Psych 100A Spring 2019: Week 2 Slides

Amanda Montoya

April 8, 2019

While we're setting up

```
NLSdata <- read.csv("http://bit.ly/NLSdata", header=TRUE)
NLSdata$HeightTotalFt <- NLSdata$HeightFt2010 +
                         NLSdata$HeightIn2010/12
```

Ask for me

What we've done so far

- What is data? How do we represent data?
- What are variables?
- Different types of variables: Quantitative vs. Qualitative
- How can we explore variables?
  - Visual summaries (histograms, bar plots)
  - Numeric summaries (min, max, mean, median, quartiles)

Learning Outcomes Today

- Explore variation using information about shape, center, and spread
- Use the DGP to discuss sampling variability
  - What it is?
  - How it is affected by sample size?

Learning in this class

Two major skills: statistical thinking (Concepts and ideas from statistics) and statistical doing (R)

- The book is focused on both thinking and doing statistics. It's self-paced so you should be learning both at the same time.
- Lecture is mostly focused on statistical thinking, but I also show you how to do what we're doing. The focus though is on the thinking.
  - If doing R at the same time as lecture is keeping you from learning about statistical thinking, stop using R.
  - Pay attention to the big ideas during lecture, work through lecture slides (to figure out the R) at home on your own pace.

Describing Shape: How many peaks?

We can describe a distribion based on how many peaks/modes it has.

Unimodal: One peak

Bimodal: Two peaks

Uniform: No peaks (all data is equally probable)

Can you come up with a word that might mean there are more than two peaks?

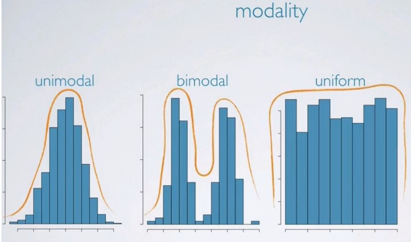


Figure 1: How many peaks/modes

Describing Shape: Skew

Skew of a distribution means that the left side and right side don't look similar.

The opposite of skewed is "symmetrical".

Skew can go in two directions "right" and "left". This corresponds to the direction of the "tail" (the part of the distribution which stretches out)

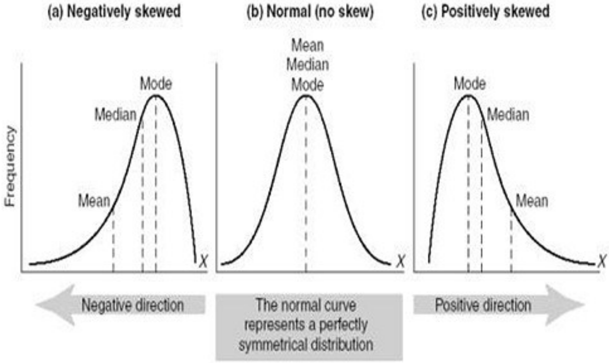


Figure 2: Distributions with Different Skew

Normal Distribution

The normal distribution can go by many names: bell-shaped, Gaussian distribution

It's very popular in statistics because it turns out that many things have a normal distribution.

In particular when we take averages of things (like heights, hours of work time, salary, etc) the averages turn out to have a normal distribution. This was proven by some very smart statisticians, and it's called the *Central Limit Theorem*. (We'll talk about this more later)

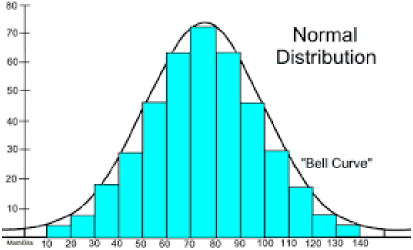


Figure 3: Normal Distribution

Center and Spread

Two characteristics that can tell us a lot about a distribution are:

Center: Where are most of the cases? Where is the middle of the distribution?

Spread: How spread out are the cases? How similar or dissimilar should we expect two cases to be?

We've already talked about a couple measures of both of these.

Center: Mean and Median, we can also add Mode (most frequent outcome)

Spread: Quartiles.

Measuring Spread

We learned that quartiles are useful for describing the data, but how do we use them to measure spread?

The **interquartile range** is a useful statistics to describe spread.

We take the 1st Quartile and subtract it from the 3rd Quartile.

Remember that the 1st Quartile is the point where 25% of the data is below this point and 75% is above. The 3rd Quartile is the point where 75% of the data is below and 25% is above.

How much of the distribution do you think will be between Q1 and Q3 (i.e., inside the interquartile range?)

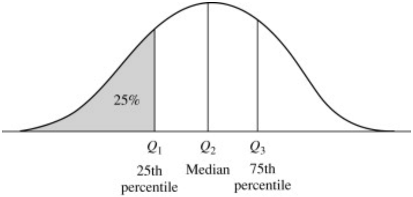


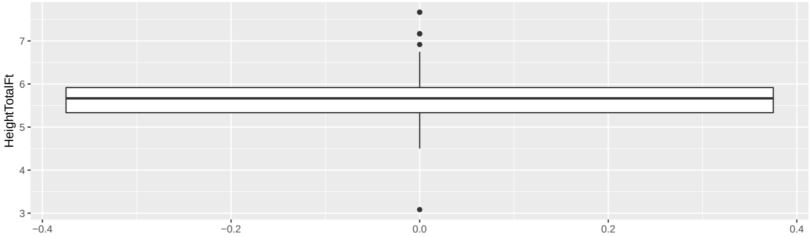
Figure 4: Interquartile Range

Box and Whisker Plots

Box and Whisker plots are a way to take the five number summary (min, Q1, Median, Q3, and max) and create a visualization.

