# 431 Class 08

Thomas E. Love

2018-09-20

# Today's Agenda

1. Assessing Normality
   - Through Visualization
   - Through Numerical Summaries
   - Through Hypothesis Testing (ugh!)
2. Transformations and the Power Ladder
3. (finally) the Kidney Cancer Maps

# NHANES, but with some new pieces. . .

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

set.seed(20180920) # new random seed

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, HealthGen) %>%
  na.omit

nh_4 <- sample_n(nh_temp, size = 500)
```

# Assessing Normality

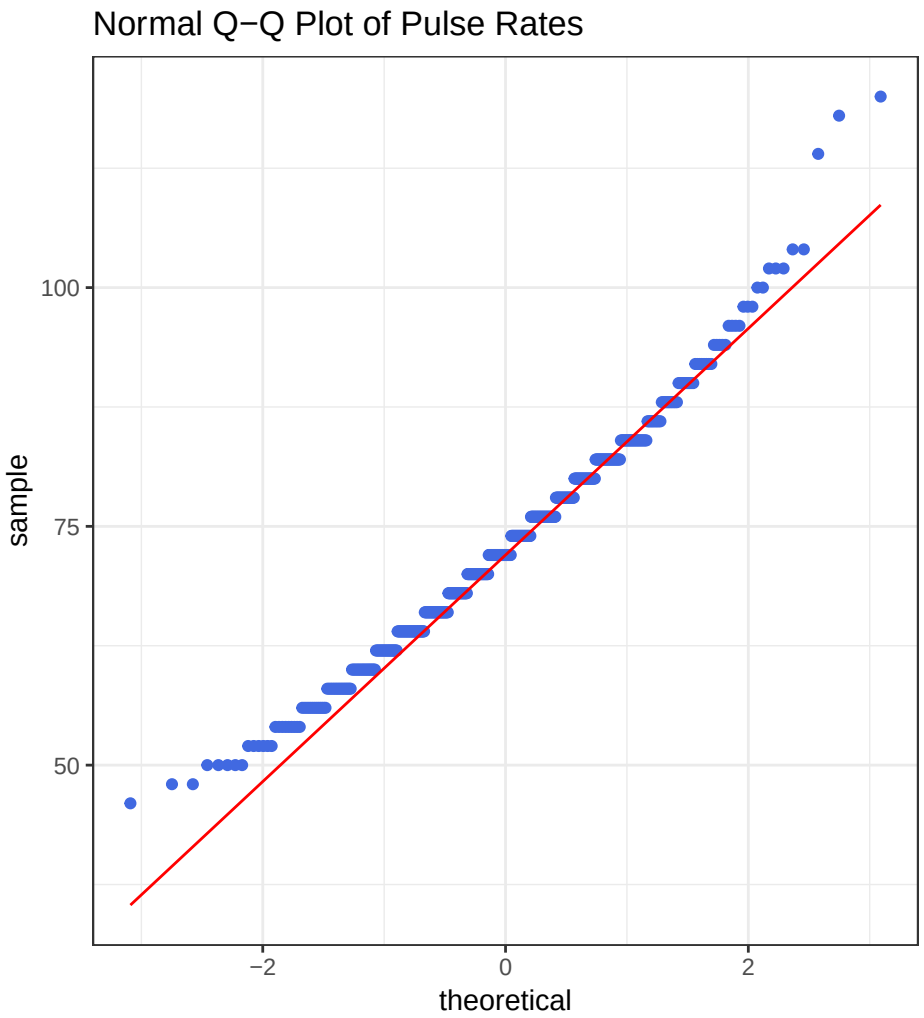
# Does a Normal model fit well for my data?

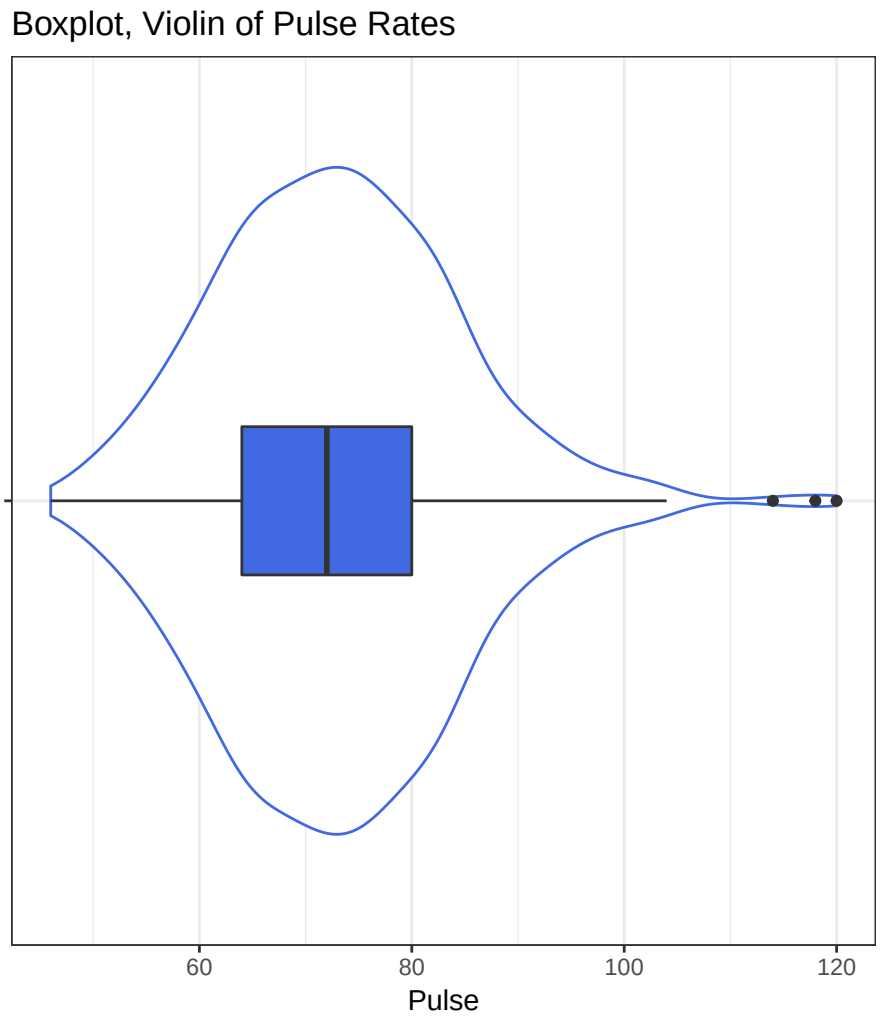
1. Is a Normal Q-Q plot showing something close to a straight line, without clear signs of skew or indications of lots of outliers (heavy-tailedness)?
2. Does a boxplot, violin plot and/or histogram also show a symmetric distribution, where both the number of outliers is modest, and the distance of those outliers from the mean is modest?
3. Do numerical measures match up with the expectations of a normal model?

# Building Descriptive Plots

