

431 Class 07

Thomas E. Love

2017-09-19

- ① Project Instructions
- ② Assignment 1 debrief (Assignment 2 due 2017-09-22)
- ③ Course Notes
 - Identifying and quantifying outliers (Ch 7)
 - Using transformations to “normalize” data (Ch 9)
- ④ Which states are in the midwest? (survey results)

R! Me hearties.



Project Instructions

All materials and information related to the project will be maintained at
<https://github.com/THOMASELOVE/431project>
Regular updates will appear there throughout the semester.

Comments on Assignments 1 and 2

Answer Sketch for Assignment 1 posted on Friday afternoon. The PDF is password protected, rather than “invalid” as Github suggests when you try to display it.

See

<https://github.com/THOMASELOVE/431homework/tree/master/HW1>.

- ➊ R Markdown for Answer Sketch is also available.
- ➋ Grading rubric and Grades for Assignment 1 went up this morning.
- ➌ Answer sketches are always longer than we expect your answers to be.
- ➍ I strongly suggest you look over Claudia's slide deck, as linked in the Class 7 README.

Our Favorite Homework Assignments

Also posted to <https://github.com/THOMASELOVE/431homework>

- 1 Submit HTML and R Markdown, without zipping the files together.
- 2 Use complete sentences, and proper English grammar, syntax and spelling.
- 3 Avoid printing things we don't need to see (like data frames) to judge your work.
- 4 Place empty lines before and after each code chunk.
- 5 Name every chunk, with no characters other than letters, numbers and underscores. No repetition!
- 6 Use headings to indicate the question being solved. Then leave a blank line before continuing with text.
- 7 Avoid repeating Dr. Love's questions verbatim.
- 8 Use the approach laid out on the next slides (and in the YOURNAME-hw_template.Rmd file on the 431 Homework page)

YOURNAME-hw_template.Rmd YAML

```
1 ---
2 title: "Assignment INSERT-NUMBER for 431"
3 author: "YOUR-NAME-GOES-HERE"
4 date: "`r Sys.Date()`"
5 output:
6   html_document:
7     toc: true
8     df_print: paged
9     code_folding: show
10 ---
11
```

YOURNAME-hw_template.Rmd Setup

```
11
12 ▾ ## Setup
13
14 ▾ ```{r setup, include=FALSE}
15 knitr::opts_chunk$set(comment = NA)
16
17 # insert other packages you need here
18 library(tidyverse)
19 ```
20
21 ▾ ```{r get_data}
22 # use read.csv("datasetname.csv") %>% tbl_df
23 ```
24
```



```
24
25 ▾ # Question 1
26
27 Answer goes here.
28
29 ▾ ```{r q01_name_of_chunk}
30 # r code goes here
31 ```
32
33 Discussion of result from code goes here.
34
35 ▾ # Question 2
36
37 etc.
```

from Class 5 Google Form

Which of the following statements best describes your current status regarding Assignment 1? *

- ☐ I have completed the Assignment.
- ☐ I have started the Assignment, but not completed it yet.
- ☐ I have not yet begun to work on the Assignment.

How confident are you in your ability to successfully complete Assignment 1? *

| | 1 | 2 | 3 | 4 | 5 | |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------|
| Not confident at all. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Very confident. |

Current Status vs. HW 1 Result

Loading tidyverse: ggplot2

Loading tidyverse: tibble

Loading tidyverse: tidyr

Loading tidyverse: readr

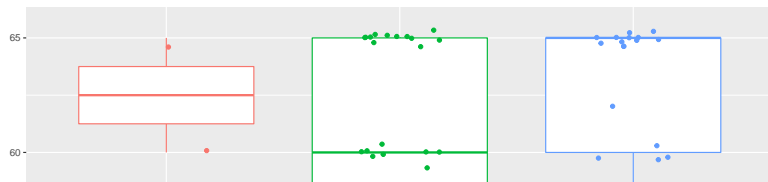
Loading tidyverse: purrr

Loading tidyverse: dplyr

Conflicts with tidy packages -----

filter(): dplyr, stats

lag(): dplyr, stats



Current Status vs. HW 1 Result

Code for prior slide.

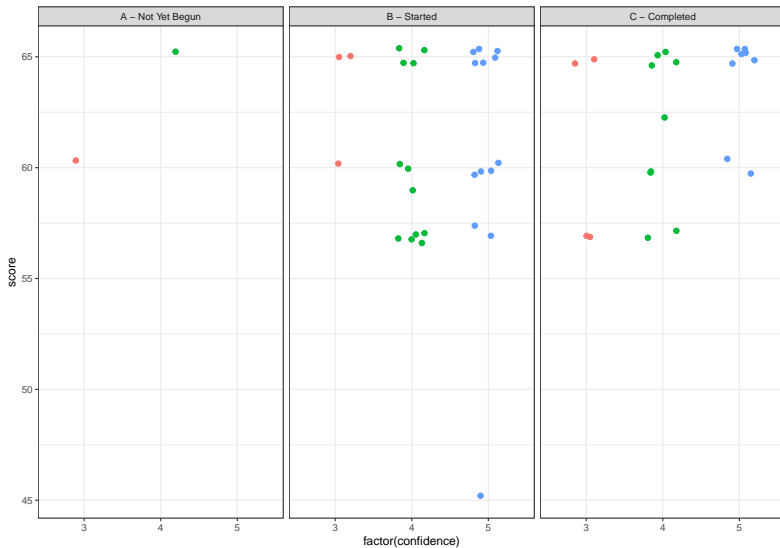
```
library(tidyverse)
hw1_f5 <- read.csv("eval_f5.csv") %>% tbl_df()
ggplot(data = hw1_f5,
       aes(x = status_class5, y = score,
           col = status_class5)) +
  geom_boxplot() +
  geom_jitter(width = 0.2) +
  guides(color = FALSE)
```


