431 Class 07

github.com/THOMASELOVE/2019-431

2019-09-17

Today's Agenda

Are these data well described by a Normal model?

1. Why should we care?
2. How can we check?
    - Histograms
    - Normal Q-Q plots
    - Boxplots with Violins
    - Numerical Approaches
3. What can we do about non-Normal data?
    - Summarize it with median and IQR, not mean and SD
    - Transform the data (perhaps a power transformation)?

Today's Packages

The R packages we're using today are NHANES, magrittr, janitor and tidyverse.

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

CWRU Colors

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

Our nh2 data set, yet again

```r
set.seed(20190910) # so we can get the same sample again

nh2 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, SurveyYr, Age, Height, Weight, BMI, Pulse,
           SleepHrsNight, BPSysAve, BPDiaAve, Gender,
           PhysActive, SleepTrouble, Smoke100,
           Race1, HealthGen, Depressed) %>%
    rename(SleepHours = SleepHrsNight, Sex = Gender,
           SBP = BPSysAve, DBP = BPDiaAve) %>%
    filter(Age > 20 & Age < 80) %>% ## ages 21-79 only
    drop_na() %>% # removes all rows with NA
    sample_n(., size = 1000) %>% # sample 1000 rows
    clean_names() # from the janitor package (snake case)
```

Today's Variables

Name	Description
`pulse`	60 second pulse rate

Name	Levels	Description
`sex`	F, M	Sex of study subject

Name	Levels	Description
`health_gen`	5	Self-reported overall general health

Building a Subset of Interest

Let's look at a subset of the `nh2` data, consisting of males who rated their general health as either Good or Very Good.

```
nh2 %>% tabyl(sex, health_gen)
```

```
    sex Excellent Vgood Good Fair Poor
 female        73   165  179   48   12
   male        71   164  204   76    8
```

How many people are we talking about?

Obtaining our Subset of Interest

```r
nh2_GVGmales <- nh2 %>%
  filter(sex == "male" &
           health_gen %in% c("Good", "Vgood"))

dim(nh2_GVGmales)
```

```
[1] 368  17
```

Let's see what we can learn about the pulse rates of these subjects.

Mean or the Median to describe center?

We're looking at pulse rates in our `nh2_GVGmales` subset of the `nh2` data, consisting of males who rated their general health as either Good or Very Good.

```
nh2_GVGmales %$% mosaic::favstats(~ pulse)
```

```
 min Q1 median Q3 max     mean       sd   n missing
  44 64     70 80 114 71.48913 11.94811 368       0
```

1. Should we choose the mean or the median to represent the center of the distribution?

How should we describe spread?

Same subset:

```
nh2_GVGmales %$% mosaic::favstats(~ pulse)
```

```
 min Q1 median Q3 max     mean       sd   n missing
  44 64     70 80 114 71.48913 11.94811 368       0
```

2. Should we choose the standard deviation or the interquartile range to describe the spread of the distribution?

```
nh2_GVGmales %$% IQR(pulse)
```

```
[1] 16
```

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:

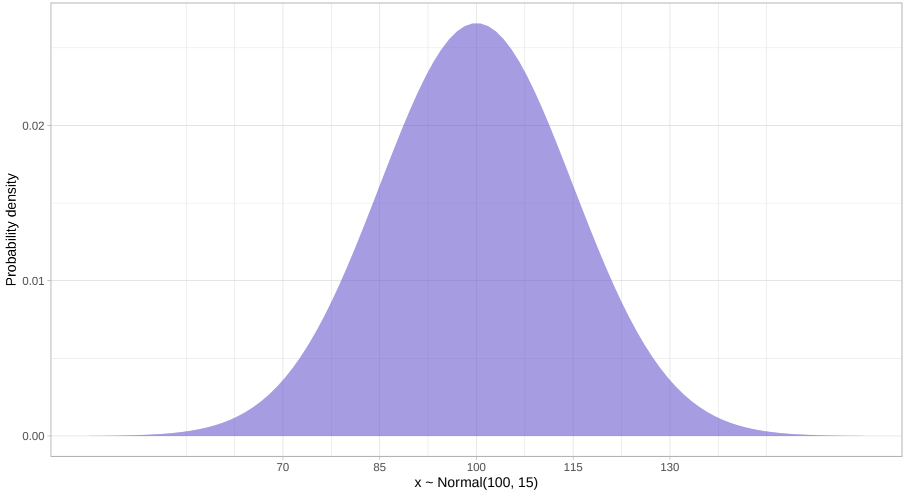
- If the data are skewed, report the median and IQR (or the three middle quantiles). You may want to include the mean and standard deviation, but you should point out why the mean and median differ. The fact that the mean and median do not agree is a sign that the distribution may be skewed. A histogram 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.

What is a Normal Model?

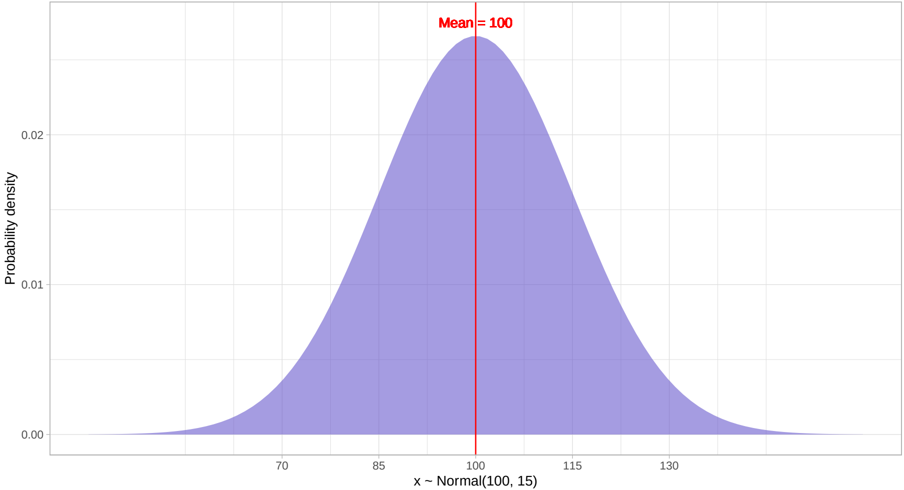
By a Normal model, we mean that the data are assumed to be the result of selecting at random from a probability distribution called the Normal (or Gaussian) distribution, which is characterized by a bell-shaped curve, and can be defined by establishing the values of two parameters: the mean and the standard deviation.

```r
mean_1 <- 100; sd_1 <- 15
x_1 <- seq(-4, 4, length = 100)*sd_1 + mean_1
y_1 <- dnorm(x_1, mean_1, sd_1)
tempdat <- tibble(x_1, y_1)

ggplot(tempdat, aes(x = x_1, y = y_1)) +
  geom_area(fill = "slateblue", alpha = 0.6) +
  theme_light() +
  labs(x = "x ~ Normal(100, 15)",
       y = "Probability density") +
  scale_x_continuous(breaks = c(70, 85, 100, 115, 130))
```

