

Data Science for Biological, Medical and Health Research: Notes for 431

Thomas E. Love, Ph.D.

Version: 2017-08-14

Contents

Introduction	5
Structure	5
Course Philosophy	6
1 Data Science	7
1.1 Why a unicorn?	7
1.2 Data Science Project Cycle	7
1.3 What Will We Discuss in 431?	9
2 Setting Up R	11
2.1 R Markdown	11
2.2 R Packages	11
2.3 Other Packages	12
Part A. Exploring Data	15
3 Visualizing Data	15
3.1 The NHANES data: Collecting a Sample	15
3.2 Age and Height	16
3.3 Subset of Subjects with Known Age and Height	17
3.4 Age-Height and Gender?	17
3.5 A Subset: Ages 21-79	21
3.6 Distribution of Heights	22
3.7 Height and Gender	24
3.8 A Look at Body-Mass Index	29
3.9 General Health Status	37
3.10 Conclusions	44
4 Data Structures and Types of Variables	45
4.1 Data require structure and context	45
4.2 A New NHANES Adult Sample	45
4.3 Types of Variables	47
5 Summarizing Quantitative Variables	51
5.1 The <code>summary</code> function for Quantitative data	51
5.2 Measuring the Center of a Distribution	52
5.3 Measuring the Spread of a Distribution	54
5.4 Measuring the Shape of a Distribution	58
5.5 More Detailed Numerical Summaries for Quantitative Variables	59
6 Summarizing Categorical Variables	63
6.1 The <code>summary</code> function for Categorical data	63

6.2	Tables to describe One Categorical Variable	64
6.3	The Mode of a Categorical Variable	65
6.4	describe in the Hmisc package	65
6.5	Cross-Tabulations	66
7	The National Youth Fitness Survey (nyfs1)	69
7.1	Looking over the Data Set	69
7.2	Summarizing the Data Set	74
7.3	Additional Summaries from favstats	75
7.4	The Histogram	75
7.5	A Note on Colors	78
7.6	The Stem-and-Leaf	78
7.7	The Dot Plot to display a distribution	81
7.8	The Frequency Polygon	82
7.9	Plotting the Probability Density Function	83
7.10	The Boxplot	84
7.11	A Simple Comparison Boxplot	86
7.12	Using describe in the psych library	89
7.13	Assessing Skew	90
7.14	Assessing Kurtosis (Heavy-Tailedness)	91
7.15	The describe function in the Hmisc library	91
7.16	xda from GitHub for numerical summaries for exploratory data analysis	93
7.17	What Summaries to Report	94
8	Assessing Normality	95
8.1	Empirical Rule Interpretation of the Standard Deviation	95
8.2	Describing Outlying Values with Z Scores	96
8.3	Comparing a Histogram to a Normal Distribution	96
8.4	Does a Normal model work well for the Ages?	98
8.5	The Normal Q-Q Plot	100
8.6	Interpreting the Normal Q-Q Plot	102
8.7	Does a Normal Distribution Fit the nyfs1 Data Well?	109
8.8	Are the Waist Circumferences Normally Distributed?	111

Introduction

These Notes provide a series of examples using R to work through issues that are likely to come up in PQHS/CRSP/MPHP 431.

While these Notes share some of the features of a textbook, they are neither comprehensive nor completely original. The main purpose is to give 431 students a set of common materials on which to draw during the course. In class, we will sometimes:

- reiterate points made in this document,
- amplify what is here,
- simplify the presentation of things done here,
- use new examples to show some of the same techniques,
- refer to issues not mentioned in this document

but what we don't do is follow these notes very precisely. We assume instead that you will read the materials and try to learn from them, just as you will attend classes and try to learn from them. We welcome feedback of all kinds on this document or anything else. Just email us at 431-help at case dot edu, or submit a pull request.

What you will mostly find are brief explanations of a key idea or summary, accompanied (most of the time) by R code and a demonstration of the results of applying that code.