```
plot_1 <- ggplot(nh_4, aes(sample = Pulse)) +
  geom_qq(col = "royalblue") + geom_qq_line(col = "red") +
  labs(title = "Normal Q-Q Plot of Pulse Rates") +
  theme_bw()

plot_2 <- ggplot(nh_4, aes(x = "", y = Pulse)) +
  geom_violin(fill = "white", col = "royalblue") +
  geom_boxplot(fill = "royalblue", width = 0.2) +
  coord_flip() +
  labs(x = "", title = "Boxplot, Violin of Pulse Rates") +
  theme_bw()

gridExtra::grid.arrange(plot_1, plot_2, ncol = 2)
```

# Pulse Rates in nh_4



# What Summaries to Report (Notes, Section 7)

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.

# Pulse rate in nh_4 - A few numerical summaries

```
nh_4 %>% select(Pulse) %>%
  summarize("n" = n(),
  #           "missing" = sum(is.na(Pulse)),
  #           "non.missing" = sum(!is.na(Pulse)),
            "mean" = mean(Pulse), "sd" = sd(Pulse),
            "median" = median(Pulse),
            "min" = min(Pulse), "max" = max(Pulse),
            "Q25" = quantile(Pulse, 0.25),
            "Q75" = quantile(Pulse, 0.75),
            "IQR" = IQR(Pulse)) %>%
  knitr::kable()
```

| n | mean | sd | median | min | max | Q25 | Q75 | IQR |
|---|---|---|---|---|---|---|---|---|
| 500 | 72.98 | 11.38045 | 72 | 46 | 120 | 64 | 80 | 16 |

# Pulse rate in nh_4 - Assessing skew numerically

```
mosaic::favstats(~ Pulse, data = nh_4)
```

```
 min Q1 median Q3 max  mean       sd   n missing
  46 64     72 80 120 72.98 11.38045 500       0
```

- What does our skew$_1$ measure suggest about symmetry?

$$skew_1 = \frac{(72.98 - 72)}{11.38} = 0.09$$

# Pulse rate in nh_4 - Assessing skew numerically

$$skew_1 = \frac{(72.98 - 72)}{11.38} = 0.09$$

- What about the skewness measure in psych::describe?

```
nh_4 %$% psych::describe(Pulse)
```

```
   vars   n  mean    sd median trimmed   mad min max
X1    1 500 72.98 11.38     72   72.58 11.86  46 120
   range skew kurtosis   se
X1    74 0.48     0.76 0.51
```

# Pulse rate in nh_4 - does Normal model fit data?

- What percentage of the Pulse data fall within one standard deviation of the sample mean?

```
nh_4 %>% count(within1sd =
                 Pulse > mean(Pulse) - sd(Pulse) &
                 Pulse < mean(Pulse) + sd(Pulse) )
```

```
# A tibble: 2 x 2
  within1sd     n
  <lgl>     <int>
1 FALSE       132
2 TRUE        368
```

368 of the 500 (73.6%) observations fall within 1 SD of the mean.

# Empirical Rule for a Normal Model

If the data followed a Normal distribution, perfectly, then:

- 68% of the data would fall within 1 standard deviation of the mean
- 95% of the data would fall within 2 standard deviations of the mean
- 99.7% of the data would fall within 3 standard deviations of the mean

Remember that, regardless of the distribution of the data:

- Half of the data will fall below the median, and half above it.
- Half of the data will fall in the Interquartile Range (IQR).

# Pulse rate in nh_4 - 2 SD Empirical Rule Check

- What percentage of the Pulse data fall within two standard deviations of the sample mean?

```
nh_4 %>% count(within2sd =
                 Pulse > mean(Pulse) - 2*sd(Pulse) &
                 Pulse < mean(Pulse) + 2*sd(Pulse) )
```

```
# A tibble: 2 x 2
  within2sd     n
  <lgl>     <int>
1 FALSE        25
2 TRUE        475
```

475 of the 500 (95%) observations fall within 2 SD of the mean.

# How the Boxplot identifies Outlier Candidates

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

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

## Example: Pulse data in nh_4

```
summary(nh_4$Pulse)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   64.00   72.00   72.98   80.00  120.00
```

```
IQR(nh_4$Pulse)
```

```
[1] 16
```

# Boxplot Identification of Outlier Candidates

- Upper fence = Q75 + 1.5 IQR
- Lower fence = Q25 - 1.5 IQR