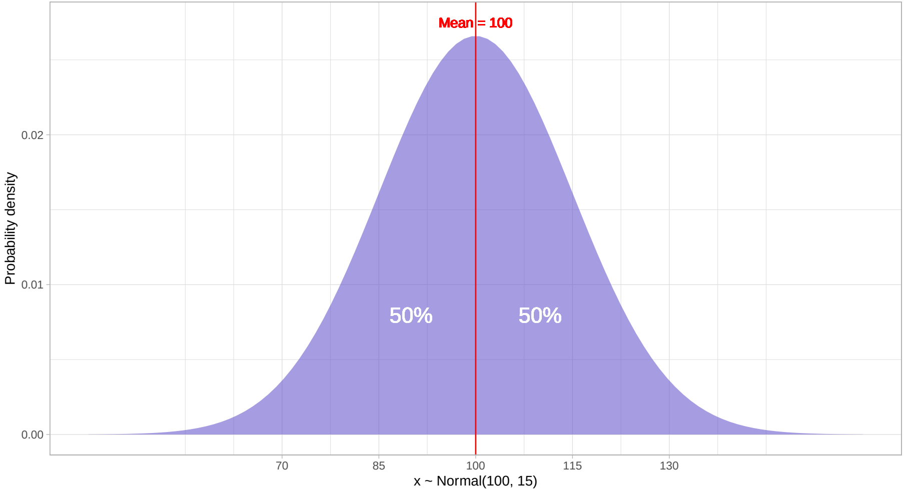
Plotting the Normal model with mean 100, sd 15



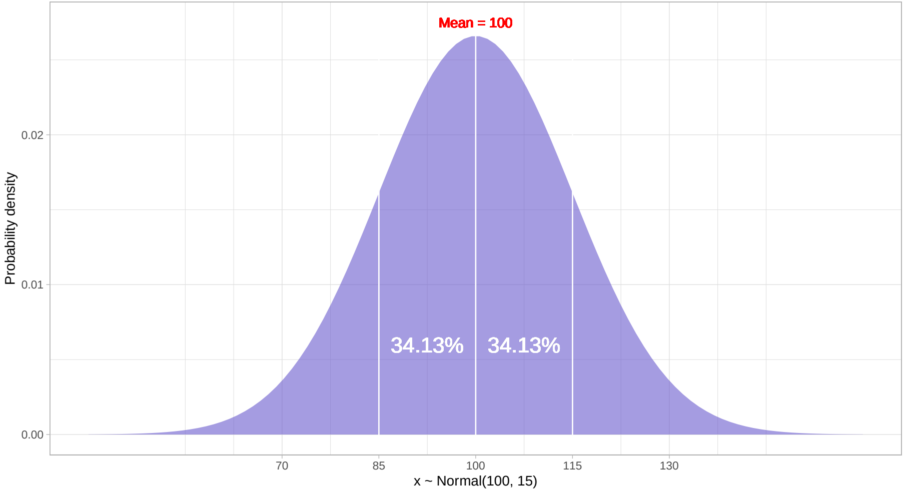
Normal Model (Mean = 100, SD = 15)



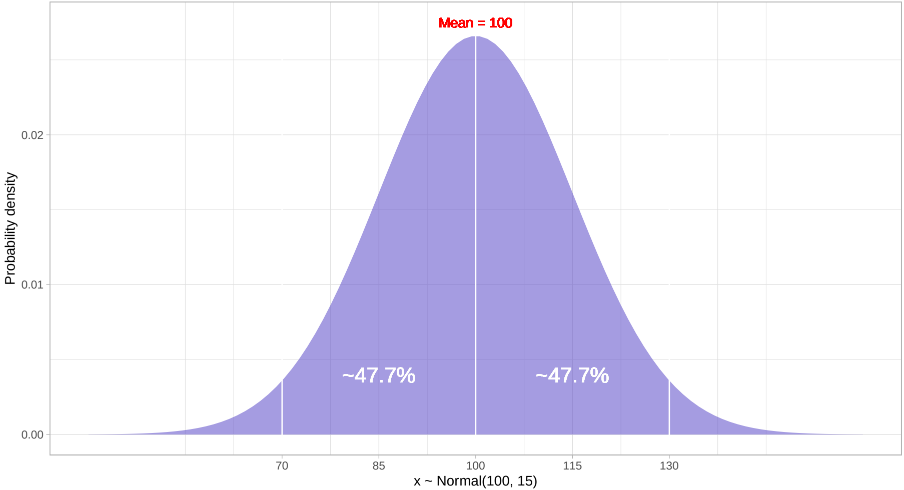
Normal Model: Mean = Median (symmetric)



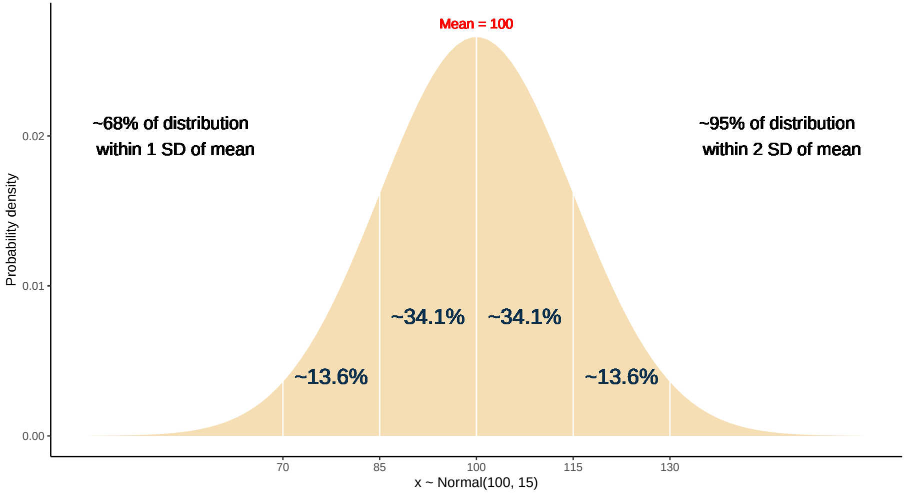
Normal Model: 68.26% of data within 1 SD of mean



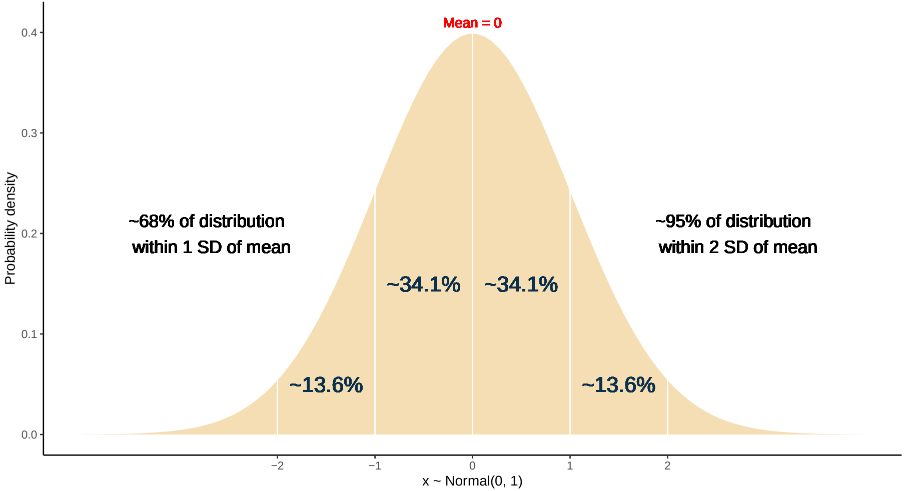
Normal Model: 95.45% of data within 2 SD of mean



What a Normal (100, 15) Model Means!



What a Normal (0, 1) Model Means!