Everything you see here is available to you as HTML or PDF. You will also have access to the R Markdown files, which contain the code which generates everything in the document, including all of the R results. We will demonstrate the use of R Markdown (this document is generated with the additional help of an R package called `bookdown`) and R Studio (the “program” which we use to interface with the R language) in class.

Structure

The Notes, like the 431 course, are split into three main parts.

Part A is about **visualizing data and exploratory data analyses**. These Notes focus on using R to work through issues that arise in the process of exploring data, managing (cleaning and manipulating) data into a tidy format to facilitate useful work downstream, and describing those data effectively with visualizations, numerical summaries, and some simple models.

Part B is about **making comparisons** with data. The Notes discuss the use of R to address comparisons of means and of rates/proportions, primarily. The main ideas include confidence intervals, the bootstrap and parametric and non-parametric tests of hypotheses. Key ideas from Part A that have an impact here include visualizations to check the assumptions behind our inferences, and cleaning/manipulating data to facilitate our comparisons.

Part C is about **building models** with data. The Notes are primarily concerned (in 431) with linear regression models for continuous quantitative outcomes, using one or more predictors. We'll see how to use

models to accomplish many of the comparisons discussed in Part B, and make heavy use of visualization and data management tools developed in Part A to assess our models.

Course Philosophy

In developing this course, we adopt a modern approach that places data at the center of our work. Our goal is to teach you how to do truly reproducible research with modern tools. We want you to be able to answer real questions using data and equip you with the tools you need in order to answer those questions well (Cetinkaya-Rundel (2017) has more on a related teaching philosophy.)

The curriculum includes more on several topics than you might expect from a standard graduate introduction to statistics.

- data gathering
- data wrangling
- exploratory data analysis and visualization
- multivariate modeling
- communication

It also nearly completely avoids formalism and is extremely applied - this is most definitely **not** a course in theoretical or mathematical statistics.

The 431 course is about **getting things done**. It's not a statistics course, nor is it a computer science course. It is instead a course in **data science**.

Chapter 1

Data Science

The definition of **data science** can be a little slippery. One current view of data science, is exemplified by Steven Geringer’s 2014 Venn diagram.

- The field encompasses ideas from mathematics and statistics and from computer science, but with a heavy reliance on subject-matter knowledge. In our case, this includes clinical, health-related, medical or biological knowledge.
- As Gelman and Nolan (2017) suggest, the experience and intuition necessary for good statistical practice are hard to obtain, and teaching data science provides an excellent opportunity to reinforce statistical thinking skills across the full cycle of a data analysis project.
- The principal form in which computer science (coding/programming) play a role in this course is to provide a form of communication. You’ll need to learn how to express your ideas not just orally and in writing, but also through your code.

1.1 Why a unicorn?

Data Science is a **team** activity. Everyone working in data science brings some part of the necessary skillset, but no one person can cover all three areas alone for excellent projects.

[The individual who is truly expert in all three key areas (mathematics/statistics, computer science and subject-matter knowledge) is] a mythical beast with magical powers who’s rumored to exist but is never actually seen in the wild.

