Psych 100A Spring 2019: Week 4 Slides

Amanda Montoya

April 23, 2019

```r
NLSdata <- read.csv("http://bit.ly/NLSdata", header = TRUE)
```

Learning Outcomes Today

- Define a Z-score
- Illustrate the use of Z-scores in understanding individual scores when there is variability present
- Use the normal distribution to describe the probability of individual scores

```
NLSdata <- read.csv("http://bit.ly/NLSdata", header = TRUE)
```

Sums of Squares

How do we know how much error there is **in total**?

One solution is to calculate the **sum of the squared residuals** (square first, then sum)

$$\sum_{i=1}^{n} e_i^2 = SS_{residual}$$

$SS_{residual}$ is a measure of how close the observed data are to the prediction. When $SS_{residual}$ is big that means there is a lot of error. But when it's small, there is only a little bit of error.

Variance

If sums of squares is a **sum** of squared error, then variance is an **average** of squared error.

$$s^2 = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-1} = sample.variance$$

Variance is the **average squared error**, so it's no longer influenced by sample size!

Standard Deviation

Variance is an average, so it's not influenced by sample size (this is good).

But when we look at means and variances, they are very hard to compare.

Variances are in squared units, so we can't compare it directly with the mean.

Standard Deviation is the square root of variance, and so is in the same units as the mean.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{SS}{n-1}}$$

Using Standard deviation

We can calculate standard deviation in R

```
sd(NLSdata$HrsSleep2009)
```

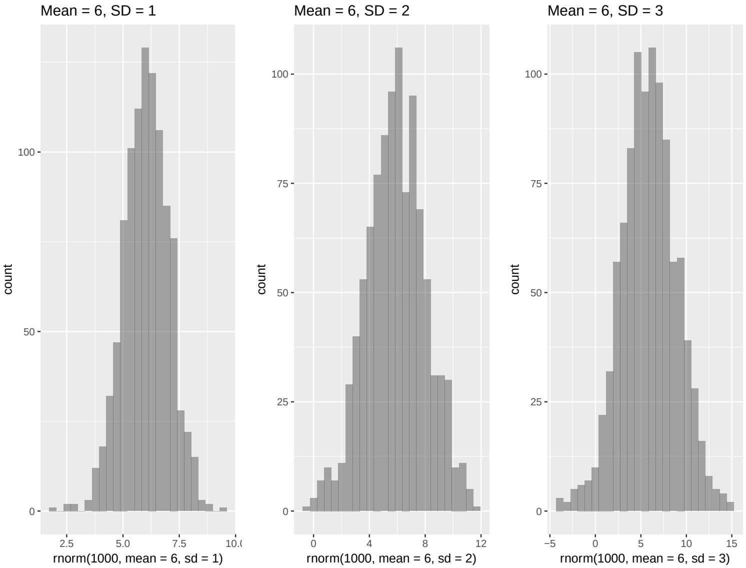
```
## [1] 1.305077
```

```
sqrt(var(NLSdata$HrsSleep2009))
```

```
## [1] 1.305077
```