Confidence Level vs. HW 1 Result (Shape)

Code for prior slide.

```
ggplot(data = hw1_f5,  
       aes(x = factor(confidence), y = score,  
           col = factor(confidence),  
           shape = status_class5)) +  
  geom_jitter(size = rel(3), width = 0.1) +  
  guides(col = FALSE)
```

Confidence Level vs. HW 1 Result (Facets)



Confidence Level vs. HW 1 Result (Facets)

Code for prior slide.

```
ggplot(data = hw1_f5,  
       aes(x = factor(confidence), y = score,  
           col = factor(confidence))) +  
  geom_jitter(size = 2, width = 0.2) +  
  guides(col = FALSE, shape = FALSE) +  
  facet_wrap(~ status_class5) +  
  theme_bw()
```


Preliminaries for Today

```
library(NHANES); library(magrittr); library(viridis)
library(gridExtra); library(tidyverse)
```

```
source("Love-boost.R")
```

```
nh_temp <- NHANES %>%
  filter(SurveyYr == "2011_12") %>%
  filter(Age >= 21 & Age < 65) %>%
  mutate(Sex = Gender, Race = Race3,
         SBP = BPSysAve, DBP = BPDiaAve) %>%
  select(ID, Sex, Age, Race, Education,
         BMI, SBP, DBP, Pulse, PhysActive,
         Smoke100, SleepTrouble, HealthGen)
```

Random Sample of 500 to get nh_adults

```
set.seed(431002)
# use set.seed to ensure that
# we all get the same random sample

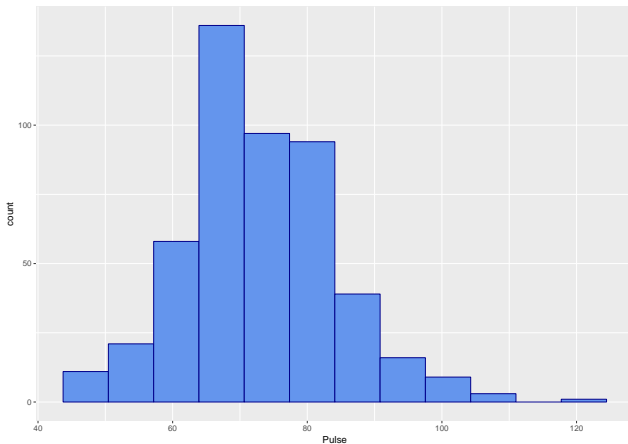
nh_adults <- sample_n(nh_temp, size = 500)
```

What Summaries to Report (from Section 7.17)

It is usually helpful to focus on the shape, center and spread of a distribution. Bock, Velleman and DeVeaux provide some useful advice:

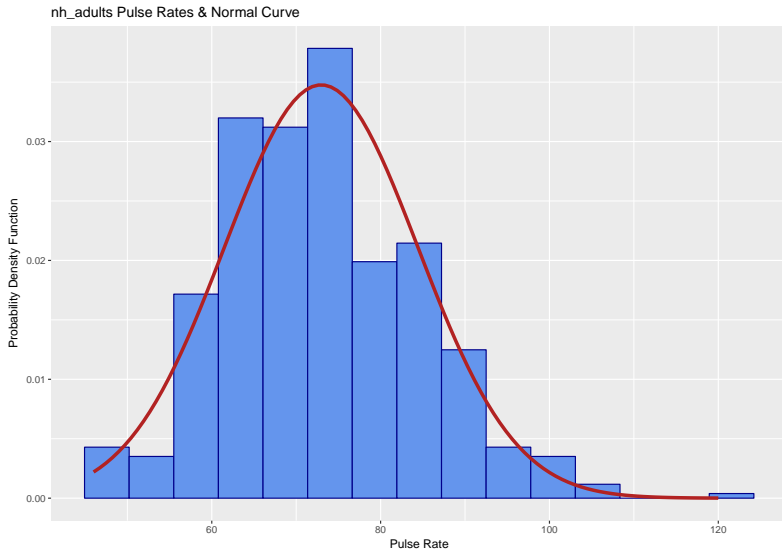
- If the data are skewed, report the median and IQR (or the three middle quantiles). You may want to include the mean and standard deviation, but you should point out why the mean and median differ. The fact that the mean and median do not agree is a sign that the distribution may be skewed. A histogram will help you make that point.
- If the data are symmetric, report the mean and standard deviation, and possibly the median and IQR as well.
- If there are clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. The median and IQR are not likely to be seriously affected by outliers.