Empirical Rule for a Normal Model

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

- 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:

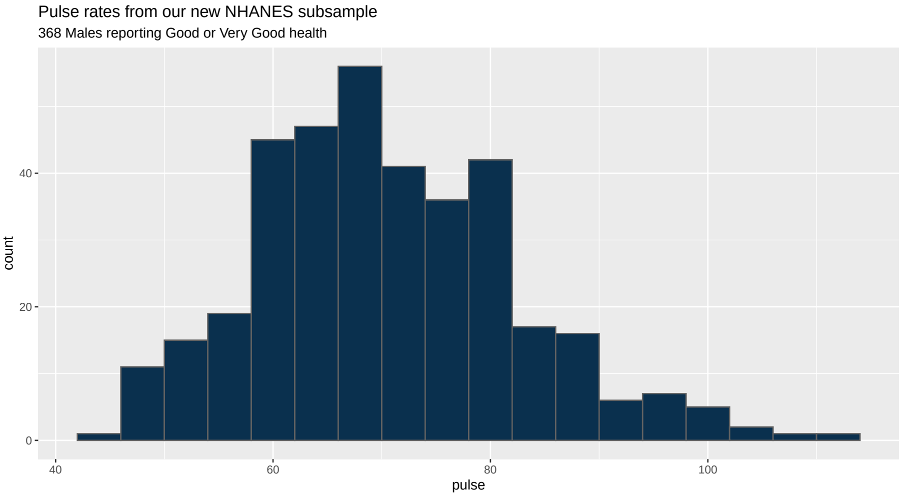
- Half of the data will fall below the median, and half above it.
- Half of the data will fall in the Interquartile Range (IQR).

Histogram of Pulse Rates in `nh2_GVGmales`

```r
ggplot(nh2_GVGmales, aes(x = pulse)) +
  geom_histogram(binwidth = 4,
                 fill = cwru.blue, col = cwru.gray) +
  labs(title = "Pulse rates from our new NHANES subsample",
       subtitle =
         paste0(nrow(nh2_GVGmales),
                " Males reporting Good or Very Good health"))
```

Plot on the next slide. Could a Normal model describe these data well?

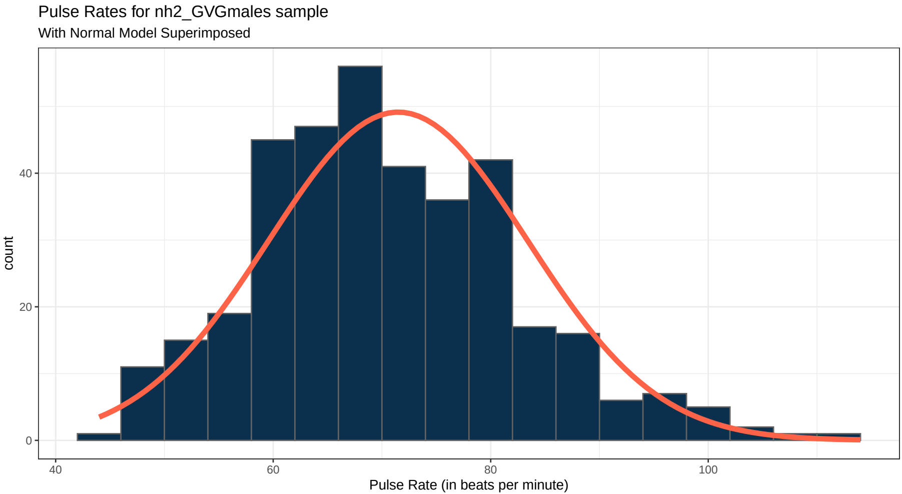
Would a Normal Model describe these rates well?



Superimposing a Normal model

```
res <- mosaic::favstats(~ pulse, data = nh2_GVGmales)
bin_w <- 4 # specify binwidth

ggplot(nh2_GVGmales, aes(x = pulse)) +
  geom_histogram(binwidth = bin_w,
                 fill = cwru.blue,
                 col = cwru.gray) +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                            sd = res$sd) * res$n * bin_w,
    col = "tomato", size = 2) +
labs(title = "Pulse Rates for nh2_GVGmales sample",
       subtitle = "With Normal Model Superimposed",
       x = "Pulse Rate (in beats per minute)")
```

Superimposing a Normal model



Boxplot Identification of Outlier Candidates

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

```
nh2_GVGmales %>% select(pulse) %>% summary()
```

```
    pulse
 Min.   : 44.00
 1st Qu.: 64.00
 Median : 70.00
 Mean   : 71.49
 3rd Qu.: 80.00
 Max.   :114.00
```

Boxplot Identification of Outlier Candidates

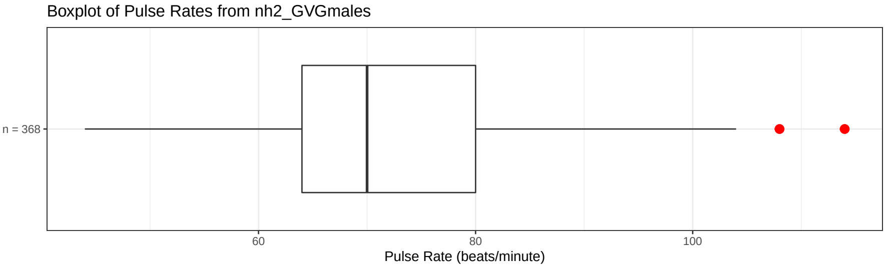
- Upper fence = Q75 + 1.5 IQR
- Lower fence = Q25 - 1.5 IQR

```
nh2_GVGmales %>% 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   366
2 TRUE  FALSE     2
```

The actual Boxplot

```
ggplot(nh2_GVGmales, aes(x = "n = 368", y = pulse)) +
  geom_boxplot(outlier.color = "red", outlier.size = 3) +
  coord_flip() +
  theme_bw() +
  labs(x = "", y = "Pulse Rate (beats/minute)",
       title = "Boxplot of Pulse Rates from nh2_GVGmales")
```



Outliers and Z scores (Notes, Section 8.2)

The maximum pulse rate in the data is 114.

```r
mosaic::favstats(~ pulse, data = nh2_GVGmales)
```

```
 min Q1 median Q3 max     mean       sd   n missing
  44 64     70 80 114 71.48913 11.94811 368       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 = 114

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

For the Pulse data, the mean = 71.5 and the standard deviation is 11.9, so we have Z score for 114 =

$$\frac{114 - 71.5}{11.9} = \frac{42.5}{11.9} = 3.57$$

.

- 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 = 3.57?

If the data really followed a Normal distribution, we could calculate the probability of obtaining as extreme a Z score as 3.57.

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 3.57 or higher, *in absolute value*. So we calculate the probability of 3.57 or more, and add it to the probability of -3.57 or less, to get an answer to the question of how likely is it to see an outlier this far away from the mean.