Examples of different standard deviations

Normal distributions



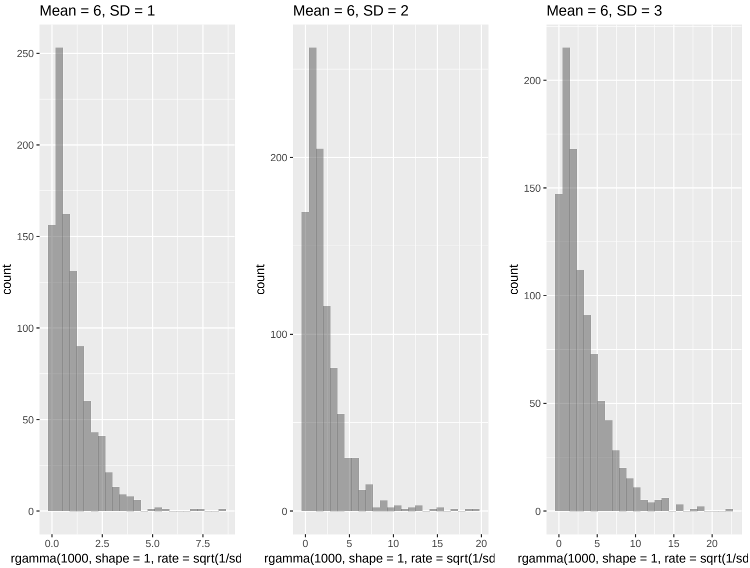
Mean = 6, SD = 1

Mean = 6, SD = 2

Mean = 6, SD = 3

Examples of different standard deviations

Gamma distributions



Mean = 6, SD = 1

Mean = 6, SD = 2

Mean = 6, SD = 3

Using standard Deviation

Regardless of the shape of the distribution, we can use standard deviation to describe how far an observation is from the mean.

Now it's easier to compare 9 to 6.7419508 using standard deviation. A person who sleep 9 hours is 1.7302035 standard deviations away from the mean. This is called a **Z-score**.

Z-score

A Z-score is a way of describing each individual, based on how many standard deviations they are away from the mean of the sample.

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

Let's make sure we understand what makes up this equation:

Describe what each element of the equation is: Y_i, $\bar{Y}$, s.

Elements of a Z-score

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

Y_i is an individuals score on an outcome variable of interest (e.g., how many hours someone sleeps)

$\bar{Y}$ is the sample average of the outcome variable (e.g., average hours people sleep)

s is the standard deviation of the outcome variable (e.g., how spread out are the individuals from the sample mean).

Breaking down a Z-score

We use Z-scores to quantify how “unusual” a score seems, compared to the rest of the distribution.

Consider the following case:

An individual sleeps 8 hours. Is that unusual, compared to others?

What proportion of people do you think sleep 8 hours or more?

Individual's score vs. the mean

Some information that might be helpful, is to compare the individual's score (i.e. $Y_i = 8$) to the mean.

```
mean(NLSdata$HrsSleep2009)
```

```
## [1] 6.741951
```

Based on this new information:

What proportion of people do you think sleep 8 hours or more?

Numerator of a Z-score

The numerator of the Z score, tells us the difference between the individual's score and the mean.

$$Y_i - \bar{Y}$$

This will tell us if the person is above the mean or below the mean, and how far away they are from the mean (in the original units of measurement)

Try it in R

We can calculate the mean based on `mean(NLSdata$HrsSleep2009)`.

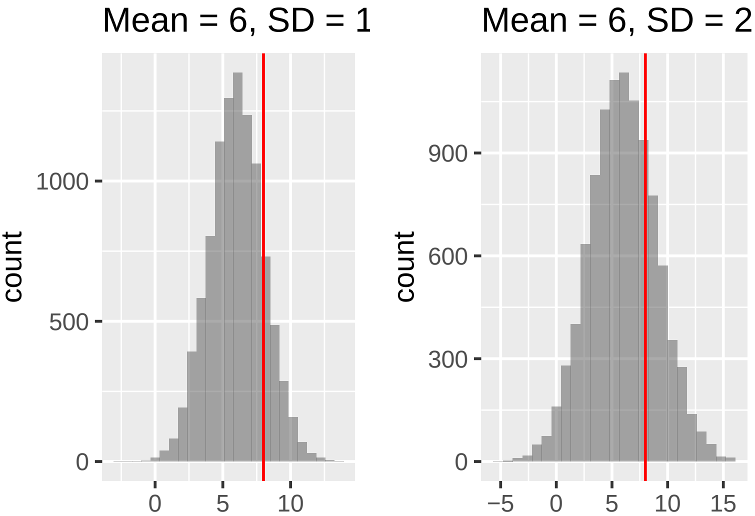
How would you create a new variable called `HrsSleep2009.num` which is the difference between someone's score on `HrsSleep2009` and the mean of `HrsSleep2009` (num stands for numerator)?

```
NLSdata$HrsSleep2009.num <-
```

Integrating standard deviation

We divide by the standard deviation, so that we can compare distributions with different spread to each other.