Do the Pulse data appear skewed? Outlier-prone?



For Pulse, $skew1 = 0.08$, and $kurtosis = 0.4$

Pulse Rates + Normal Model

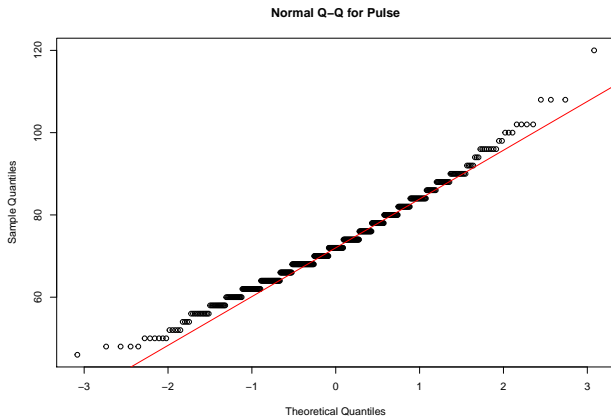


Pulse Rates + Normal Model (Code)

```
nh_adults %>% filter(!is.na(Pulse)) %>%  
ggplot(., aes(x = Pulse)) +  
  geom_histogram(aes(y = ..density..), bins=15,  
                 fill = "cornflowerblue",  
                 color = "blue4") +  
  stat_function(fun = dnorm,  
               args = list(  
                 mean = mean(nh_adults$Pulse, na.rm=TRUE),  
                 sd = sd(nh_adults$Pulse, na.rm=TRUE)),  
               lwd = 1.5, col = "firebrick") +  
  labs(title = "nh_adults Pulse Rates & Normal Curve",  
       x = "Pulse Rate",  
       y = "Probability Density Function")
```

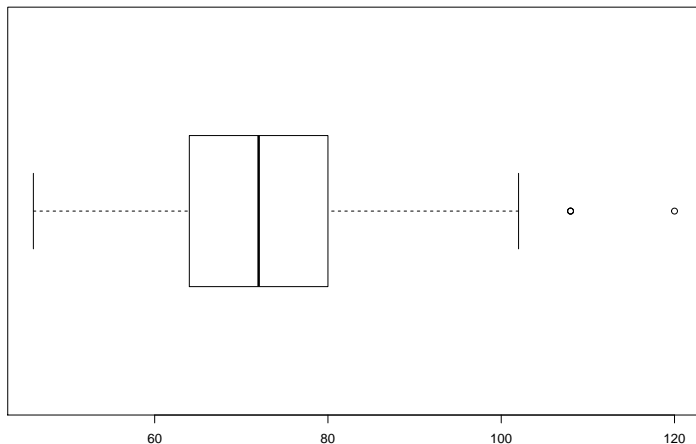
Normal Q-Q plot for Pulse

```
qqnorm(nh_adults$Pulse, main = "Normal Q-Q for Pulse")  
qqline(nh_adults$Pulse, col = "red")
```



Identifying outliers (with a boxplot)

```
boxplot(nh_adults$Pulse, horizontal = TRUE)
```



How the Boxplot identifies Outlier Candidates

Calculate the upper and lower (inner) fences. Points outside that range are candidate outliers. If $IQR = 75^{\text{th}} \text{ percentile} - 25^{\text{th}} \text{ percentile}$, then

- Upper fence = $75^{\text{th}} \text{ percentile} + 1.5 \text{ IQR}$
- Lower fence = $25^{\text{th}} \text{ percentile} - 1.5 \text{ IQR}$

Let us consider the SBP data again. There, we have

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|-------|---------|-------|
| 84.0 | 109.0 | 118.0 | 118.6 | 127.0 | 202.0 |
| NA's | | | | | |
| 15 | | | | | |

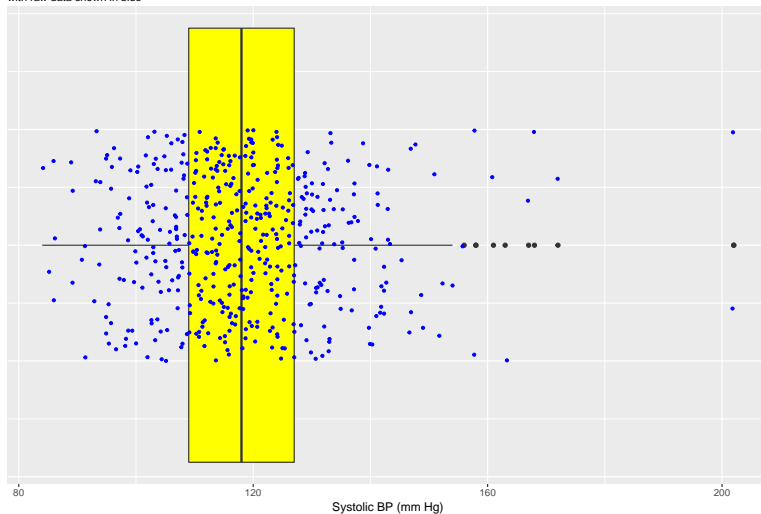
Identifying outliers (with ggplot and a boxplot)