```
summary(nh_4$Pulse)
```

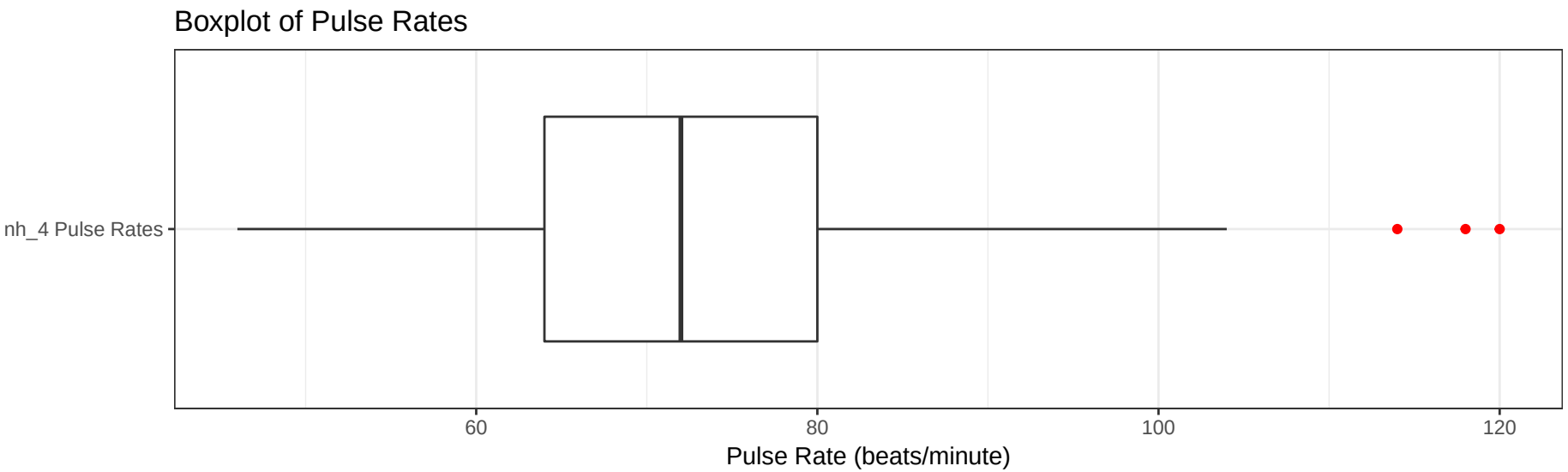
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   64.00   72.00   72.98   80.00  120.00
```

```
nh_4 %>% count("high" = Pulse > 80 + (1.5*16),
               "low" = Pulse < 64 - (1.5*16))
```

```
# A tibble: 2 x 3
  high  low       n
  <lgl> <lgl> <int>
1 FALSE FALSE   497
2 TRUE  FALSE     3
```

# The actual Boxplot

```
ggplot(nh_4, aes(x = "nh_4 Pulse Rates", y = Pulse)) +
  geom_boxplot(outlier.color = "red") +
  coord_flip() +
  theme_bw() +
  labs(x = "", y = "Pulse Rate (beats/minute)",
       title = "Boxplot of Pulse Rates")
```



# Outliers and Z scores (Notes, Section 8.2)

The maximum pulse rate in the data is 120.

```
mosaic::favstats(~ Pulse, data = nh_4)
```

```
 min Q1 median Q3 max  mean       sd   n missing
  46 64     72 80 120 72.98 11.38045 500       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 Pulse = 120

$$Z = \frac{value - mean}{sd}.$$

For the Pulse data, the mean = 72.98 and the standard deviation is 11.38, so we have Z score for 120 =

$$\frac{120 - 72.98}{11.38} = \frac{47.02}{11.38} = 4.13$$

.

- 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 = 4.13?

If the data really followed a Normal distribution, we could calculate the probability of obtaining as extreme a Z score as 4.13.

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 4.13 or higher, *in absolute value*. So we calculate the probability of 4.13 or more, and add it to the probability of -4.13 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 = 4.13, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 1.813816e-05
```

```
pnorm(q = -4.13, mean = 0, sd = 1, lower.tail = TRUE)
```

```
[1] 1.813816e-05
```

# But the Normal distribution is symmetric

```
2*pnorm(q = 4.13, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 3.627632e-05
```

The probability that a single draw from a Normal distribution with mean 0 and standard deviation 1 will produce a value as extreme as 4.13 is 0.000036

The probability that a single draw from a Normal distribution with mean 72.98 and standard deviation 11.38 will produce a value as extreme as 120 is also 0.000036, since the Normal distribution is completely characterized by its mean and standard deviation.

So, is 120 an outlier here? Do the Pulse data look like they come from a Normal distribution by this metric?

# Hypothesis Testing to assess Normality

Don't. Graphical approaches are **far** better than hypothesis tests.

```
shapiro.test(nh_4$Pulse)
```

```
    Shapiro-Wilk normality test

data:  nh_4$Pulse
W = 0.98309, p-value = 1.466e-05
```

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

# Why not test for Normality?

There are multiple hypothesis testing schemes (Kolmogorov-Smirnov, etc.) and each looks for one specific violation of a Normality assumption. None can capture the wide range of issues our brains can envision, and none by itself is great at its job.

- With any sort of reasonable sample size, 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.
- And without a reasonable sample size, the test is essentially useless.

Whenever you can avoid hypothesis testing and instead actually plot the data, you should plot the data.

# 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

# Power Transformations Ladder (Notes: Section 9)

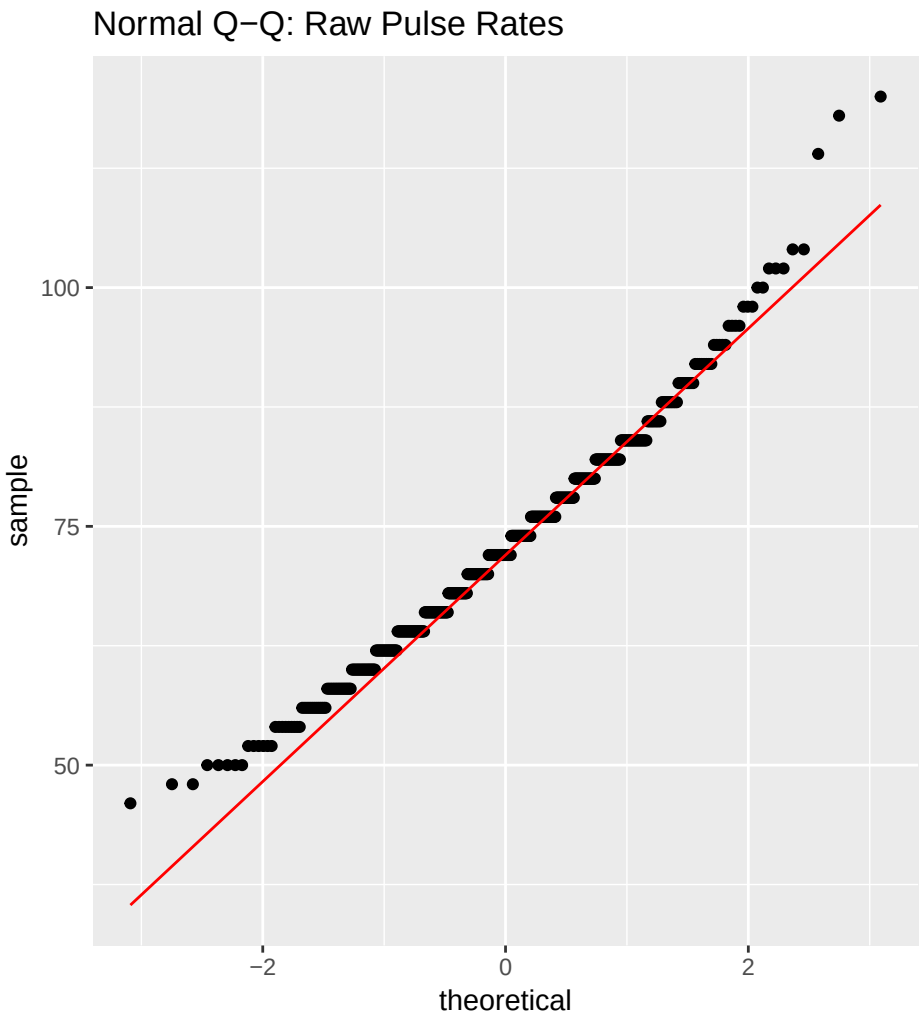
The key notion in re-expression of a single variable to obtain a better fit to a Normal model, is that of a **ladder of power transformations**, which can apply to any unimodal data.