A score of 8 may be “unusual” for one distribution, but not for another (even if they have the same mean!).



Standard Deviation with NLSdata

The standard deviation in the NLSdata for `HrsSleep2009` is 1.3050773.

On average individuals are about 1.3 hours away from the mean (6.7).

It's pretty typical for people to be 1.3 hours away from the mean.

Remember that standard deviation captures individuals who are both below and above the mean.

Based on this information: What proportion of people do you think sleep 8 hours or more?

Understanding the Unusual

In order to understand how unusual a case is, we need information about the value for a specific case, the mean of the variable, and the standard deviation of the variable.

A *z-score* integrates the information about all three of these things together in one score.

We can say that 8 is 0.9639653 standard deviations away from the mean.

A this type of score is appropriate to compare across distributions.

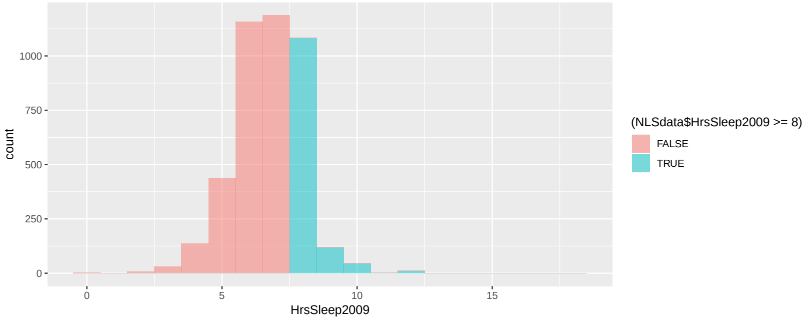
Understanding the Unusual

In our data, what was the actual proportion of people at 8 or above?

```
tally(~ HrsSleep2009 >= 8, data = NLSdata, format = "proportion")
```

```
## HrsSleep2009 >= 8
##      TRUE     FALSE
## 0.2987689 0.7012311
```

```
gf_histogram(~HrsSleep2009, data = NLSdata,
             fill = ~(NLSdata$HrsSleep2009 >= 8), binwidth = 1)
```



Creating Z-scores for Everyone

How would you create a variable in `R` which saves a z-score for everyone?

Give it a try!

We've already created the numerator, so you can either use that variable, or start from scratch!

```
NLSdata$HrsSleep2009.z <-
```

What is the z-score for the 5th person in the data? Write a one sentence interpretation of what that number **means**

Integrating the Normal Curve

In your homework you learned about the Empirical Rule.

Based on how many standard deviations away from the mean a score is, if the distribution is normal we can know exactly what proportion of the data should be above or below that score.