Here is the code (adapted from Course Notes 7.10.1)

```
nh_adults %>%  
  filter(!is.na(SBP)) %>%  
  ggplot(., aes(x = 1, y = SBP)) +  
  geom_boxplot(fill = "yellow") +  
  geom_point(col = "blue", size = 0.4) +  
  coord_flip() +  
  theme(axis.text.y = element_blank(),  
        axis.ticks.y = element_blank()) +  
  labs(title = "Boxplot of SBP for nh_adults",  
        subtitle = "with raw data shown in blue",  
        x = "", y = "Systolic BP (mm Hg)")
```

The Resulting Boxplot

Boxplot of SBP for nh_adults
with raw data shown in blue



Identifying outliers (with Z scores) (Section 8.2)

The maximum systolic blood pressure in the data is NA.

```
nh_adults %>%  
  filter(!is.na(SBP)) %$%  
  mosaic::favstats(SBP)
```

| min | Q1 | median | Q3 | max | mean | sd | n | missing |
|-----|-----|--------|-----|-----|----------|----------|-----|---------|
| 84 | 109 | 118 | 127 | 202 | 118.5918 | 15.30267 | 485 | 0 |

But how unusual is that value? One way to gauge how extreme this is (or how much of an outlier it is) uses that observation's **Z score**, the number of standard deviations away from the mean that the observation falls.

Z score for SBP = 202

Z score =

$$\frac{\text{value} - \text{mean}}{sd}$$

.

For the SBP data, the mean = 118.6 and the standard deviation is 15.3, so we have Z score for 202 =

$$\frac{202 - 118.6}{15.3} = 83.415.3 = 5.45$$

.

- A negative Z score indicates a point below the mean
- A positive Z score indicates a point above the mean
- The Empirical Rule suggests that for a variable that followed a Normal distribution, about 95% of observations would have a Z score in (-2, 2) and about 99.7% would have a Z score in (-3, 3).

How unusual is a value as extreme as $Z = 5.45$?

If the data really followed a Normal distribution, we could calculate the probability of obtaining as extreme a Z score as 5.45.

A Standard Normal distribution, with mean 0 and standard deviation 1, is what we want, and we want to find the probability that a random draw from such a distribution would be 5.45 or higher, *in absolute value*. So we calculate the probability of 5.45 or more, and add it to the probability of -5.45 or less, to get an answer to the question of how likely is it to see an outlier this far away from the mean.

```
pnorm(q = 5.45, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 2.518491e-08
```

```
pnorm(q = -5.45, mean = 0, sd = 1, lower.tail = TRUE)
```

```
[1] 2.518491e-08
```

But the Normal distribution is symmetric

```
2*pnorm(q = 5.45, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 5.036982e-08
```

The probability that a single draw from a Normal distribution with mean 0 and standard deviation 1 will produce a value as extreme as 5.45 is 0.00000005

The probability that a single draw from a Normal distribution with mean 118.6 and standard deviation 15.3 will produce a value as extreme as 202 is also 0.00000005, since the Normal distribution is completely characterized by its mean and standard deviation.

So, is 202 an outlier here? Do the SBP data look like they come from a Normal distribution?

Fences and Z Scores

Note the relationship between the fences (Tukey's approach to identifying points which fall within the whiskers of a boxplot, as compared to candidate outliers) and the Z scores.

| min | Q1 | median | Q3 | max | mean | sd | n | missing |
|-----|-----|--------|-----|-----|----------|----------|-----|---------|
| 84 | 109 | 118 | 127 | 202 | 118.5918 | 15.30267 | 485 | 0 |

For the SBP data, the IQR is $127 - 109 = 18$, so

- the upper inner fence is at $127 + 1.5*(18)$, or 154, and
- the lower inner fence is at $109 - 1.5*(18)$, or 82.
- Since the mean is 118.6 and the standard deviation is 15.3,
 - the Z score for the upper inner fence is 2.31, and
 - the Z score for the lower inner fence is -2.39
- It is neither unusual nor inevitable for the inner fences to fall at Z scores near -2.0 and +2.0.

Summing Up: Does a Normal Model fit well?

If a Normal model fits our data well, then we should see the following graphical indications:

- 1 A histogram that is symmetric and bell-shaped.
- 2 A boxplot where the box is symmetric around the median, as are the whiskers, without a serious outlier problem.
- 3 A normal Q-Q plot that essentially falls on a straight line.

As for numerical summaries, we'd like to see

- 4 The mean and median within 0.2 standard deviation of each other.
- 5 No real evidence of too many outlier candidates (more than 5% starts to get us concerned about a Normal model)
- 6 No real evidence of individual outliers outside the reasonable range for the size of our data (we might expect about 3 observations in 1000 to fall more than 3 standard deviations away from the mean.)

Should our data not be well-modeled by the Normal, what can we do?