```r
pnorm(q = 3.57, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 0.0001784906
```

```r
pnorm(q = -3.57, mean = 0, sd = 1, lower.tail = TRUE)
```

```
[1] 0.0001784906
```

But the Normal distribution is symmetric

```r
2*pnorm(q = 3.57, mean = 0, sd = 1, lower.tail = FALSE)
```

```
[1] 0.0003569812
```

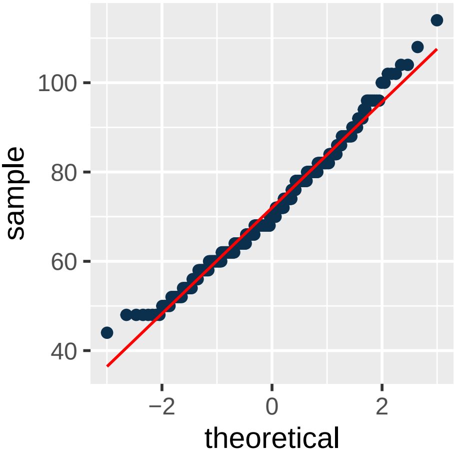
The probability that a single draw from a Normal distribution with mean 0 and standard deviation 1 will produce a value as extreme as 3.57 is 0.00036

The probability that a single draw from a Normal distribution with mean 71.5 and standard deviation 11.9 will produce a value as extreme as 114 is also 0.00036, since the Normal model is completely characterized by its mean and standard deviation.

So, is 114 an outlier here? Do the `pulse` data in this sample look like they come from a Normal distribution by this metric?

Normal Q-Q plot for these Pulse Rates

