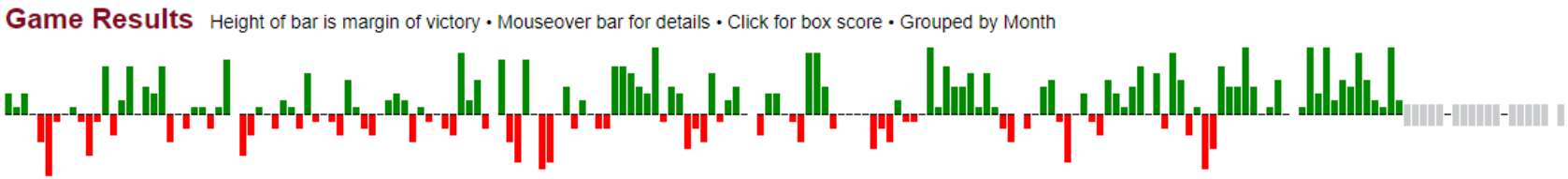
431 Class 06

Thomas E. Love

2017-09-14



Source: https://www.baseball-reference.com/teams/CLE/2017.shtml

Comments on the Chapter 5 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.	○	○	○	○	○	Very confident.

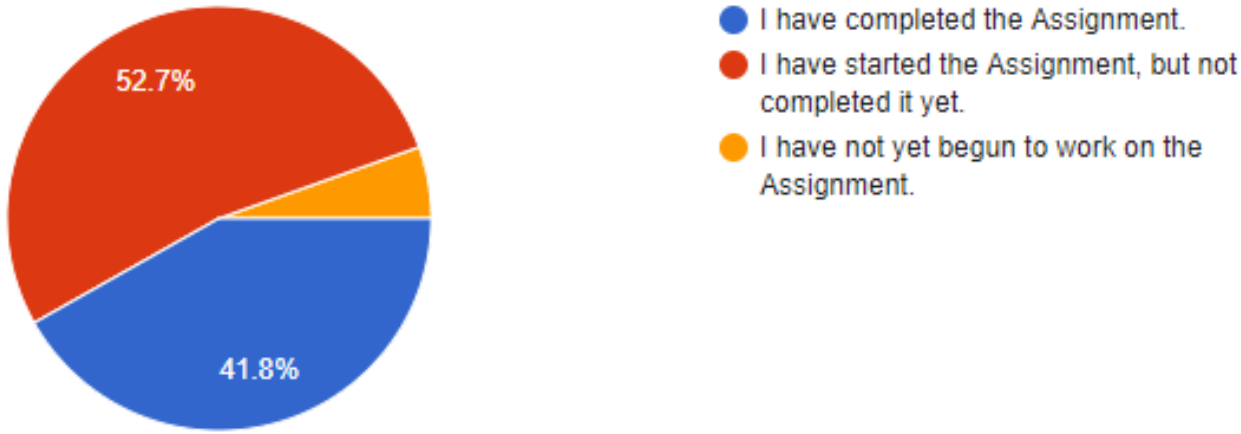
Please describe something that you would like to be able to better predict, where using data, or using it better, seems likely to be effective. *

Please take into account what you have read in The Signal and the Noise (Introduction and Chapter 1) for our class.

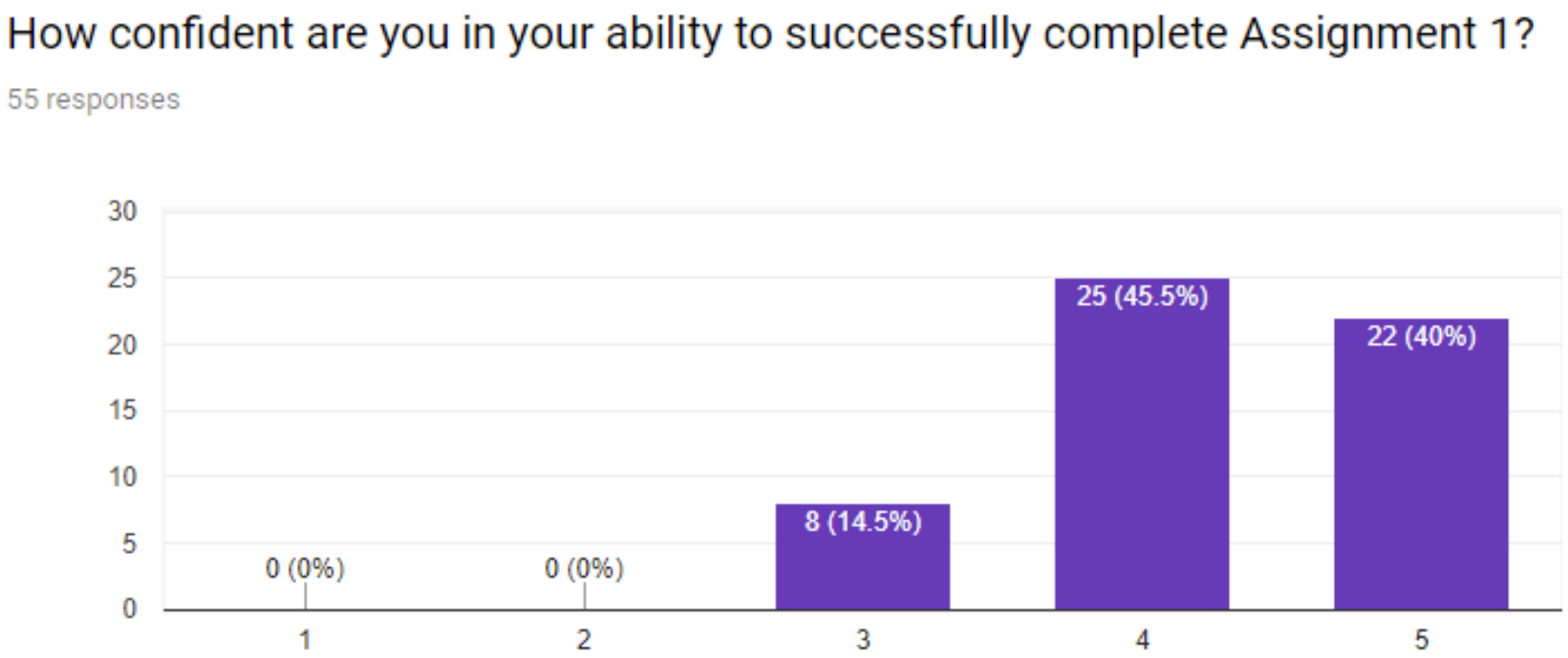
Chapter 5 Form, Question 1

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

55 responses



Chapter 5 Form, Question 2



Question 3

See the Google Doc at

https://goo.gl/MbqRy2

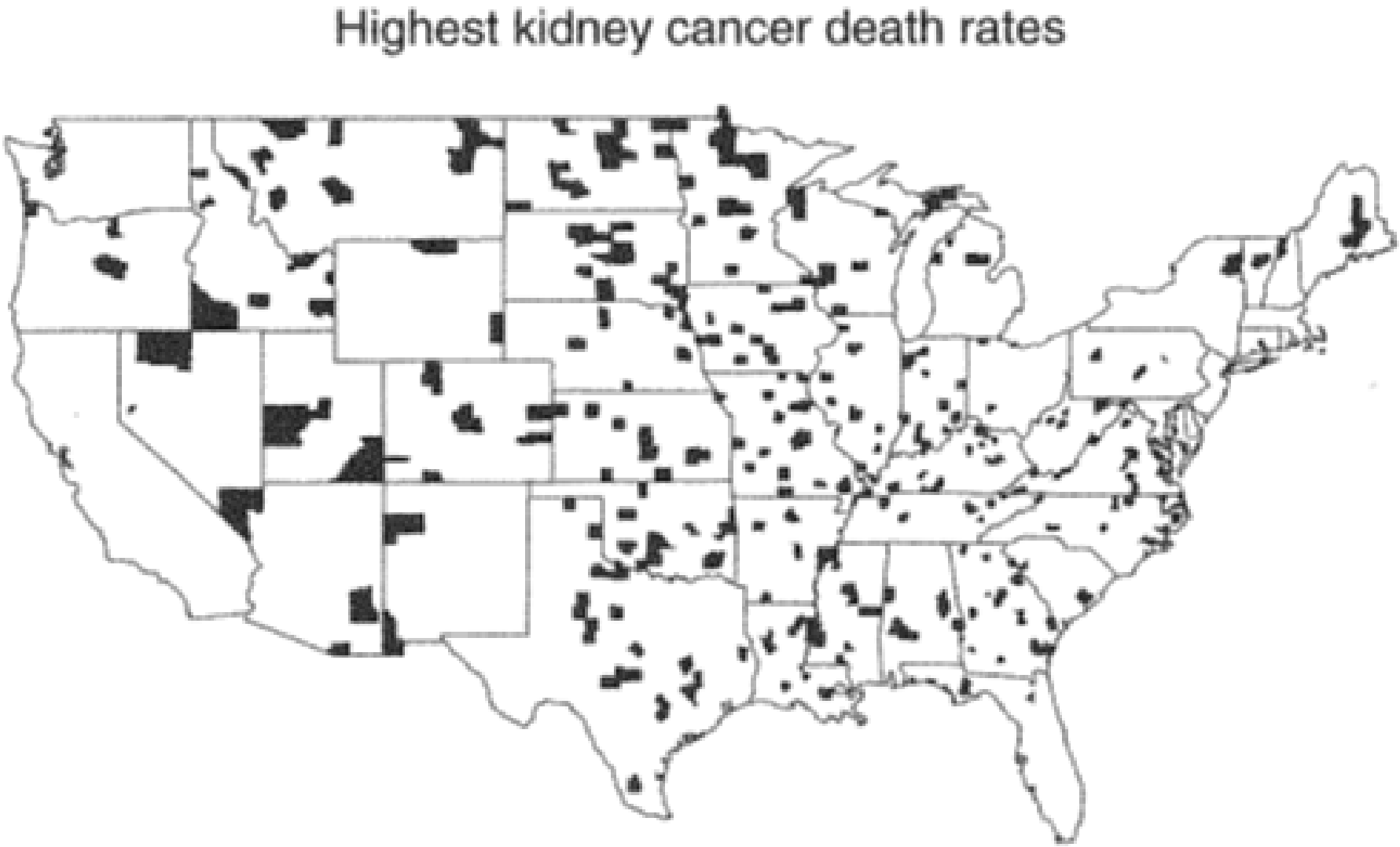
for details.

Kidney Cancer Death Rates

Your map shows U.S. counties.

- The shaded counties are in the top 10% of age-standardized rates for death due to cancer of the kidney/ureter for white males, in 1980-1989.