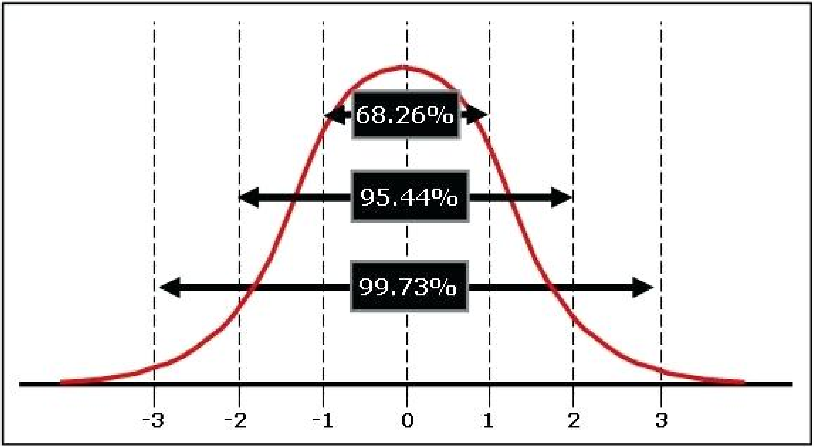
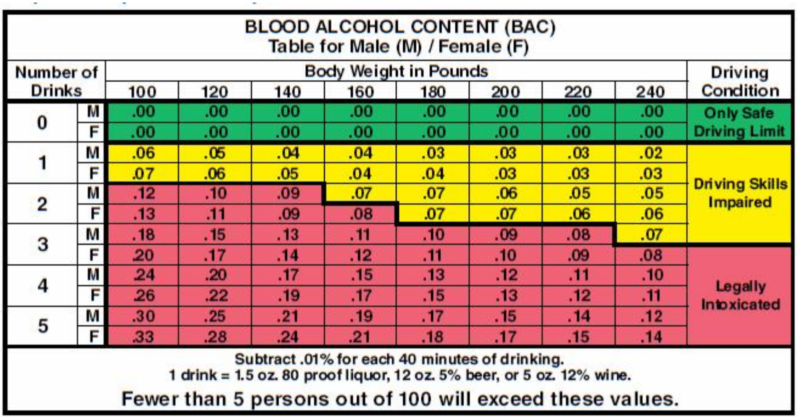


Figure 1: Empirical Rule

Applying the Normal Curve

The TAs and I are passing out a worksheet, which has a table disseminated by the DMV about Blood Alcohol Content (BAC) after drinking alcohol given sex, weight, and number of drinks.



BLOOD ALCOHOL CONTENT (BAC)
Table for Male (M) / Female (F)

Number of Drinks		Body Weight in Pounds								Driving Condition
		100	120	140	160	180	200	220	240	
0	M	.00	.00	.00	.00	.00	.00	.00	.00	Only Safe Driving Limit
	F	.00	.00	.00	.00	.00	.00	.00	.00	
1	M	.06	.05	.04	.04	.03	.03	.03	.02	Driving Skills Impaired
	F	.07	.06	.05	.04	.04	.03	.03	.03	
2	M	.12	.10	.09	.07	.07	.06	.05	.05	
	F	.13	.11	.09	.08	.07	.07	.06	.06	
3	M	.18	.15	.13	.11	.10	.09	.08	.07	
	F	.20	.17	.14	.12	.11	.10	.09	.08	Legally Intoxicated
4	M	.24	.20	.17	.15	.13	.12	.11	.10	
	F	.26	.22	.19	.17	.15	.13	.12	.11	
5	M	.30	.25	.21	.19	.17	.15	.14	.12	
	F	.33	.28	.24	.21	.18	.17	.15	.14	

Subtract .01% for each 40 minutes of drinking.
1 drink = 1.5 oz. 80 proof liquor, 12 oz. 5% beer, or 5 oz. 12% wine.
Fewer than 5 persons out of 100 will exceed these values.

Figure 2: BAC Levels

Goals

The goal of this exercise is to:

- Practice working with Z-scores
- Apply what you've learned about normal distributions and Z-scores to a new example
- Use the normal distribution to describe the probability of individual scores

Before Next Time

- Quiz, Friday 4/26
- Bring questions you have about Chapters 5 & 6 to class Thursday

Thursday 4/25/19

- Go over Z-score Worksheet
- Practice Questions for Quiz
- Introduction to Explanatory Models

Quiz on Friday

- Bring your student ID, a **charged laptop**, 1 page of notes **hand-written** 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. What is the connection between *model* and *error*?

- A. Model is what we predict for an individual person, and error is how far off that prediction is
- B. We use the model to predict new cases, but we use the errors to predict old cases
- C. Models are for categorical variables and errors are for continuous variables
- D. None of the above

2. Which two commands could I use to calculate the "model" in a simple/null model which minimizes squared error?

- A. median(), lm()
- B. mean(), lm()
- C. median(), favstats()
- D. zscore(), favstats()

Practice Questions

3. Marshawn Lynch (NFL Football Player) averaged 4.2 yards per carry in the 2018 season. His career average was 4.3 (average yards per carry across all his years of football). The standard deviation of his yards per carry is 0.4. Calculate a Z-score for Marshawn's yards per carry in 2018.