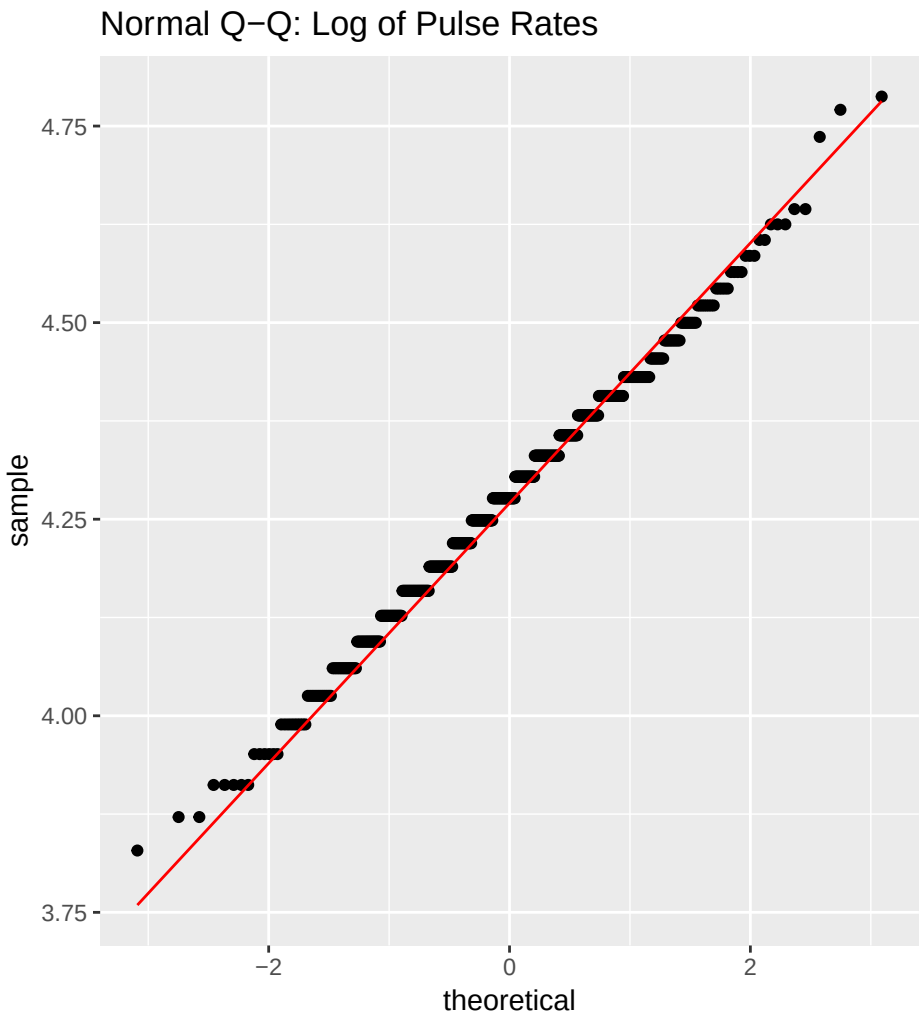
| Power | Transformation |
|---|---|
| 3 | $x^3$ |
| 2 | $x^2$ |
| 1 | x (unchanged) |
| 0.5 | $x^{0.5} = \sqrt{x}$ |
| 0 | ln x |
| -0.5 | $x^{-0.5} = 1/\sqrt{x}$ |
| -1 | $x^{-1} = 1/x$ |
| -2 | $x^{-2} = 1/x^2$ |

# nh_4 Pulse Rates, and their Natural Logarithms

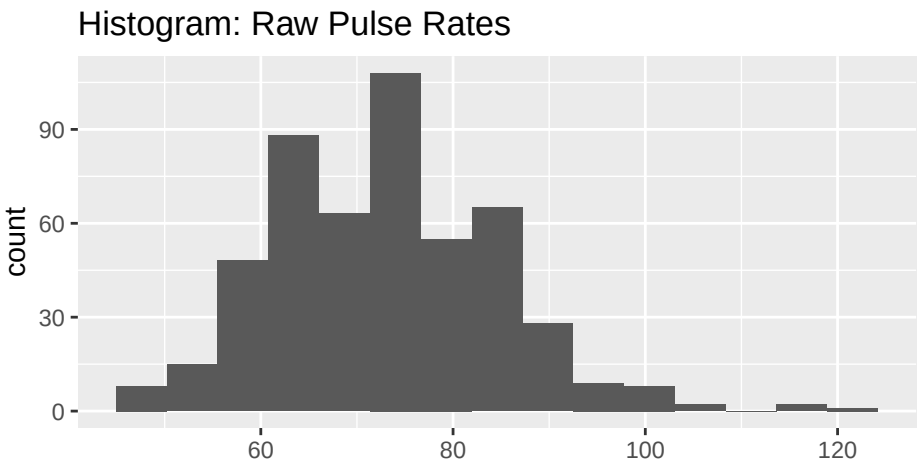
```
p1 <- ggplot(data = nh_4, aes(sample = Pulse)) +
  geom_qq() + geom_qq_line(col = "red") +
  labs(title = "Normal Q-Q: Raw Pulse Rates")

p2 <-  ggplot(data = nh_4, aes(sample = log(Pulse))) +
  geom_qq() + geom_qq_line(col = "red") +
  labs(title = "Normal Q-Q: Logarithm of Pulse Rates")

gridExtra::grid.arrange(p1, p2, ncol = 2)
```

