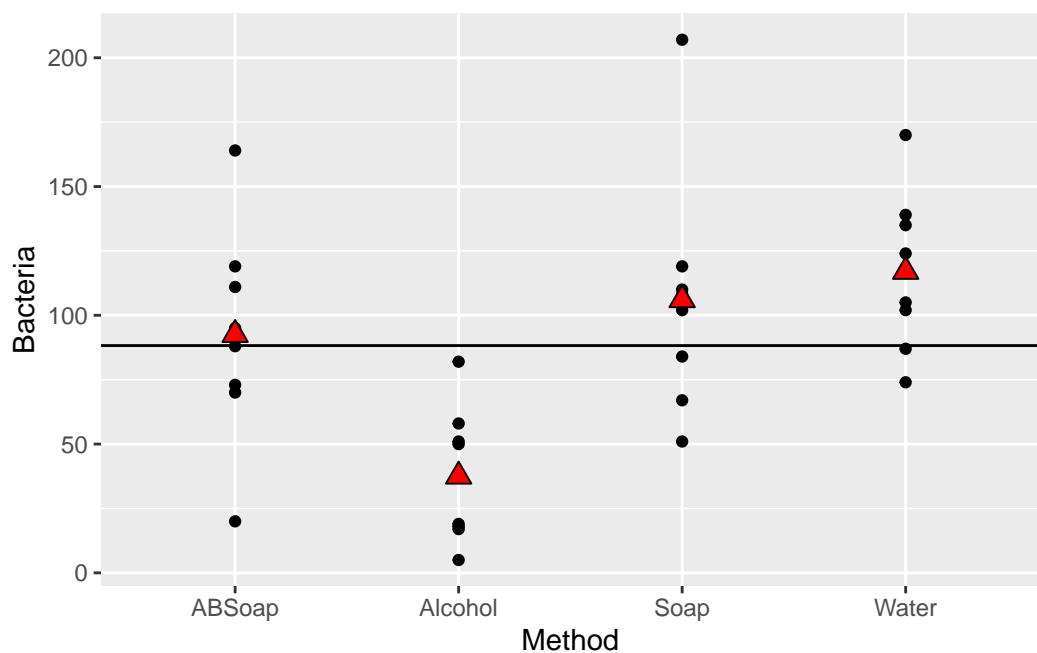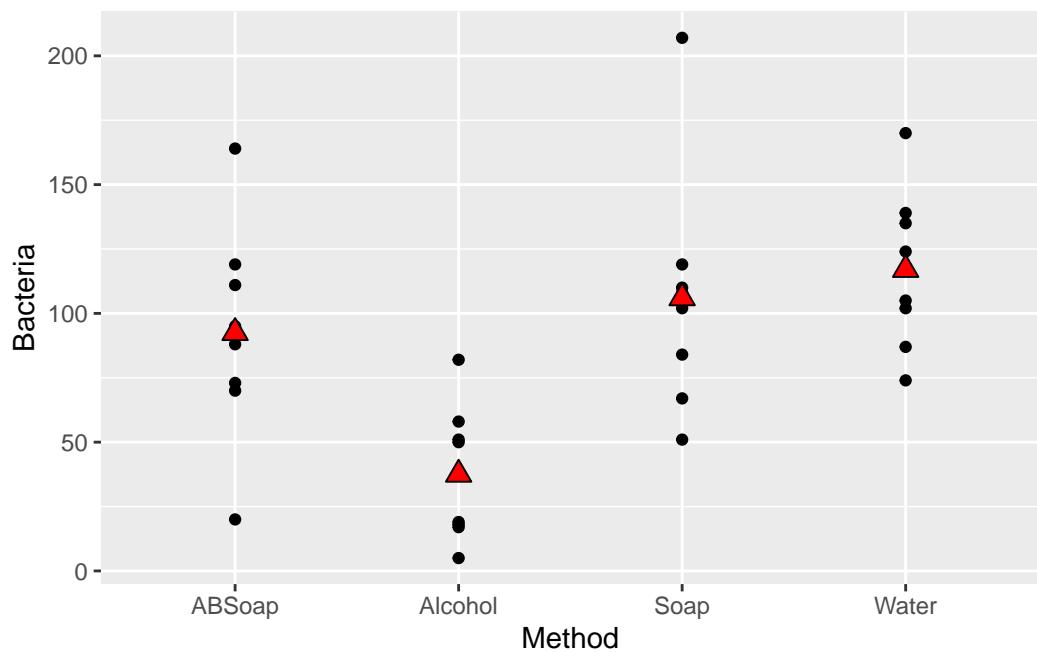
$$
\begin{aligned}
Y_{i\cdot} &= \sum_{j=1}^{n_i} Y_{ij} = i^{th} \text{ sample total} \\
\bar{Y}_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = i^{th} \text{ sample mean} \\
Y_{\cdot\cdot} &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij} = \text{ grand total} \\
\bar{Y}_{\cdot\cdot} &= \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij} = \text{ grand mean } \left(N = \sum_{i=1}^{n_i} n_i\right)
\end{aligned}
$$

Example: A student carried out an experiment to investigate handwashing methods: water only, regular soap, antibacterial soap, and alcohol spray. Each treatment was replicated 8 times, and bacteria count was observed. The data are in 'handwash.csv'.


```
  Group.1     x
1  ABSoap  92.5
2 Alcohol  37.5
3    Soap 106.0
4   Water 117.0
```

```
  Group.1        x
1  ABSoap 41.96257
2 Alcohol 26.55991
3    Soap 46.95895
4   Water 31.13106
```

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.

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 $j = 1$ 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

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

We often summarize our calculations in a table (df 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 can 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:

We still have assumptions: - Independence between and among groups - Responses/errors 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:

- 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.

- 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 `pf()` in R to find p-values

```
pf(93.2,df1=1,df2=12,lower.tail=FALSE)
```

```
[1] 5.232224e-07
```

The p-value typically gets added to the table as well:

Source	df	SS	MS	F	p-value
Model	1	171386	171386	93.2	0.0000005
Error	12	22073	1839		
Total	13	193459			

This is the only time we'll do an ANOVA by hand! Let's do the same in R.

```
anova(lm(Time~Envr, data=mice))
```

```
Analysis of Variance Table

Response: Time
          Df Sum Sq Mean Sq F value   Pr(>F)
Envr       1 171386  171386  93.173 5.24e-07 ***
Residuals 12  22073    1839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Let's now carry out the ANOVA on the handwashing data. We'll start by writing the model and sketching the ANOVA table.

```
anova(lm(Bacteria~Method,data=handwash))
```

```
Analysis of Variance Table

Response: Bacteria
          Df Sum Sq Mean Sq F value   Pr(>F)
Method     3  29882  9960.7  7.0636 0.001111 **
Residuals 28  39484  1410.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We'll do some more examples, focusing on sketching the ANOVA table.

Example: A teacher takes a random sample of 30 student GPAs, along with where they chose to sit in a classroom (front, middle, back). We want to see if mean GPA differs based on where a student sits.

Example: Baseball run time. The data gives run time in seconds for 50 yards for 29 players at three different positions (OF, IF, C).

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 took a random sample of 40 students, and they grouped majors into four categories: applied science (as), natural science (ns), social science (ss), and arts/humanities (ah).