- A. Z = 4.2-4.3/2 = 2.05
- B. Z = 4.2+4.3/2 = 6.35
- C. Z = (4.2+4.3)/2 = 4.25
- D. Z = (4.2-4.3)/2 = -0.05

4. Todd Gurley averaged 4.9 yards per carry in 2018. His career average is 4.4 with a standard deviation of 0.8. Who had a more unusual year in 2018: Todd Gurley or Marshawn Lynch? (i.e., who's score in 2018 is furthest away from the mean taking into account standard deviation).

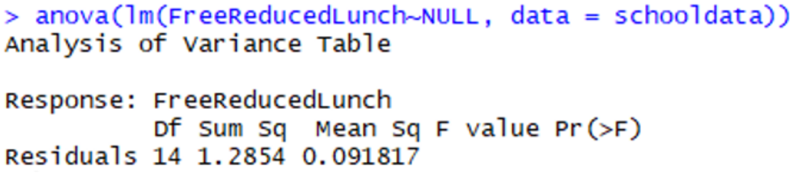
- A. Marshawn Lynch
- B. Todd Gurley
- C. It's not appropriate to compare the scores because one is positive and one is negative
- D. It's not appropriate to compare the scores because the distributions have different standard deviations

Practice Questions

Below is the R output for analyzing the variable “Proportion of Students with Free Reduced Lunch” in a dataset where the cases are schools.

5. Which of the following is true based on the output:

- A. The average deviation from the mean is 0.09, but the average squared deviation from the mean is 1.28.
- B. The average squared deviation from the mean is 0.09, but the average deviation from the mean is 1.28.
- C. The sample variance is 0.09, and there are 15 schools in the dataset
- D. The sample variance is 1.26, and there are 15 schools in the dataset



```
> anova(lm(FreeReducedLunch~NULL, data = schooldata))
Analysis of Variance Table

Response: FreeReducedLunch
          Df Sum Sq  Mean Sq F value Pr(>F)
Residuals 14 1.2854 0.091817
```

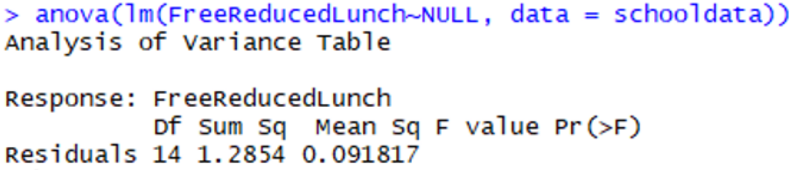
Figure 3: R Output

Practice Questions

Below is the R output for analyzing the variable “Proportion of Students with Free Reduced Lunch” in a dataset where the cases are schools.

6. What’s an alternative way to calculate Sum Sq and Mean Sq from the ANOVA table? Imagine there is an object which is the null model fit with the schools data: emptymodel <- lm(FreeReducedLunch~NULL, data = schooldata)

- A. Sum Sq: sum(resid(emptymodel)ˆ2), Mean Sq: var(schooldata$FreeReducedLunch)
- B. Sum Sq: sum(resid(emptymodel))ˆ2, Mean Sq: var(schooldata$FreeReducedLunch)
- C. Sum Sq: sum(resid(emptymodel)ˆ2), Mean Sq: sum(resid(emptymodel)ˆ2)/n
- D. Sum Sq: sum(resid(emptymodel))ˆ2, Mean Sq: sum(resid(emptymodel))ˆ2/n



```
> anova(lm(FreeReducedLunch~NULL, data = schooldata))
Analysis of Variance Table

Response: FreeReducedLunch
          Df Sum Sq  Mean Sq F value Pr(>F)
Residuals 14 1.2854 0.091817
```

Questions about the Quiz

Adding Explanatory Variables to Models

Up to this point we've really only looked at variation.