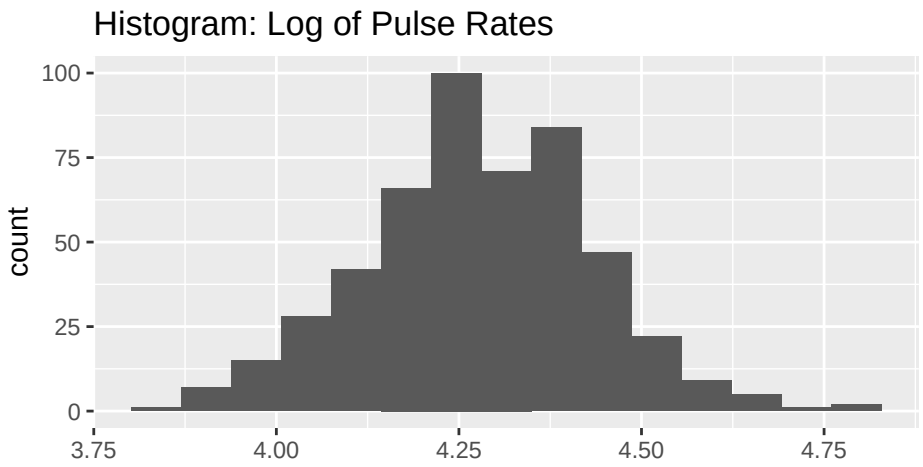
# nh_4 Pulse Rates, and their Natural Logarithms