These plots are particularly ideal for spotting unusual cases (i.e., outliers)

```
gf_boxplot(~HeightTotalFt, data = NLSdata)
```



```
summary(NLSdata$HeightTotalFt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.083   5.333   5.667   5.650   5.917   7.667
```

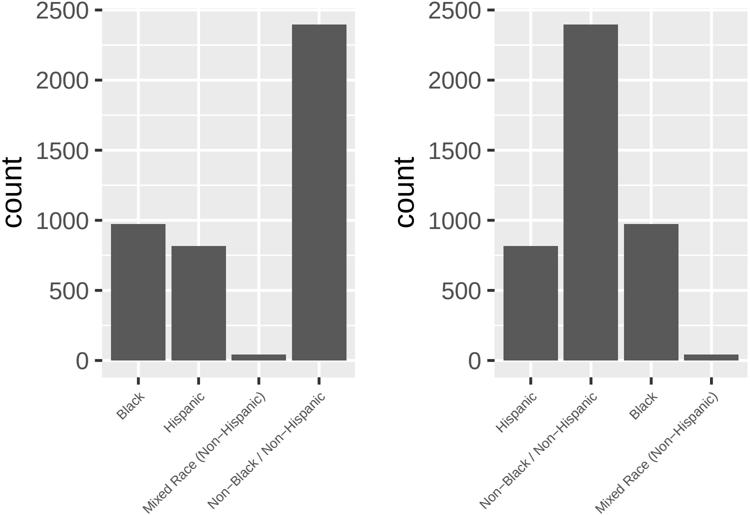
Thinking about Categorical Data

We've just learned how to describe the shape of a distribution of a quantitative variable, but what if the variable is categorical/qualitative. Does it make sense to talk about the shape and skew of a distribution of qualitative observations?

Thinking about Categorical Data

Does it make sense to talk about the shape and skew of a distribution of qualitative observations?

The ordering of the categories is arbitrary, so we could reorganize the distribution and it would look totally different!



Big Picture vs. Finer Detail

Imagine you work for a car insurance company as a data-specialist. Below are four requests. Choose the option that is most appropriate for each request:

1. We think there are generally two groups of people: people who drive a lot and people who drive very little. Can you look at our “hours of driving” data to tell us if that’s the case?
2. When do most people get their driver’s license, is there a range that’s typical for people to get their first drivers license? Can we use the data about what age people were when they got their license?
3. What is the average amount of money we spend when someone files a claim? Can you pull the data and give me your best estimate?
4. Laura in the Wisconsin office is curious what proportion of car crashes happen on the freeway as compared to arterials and other roadways. Could you give her some information about where crashes happen?

Options:

1. Interquartile Range
2. Mean
3. Histogram
4. Bar Graph with Densities

The Data Generating Process

The Data Generating Process refers to what a population (all the cases we could ever want) looks like.

Most of the time we don't know the DGP, it's the things we want to learn about.

We take small groups of cases (samples) randomly from the population, and these small groups can give us some information about what the DGP looks like (inference).

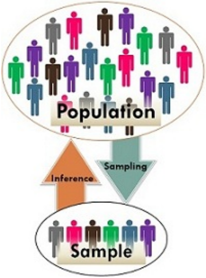


Figure 5: Population to Sample

The Data Generating Process

I'm interested in knowing what proportion of you are from the different majors on campus.

We did a PollEverywhere Question about this on the first day.

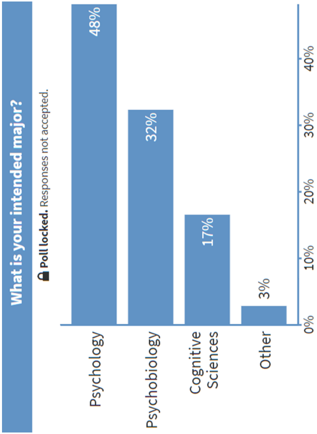


Figure 6: PollEverywhere Major

Bottom up vs. Top Down

Most of the time in statistics, we won't know what the Data Generating Process is in the population.

Without the PollEverywhere information, I might have to rely on asking a small group of students and assuming that based on the results I get from the students, I can make some conclusions about the entire class. (Bottom Up)

If I know what the DGP is (PollEverywhere Information), I can think about what samples of students might look like if I still asked small groups (Top down).

You'll notice that the Top-Down approach doesn't make much sense on it's own. Most of the time we **have to go Bottom-Up**. But by thinking about the Top-Down approach, we'll train your intuition for the Bottom Up approach.