We haven't done much to **explain** it!

But all of the tools that we've developed for quantifying variation will come in handy when we're explaining variation.

Tools In Your Belt:

- Visualizing Variability: Histograms, Box Plots, Dotplots
- Fitting a linear model
- Visualizing a model: Adding a prediction to a histogram
- Predicting Values
- Quantifying Residuals
- Quantifying Total Error: Sums of Squares
- Describing Variability: Variance and Standard Deviation

Hours of Sleep 2009: A simple model

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
```

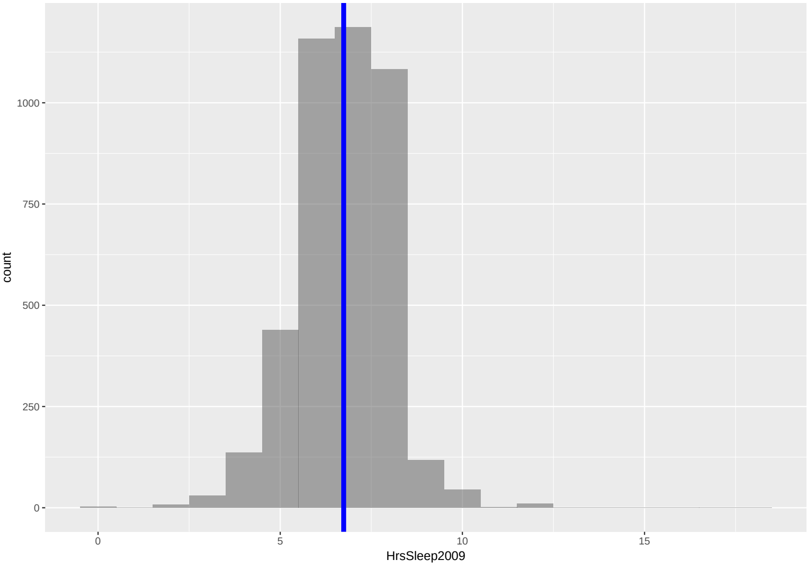
```
emptymodel <- lm(HrsSleep2009~NULL, data = NLSdata)
emptymodel
```

```
## 
## Call:
## lm(formula = HrsSleep2009 ~ NULL, data = NLSdata)
## 
## Coefficients:
## (Intercept)  
##       6.742
```

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)
```



The Two Group Model

We've spent a lot of time talking about Hours of Sleep, but what exactly predicts how many hours of sleep one gets?

Let's consider an empirical question: Does living with your significant other predict how many hours of sleep you get?

Representing a Model with a Word Equation:

$$Outcome = Model + Error$$

$$Sleep = Cohab + Error$$

Representing a model with GLM notation

$$Y_i = b_0 + b_1 X_i + e_i$$

Here Y_i is person i's hours of sleep. X_i is person i's cohabitation status ($0 =$ No, $1 =$ Yes). e_i is the difference between the prediction and the actual value of Y_i

We'll talk about b_0 and b_1 shortly

Fitting Separate Models (No Cohab)

Imagine we only have the data for the no-cohab group.

```
nocohab <- subset(NLSdata, Cohab2009 == 0)
head(select(nocohab, HrsSleep2009, Cohab2009))
```

```
##    HrsSleep2009 Cohab2009
## 1             7     FALSE
## 2             8     FALSE
## 5             6     FALSE
## 7             7     FALSE
## 8             6     FALSE
## 12            7     FALSE
```