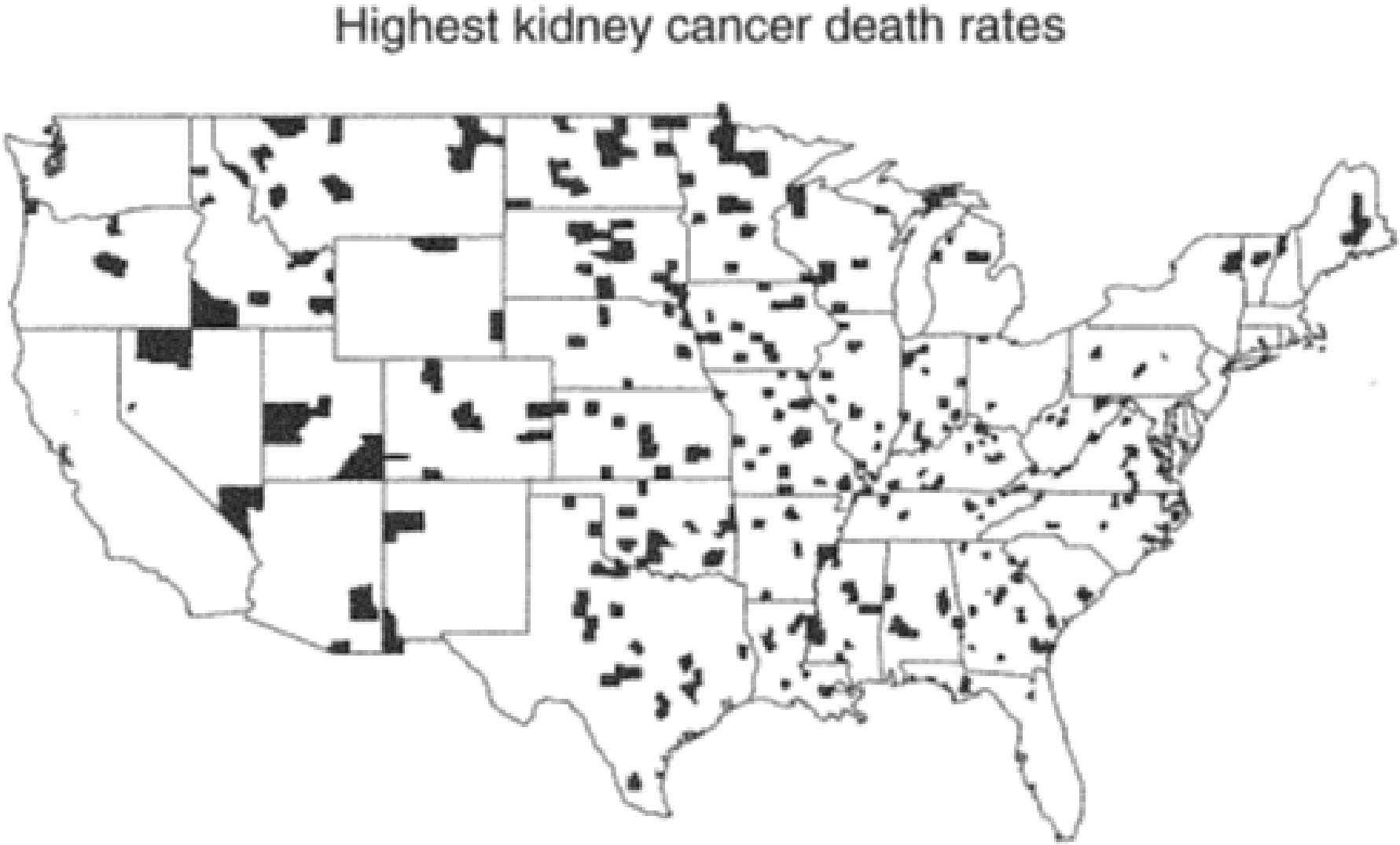
Kidney Cancer Death Rates



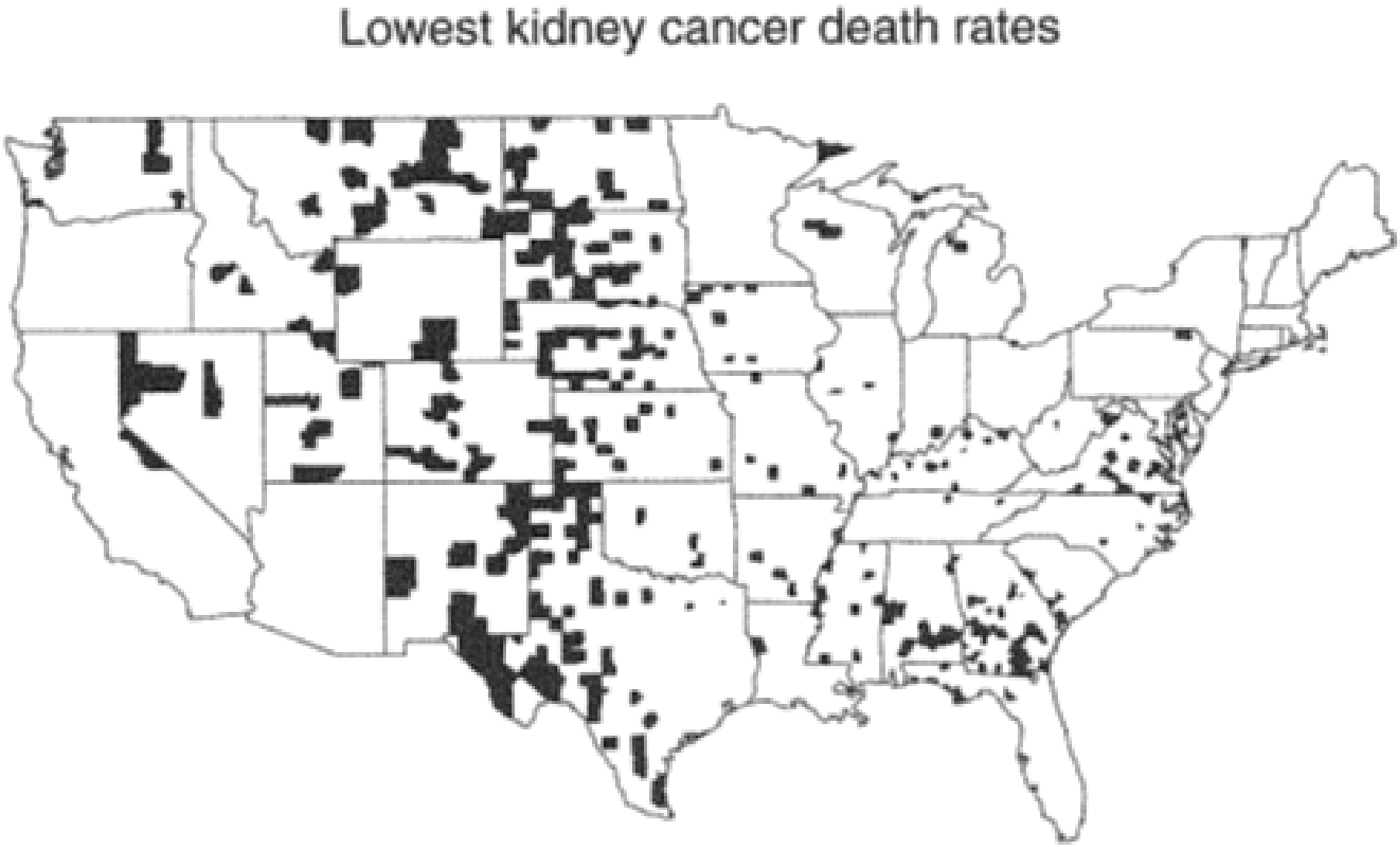
Your Tasks

1. Describe the patterns you see in the map.
2. Speculate as to the cause of these patterns.

The Map, Again



Another Map



Assignment 1 Common Issues

1. In lines 2 and 3, you haven't edited the generic names to replace YOURNAME with, you know, your actual name.
2. You have chunk labels that repeat. They cannot. Every chunk must have a unique name.
3. Don't be afraid to hit the enter key.

```
# Question 4 My answer is produces yelling.
```

vs.

```
# Question 4

My answer is ...
```

4. The variables are described in the help file for each data set. To see those files, type ?faithful and ?MASS::geyser in the R Console. Check the descriptions there carefully.

Preliminaries for Today

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

cwru.blue <- '#0a304e'
cwru.gray <- '#626262'

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)
```

So far „ ,

1. In exploring data, good visualizations beat good numerical summaries, essentially every time.
2. The false sense of security one gets by placing a number on a description of a data set, rather than a picture, is in fact usually false. Exploratory data analysis is not about appealing to a higher authority for the "correct" answer,
3. Thinking about the distribution of a quantitative variable requires effort to understand the center, spread, and shape. Shape includes assessments of symmetry, outliers and number of modes, at the least.

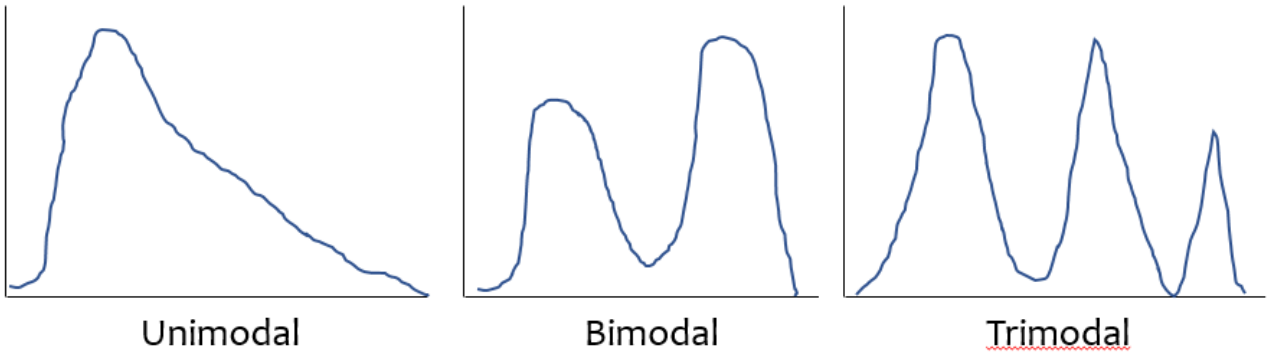
Measuring/Assessing the Shape of a Distribution

When considering the shape of a distribution, one is often interested in three key points.

- The number of modes in the distribution, which I always assess through plotting the data.
- The **skewness**, or symmetry that is present, which I typically assess by looking at a plot of the distribution of the data, but if required to, will summarize with a non-parametric measure of **skewness**.
- The **kurtosis**, or heavy-tailedness (outlier-proneness) that is present, usually in comparison to a Normal distribution. Again, this is something I nearly inevitably assess graphically, but there are measures.

A Normal distribution has a single mode, is symmetric and, naturally, is neither heavy-tailed or light-tailed as compared to a Normal distribution (we call this mesokurtic).

Multimodal vs. Unimodal distributions



Truly multimodal distributions are usually described that way in terms of shape.

Skew

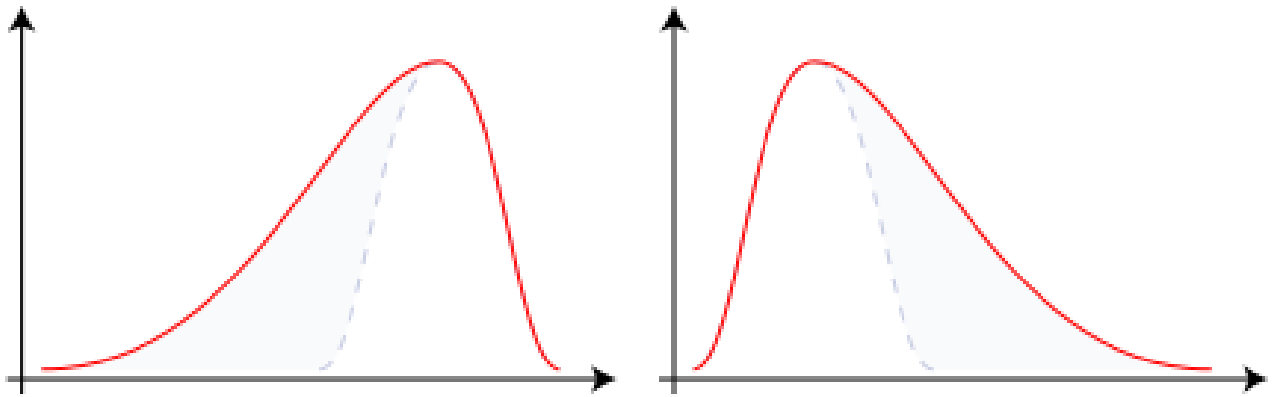
For unimodal distributions, skewness and kurtosis become useful ideas. Whether or not a distribution is approximately symmetric is an important consideration in describing its shape. Graphical assessments are always most useful in this setting, particularly for unimodal data. My favorite measure of skew, or skewness if the data have a single mode, is:

$$skew_1 = \frac{\text{mean} - \text{median}}{\text{standard deviation}}$$

- Symmetric distributions generally show values of $skew_1$ near zero. If the distribution is actually symmetric, the mean should be equal to the median.
- Distributions with $skew_1$ values above 0.2 in absolute value generally indicate meaningful skew.

Measuring Skew

- Positive skew (mean $>$ median if the data are unimodal) is also referred to as *right skew*.
- Negative skew (mean $<$ median if the data are unimodal) is referred to as *left skew*.



$skew_1$ for SBP in our NHANES data