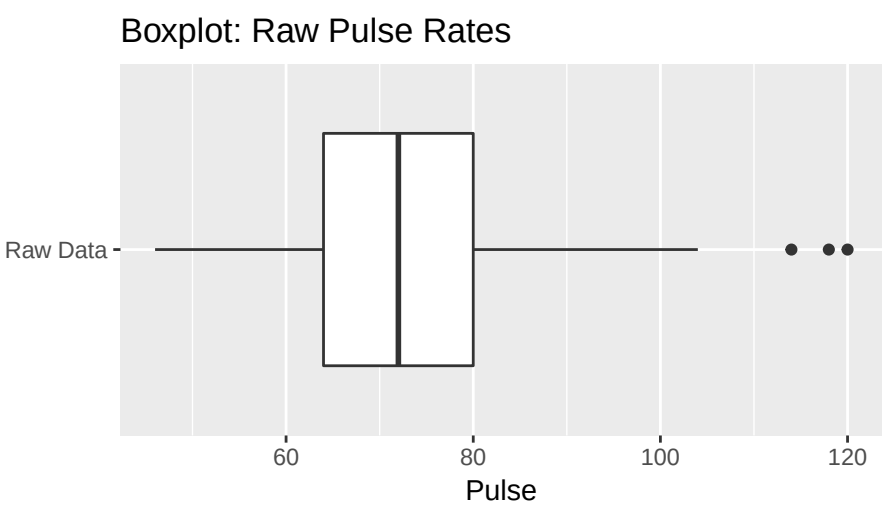


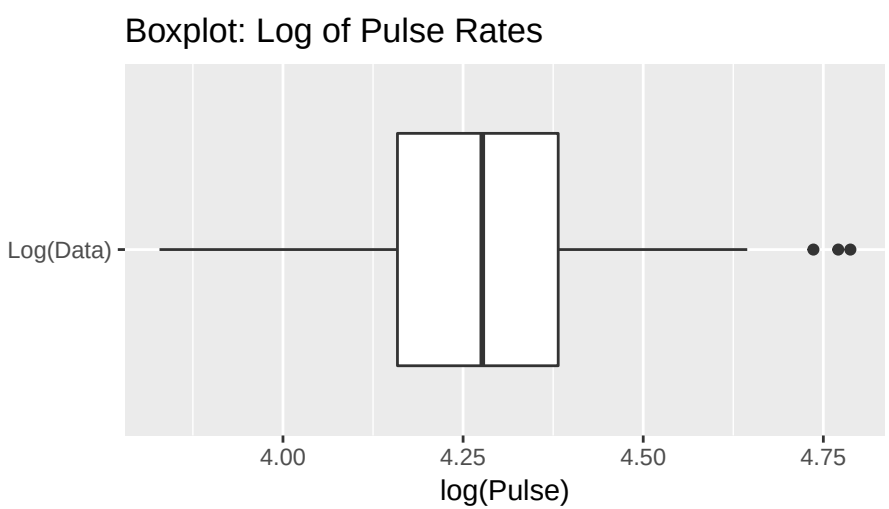



# nh_4 Pulse Rates, and their Natural Logarithms



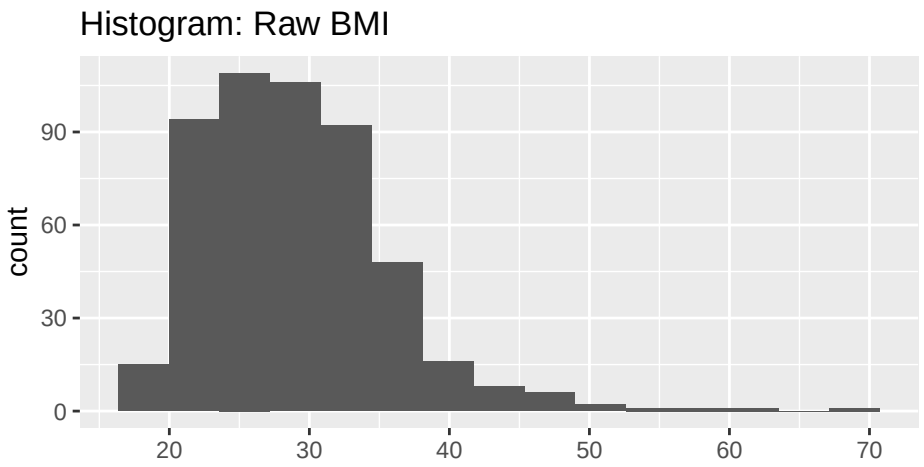



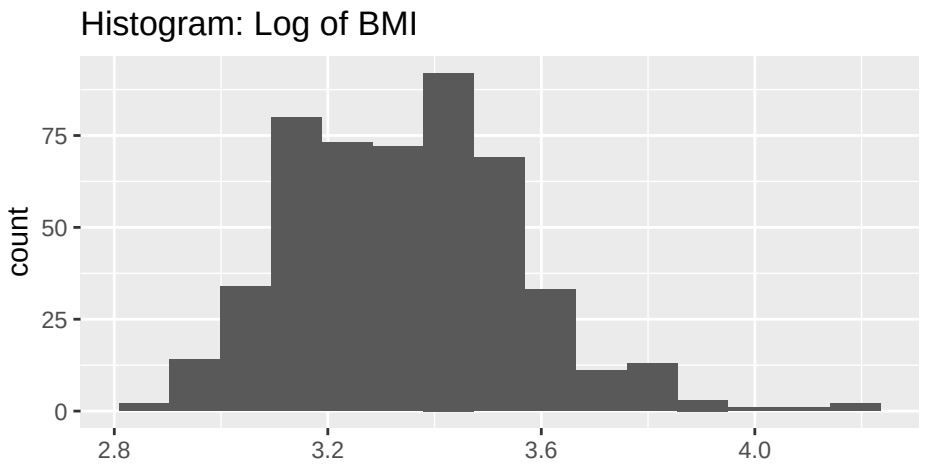
# Using the Ladder

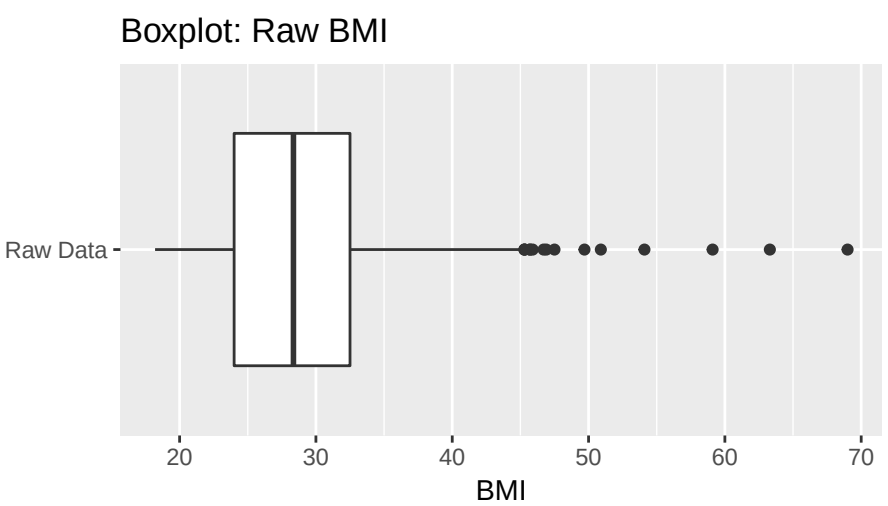
- The ladder is most useful for strictly positive, ratio variables.
- Sometimes, if 0 is a value in the data set, we will add 1 to each value before applying a transformation like the logarithm.
- Interpretability is often an important criterion, although back-transformation at the end of an analysis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

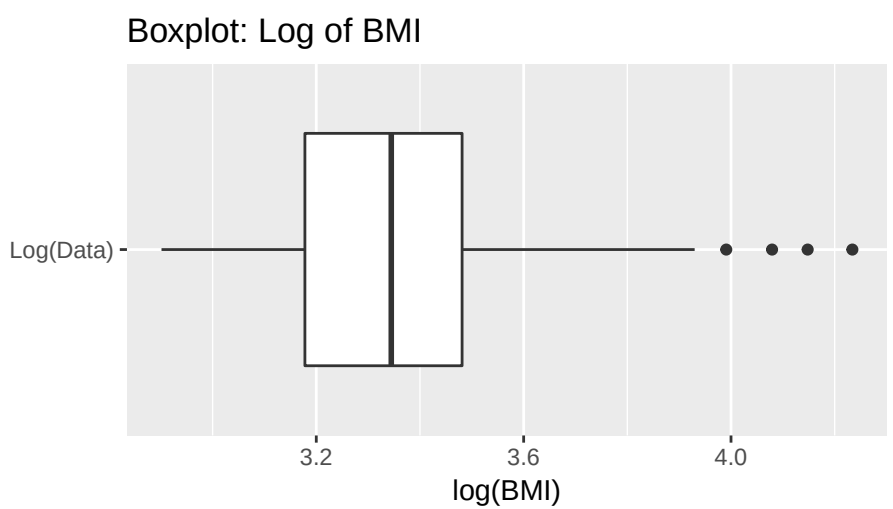
| Power | -2 | -1 | -0.5 | 0 | 0.5 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|
| Transformation | $1/x^2$ | $1/x$ | $1/\sqrt{x}$ | ln x | $\sqrt{x}$ | x | $x^2$ | $x^3$ |