```
ggplot(nh2_GVGmales, aes(sample = pulse)) +
  geom_qq(col = cwru.blue) + geom_qq_line(col = "red")
```



What is a Normal Q-Q Plot?

- The y-axis shows the data in our sample.
- The x-axis shows the “theoretical” values (Z scores) that we would observe in a Normal distribution with the same number of observations as our data.
- A diagonal line is drawn as a reference, as determined by the mean and standard deviation of the data.

What is a Z score?

Take a value, x, drawn from a distribution with a known mean and standard deviation. Then

$$Z = \frac{x - mean}{sd}$$

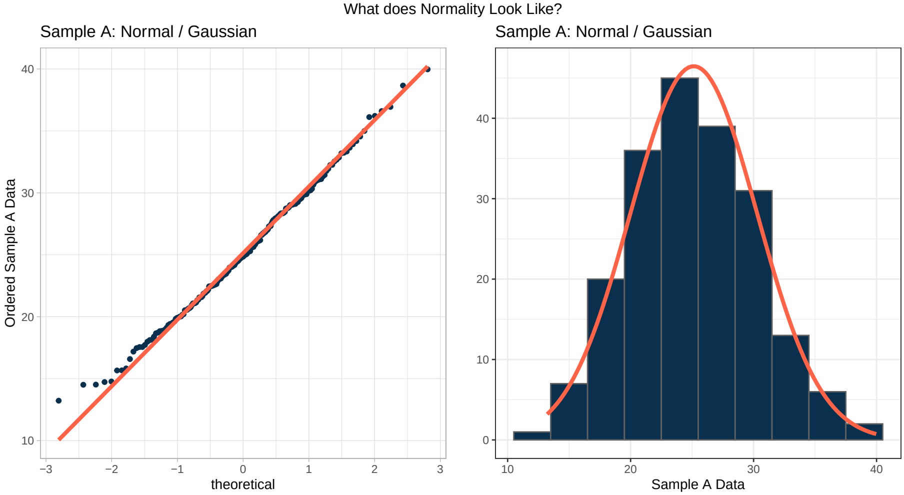
Interpreting the Normal Q-Q plot?

The Normal Q-Q plot can help us identify data as well approximated by a Normal distribution, or not, because of:

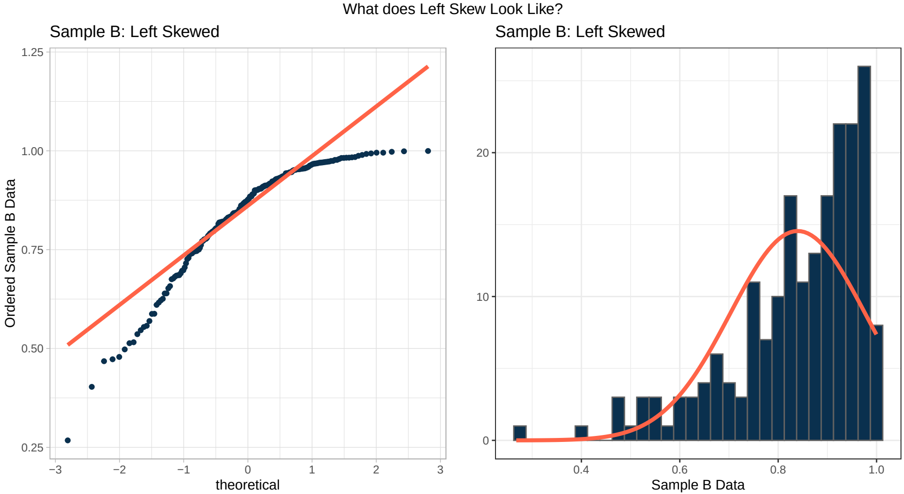
- skew (including distinguishing between right skew and left skew)
- behavior in the tails (which could be heavy-tailed [more outliers than expected] or light-tailed)

1. Normally distributed data would be indicated by close adherence of the points to the diagonal reference line.
2. Skew is indicated by substantial curving (on both ends of the distribution) in the points away from the reference line (if both ends curve up, we have right skew; if both ends curve down, this indicates left skew)
3. An abundance or dearth of outliers (as compared to the expectations of a Normal model) are indicated in the tails of the distribution by an “S” shape or reverse “S” shape in the points.

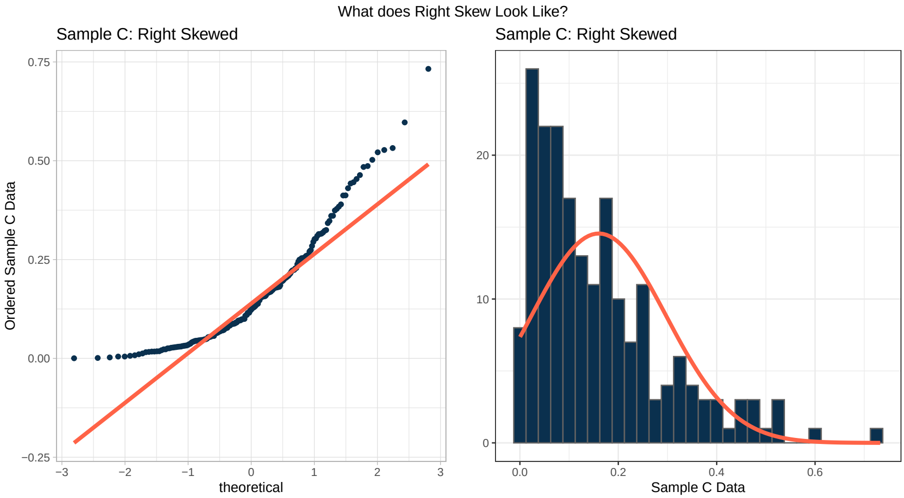