Creating a dataset which is the population

I can use R to create a dataset where each observation is one of you!

The variable we'll look at is major.

```r
rep("Psychology", times = 5)
```

```
## [1] "Psychology" "Psychology" "Psychology" "Psychology" "Psychology"
```

```r
popdata <- data.frame(major =
          c(rep("Psychology", times = .48*200),
            rep("Psychobiology", times = .32*200),
            rep("Cognitive Sciences", times = .17*200),
            rep("Other", times = .03*200)))
```

Creating the population dataset

Let's look at what we've made and make sure it makes sense.

```
c(head(popdata),tail(popdata))
```

```
## $major
## [1] Psychology Psychology Psychology Psychology Psychology Psycholog
## Levels: Cognitive Sciences Other Psychobiology Psychology
##
## $major
## [1] Other Other Other Other Other Other
## Levels: Cognitive Sciences Other Psychobiology Psychology
```

Creating the population dataset

Let's look at what we've made and make sure it makes sense.

```
tally(~major, data = popdata)
```

```
## major
## Cognitive Sciences              Other      Psychobiology
##                 34                  6                 64
##         Psychology
##                 96
```

```
tally(~major, data = popdata, format = "proportion")
```

```
## major
## Cognitive Sciences              Other      Psychobiology
##               0.17               0.03               0.32
##         Psychology
##               0.48
```

Sampling from the population

What if I randomly called on 20 students to tell me their major?

Do you think this group will have similar proportions as the whole class?

If I called on a different group of 20 students, do you think the proportion of majors would be **exactly the same**?

Sampling Variability

Sampling variability is a term we use to mean the differences that we typically find between samples of the same size from the same population.

Not all samples will be exactly the same. I could have to ask 20 students and they might all be psychology majors. Or I could ask another group and we might get 10 Other majors.

Across each sample the distribution of majors is going to look a little bit different.

Examples of sampling variability

I'm going to randomly choose 20 observations from my dataset (this is like calling on people but a little bit faster). Based on that sample I'm going to calculate the proportion of each major from that sample.

```r
randompeople <- sample(1:200, 20)
randompeople
```

```
##  [1]  90  74  51 150  96  41 122 157 102 142  43  57  72 171 115  80
## [18] 101  76 158
```

```r
sample1 <- popdata[randompeople,]
tally(~sample1)
```

```
## sample1
## Cognitive Sciences              Other      Psychobiology
##                  1                  0                  8
##         Psychology
##                 11
```

Getting another sample

```r
sample2 <- popdata[sample(1:200, 20),]
tally(~sample2)
```

```
## sample2
## Cognitive Sciences              Other      Psychobiology
##                  4                  1                  3
##         Psychology
##                 12
```

Many many times

Let's do this many many times, and plot the proportion of the different groups.

```
manysamples <- do(1000)*
  tally(~popdata[sample(1:200, 20),], format = "proportion")
head(manysamples)
```

```
##   Cognitive.Sciences Other Psychobiology Psychology
## 1               0.25   0.1          0.40       0.25
## 2               0.10   0.0          0.45       0.45
## 3               0.20   0.0          0.25       0.55
## 4               0.25   0.0          0.30       0.45
## 5               0.25   0.0          0.45       0.30
## 6               0.20   0.0          0.30       0.50
```

Thursday 4/11/2019

Setting up R

```
NLSdata <- read.csv("http://bit.ly/NLSdata", header=TRUE)
NLSdata$HeightTotalFt <- NLSdata$HeightFt2010 +
                         NLSdata$HeightIn2010/12

popdata <- data.frame(major =
          c(rep("Psychology", times = .48*200),
            rep("Psychobiology", times = .32*200),
            rep("Cognitive Sciences", times = .17*200),
            rep("Other", times = .03*200)))
manysamples <- do(1000)*
  tally(~popdata[sample(1:200, 20),], format = "proportion")
```

Quiz on Friday

- Bring your student ID, a **charged laptop**, 1 page of notes single sided 8.5’‘x 11.5” (with your name on it)
- We will give you the Rcheatsheet
- Tonight you will get an email with a survey link, this link is yours and unique to you
- There is a password to open the quiz, we will give you the password in section on Friday
- Some questions are just **statistical thinking**, some are **statistical doing** (`R`), and some are a combination.

Practice Questions

1. Which of the following is not a measure of center of a distribution?

- A. Mean
- B. Median
- C. Interquartile Range
- D. Mode

2. If I'd like to make a smaller version of a data set where I only keep some of the variables, which of the following functions would I use?

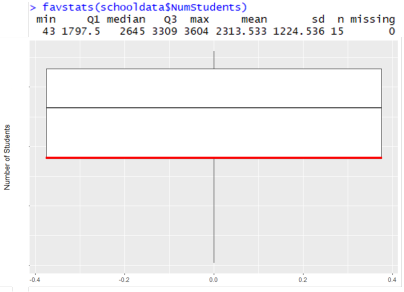
- A. select()
- B. filter()
- C. tally()
- D. recode()

Practice Questions

Below is the R output for analyzing the variable “Number of Students” in a dataset where the cases are schools. The labels on the Y-axis of the plot have been removed.