# nh_4 Body-Mass Index Data (Raw data and Log)

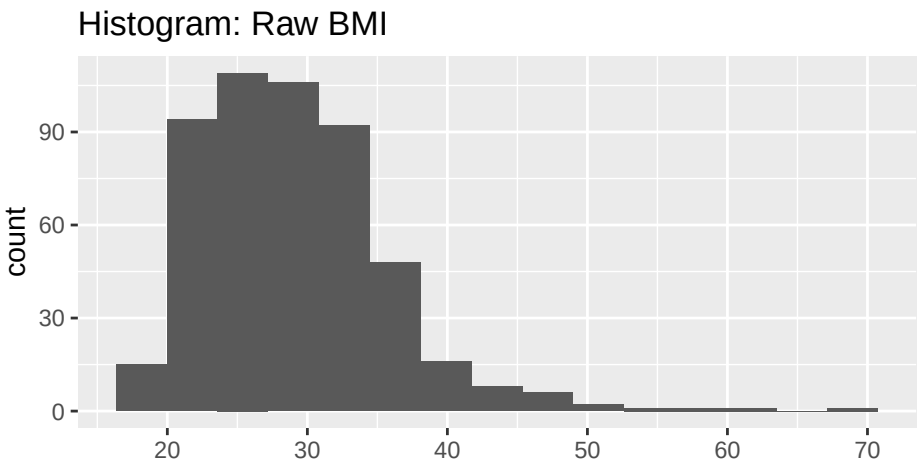


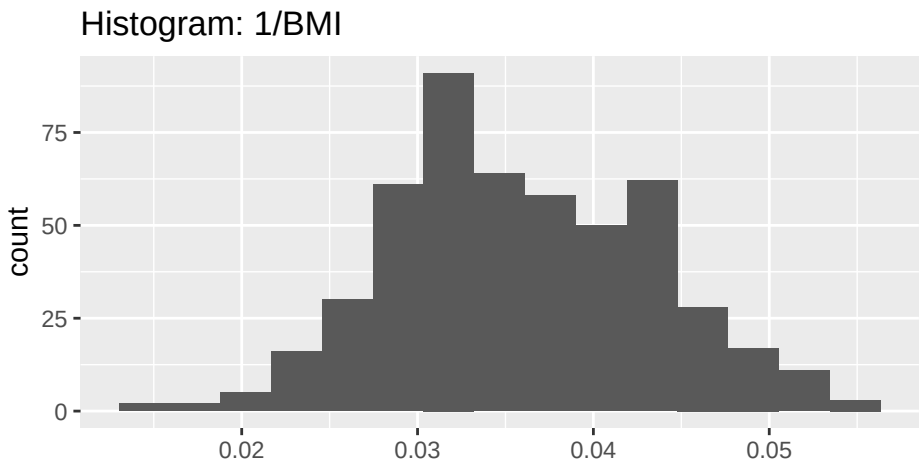



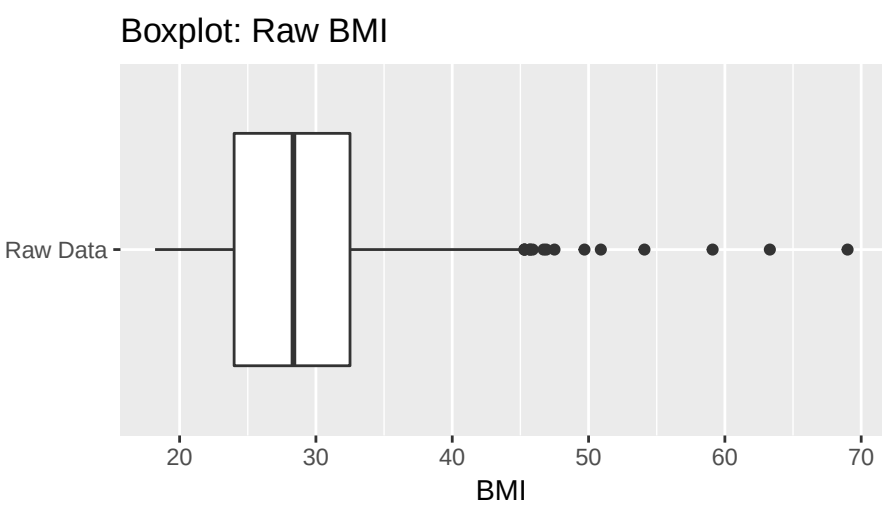



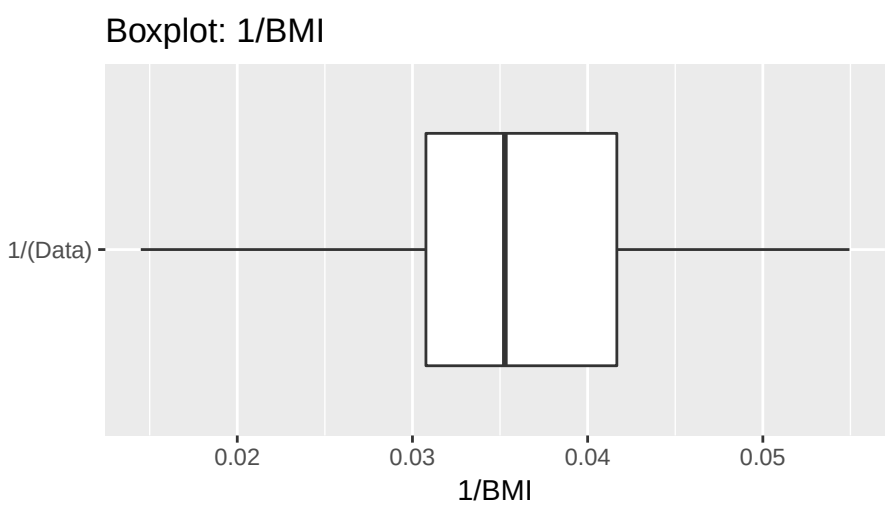



# nh_4 BMI - move down the ladder to 1/BMI?

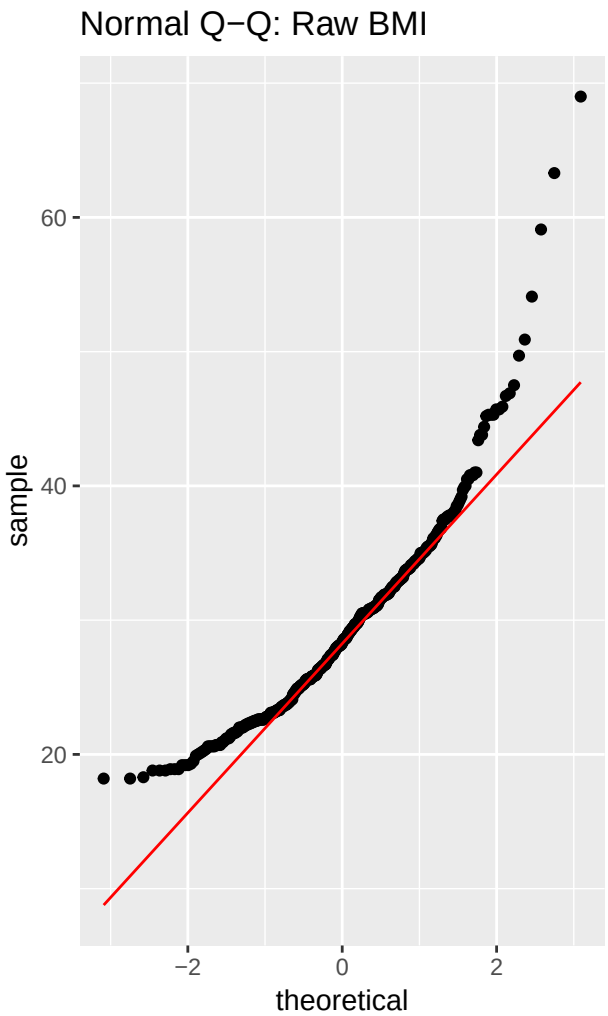


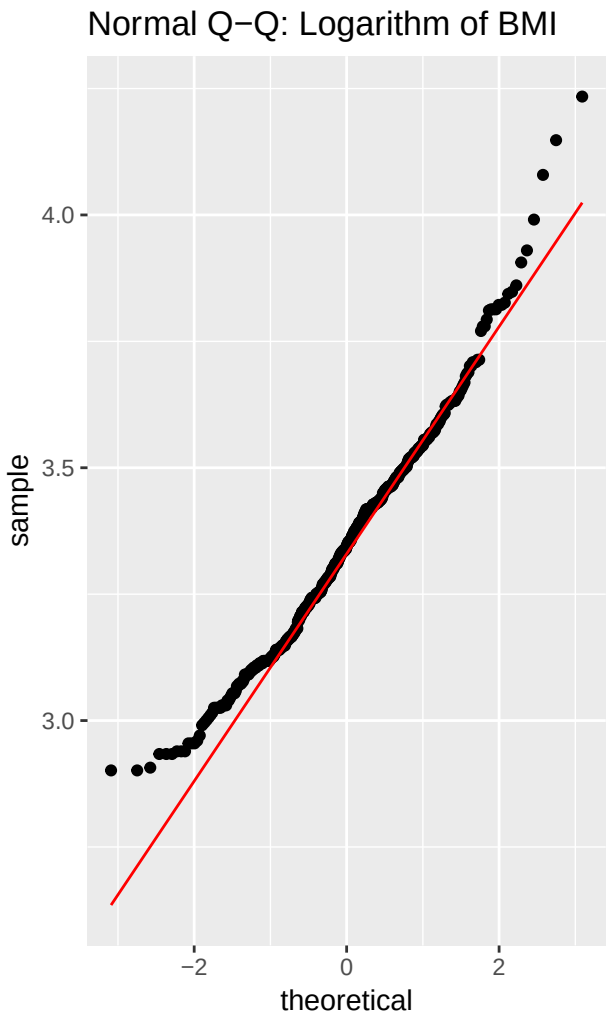



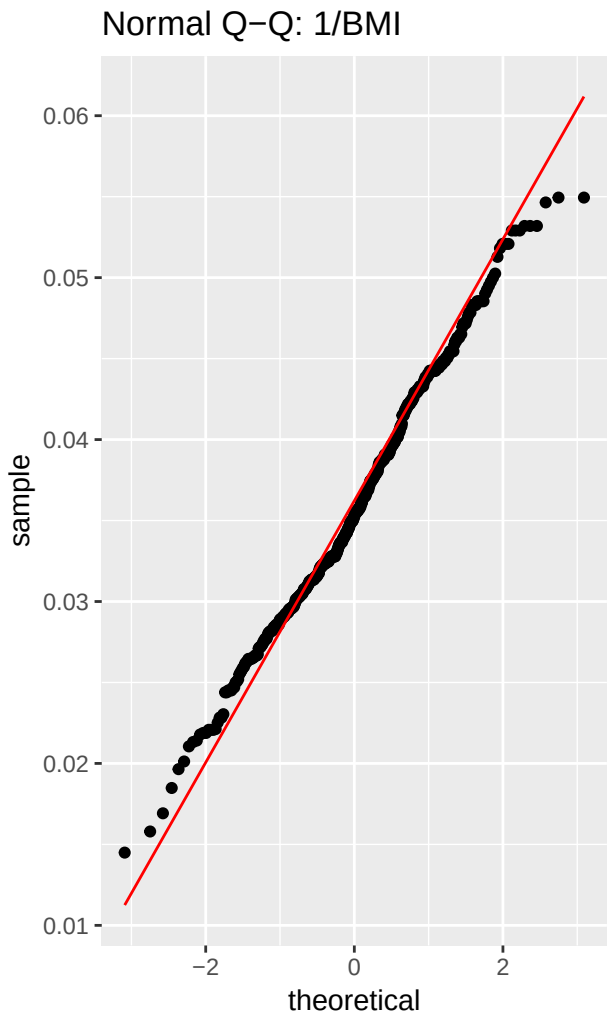



# Normal Q-Q plots for BMI



# Conclusions

This is just a small piece of the toolbox for visualizations that we'll create in this class. Many additional tools are on the way, but the main idea won't change. Using the `ggplot2` package, we can accomplish several critical tasks in creating a visualization, including:

- Identifying (and labeling) the axes and titles
- Identifying a type of `geom` to use, like a point, bar or histogram
- Changing fill, color, shape, size to facilitate comparisons
- Building "small multiples" of plots with faceting

Good data visualizations make it easy to see the data, and `ggplot2`'s tools make it relatively difficult to make a really bad graph.

# Group Task: 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.

## Your Tasks

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

# Highest kidney cancer death rates

## Lowest kidney cancer death rates

# Reminders

- Homework 3 is due Friday at Noon.
- Project Groups formed Tuesday in class.
- So far, we've covered most of the material in Chapters 1-10 of our Course Notes.
  - Part A covers Chapters 1-14.
  - Next Time: Studying Association with Scatterplots and Correlations (Chapters 11-12)

# 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

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.