```
nh_adults %>%
  filter(!is.na(SBP)) %>%
  summarize(mean = mean(SBP), median = median(SBP),
            sd = sd(SBP), skew1 = (mean - median)/sd)
```

```
# A tibble: 1 x 4
      mean median       sd      skew1
     <dbl>  <int>    <dbl>      <dbl>
1 118.5918    118 15.30267 0.03866988
```

Using the skew1 function in Love-boost.R

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

temp <- filter(nh_adults, !is.na(SBP))
skew1(temp$SBP)
```

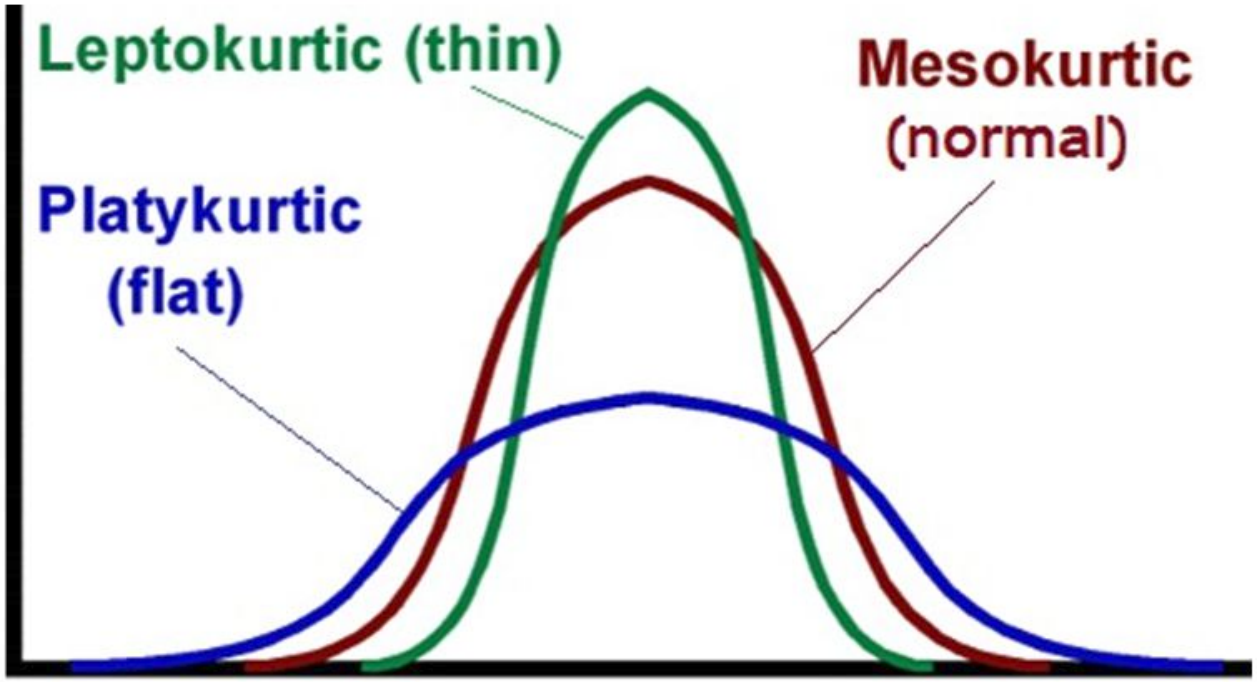
```
[1] 0.03866988
```

```
nh_adults %>%
  filter(complete.cases(DBP)) %$%
  skew1(DBP)
```

```
[1] 0.02265248
```

Kurtosis

When we have a unimodal distribution that is symmetric, we will often be interested in the behavior of the tails of the distribution, as compared to a Normal distribution with the same mean and standard deviation.



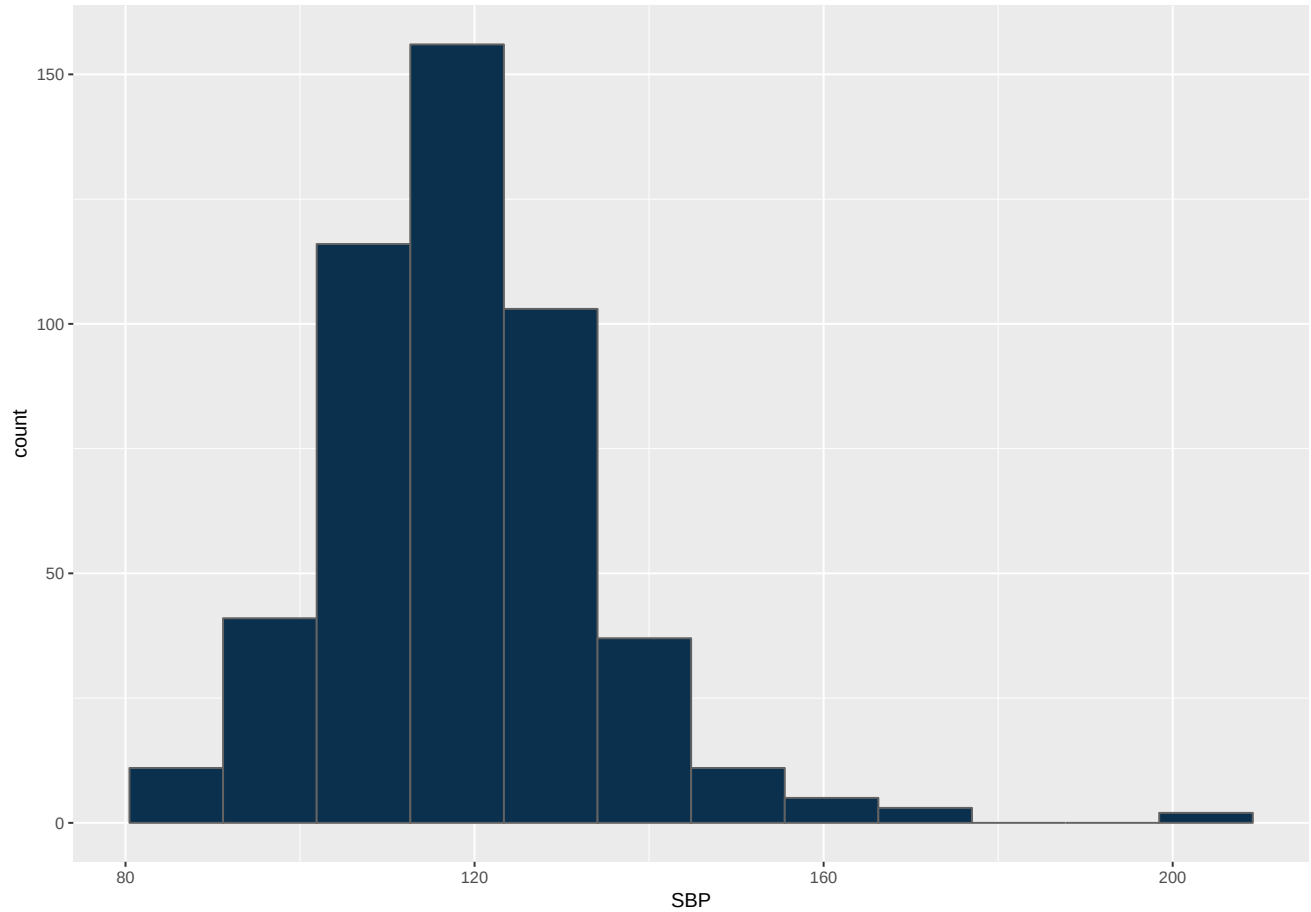
The describe function in the psych package

```
psych::describe(nh_adults %>% select(Age, SBP))
```

```
    vars   n   mean    sd median trimmed   mad min max
Age    1 500  42.10 12.54     42   42.11 16.31  21  64
SBP    2 485 118.59 15.30    118  117.79 13.34  84 202
    range  skew kurtosis   se
Age    43 -0.03    -1.23 0.56
SBP   118  1.00     3.44 0.69
```