3. Based on the favstats() output, what number does the bottom of the box in the boxplot represent. This line is highlighted in red.

- A. min = 43
- B. Q1 = 1797.5
- C. median = 2645
- D. Q3 = 3309
- E. max = 3604
- F. mean = 2313.533



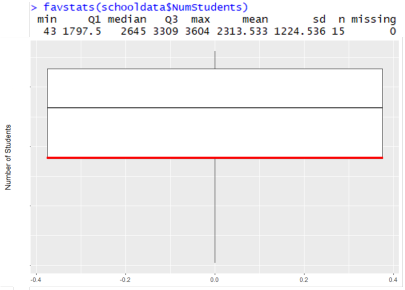
```
> favstats(schooldata$NumStudents)
 min     Q1 median   Q3  max     mean       sd  n missing
  43 1797.5   2645 3309 3604 2313.533 1224.536 15       0
```

Practice Questions

Below is the R output for analyzing the variable “Number of Students” in a dataset where the cases are schools. The labels on the Y-axis of the plot have been removed.

4. How would we interpret the bold line in the middle of the box?

- A. This is the numeric average of all the observations
- B. 50% of the data falls below this point and 50% of the data falls above this point
- C. Most of the data is stacked on this exact point
- D. None of the above



```
> favstats(schooldata$NumStudents)
 min     Q1 median   Q3  max     mean       sd  n missing
  43 1797.5   2645 3309 3604 2313.533 1224.536 15       0
```

Figure 8: R Output

Questions about the Quiz

Recapping: Sampling Variability

I'm interested in knowing what proportion of you are from the different majors on campus.

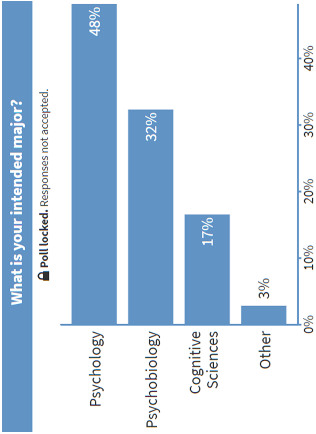


Figure 9: PollEverywhere Major

What if instead of using the population data, I took a sample? How good is the sample at approximating the population?

Our sample from class

Population

Psychology: 48%, Psychobiology: 32%, Cognitive Science: 17%, Other: 3%

Sample

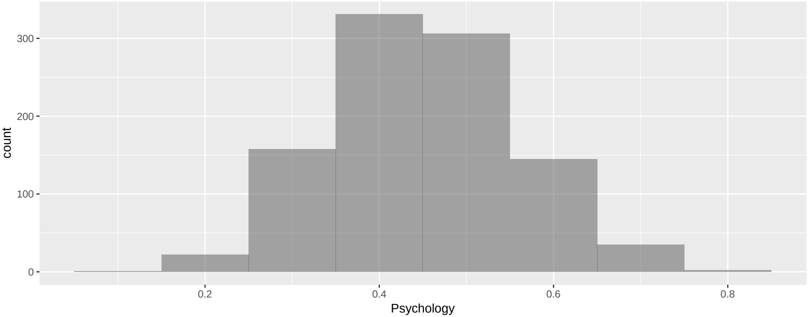
Psychology: 40%. Psychobiology: 25%, Cognitive Science: 30%, Other 5%

Sample is close, but it is not perfect! What if we took many samples? How close are samples typically to the population?

Visualizing Sampling Variability

Let's look at the different proportions that get estimated

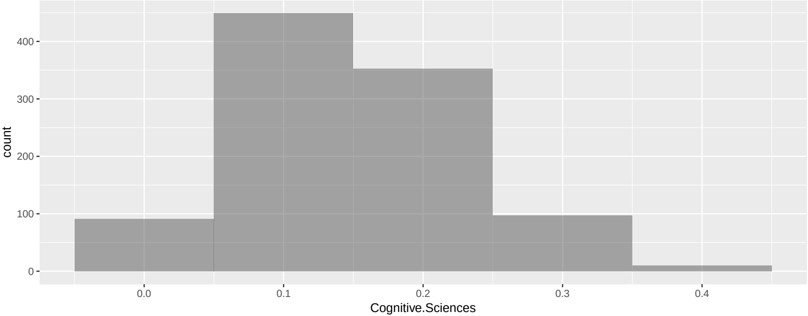
```
gf_histogram(~Psychology, data = manysamples, binwidth = .1)
```



Visualizing Sampling Variability

Let's look at the different proportions that get estimated

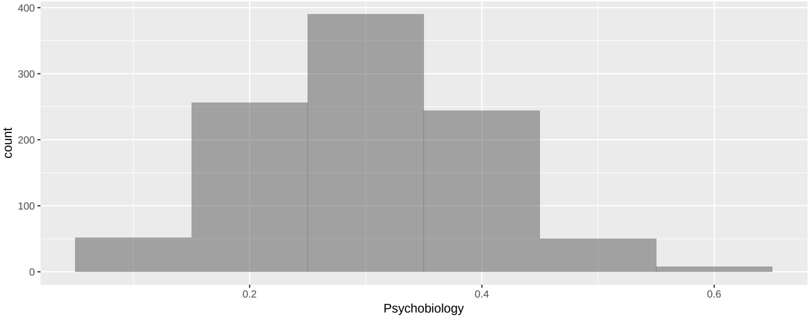
```
gf_histogram(~Cognitive.Sciences, data = manysamples, binwidth = .1)
```