The Ladder of Power Transformations (Section 9)

The key notion in re-expression of a single variable to obtain a more normal distribution or re-expression of an outcome in a simple regression model is that of a **ladder of power transformations**, which can apply to any unimodal data.

| Power | Transformation |
|-------|-------------------------|
| 3 | x^3 |
| 2 | x^2 |
| 1 | x (unchanged) |
| 0.5 | $x^{0.5} = \sqrt{x}$ |
| 0 | $\ln x$ |
| -0.5 | $x^{-0.5} = 1/\sqrt{x}$ |
| -1 | $x^{-1} = 1/x$ |
| -2 | $x^{-2} = 1/x^2$ |

Using the Ladder

- The ladder is most useful for strictly positive, ratio variables.
- Sometimes, if 0 is a value in the data set, we will add 1 to each value before applying a transformation like the logarithm.
- Interpretability is often an important criterion, although back-transformation at the end of an analysis is usually a sensible strategy.

| Power | -2 | -1 | -0.5 | 0 | 0.5 | 1 | 2 | 3 |
|----------------|---------|-------|--------------|---------|------------|-----|-------|-------|
| Transformation | $1/x^2$ | $1/x$ | $1/\sqrt{x}$ | $\ln x$ | \sqrt{x} | x | x^2 | x^3 |

The nyfs1 data (see Chapter 7 of our Notes)

The nyfs1.csv data come from the 2012 National Youth Fitness Survey.

```
## first, we'll import the data into the nyfs1 data frame  
nyfs1 <- read.csv("nyfs1.csv") %>% tbl_df()
```

```
dim(nyfs1)
```

```
[1] 1416    7
```

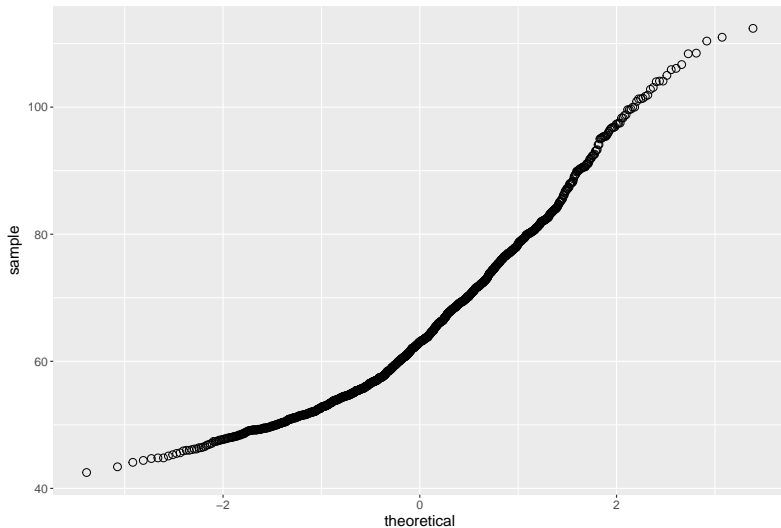
Structure of our Tibble

```
str(nyfs1)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':  1416 obs. of  7 variables:
 $ subject.id      : int  71918 71919 71921 71922 71923 71924 ...
 $ sex             : Factor w/ 2 levels "Female","Male": 1 1 2 ...
 $ age.exam        : int   8 14 3 12 12 8 7 8 3 9 ...
 $ bmi             : num  22.3 19.8 15.2 25.9 22.5 14.4 15.9 1 ...
 $ bmi.cat         : Factor w/ 4 levels "1 Underweight",...: 4 ...
 $ waist.circ      : num  71.9 79.4 46.8 90 72.3 56.1 54.5 59. ...
 $ triceps.skinfold: num   19.9 15 8.6 22.8 20.5 12.9 6.9 8.8 1 ...
```

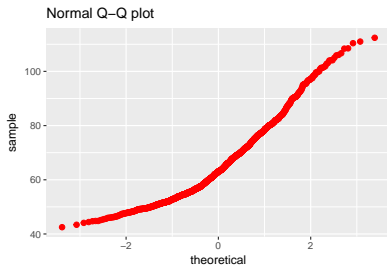
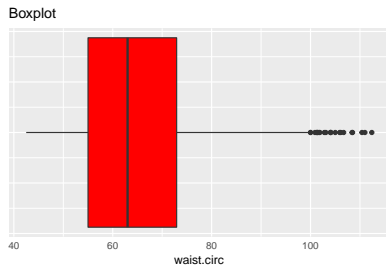
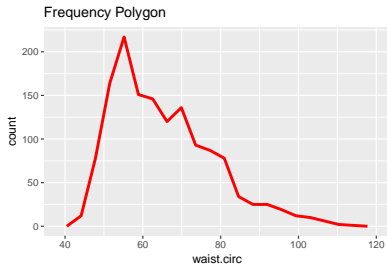
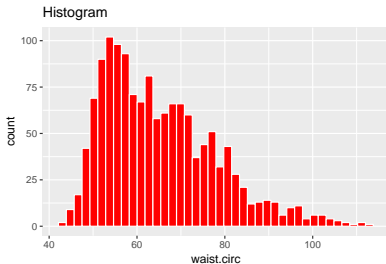
Normal Q-Q Plot of Waist Circumferences