Simulated Data from a Normal Model



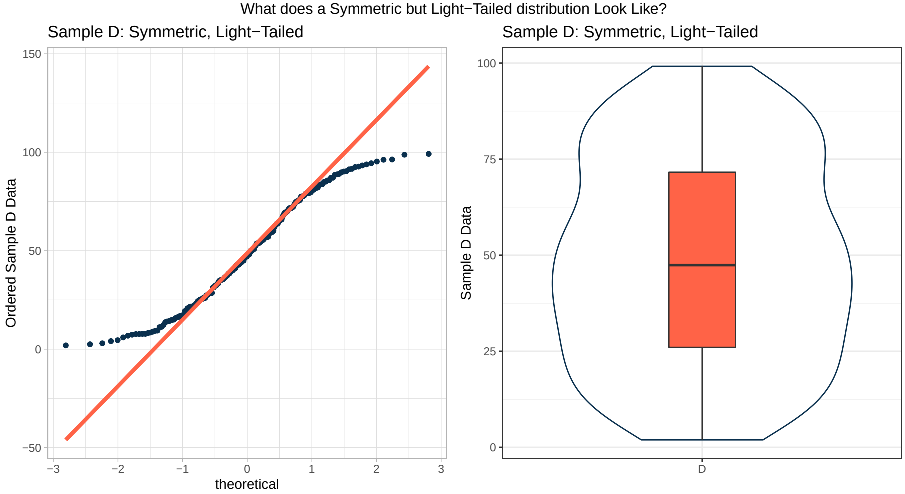
Simulated Left-Skewed Data



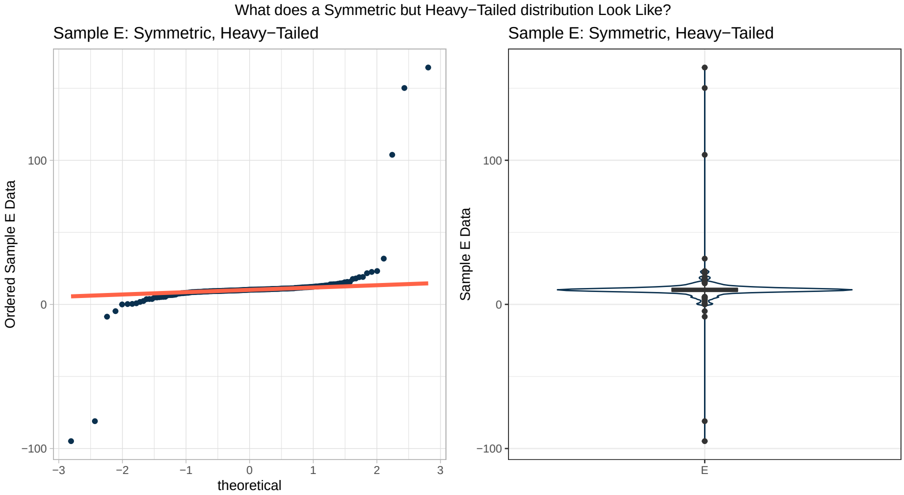
Simulated Right-Skewed Data



Simulated Data from a Symmetric, Light-Tailed Distribution



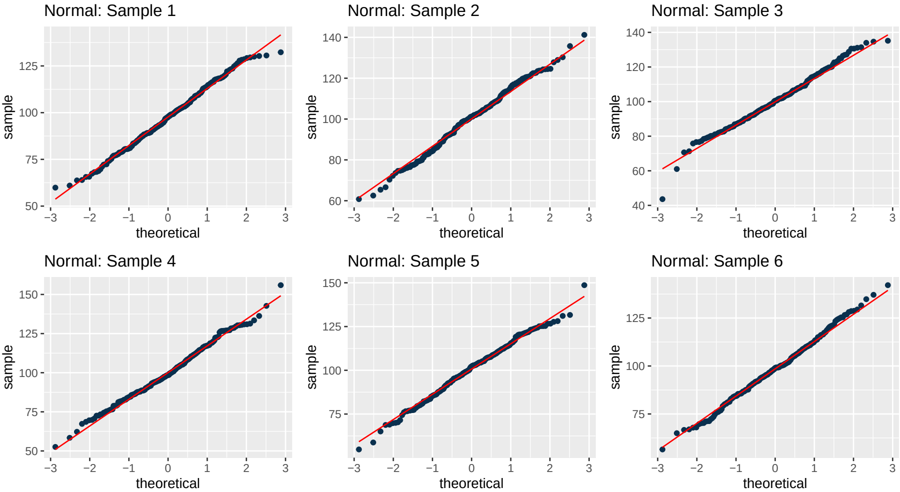
Simulated Data from a Symmetric, Heavy-Tailed Distribution



Note: This is where we stopped at the end of Class 07. We'll continue in Class 08 with the rest of this material.

6 Normal Q-Q plots: Simulated Normal Data

Six simulations from a Normal distribution.



Normal: Sample 1

Normal: Sample 2

Normal: Sample 3

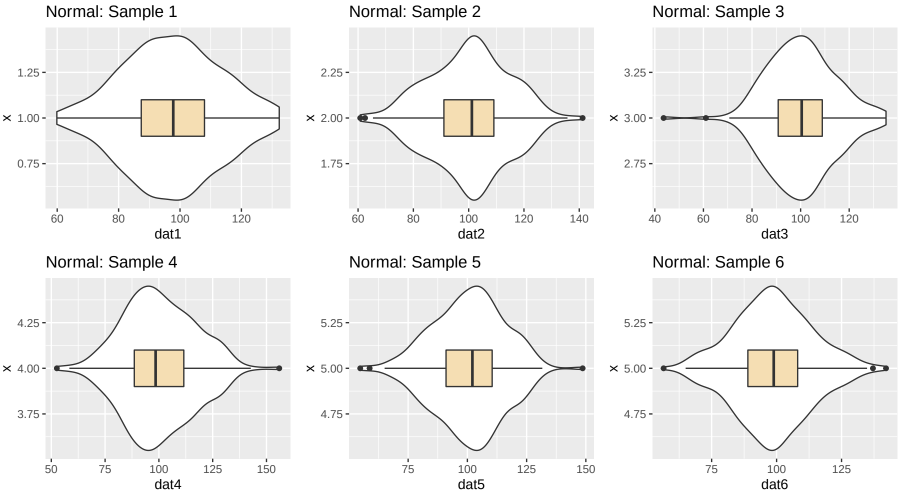
Normal: Sample 4

Normal: Sample 5

Normal: Sample 6

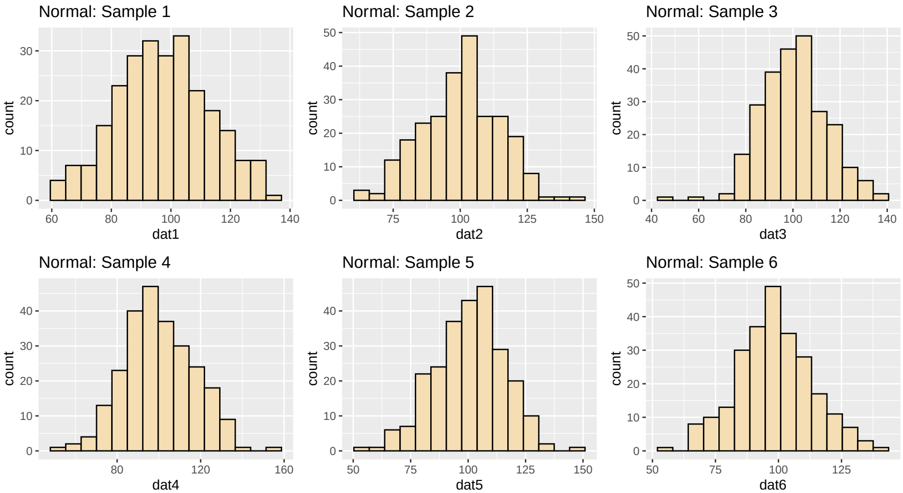
Same Six Simulations, in Box + Violin Plots

Six simulations from a Normal distribution.



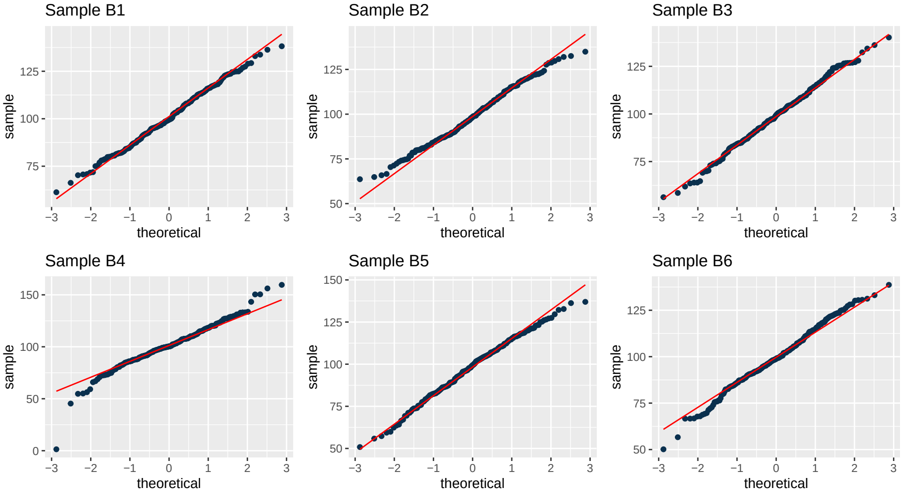
Same Six Simulations, in Histograms

Six simulations from a Normal distribution.

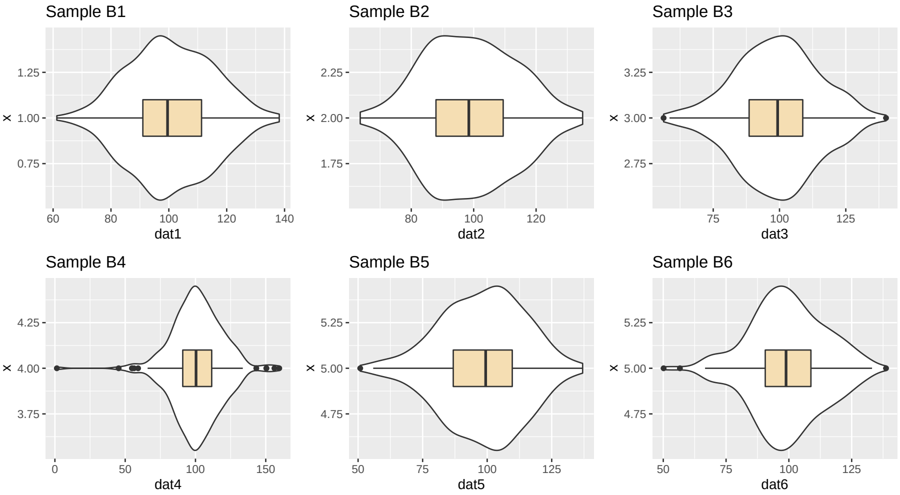


One of these things is not like the others

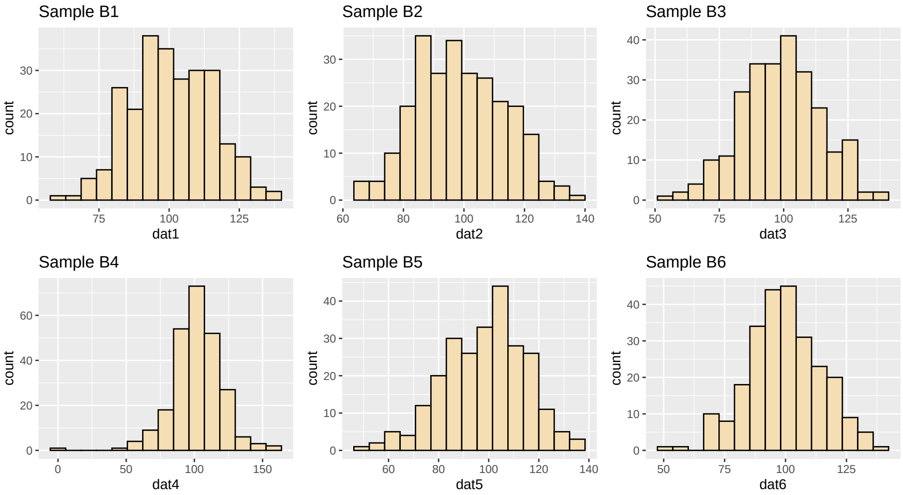
5 simulations of the Normal distribution, one of a heavy-tailed distribution.



Box + Violin Plots of these 6 Samples

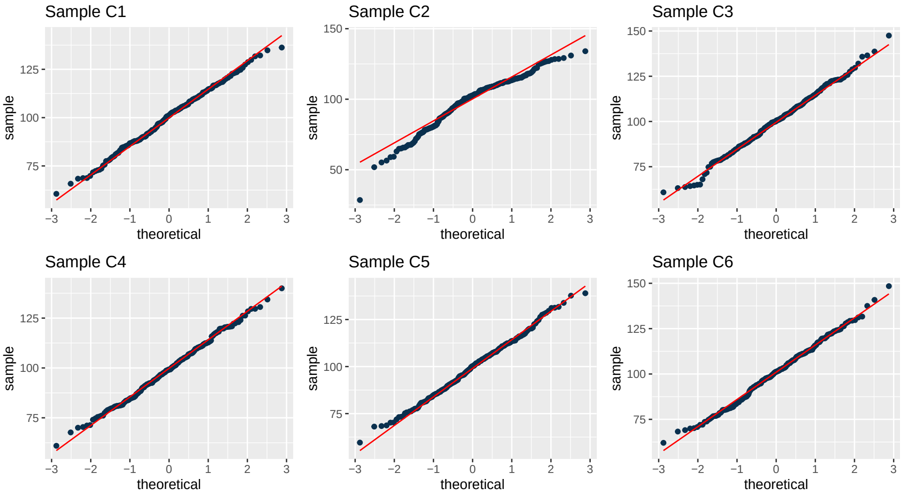


Same Six Simulations, in Histograms

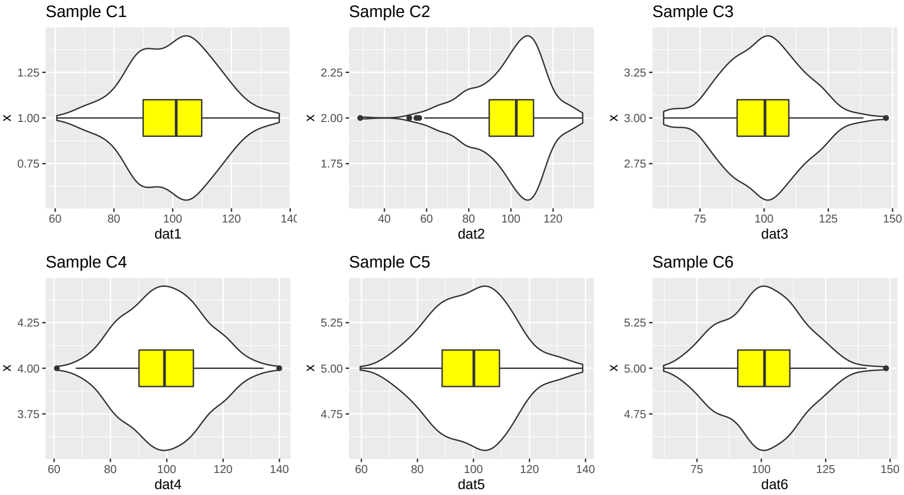


Again, one of these is not like the others

5 simulations of the Normal distribution, one of a left-skewed distribution.



Box + Violin Plots of these 6 Samples

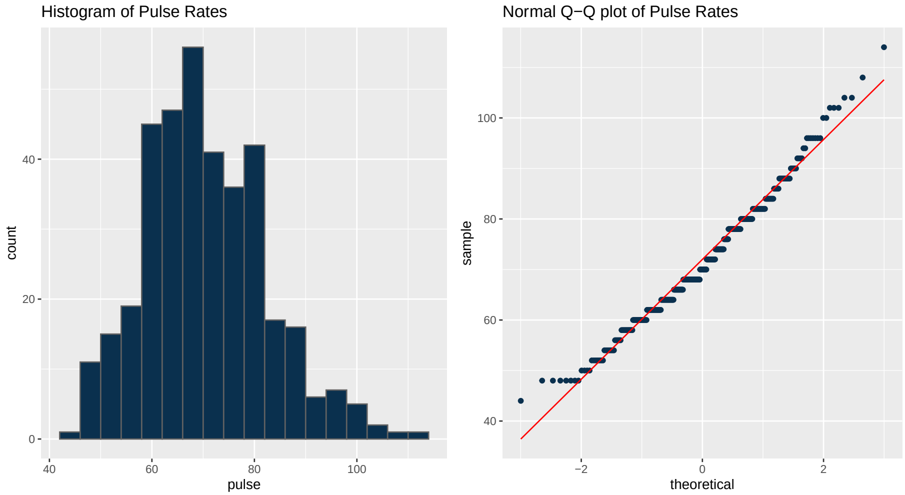


Two plots, side by side