Visualizing Sampling Variability

Let's look at the different proportions that get estimated

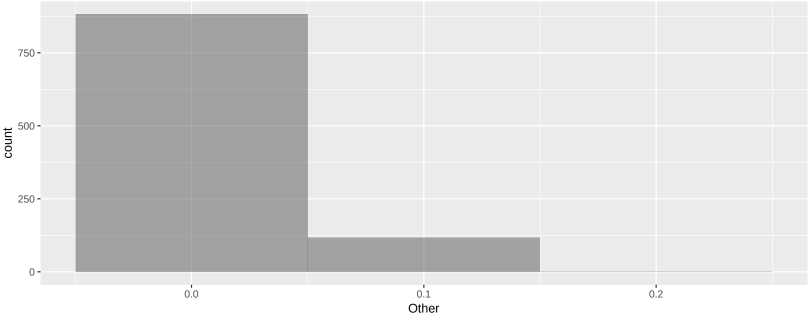
```
gf_histogram(~Psychobiology, data = manysamples, binwidth = .1)
```



Visualizing Sampling Variability

Let's look at the different proportions that get estimated

```
gf_histogram(~Other, data = manysamples, binwidth = .1)
```



Sample size and sampling variability

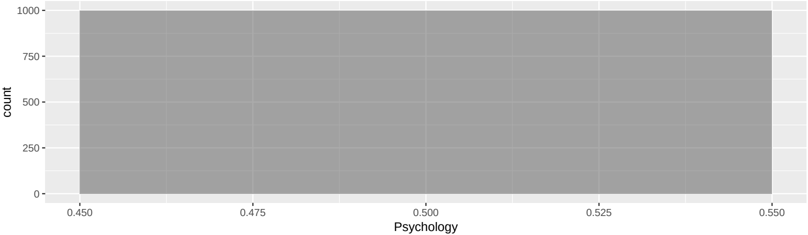
Imagine I took a sample of 200 students (Everyone in the class) what do you think the distribution of `manysamples` for Psychology would look like?

Try to use some of the words that we've learned to describe the shape of a distribution: center, spread, mean, interquartile range (IQR)

Sample size and sampling variability

Imagine I took a sample of 200 students (Everyone in the class) what do you think the distribution of `manysamples` for Psychology would look like?

```
manysamples <- do(1000)*
  tally(~popdata[sample(1:200, 200),], format = "proportion")
gf_histogram(~Psychology, data = manysamples, binwidth = .1)
```



```
summary(manysamples$Psychology)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.48    0.48    0.48    0.48    0.48    0.48
```

Sample size and sampling variability

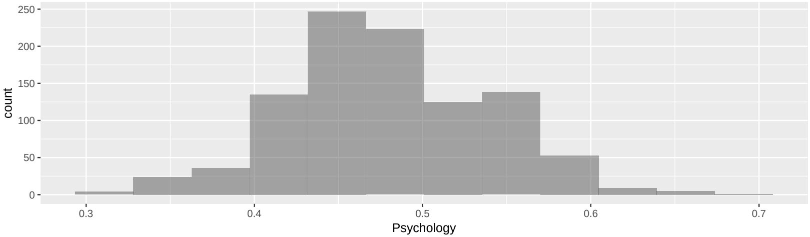
Imagine I took a sample bigger than 20, but not quite 200, let's say 50. What do you think would happen to the distribution of `many samples` for Psychology?

Try to use some of the words that we've learned to describe the shape of a distribution: center, spread, mean, interquartile range (IQR)

Sample size and sampling variability

Imagine I took a sample bigger than 20, but not quite 200, let's say 50. What do you think would happen to the distribution of `many samples` for Psychology?

```
manysamples <- do(1000)*
  tally(~popdata[sample(1:200, 50),], format = "proportion")
gf_histogram(~Psychology, data = manysamples, bins = 12)
```



```
summary(manysamples$Psychology)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.300   0.440   0.480   0.481   0.520   0.680 
```

Sample size and sampling variability

The more samples we get, the less variable the sampling distribution will be (smaller spread). The center of the distribution will also get closer to what the true value is from the DGP!

So a bigger sample will always give us a better guess as to what's going on in the population.

But! Observations cost money and time to collect, so we may want to collect a pretty big sample, but not too big.

How many samples do we need?

Use the do() function to create a sampling distribution of proportions while changing the sample size.

How many people do we need (approximately) to make a pretty good guess about the proportion of Psychobiology majors in the class.

Change N in the code below a few times, start low and get higher. Stop when the histogram is mostly around 32% (let's say most of the observations are between 30% and 34%).