Normal Q-Q for Waist Circumference



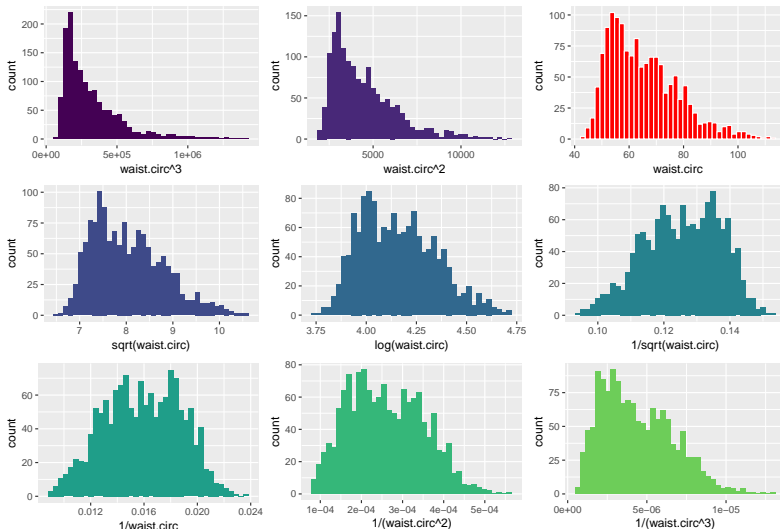
The Waist Circumference Data

NYFS1 Waist Circumference Distribution



Waist Circumference Histograms: Ladder

Ladder of Power Transformations



skew₁ and Power Transformations

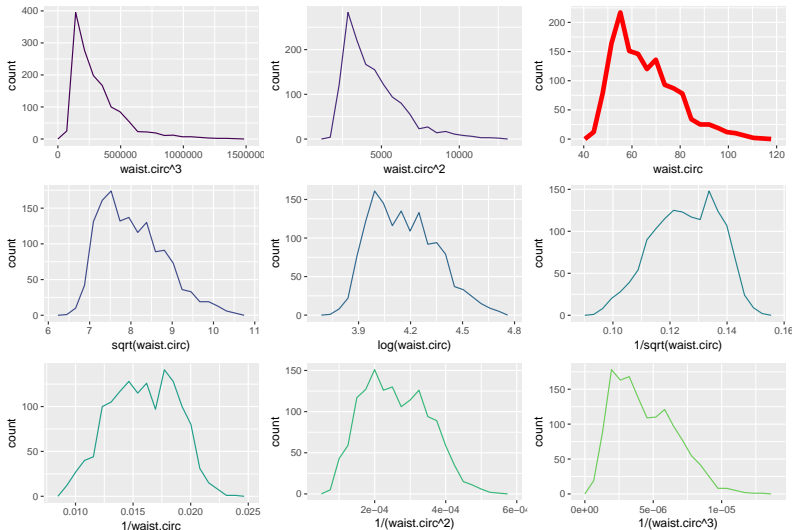
Let's add in the skew₁ values we observe for each of the available transformations of the waist circumference data.

For waist circumference,

| Power | Transformation | skew ₁ |
|-------|----------------|-------------------|
| 3 | x^3 | 0.31 |
| 2 | x^2 | 0.25 |
| 1 | x | 0.18 |
| 0.5 | \sqrt{x} | 0.14 |
| 0 | $\ln x$ | 0.09 |
| -0.5 | $1/\sqrt{x}$ | -0.05 |
| -1 | $1/x$ | 0 |
| -2 | $1/x^2$ | 0.09 |
| -3 | $1/x^3$ | 0.17 |

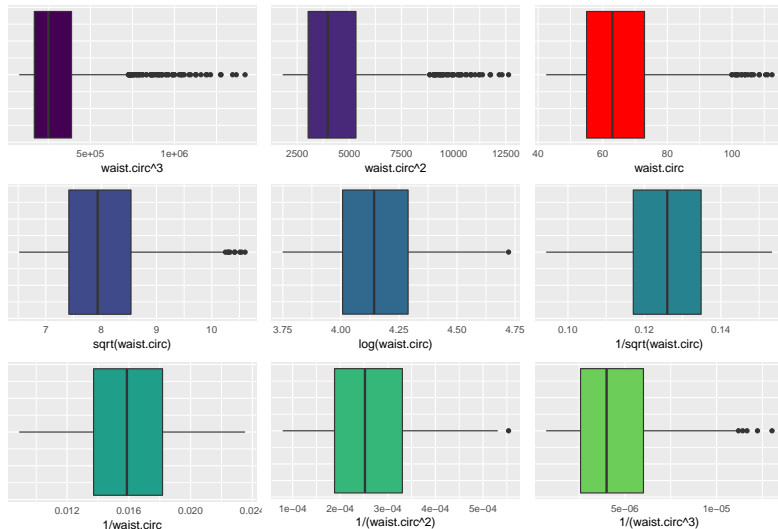
But we should be looking at Frequency Polygons...