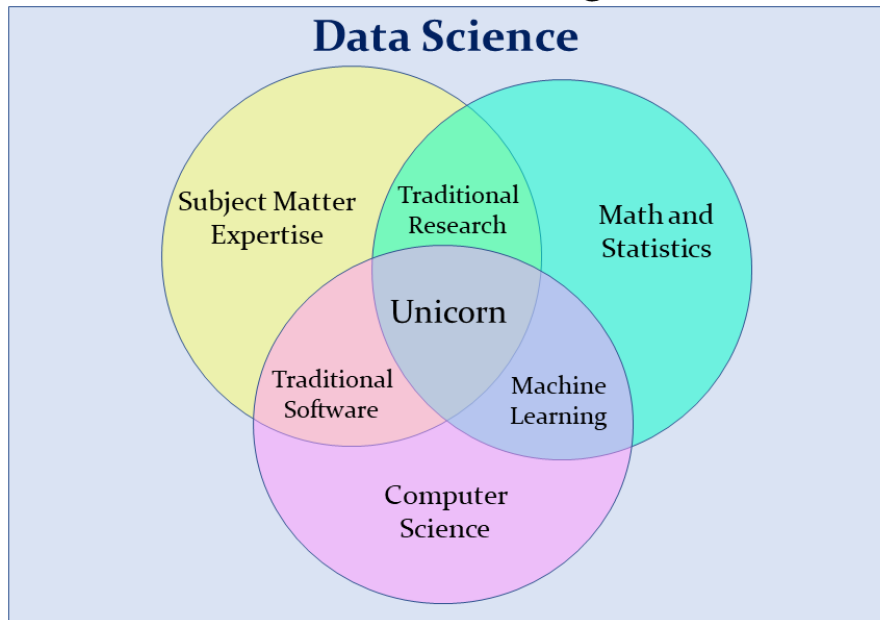
<http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html>

1.2 Data Science Project Cycle

A typical data science project can be modeled as follows, which comes from the introduction to the amazing book **R for Data Science**, by Garrett Golemund and Hadley Wickham, which is a key text for this course (Golemund and Wickham 2017).

This diagram is sometimes referred to as the Krebs Cycle of Data Science. For more on the steps of a data science project, we encourage you to read the Introduction of Golemund and Wickham (2017).

Data Science Venn Diagram 2.0



Original Image Copyright © 2014 by Steven Geringer, Raleigh NC.
 Permission is granted to use, distribute or modify this image, provided that this copyright notice remains intact.

Figure 1.1: Data Science Venn Diagram from Steven Geringer

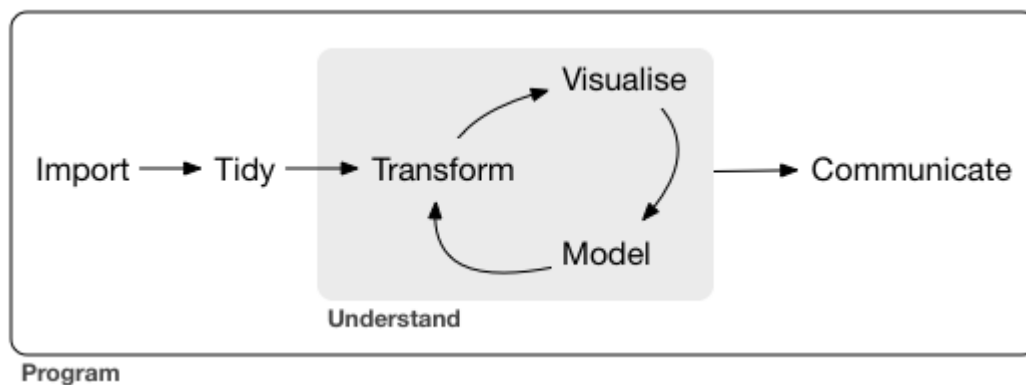


Figure 1.2: Source: R for Data Science: Introduction

1.3 What Will We Discuss in 431?

We'll discuss each of these elements in the 431 course, focusing at the start on understanding our data through transformation, modeling and (especially in the early stages) visualization. In 431, we learn how to get things done.