Click in with your sample size

```r
N <- 20
manysamples2 <- do(1000)*
  tally(~popdata[sample(1:200, N),], format = "proportion")
gf_histogram(~Psychobiology, data = manysamples2)
summary(manysamples2$Psychobiology)
```

Explaining Variability

Does knowing someone's value on an explanatory variable, give us information about their value on the outcome variable?

An example of explaining variability

Consider our NLSdata. If our outcome variable is number of hours of sleep in 2009 which variables might explain some of the variance in HrsSleep2009.

```r
names(NLSdata)
```

```
##  [1] "ID"              "HrsSleep2009"  "Sex"           "BdayMonth"
##  [5] "YearBorn"        "Ethnicity"     "ASVAB"         "LifeSat2008"
##  [9] "Income2008"      "HeightFt2010"  "HeightIn2010"  "Weight2010"
## [13] "Computer2010"    "HrsSleep2010"  "Cohab2009"     "HeightTotalFt"
```

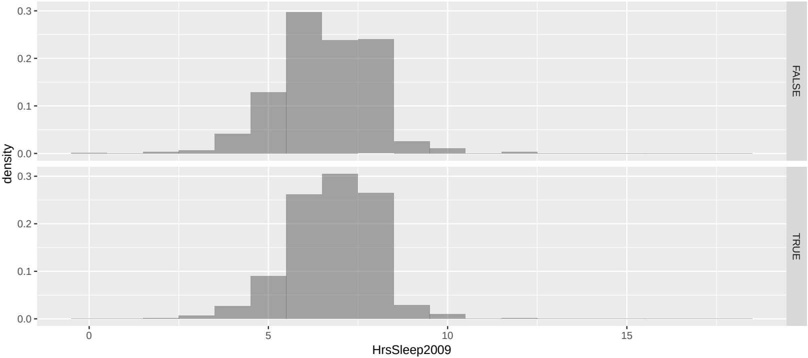
Visually exploring

Let's try Cohab2009 as an explanatory variable.

Does living with your romantic partner explain how much sleep you get?

Using a word equation: HrsSleep2009 = Cohab2009 + Otherstuff

```
gf_dhistogram(~HrsSleep2009, data = NLSdata, binwidth = 1)%>%
gf_facet_grid(Cohab2009 ~ .)
```



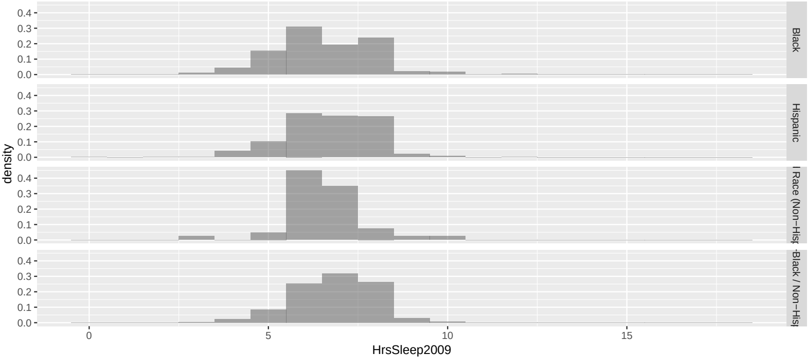
Visually exploring

Let's try Ethnicity as an explanatory variable.

People of certain ethnicities get more sleep?

Using a word equation: HrsSleep2009 $=$ Ethnicity $+$ Otherstuff

```
gf_dhistogram(~HrsSleep2009, data = NLSdata, binwidth = 1)%>%
gf_facet_grid(Ethnicity ~ .)
```



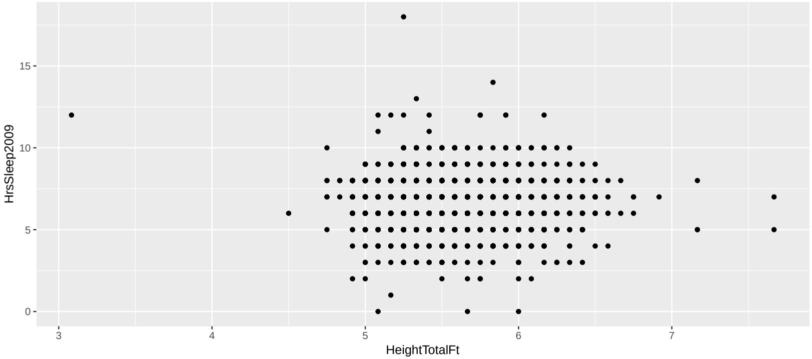
Visually exploring

Let's try HeightTotalFt as an explanatory variable.

Do taller or shorter people get more sleep?

Using a word equation: HrsSleep2009 = HeightTotalFt + Otherstuff

```
gf_point(HrsSleep2009~HeightTotalFt, data = NLSdata)
```



Visual Explorations are Imperfect

- Human brains are good at finding patterns when they aren’t there
- We want to make sure we’re making solid conclusions
- Creating statistical models can help us avoid ambiguity

Why do we need a simple model?

Ultimately, we want to know if an explanatory variable predicts an outcome variable. We can create a model, that allows this to happen, but then we need to know if it's a good model!