Ladder of Power Transformations



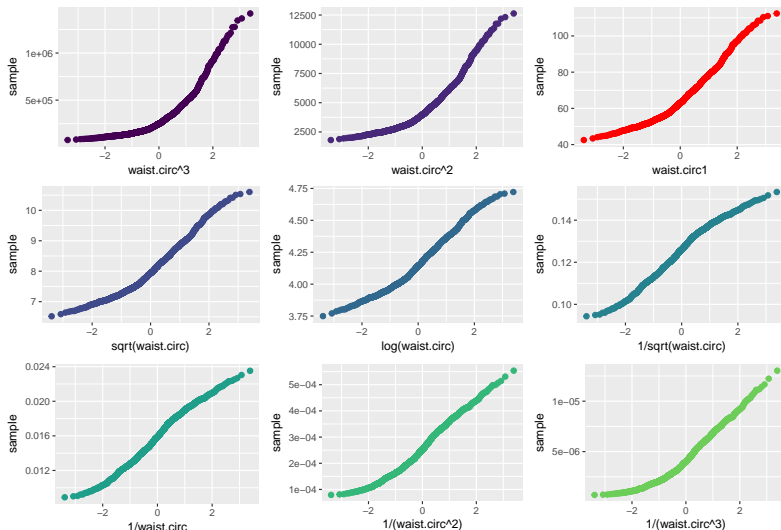
Waist Circumference Boxplots: Ladder

Ladder of Power Transformations



Waist Circumference Normal Q-Q: Ladder

Ladder of Power Transformations



Can we test to see if our data follow a Normal model?

Yes, but don't. Graphical approaches are far better.

What would such a test look like?

```
shapiro.test(nyfs1$waist.circ)
```

Shapiro-Wilk normality test

```
data:  nyfs1$waist.circ  
W = 0.94391, p-value < 2.2e-16
```

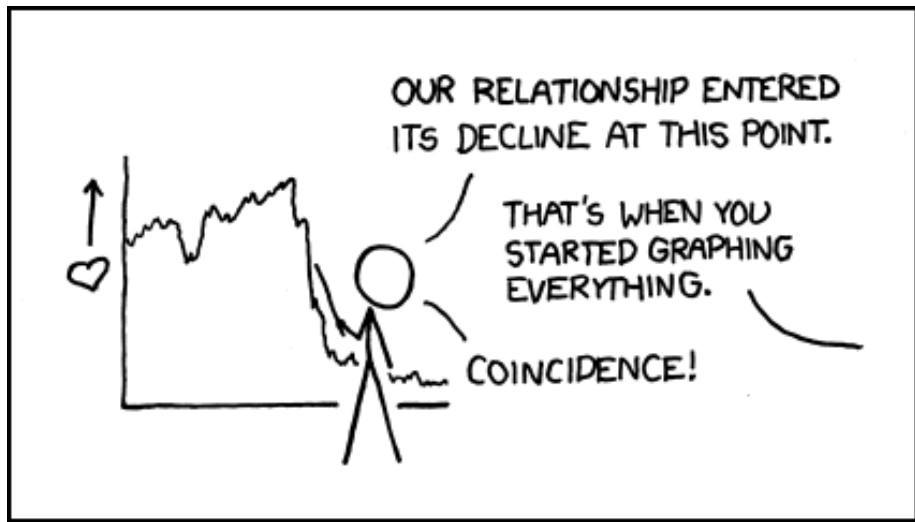
The very small p value indicates that the test finds strong indications **against** adopting a Normal model.

Why not test?

Because the sample size is so large, and the test is so poor at detecting non-normality compared to our eyes, that it finds problems we don't care about, and ignores problems we do care about. For waist circumference. . .

| Power | Transformation | Shapiro-Wilk p value |
|-------|----------------|------------------------|
| 3 | x^3 | $< 2.2 \text{ e-16}$ |
| 2 | x^2 | $< 2.2 \text{ e-16}$ |
| 1 | x | $< 2.2 \text{ e-16}$ |
| 0.5 | \sqrt{x} | $< 2.2 \text{ e-16}$ |
| 0 | $\ln x$ | 1.5 e-14 |
| -0.5 | $1/\sqrt{x}$ | $< 2.2 \text{ e-16}$ |
| -1 | $1/x$ | 1.8 e-09 |
| -2 | $1/x^2$ | 1.4 e-12 |
| -3 | $1/x^3$ | $< 2.2 \text{ e-16}$ |

DON'T DO THIS - instead, graph everything!



Which states are in the Midwest?

The Midwest Survey (prior to 431 Class 7)

Please complete the survey to the best of your ability. A quick response is desirable. Thank you! Dr. Love

* Required

I have a strong opinion about which U.S. states are part of the Midwest. *

Please indicate the degree to which you agree with the statement above.

1 2 3 4 5
Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

Map of the United States



For each of the 25 states listed below, please indicate whether or not you consider it to be part of the Midwest. *

| | Yes, in Midwest | No, not in Midwest |
|-------------------|--------------------------|--------------------------|
| Missouri (MO) | <input type="checkbox"/> | <input type="checkbox"/> |
| California (CA) | <input type="checkbox"/> | <input type="checkbox"/> |
| Colorado (CO) | <input type="checkbox"/> | <input type="checkbox"/> |
| Arkansas (AR) | <input type="checkbox"/> | <input type="checkbox"/> |
| Pennsylvania (PA) | <input type="checkbox"/> | <input type="checkbox"/> |
| Wisconsin (WI) | <input type="checkbox"/> | <input type="checkbox"/> |
| Iowa (IA) | <input type="checkbox"/> | <input type="checkbox"/> |
| Indiana (IN) | <input type="checkbox"/> | <input type="checkbox"/> |
| Ohio (OH) | <input type="checkbox"/> | <input type="checkbox"/> |

FiveThirtyEight

Politics Sports Science & Health Economics Culture

APR. 29, 2014 AT 9:26 AM

Which States Are in the Midwest?