```
plot_a <- ggplot(nh2_GVGmales, aes(x = pulse)) +
  geom_histogram(binwidth = 4,
                 fill = cwru.blue, col = cwru.gray) +
  labs(title = "Histogram of Pulse Rates")

plot_b <- ggplot(nh2_GVGmales, aes(sample = pulse)) +
  geom_qq(col = cwru.blue) + geom_qq_line(col = "red") +
  labs(title = "Normal Q-Q plot of Pulse Rates")

gridExtra::grid.arrange(plot_a, plot_b, ncol = 2)
```

Resulting plot on the next slide. . .

Would a Normal model work well here?

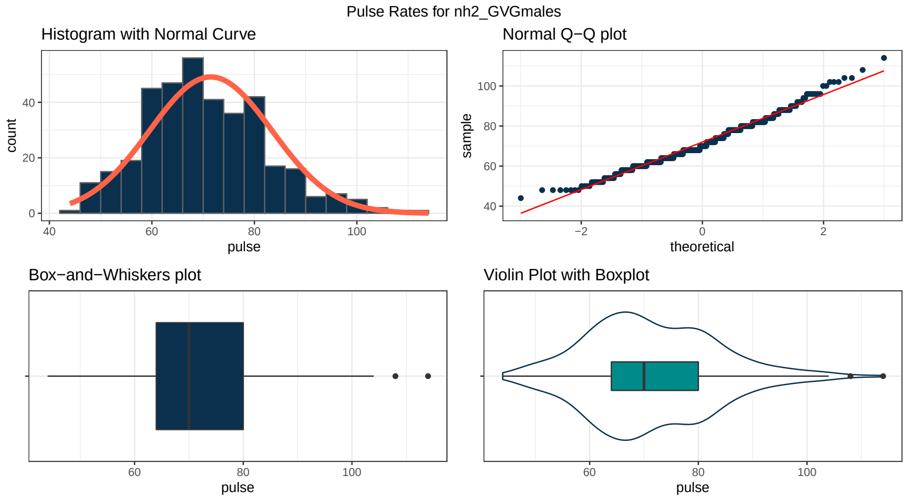


Does a Normal model fit well for my data?

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?

Let's start by looking at 1 and 2.

Four (potentially) Useful Plots



Pulse Rates for nh2_GVGmales

Histogram with Normal Curve

Normal Q−Q plot

Box−and−Whiskers plot

Violin Plot with Boxplot

Does a Normal model fit well for my data?

3. Do numerical measures match up with the expectations of a normal model?
   - Is the mean close to the median (perhaps so that $skew_1$ is less than 0.2 in absolute value)?
   - In a Normal model, mean $\pm$ 1 standard deviation covers 68% of the data.
   - In a Normal model, mean $\pm$ 2 standard deviations covers 95% of the data.
   - In a Normal model, mean $\pm$ 3 standard deviations covers 99.7% of the data.

Normal model for pulse rates of nh2_GVGmales?

```
mosaic::favstats(~ pulse, data = nh2_GVGmales)
```

```
 min Q1 median Q3 max     mean       sd   n missing
  44 64     70 80 114 71.48913 11.94811 368       0
```

What is *skew*$_1$ here?

```
nh2_GVGmales %>%
  summarize(skew1 = (mean(pulse) - median(pulse))/sd(pulse))
```

```
# A tibble: 1 x 1
  skew1
  <dbl>
1 0.125
```

How many of the observations are within 1 SD of the mean?

```
nh2_GVGmales %>%
  count(pulse > mean(pulse) - sd(pulse),
        pulse < mean(pulse) + sd(pulse))
```

```
# A tibble: 3 x 3
  `pulse > mean(pulse) -~ `pulse < mean(pulse) +~      n
  <lgl>                   <lgl>                    <int>
1 FALSE                   TRUE                        46
2 TRUE                    FALSE                       55
3 TRUE                    TRUE                       267
```

So 267 of the 368 (72.6%) observations are within 1 SD of the mean. How does this compare to the expectation under a Normal model?

How about the mean ± 2 standard deviations rule?

The total sample size here is 368.

```
nh2_GVGmales %>%
  count(pulse > mean(pulse) - 2*sd(pulse),
        pulse < mean(pulse) + 2*sd(pulse))
```

```
# A tibble: 3 x 3
  `pulse > mean(pulse) -~ `pulse < mean(pulse) +~     n
  <lgl>                   <lgl>                   <int>
1 FALSE                   TRUE                        1
2 TRUE                    FALSE                      16
3 TRUE                    TRUE                      351
```

So 351 of the 368 (95.4%) observations are within 2 SD of the mean. How does this compare to the expectation under a Normal model?

Hypothesis Testing to assess Normality

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