To know if the explanatory model is good, we need something to compare it to.

We'll compare the explanatory model to a **simple model** (AKA **Null model**)

The Null Model is a very basic model: The best guess for the outcome for any observation is the average (mean) of the outcome.

Hours of Sleep 2009

Let's look at what this model would be for Hours of Sleep in 2009

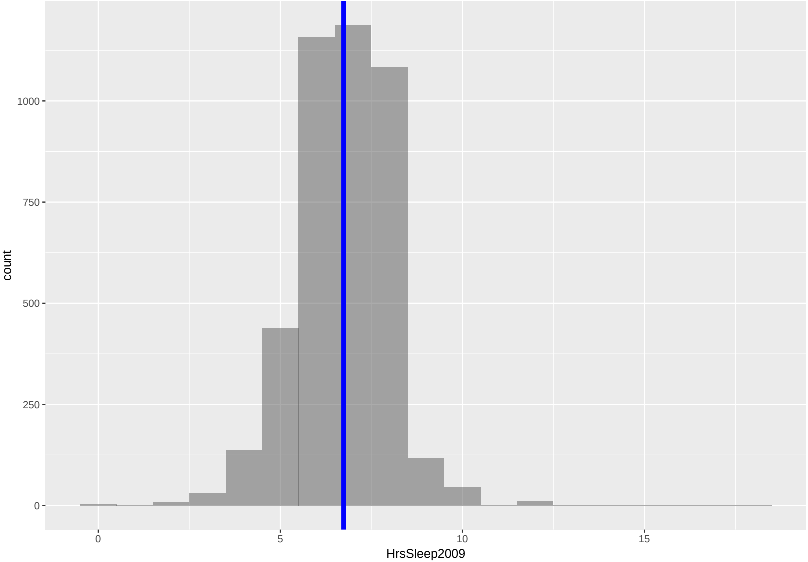
```
SleepStats <- favstats(NLSdata$HrsSleep2009)
print(SleepStats)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##    0  6      7  8  18 6.741951 1.305077 4224       0
```

Regardless of any explanatory variables, the average hours of sleep in 2009 is 6.7419!

Adding prediction to the plot

```
gf_histogram(~HrsSleep2009, data = NLSdata, binwidth = 1) %>%
  gf_vline(xintercept=~mean,data =SleepStats,color="blue",size=2)
```



Explanatory Variable: Two groups

When there are two groups, we can use the mean from each group as the prediction, and we can compare this to the simple model

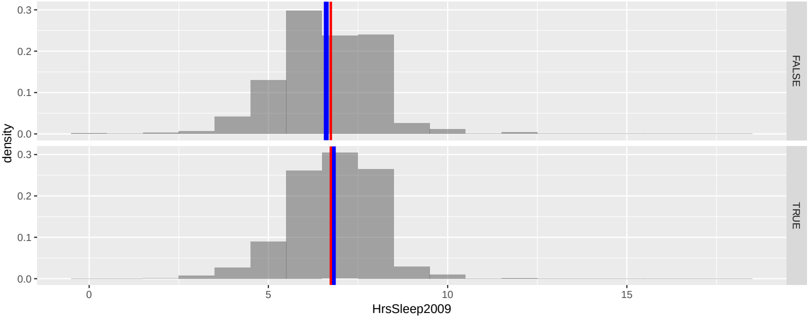
```
SleepCohabStats <- favstats(HrsSleep2009~Cohab2009, data = NLSdata)
SleepCohabStats
```

```
##   Cohab2009 min Q1 median Q3 max     mean       sd    n missing
## 1     FALSE   0  6      7  8  12 6.613682 1.361868 1491       0
## 2      TRUE   0  6      7  8  18 6.811928 1.267820 2733       0
```

Explanatory Variable: Two groups

When there are two groups, we can use the mean from each group as the prediction, and we can compare this to the simple model

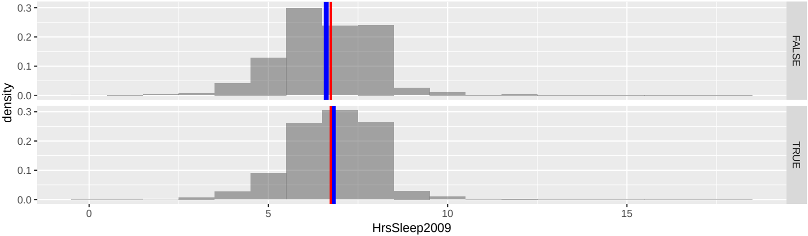
```
##   Cohab2009 min Q1 median Q3 max     mean       sd    n missing
## 1     FALSE   0  6      7  8  12 6.613682 1.361868 1491       0
## 2      TRUE   0  6      7  8  18 6.811928 1.267820 2733       0
```



Explanatory Variable: Two groups

When there are two groups, we can use the mean from each group as the prediction, and we can compare this to the simple model

```
gf_dhistogram(~HrsSleep2009, data = NLSdata, binwidth = 1)%>%
gf_facet_grid(Cohab2009 ~ .) %>%
gf_vline(xintercept=~mean,data=SleepCohabStats,color="blue",size=2)%>%
gf_vline(xintercept=~mean,data=SleepStats,color = "red",size=1)
```