By Walt Hickey

Filed under Regionalism

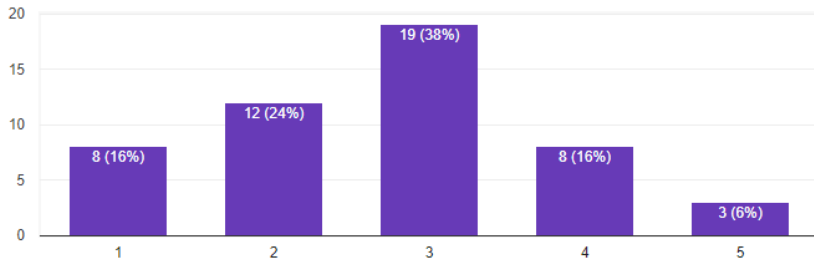
Get the data on GitHub



Did we have strong opinions?

I have a strong opinion about which U.S. states are part of the Midwest.

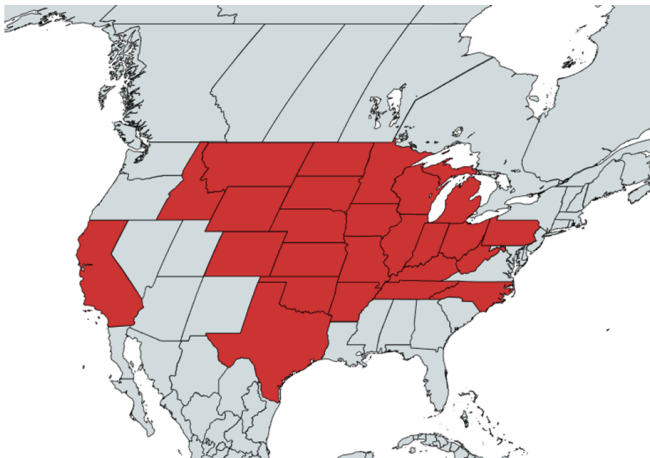
50 responses



Which states got the most votes in our poll?

25 states I asked about
are **shown in red**...

1. Arkansas (AR)
2. California (CA)
3. Colorado (CO)
4. Idaho (ID)
5. Illinois (IL)
6. Indiana (IN)
7. Iowa (IA)
8. Kansas (KS)
9. Kentucky (KY)
10. Michigan (MI)
11. Minnesota (MN)
12. Missouri (MO)
13. Montana (MT)
14. Nebraska (NE)
15. North Carolina (NC)
16. North Dakota (ND)
17. Ohio (OH)
18. Oklahoma (OK)
19. Pennsylvania (PA)
20. South Dakota (SD)
21. Tennessee (TN)
22. Texas (TX)
23. West Virginia (WV)
24. Wisconsin (WI)
25. Wyoming (WY)



Our 50 responses

For each of the 25 states listed below, please indicate whether or not you consider it to be part of the Midwest.

44 Iowa

43 Indiana 42 Illinois, Michigan

40 Wisconsin 39 Minnesota, Missouri

38 Ohio 33 Kansas

27 Nebraska 25 North Dakota

23 South Dakota, Oklahoma

22 Kentucky 20 Arkansas

18 Wyoming 16 Idaho, Tennessee 15 Colorado

14 Montana 12 West Virginia 11 Pennsylvania

5 North Carolina 4 Texas 1 California

Which states are in the Midwest?

FiveThirtyEight

Politics Sports Science & Health Economics Culture

APR. 29, 2014 AT 9:26 AM

Which States Are in the Midwest?

By [Walt Hickey](#)

Filed under [Regionalism](#)

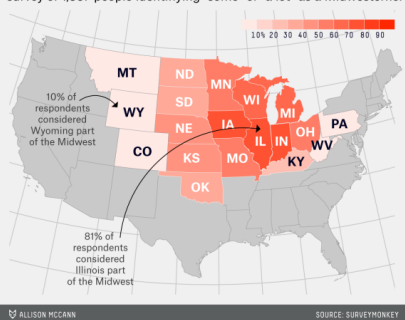
Get the data on [GitHub](#)



To get this broad-based view, we asked SurveyMonkey Audience to ask self-identified Midwesterners which states make the cut. We ran a national survey that targeted the Midwest from March 12 to March 17, with 2,778 respondents. Of those, 1,357 respondents identified “a lot” or “some” as a Midwesterner. We then asked this group to identify the states they consider part of the Midwest.

‘Which States Do You Consider Part of the Midwest?’

Percentage classifying each state as part of the Midwest, from a survey of 1,357 people identifying “some” or “a lot” as a Midwesterner



<https://fivethirtyeight.com/datalab/what-states-are-in-the-midwest/>

Link to today's Google Form

You'll find the link at today's README (Class 7). This is different than the Midwest Survey that I emailed to you.

Please submit the form by 11 AM Thursday.