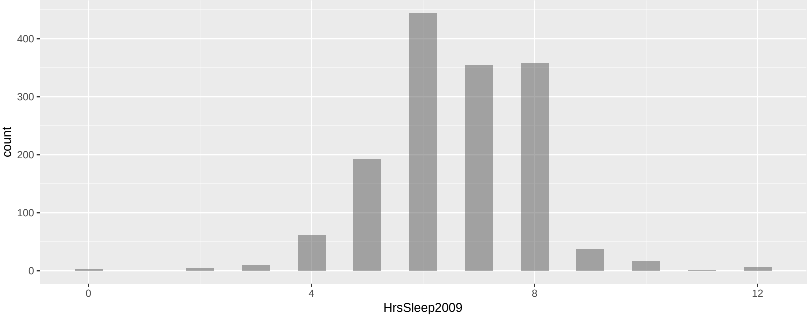
Fitting Separate Models (No Cohab)

Imagine we only have the data for the no-cohab group.

```
favstats(~HrsSleep2009, data = nocohab)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##    0  6      7  8  12 6.613682 1.361868 1491       0
```

```
gf_histogram(~HrsSleep2009, data = nocohab)
```



Fitting Separate Models (No Cohab)

If I wanted to know what the typical Hours of Sleep in 2009 is for the nocohab group, I might decide to fit a simple model.

```r
nocohabmodel <- lm(HrsSleep2009~NULL, data = nocohab)
nocohabmodel
```

```
##
## Call:
## lm(formula = HrsSleep2009 ~ NULL, data = nocohab)
##
## Coefficients:
## (Intercept)
##       6.614
```

This model would predict the mean hours of sleep (not for all people) but the average across all people who do not live with their partner.

Fitting Separate Models (No Cohab)

I can examine how well the model fits by looking at Sums of Squares and variance/MSE

```
anova(nocohabmodel)
```

```
## Analysis of Variance Table
##
## Response: HrsSleep2009
##              Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 1490 2763.5  1.8547
```

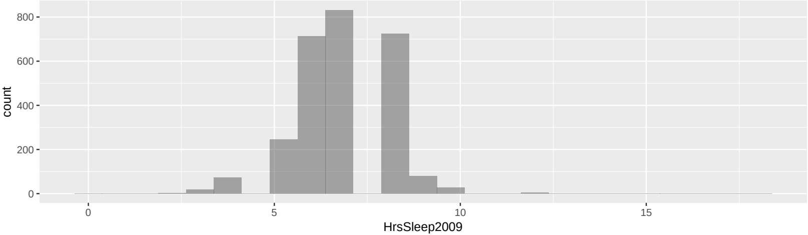
Fitting Separate Models (Cohab)

We can do something similar for those who do live with their partner

```
cohab <- subset(NLSdata, Cohab2009 == 1)
favstats(~HrsSleep2009, data = cohab)
```

```
##  min Q1 median Q3 max     mean      sd    n missing
##    0  6      7  8  18 6.811928 1.26782 2733       0
```

```
gf_histogram(~HrsSleep2009, data = cohab)
```



Fitting Separate Models

I could also fit a simple model to estimate typical hours of sleep for those who do cohabitate with their romantic partner.

```
cohabmodel <- lm(HrsSleep2009~NULL, data = cohab)
cohabmodel
```

```
##
## Call:
## lm(formula = HrsSleep2009 ~ NULL, data = cohab)
##
## Coefficients:
## (Intercept)
##       6.812
```

```
anova(nocohabmodel)
```

```
## Analysis of Variance Table
##
## Response: HrsSleep2009
##             Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 1490 2763.5  1.8547
```

Prediction?

If having information about whether someone lives with their partner helps us know better how long someone sleeps, we would say that cohabitation predicts sleep!

But to know whether cohabitation predicts sleep we need to fit the model with all people together.