Explanatory Variable: More groups

When there are more groups, we can use the mean from each group as the prediction, and we can compare this to the simple model

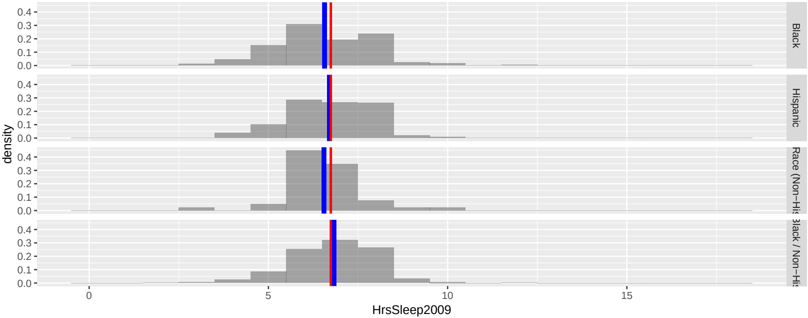
```
SleepEthnStats <- favstats(HrsSleep2009~Ethnicity, data = NLSdata)
select(SleepEthnStats, Ethnicity, mean)
```

```
##                   Ethnicity     mean
## 1                     Black 6.566290
## 2                  Hispanic 6.702206
## 3 Mixed Race (Non-Hispanic) 6.550000
## 4  Non-Black / Non-Hispanic 6.830063
```

Explanatory Variable: More groups

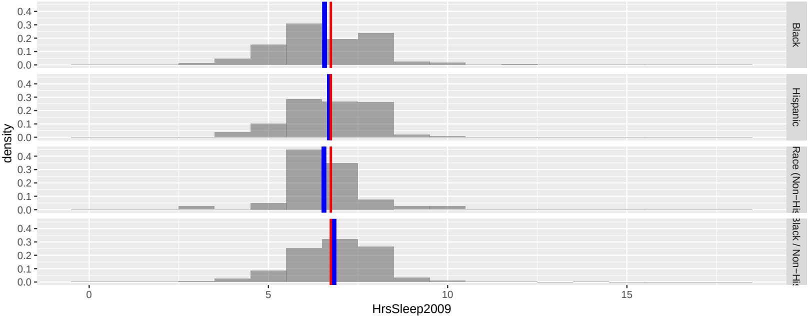
When there are more groups, we can use the mean from each group as the prediction, and we can compare this to the simple model

```
##                        Ethnicity     mean
## 1                          Black 6.566290
## 2                       Hispanic 6.702206
## 3 Mixed Race (Non-Hispanic) 6.550000
## 4  Non-Black / Non-Hispanic 6.830063
```



Explanatory Variable: More groups

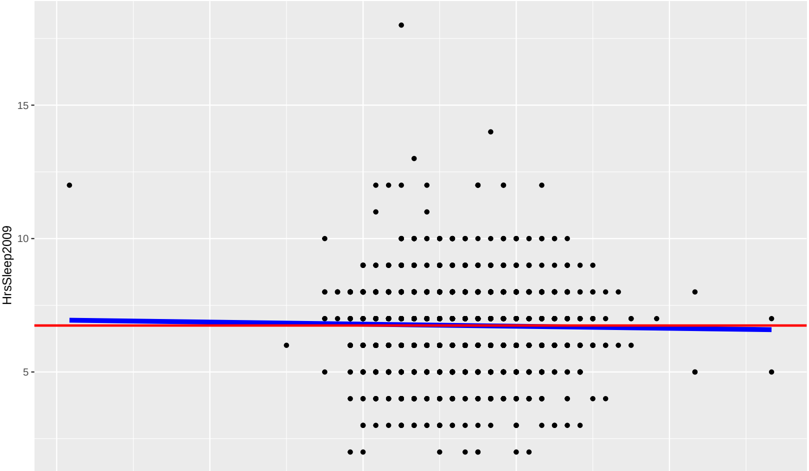
```
gf_dhistogram(~HrsSleep2009, data = NLSdata, binwidth = 1)%>%
gf_facet_grid(Ethnicity ~ .) %>%
gf_vline(xintercept=~mean,data=SleepEthnStats,color="blue",size=2)%>%
gf_vline(xintercept=~mean,data=SleepStats,color="red",size = 1)
```



Explanatory Variable: Continuous Explanatory Variable

When there are more groups, we can use the mean from each group as the prediction, and we can compare this to the simple model

```
gf_point(HrsSleep2009~HeightTotalFt, data = NLSdata)%>%
gf_lm(color = "blue", size = 2 ) %>%
  gf_hline(yintercept=~mean,data=SleepStats,color="red",size=1)
```



Before Next Time

- Study for Quiz: Chapter 1 - 4
- Chapter 5 Due: Monday 5/15 11:59pm, no more extensions
- Try the samples exercise with estimating the Cognitive Sciences major, do we need similar sample sizes for major? What about Psychology? What about Other?