```
shapiro.test(nh2_GVGmales$pulse)
```

```

    Shapiro-Wilk normality test

data:  nh2_GVGmales$pulse
W = 0.98244, p-value = 0.0001868
```

The very small p value indicates that the test finds some 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

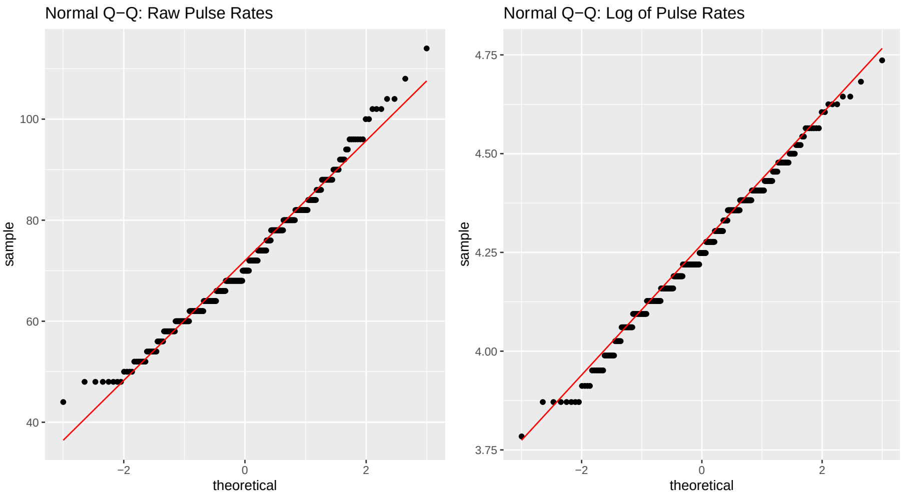
The key notion in re-expression of a single variable to obtain a better fit to a Normal model, is that of a **ladder of power transformations**, which can apply to any unimodal data.

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	$x^{-2} = 1/x^2$

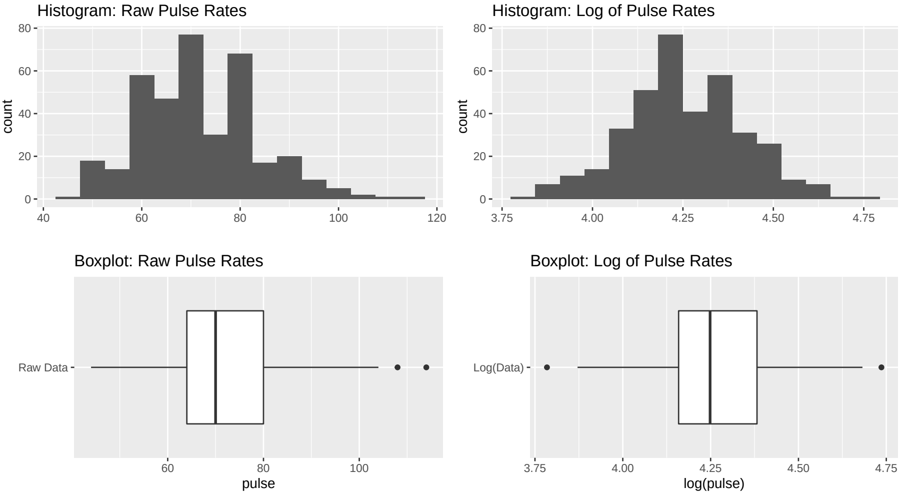
nh2_GVGmales Pulse Rates, and their Natural Logarithms

```r
p1 <- ggplot(data = nh2_GVGmales, aes(sample = pulse)) +
  geom_qq() + geom_qq_line(col = "red") +
  labs(title = "Normal Q-Q: Raw Pulse Rates")

p2 <-  ggplot(data = nh2_GVGmales, 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)
```

`nh2_GVGmales` Pulse Rates, and their Natural Logarithms



`nh2_GVGmales` Pulse Rates, and their Natural Logarithms

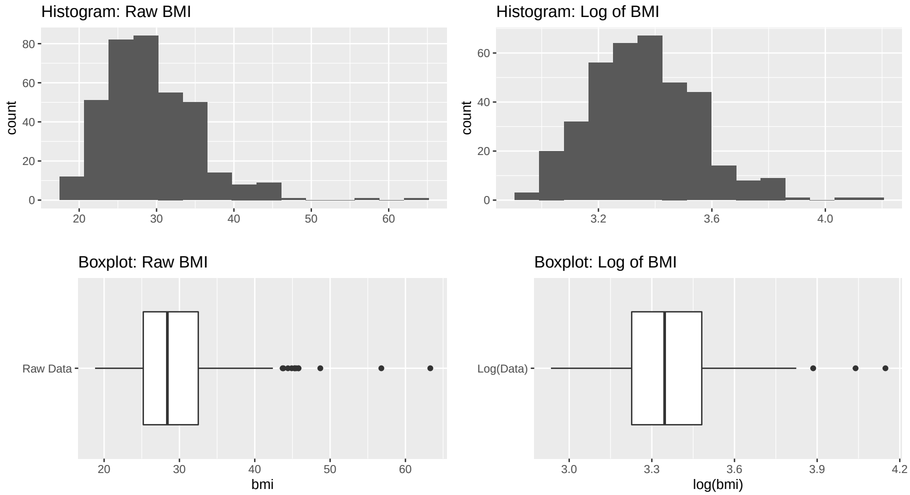


Using the Ladder

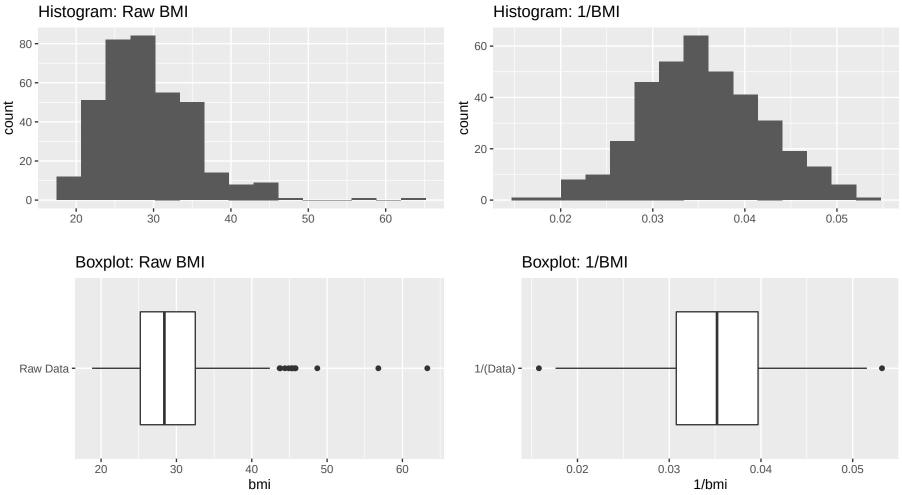
- The ladder is most useful for strictly positive, ratio variables.
- Sometimes, if 0 is a value in the data set, we will add 1 to each value before applying a transformation like the logarithm.
- Interpretability is often an important criterion, although back-transformation at the end of an analysis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

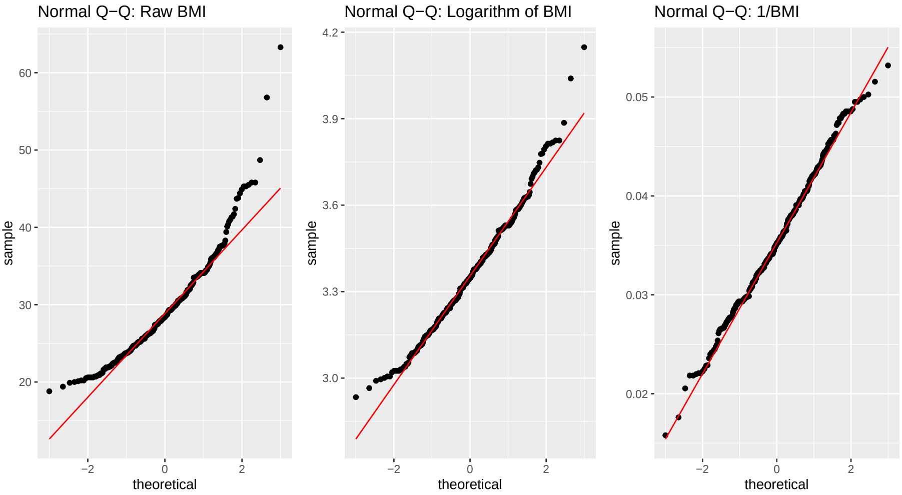
nh2_GVGmales BMI Data (Raw data and Log)



nh2_GVGmales BMI - down the ladder to 1/BMI?



Normal Q-Q plots for BMI



Again, does a Normal Model fit our data?

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.)