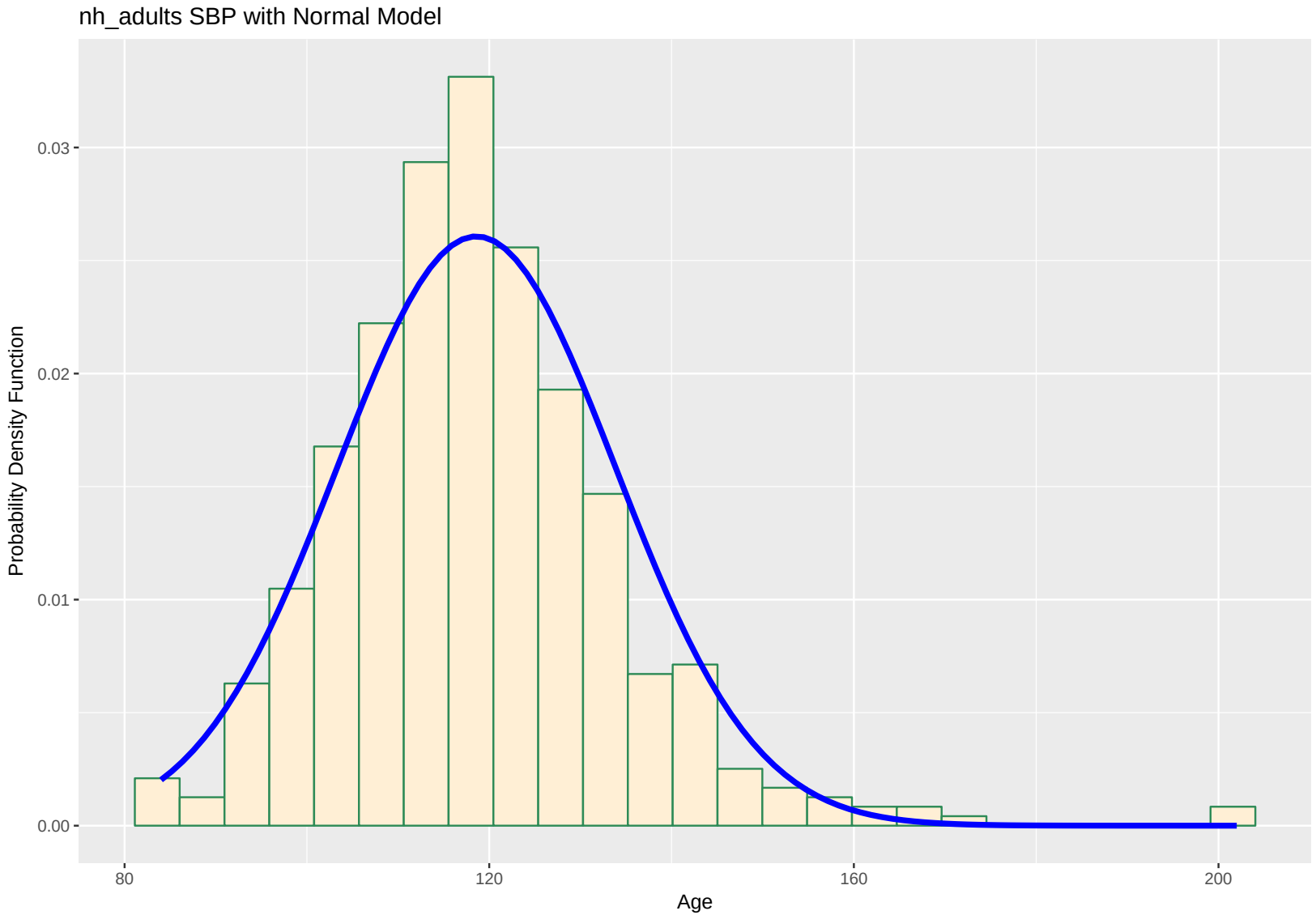
1. This `skew` is not our $skew_1$ measure.
2. Interpret `kurtosis` with care. (Near 3 = "Normalish.") A plot is more useful.
3. Meaning of `trimmed`, `mad` in Course Notes, Section 5
4. `se` = standard error of the mean = $$\frac{sd}{\sqrt{n}}$$

Do the SBP data appear skewed? Outlier-prone?



*skew*1 was 0.04, while kurtosis = 3.4

SBP + Normal Model (Section 8.3)



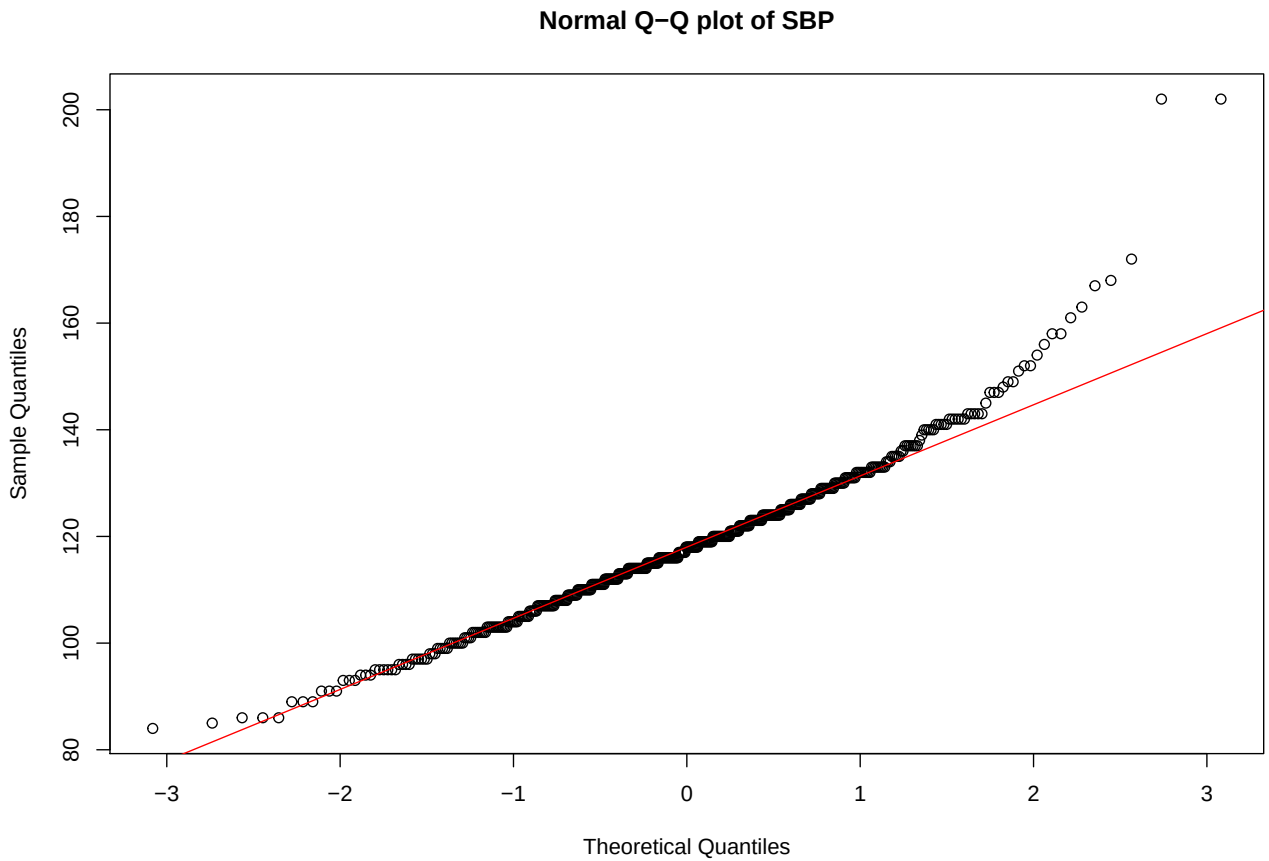
Code for Prior Histogram

```
temp <- nh_adults %>% filter(!is.na(SBP))

ggplot(temp, aes(x = SBP)) +
    geom_histogram(aes(y = ..density..), bins=25,
                   fill = "papayawhip",
                   color = "seagreen") +
    stat_function(fun = dnorm,
                  args = list(mean = mean(temp$SBP),
                              sd = sd(temp$SBP)),
                  lwd = 1.5, col = "blue") +
    labs(title = "nh_adults SBP with Normal Model",
         x = "Age", y = "Probability Density Function")
```

Normal Q-Q plot for SBP

```
qqnorm(nh_adults$SBP, main = "Normal Q-Q plot of SBP")
qqline(nh_adults$SBP, col = "red")
```



Interpreting the Normal Q-Q plot (Notes section 8.6)

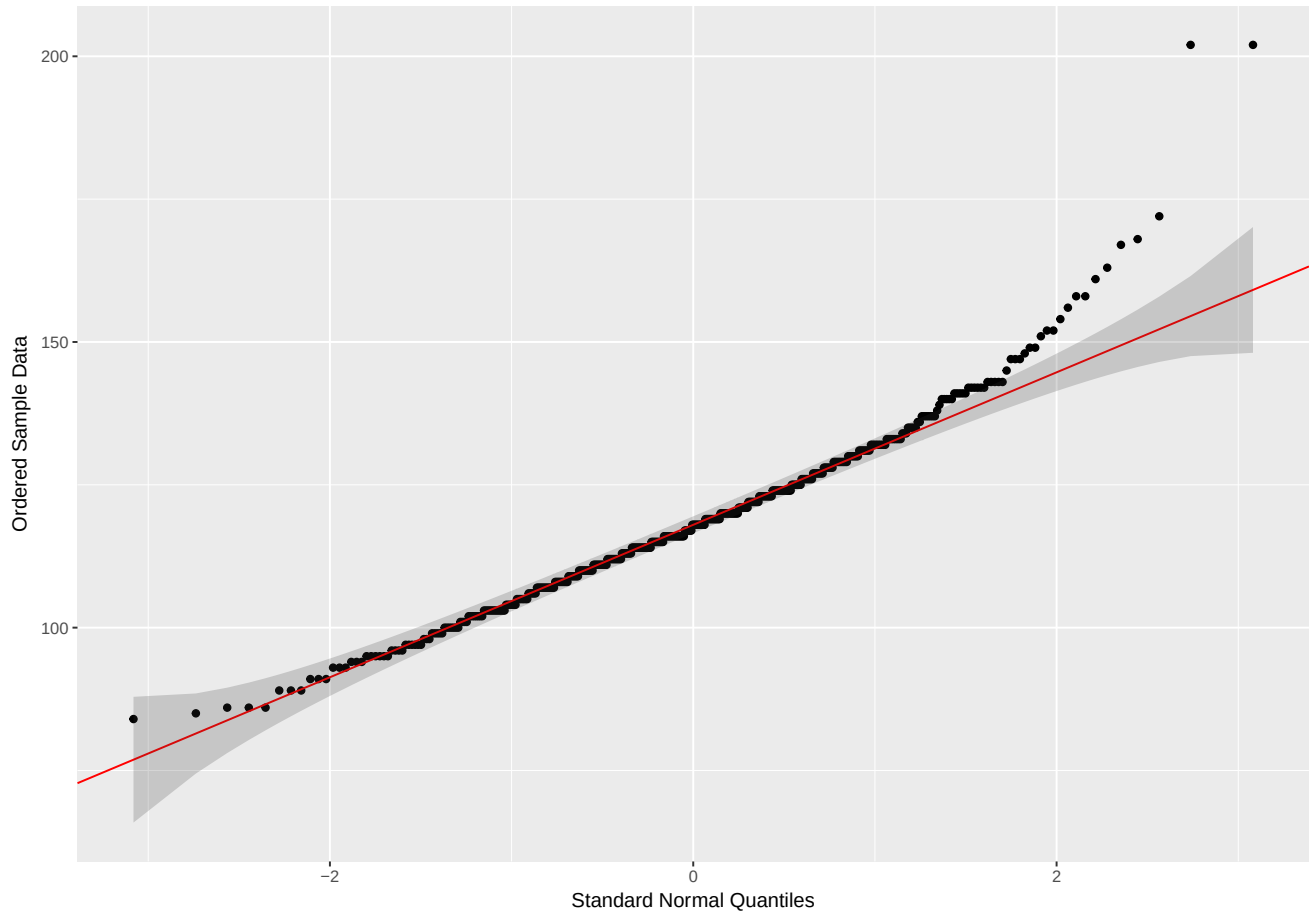
The purpose of a Normal Q-Q plot is to help point out distinctions from a Normal distribution. A Normal distribution is symmetric and has certain expectations regarding its tails. The Normal Q-Q plot can help us identify data as - well approximated by a Normal distribution, or not because of - skew (including distinguishing between right skew and left skew) - behavior in the tails (which could be heavy-tailed [more outliers than expected] or light-tailed.) Normal Q-Q plots work well with unimodal data, not so well with multiple modes.