Explanatory Variable: Two groups

When there are two groups, we can use a general linear model to fit the data, which will predict the group mean for individuals in each group (i.e., $\bar{Y}_{cohab}$ or $\bar{Y}_{nocohab}$ depending on which group someone is in)

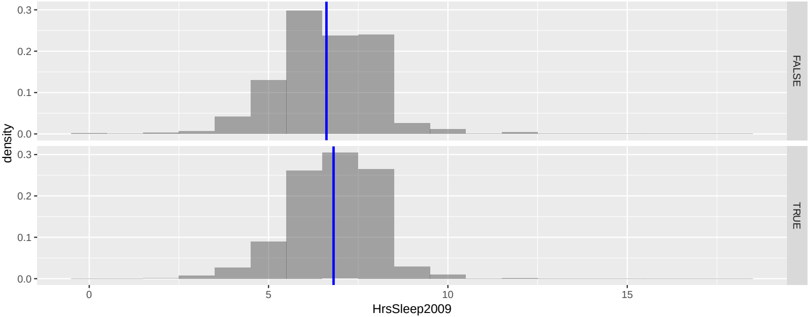
```
SleepCohabStats <- favstats(HrsSleep2009~Cohab2009, data = NLSdata)
SleepCohabStats
```

```
##   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

```
##   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
```



Fitting a linear model

We can fit a linear model using Cohab2009 as a predictor using R

```r
Cohab.model <- lm(HrsSleep2009~Cohab2009, data = NLSdata)
Cohab.model
```

```
##
## Call:
## lm(formula = HrsSleep2009 ~ Cohab2009, data = NLSdata)
##
## Coefficients:
##   (Intercept)  Cohab2009TRUE
##        6.6137         0.1982
```

These coefficients correspond to our general linear model notation

$$\hat{Y} = b_0 + b_1 X_i$$

b_0: Intercept b_1: Coefficient for X_i

$$\hat{Y} = 6.614 + 0.198 X_i$$

Predicting Y

X_i is an indicator which says whether or not someone lives with their partner.

Let's think about someone who does not live with their partner: $X_i = 0$

$$\hat{Y} = b_0 + b_1 X_i$$

$$\hat{Y} = 6.614 + 0.198 \times 0 = 6.614$$

The predicted Y for someone who does not live with their partner is 6.614.

The intercept will always be the predicted Y_i for individuals with a score of 0 on X_i.

Predicting Y

Let's think about someone who does live with their partner: $X_i = 1$

$$\hat{Y} = b_0 + b_1 X_i$$

$$\hat{Y} = 6.614 + 0.198 \times 1 = 6.614 + 0.198 = 6.812$$

The predicted Y for someone who does not live with their partner is 6.6812.

For the two group model, the coefficient for X_i (b_1) will always be the difference between the Group coded as 0 and the Group coded as 1.

In general, b_1 will always be the change in predicted Y_i with a one unit increase in X_i.

Calculating Predictions

```r
NLSdata$CohabPred <- predict(Cohab.model)

head(select(NLSdata, HrsSleep2009, Cohab2009, CohabPred))
```

```
##   HrsSleep2009 Cohab2009 CohabPred
## 1            7     FALSE  6.613682
## 2            8     FALSE  6.613682
## 3            6      TRUE  6.811928
## 4            5      TRUE  6.811928
## 5            6     FALSE  6.613682
## 6            5      TRUE  6.811928
```

Calculating Residuals

```r
NLSdata$CohabResid <- resid(Cohab.model)

head(select(NLSdata, HrsSleep2009, Cohab2009, CohabPred, CohabResid))
```

```
##   HrsSleep2009 Cohab2009 CohabPred CohabResid
## 1            7     FALSE  6.613682  0.3863179
## 2            8     FALSE  6.613682  1.3863179
## 3            6      TRUE  6.811928 -0.8119283
## 4            5      TRUE  6.811928 -1.8119283
## 5            6     FALSE  6.613682 -0.6136821
## 6            5      TRUE  6.811928 -1.8119283
```

Next Time

- Quantifying Error in the Two Group Model (Sums of Squares)
- Evaluating Model: Proportional Reduced Error
- More than Two Groups

Before Next Time

- Quiz Friday 4/26 10am Rolfe 1200
- Chapter 7 Homework Due Monday 4/29 11:59pm