What to look for in a Normal Q-Q plot

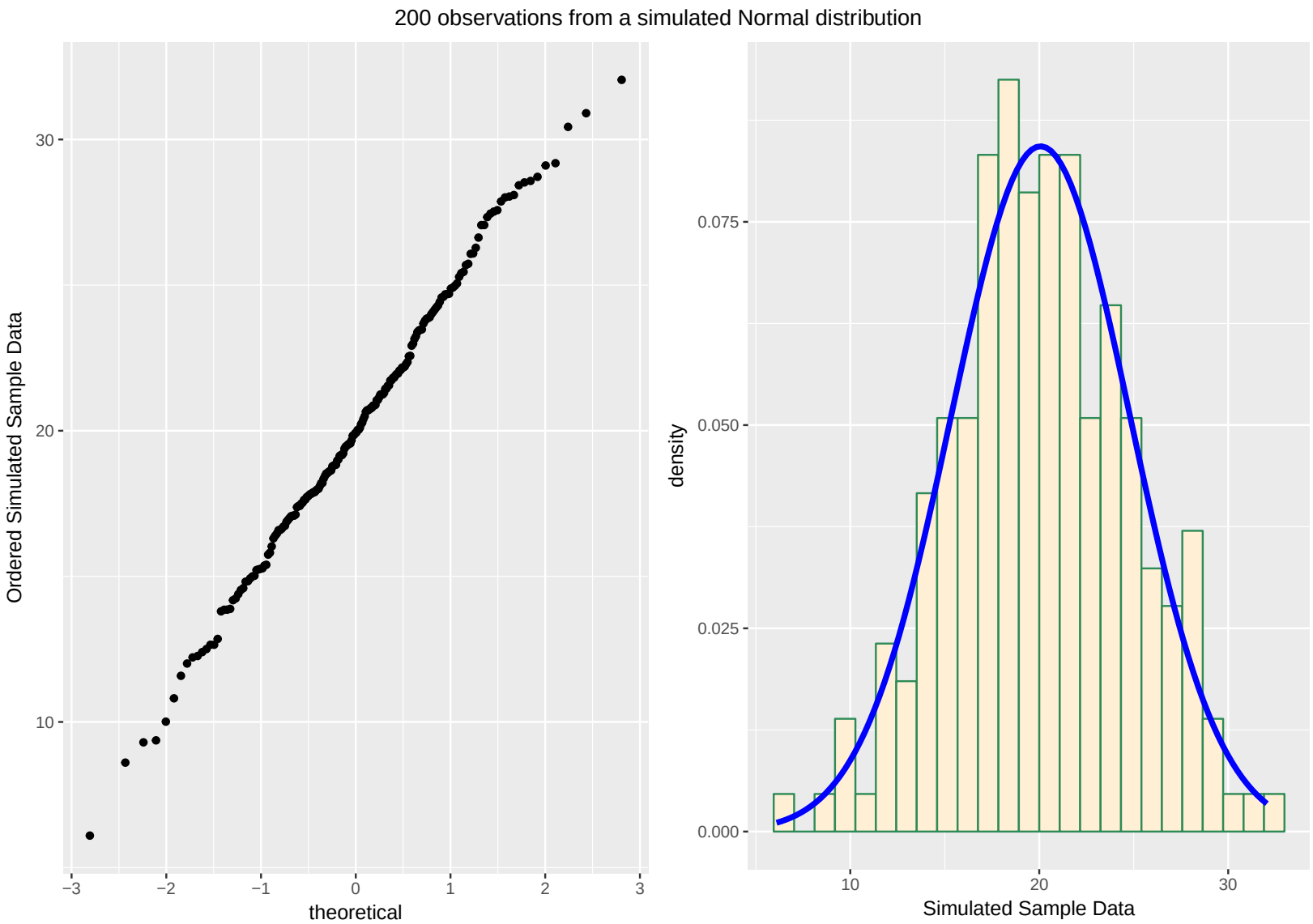
- Data from a Normal distribution shows up as a straight line in a Normal Q-Q plot
- Skew is indicated by monotonic curves in the Normal Q-Q plot
- Outlier-proneness is indicated by "s-shaped" curves in a Normal Q-Q plot
    - Heavy-tailed but symmetric distributions are indicated by reverse "S"-shapes
    - Light-tailed but symmetric distributions are indicated by "S" shapes

Using gg_qq from Love-boost.R

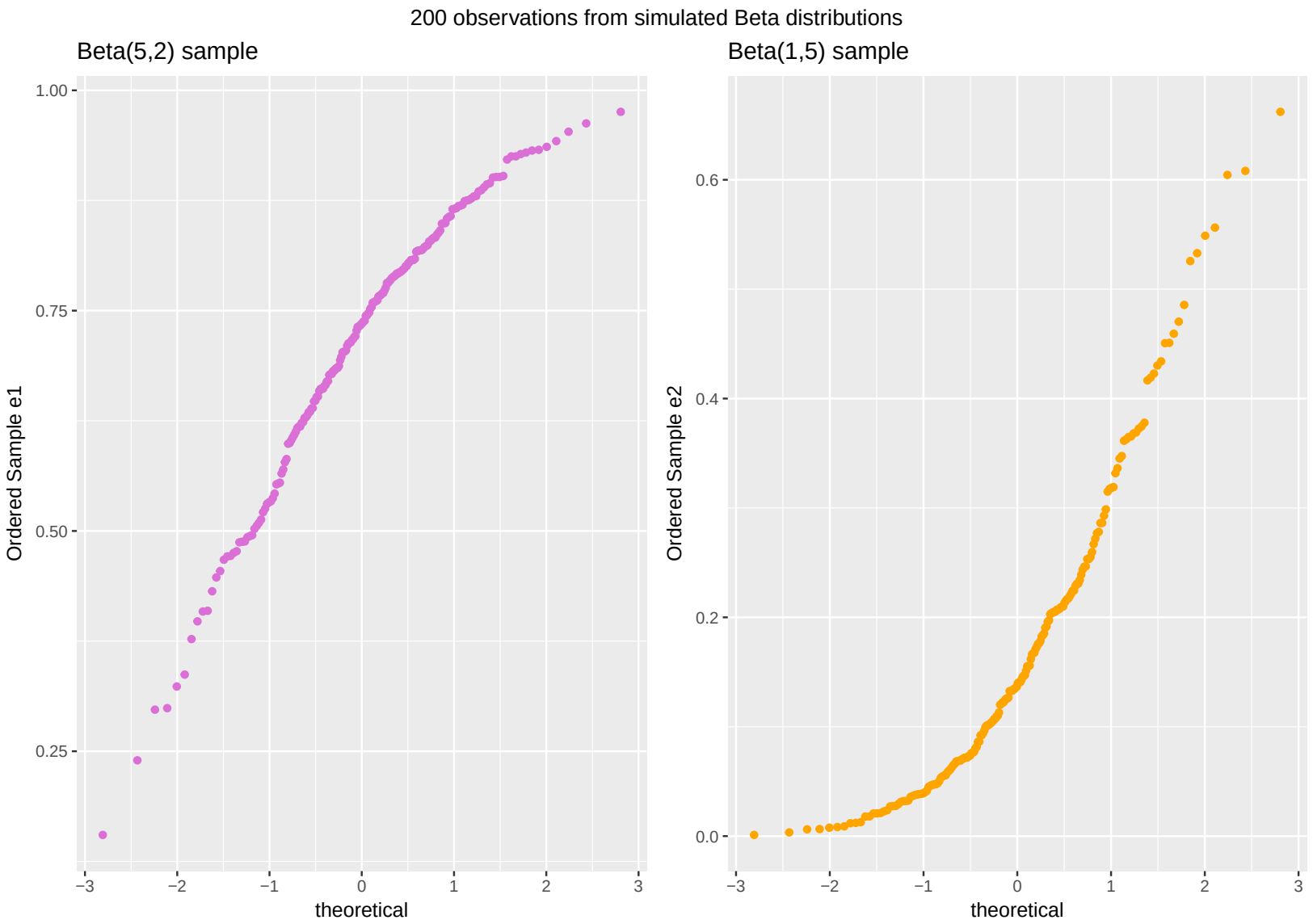
```
gg_qq(nh_adults$SBP)
```



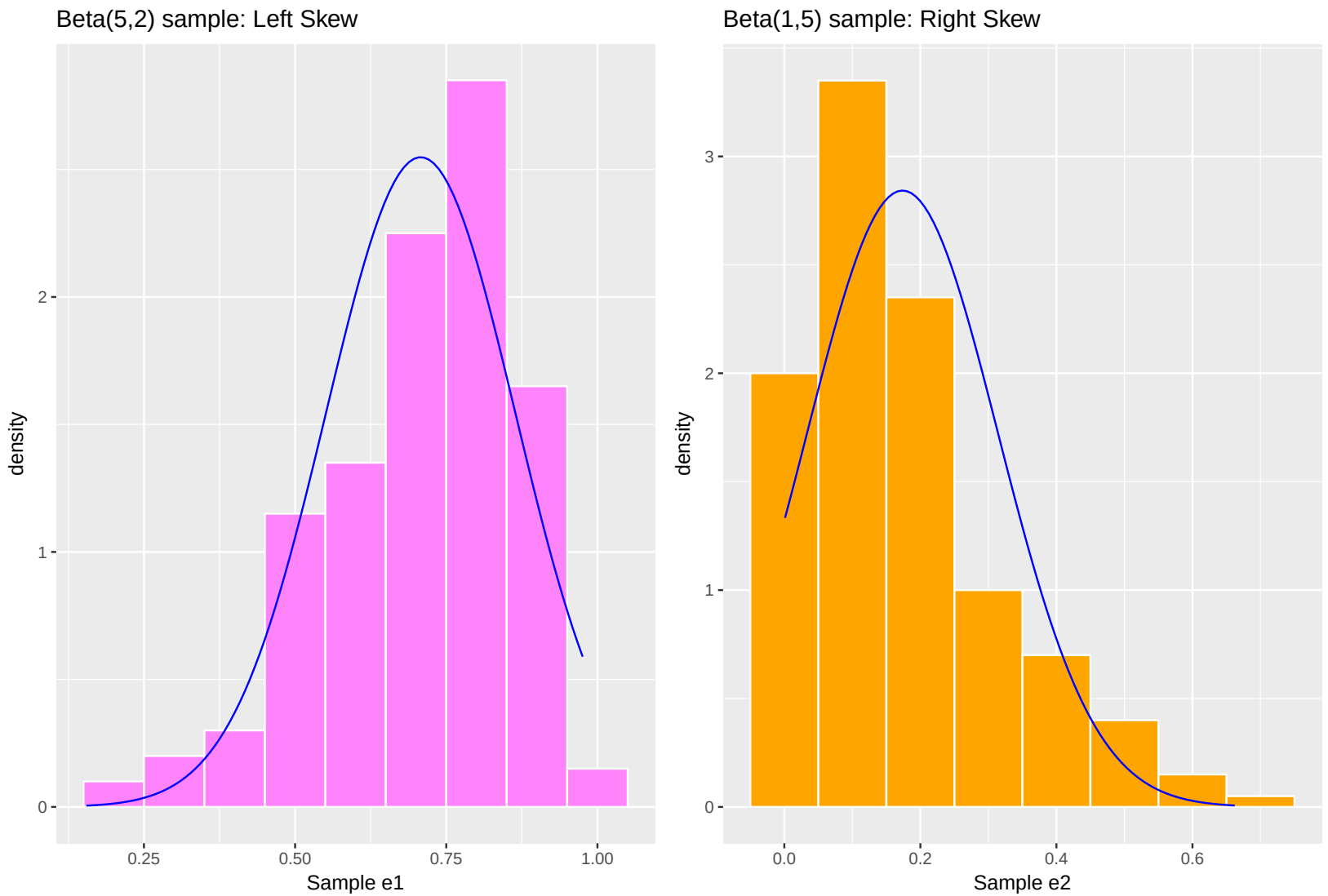
A Simulated Example (Course Notes 8.6.1)



Demonstrating Skew (Section 8.6.2)

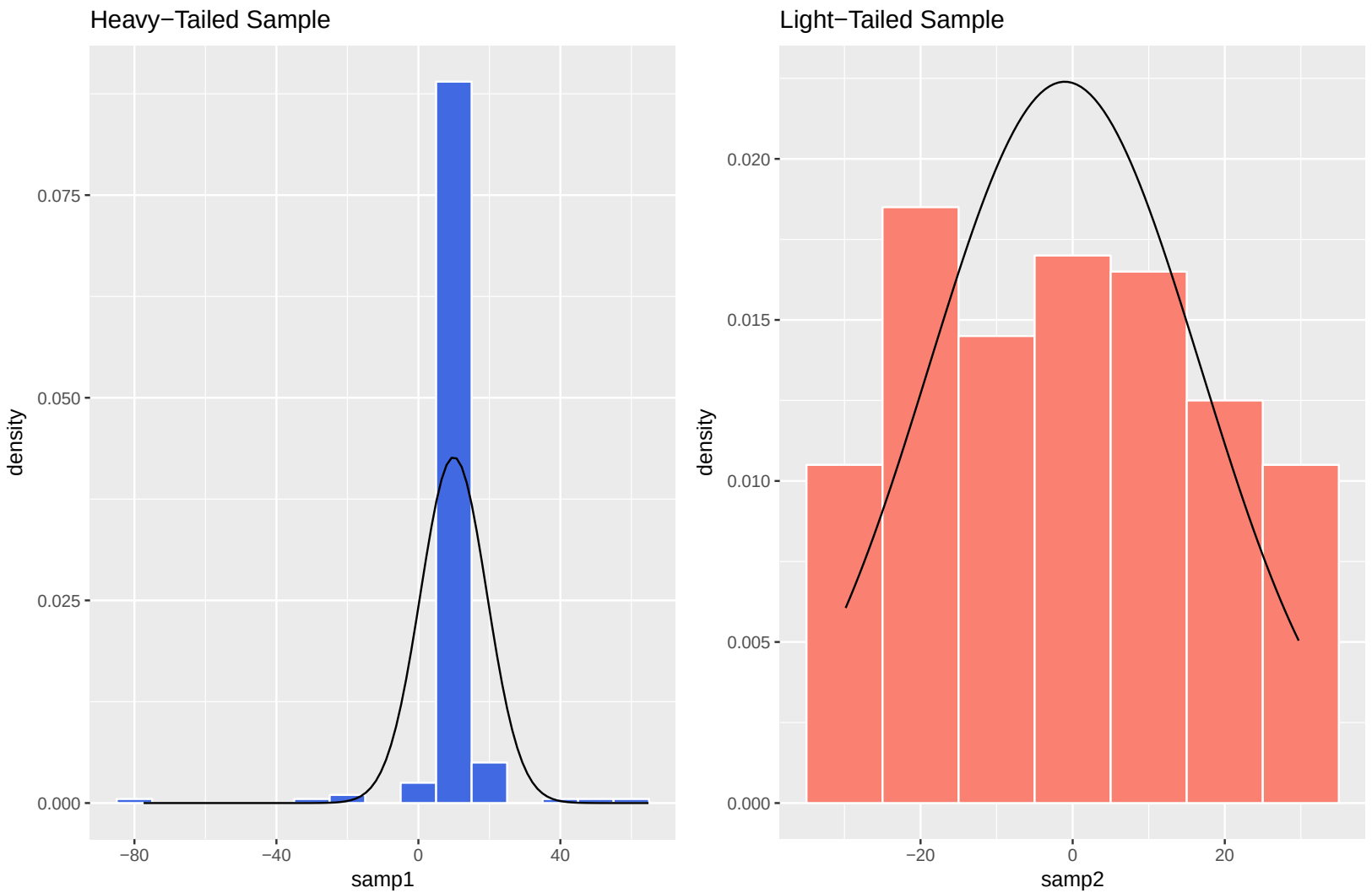


Same simulation as prior slide, but with histograms



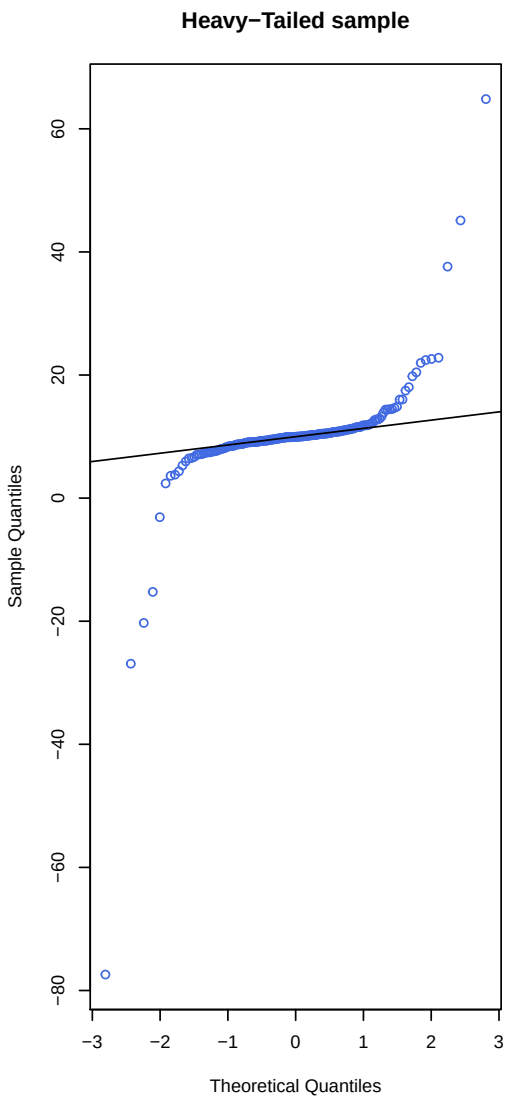
Histograms with Normal curve superimposed

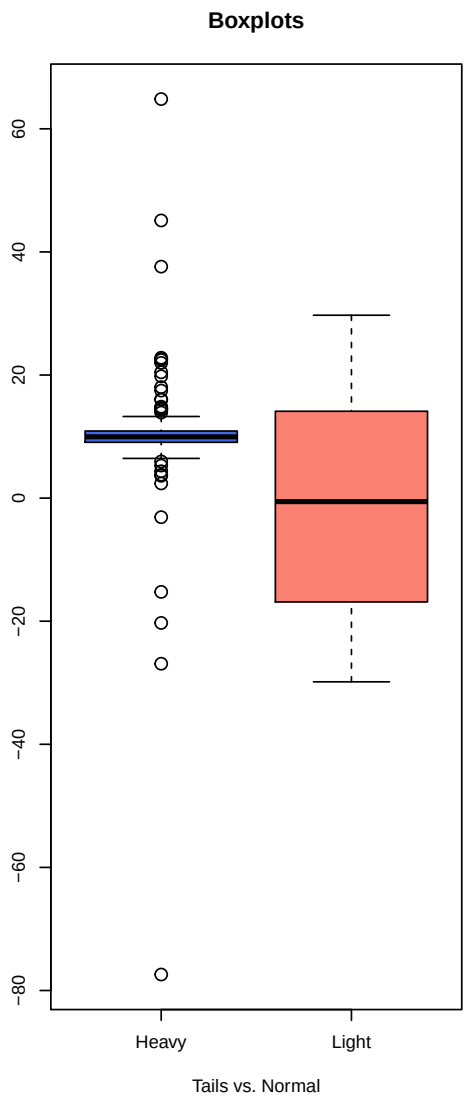
Simulating Heavy-Tailed and Light-Tailed Data

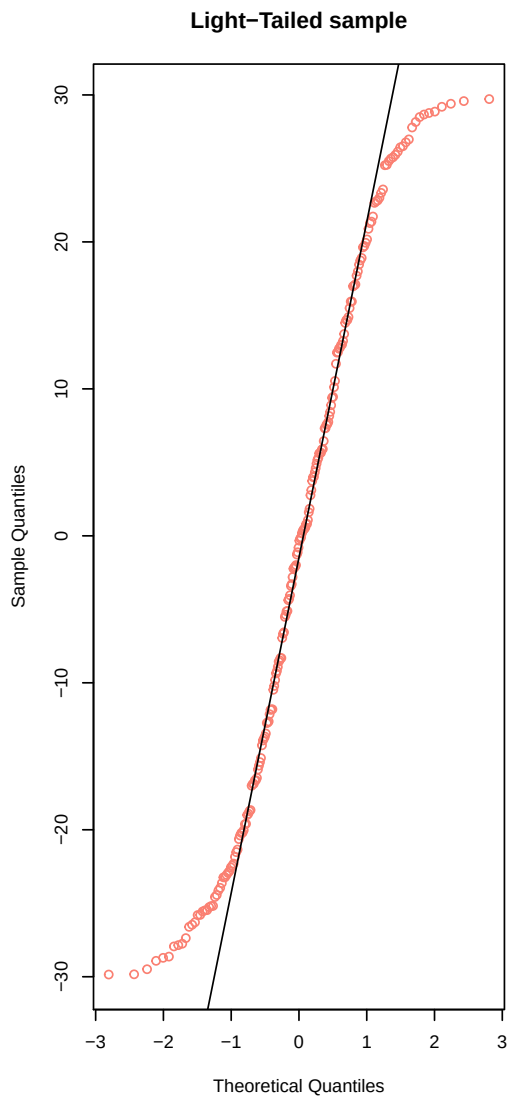


Heavy− and light−tailed distributions
with superimposed Normal models

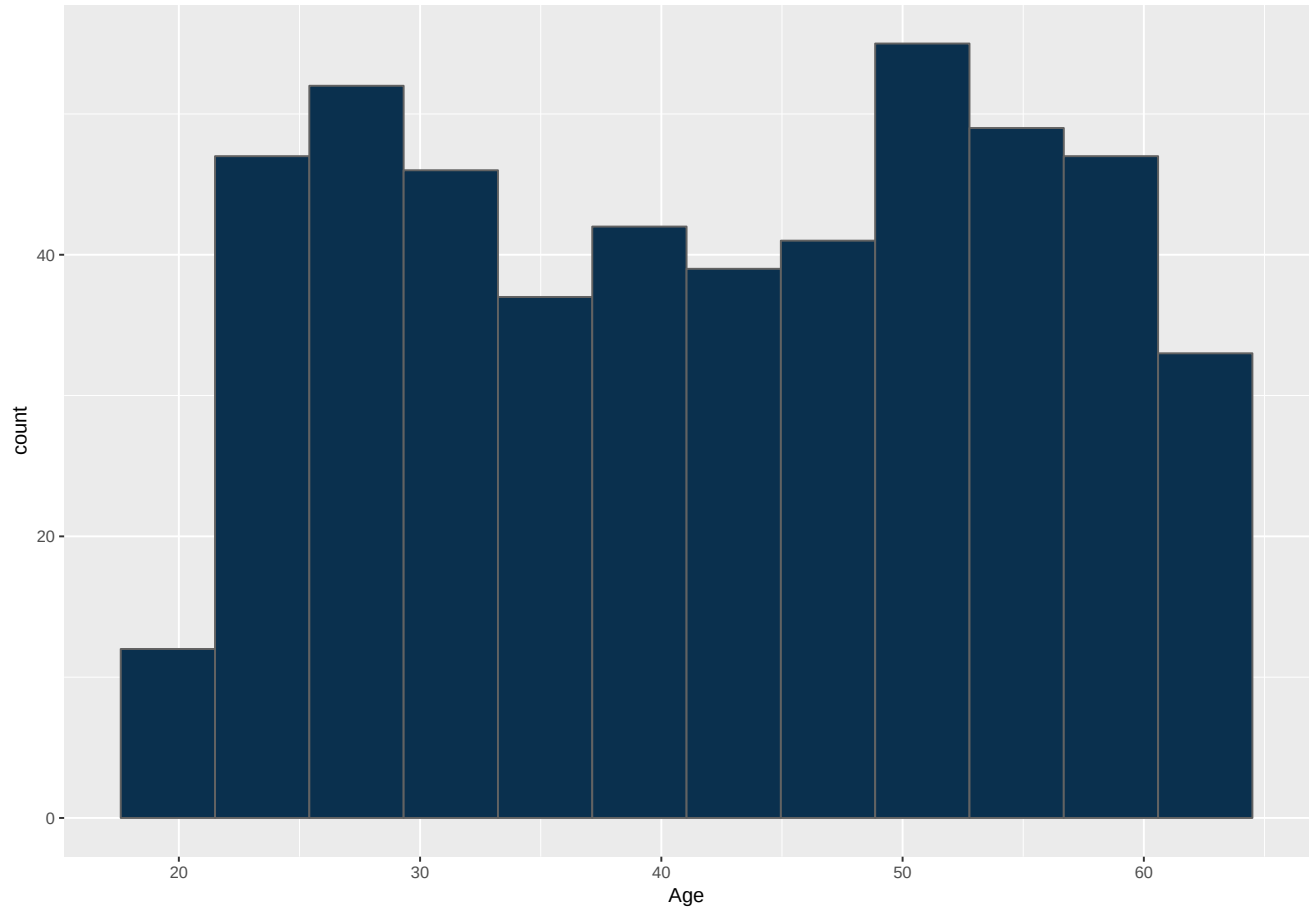
Normal Q-Q Plots for the Prior Slide's Simulations





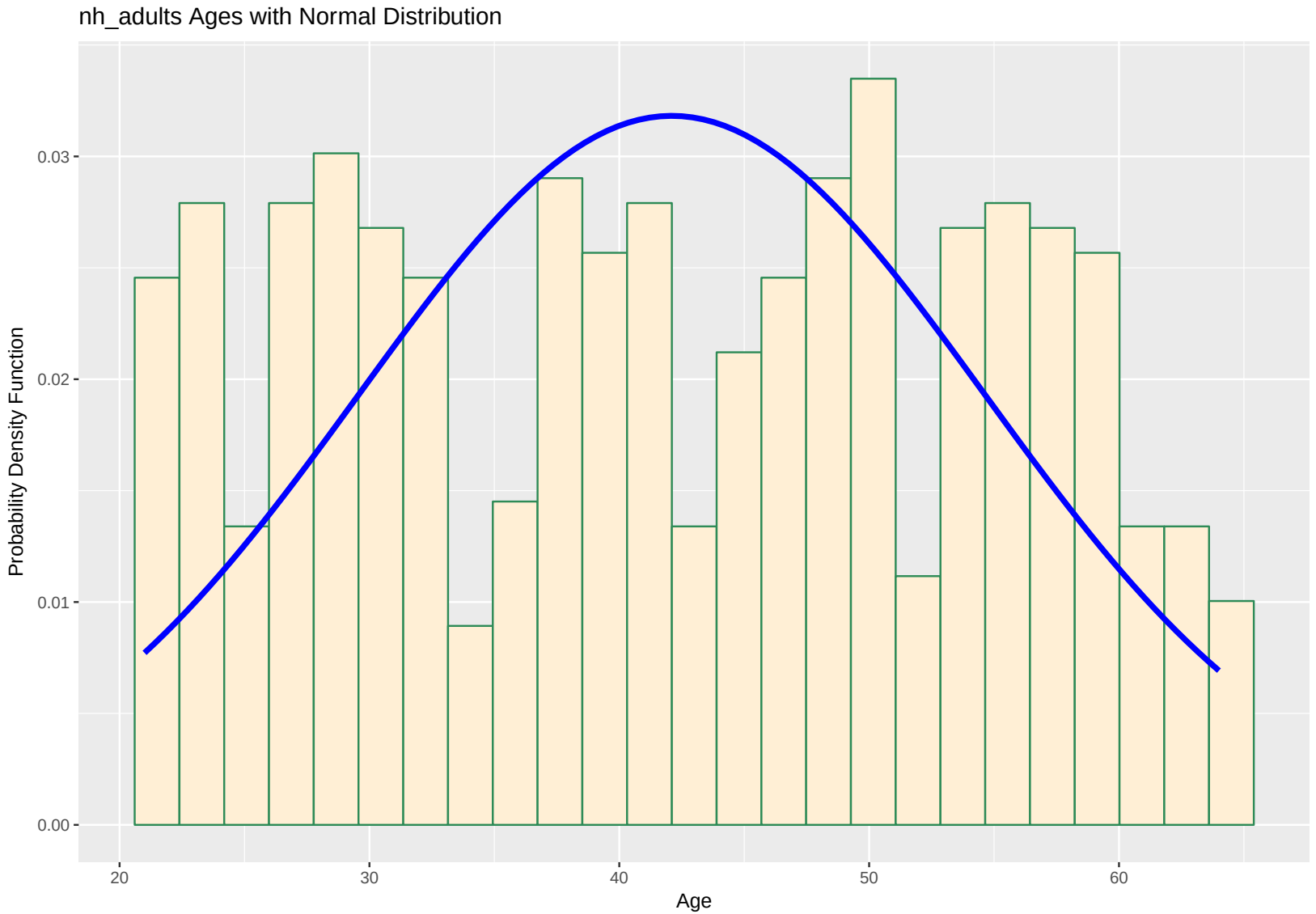


Do the Age data appear skewed? Outlier-prone?



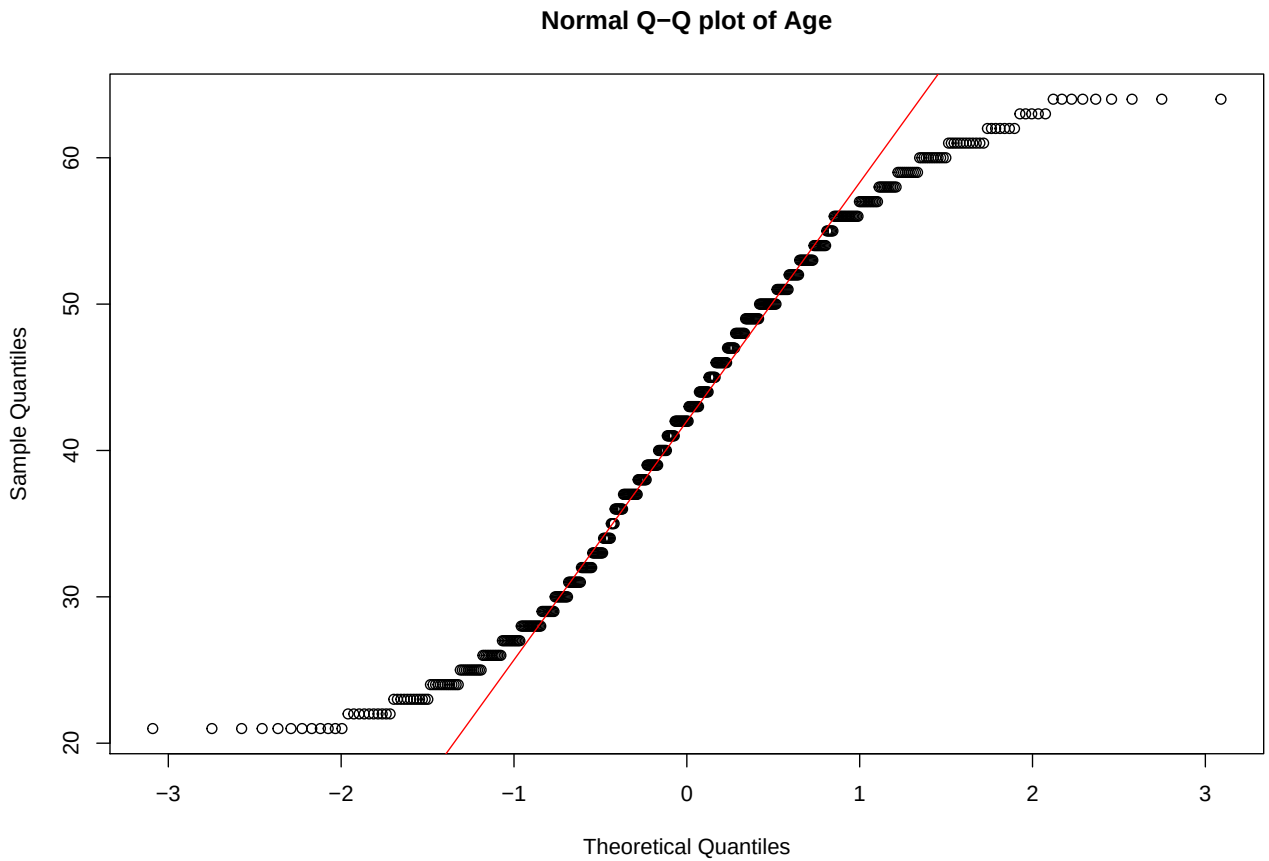
For Age, *skew*1 = 0.01, and kurtosis = -1.2

Ages + Normal Model



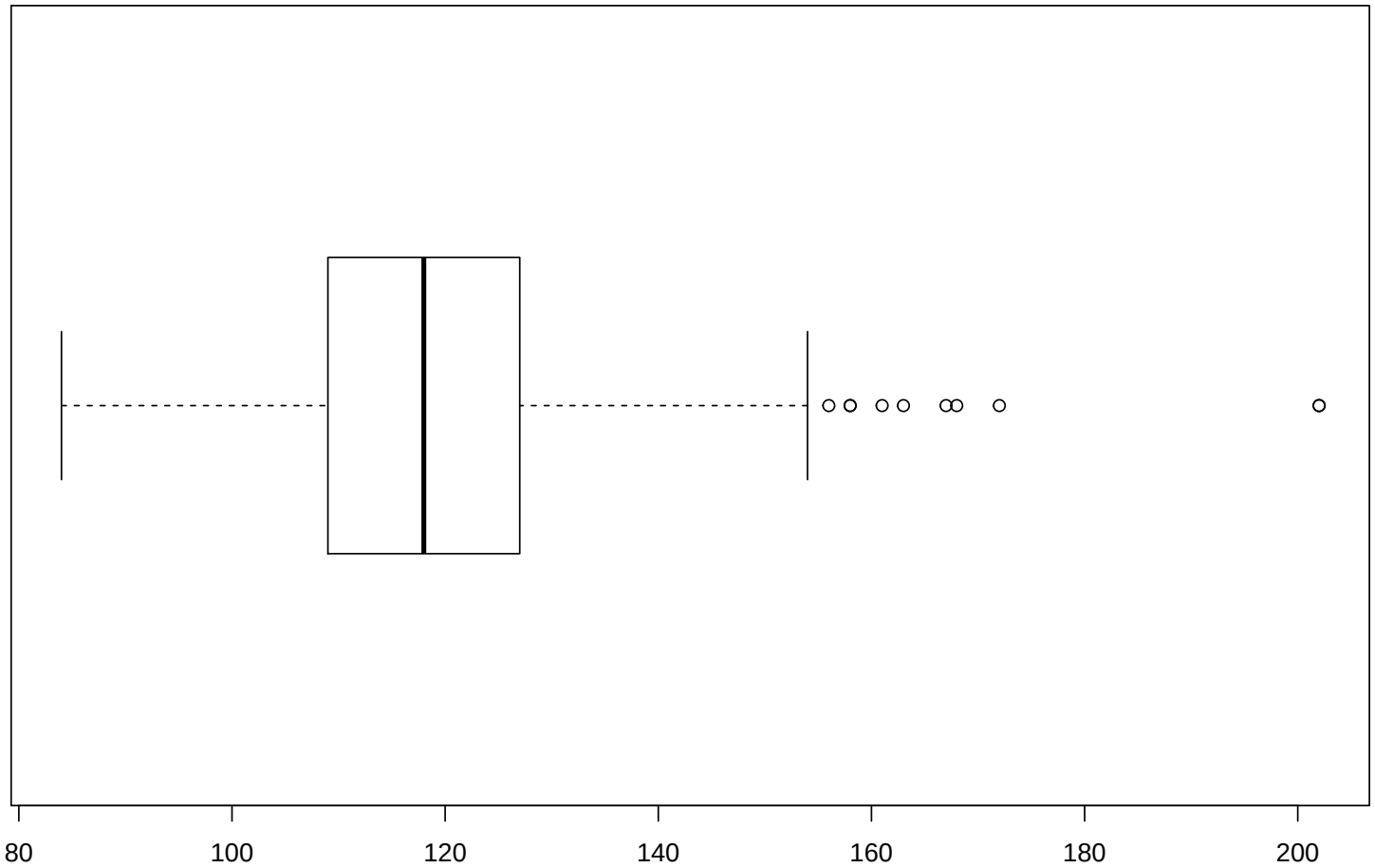
Normal Q-Q plot for Age

```
qqnorm(nh_adults$Age, main = "Normal Q-Q plot of Age")
qqline(nh_adults$Age, col = "red")
```



Identifying outliers (with a boxplot)

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



How the Boxplot identifies Outlier Candidates

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

- Upper fence = 75th percentile + 1.5 IQR
- Lower fence = 25th percentile - 1.5 IQR

For the SBP data, we have

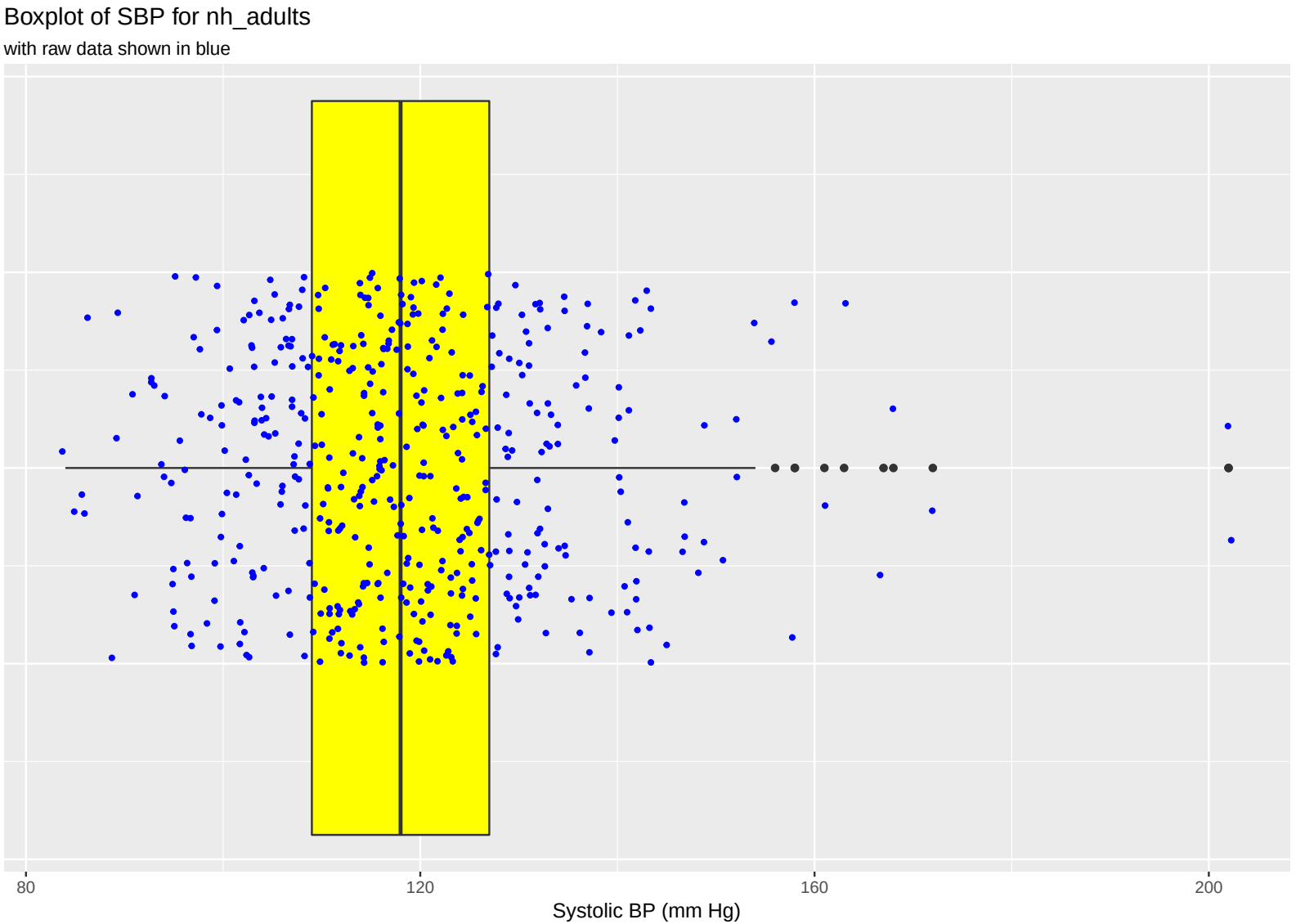
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   84.0   109.0   118.0   118.6   127.0   202.0
   NA's
     15
```

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



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 =

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

.

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.

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.

Coming Next Week

1. Project Instructions
2. Assignment 1 debrief (Assignment 2 due 2017-09-22)
3. Course Notes Chapters 9-11 (plus examples in 12-13)
    - Using transformations to "normalize" data (Ch 9)
    - Summaries withing subgroups (Ch 10)
    - Associations, Using Linear Models

Notes on the Kidney Cancer example, 1

I first asked you what you noticed about the map, in the hope that someone would point out the obvious pattern, which is that many of the countries in the Great Plains but relatively few near the coasts are shaded.

- Why might that be? Could these be the counties with more old people? Ah, but these rates are age-adjusted.
- They're mostly in rural areas: could the health care there be worse than in major cities? Or perhaps people living in rural areas have less healthy diets, or are exposed to more harmful chemicals? Maybe, but the confusing fact is that the highest 10% and the lowest 10% each show disproportionately higher rates in those Great Plains counties.

Notes on the Kidney Cancer example, 2

- Consider a county with 100 white males. If it has even one kidney death in the 1980s, its rate is 1 per thousand per year, which is among the highest in the nation. If it has no such deaths, its rate will be 0, which is the lowest in the nation.
- The observed rates for smaller counties are *much* more variable, and hence they are more likely to be shaded, even if nothing special is truly going on.
- If a small county has an observed rate of 1 per thousand per year, it's probably random fluctuation. But if a large county (like Cuyahoga) has a very high rate, it is probably a real phenomenon.

Source for Kidney Cancer example

My source for this example was Andrew Gelman and Deborah Nolan's book *Teaching Statistics: a bag of tricks* which is the source of a number of things we'll see in the course, including some of the "age guessing" example we've previously done.