- We get people working with R and R Studio and R Markdown, even if they are completely new to coding. A gentle introduction is provided at Ismay and Kim (2017)
- We learn how to use the **tidyverse** (<http://www.tidyverse.org/>), an array of tools in R (mostly developed by Hadley Wickham and his colleagues at R Studio) which share an underlying philosophy to make data science faster, easier, more reproducible and more fun. A critical text for understanding the tidyverse is Golemund and Wickham (2017). Tidyverse tools facilitate:
 - **importing** data into R, which can be the source of intense pain for some things, but is really quite easy 95% of the time with the right tool.
 - **tidying** data, that is, storing it in a format that includes one row per observation and one column per variable. This is harder, and more important, than you might think.
 - **transforming** data, perhaps by identifying specific subgroups of interest, creating new variables based on existing ones, or calculating summaries.
 - **visualizing** data to generate actual knowledge and identify questions about the data - this is an area where R really shines, and we'll start with it in class.
 - **modeling** data, taking the approach that modeling is complementary to visualization, and allows us to answer questions that visualization helps us identify.
 - and last, but definitely not least, **communicating** results, models and visualizations to others, in a way that is reproducible and effective.
- Some programming/coding is an inevitable requirement to accomplish all of these aims. If you are leery of coding, you'll need to get past that, with the help of this course and our stellar teaching assistants. Getting started is always the most challenging part, but our experience is that most of the pain of developing these new skills evaporates by early October.
- Having completed some fundamental work in Part A of the course, we then learn how to use a variety of R packages and statistical methods to accomplish specific inferential tasks (in Part B, mostly) and modeling tasks (in Part C, mostly.)

Chapter 2

Setting Up R

These Notes make extensive use of

- the statistical software language R, and
- the development environment R Studio

both of which are free, and you'll need to install them on your machine. Instructions for doing so are in found in the course syllabus.

If you need an even gentler introduction, or if you're just new to R and RStudio and need to learn about them, we encourage you to take a look at <http://moderndive.com/>, which provides an introduction to statistical and data sciences via R at Ismay and Kim (2017).

2.1 R Markdown

These notes were written using R Markdown. R Markdown, like R and R Studio, is free and open source.

R Markdown is described as an *authoring framework* for data science, which lets you

- save and execute R code
- generate high-quality reports that can be shared with an audience

This description comes from <http://rmarkdown.rstudio.com/lesson-1.html> which you can visit to get an overview and quick tour of what's possible with R Markdown.

Another excellent resource to learn more about R Markdown tools is the Communicate section (especially the R Markdown chapter) of Golemund and Wickham (2017).

2.2 R Packages

To start, I'll present a series of commands I run at the beginning of these Notes. These particular commands set up the output so it will look nice as either an HTML or PDF file, and also set up R to use several packages (libraries) of functions that expand its capabilities. A chunk of code like this will occur near the top of any R Markdown work.

```
knitr::opts_chunk$set(comment = NA)

library(boot); library(devtools); library(forcats)
library(grid); library(knitr); library(pander)
```

```
library(pwr); library(viridis); library(NHANES)
library(tidyverse)

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

I have deliberately set up this list of loaded packages/libraries to be relatively small, and will add some other packages later, as needed. You only need to install a package once, but you need to reload it every time you start a new session.

2.3 Other Packages

I will also make use of functions in the following packages/libraries, but when I do so, I will explicitly specify the package name, using a command like `Hmisc::describe(x)`, rather than just `describe(x)`, so as to specify that I want the Hmisc package's version of `describe` applied to whatever `x` is. Those packages are:

- `aplpack` which provides `stem.leaf` and `stem.leaf.backback` for building fancier stem-and-leaf displays
- `arm` which provides a set of functions for model building and checking that are used in Gelman and Hill (2007)
- `car` which provides some tools for building scatterplot matrices, but also many other functions described in Fox and Weisberg (2011)
- `Epi` for 2x2 table analyses and materials for classical epidemiology: <http://BendixCarstensen.com/Epi/>
- `GGally` for scatterplot and correlation matrix visualizations: <http://ggobi.github.io/ggally/>
- `gridExtra` which includes a variety of functions for manipulating graphs: <https://github.com/baptiste/gridextra>
- `Hmisc` from Frank Harrell at Vanderbilt U., for its version of `describe` and for many regression modeling functions we'll use in 432. Details on Hmisc are at <http://biostat.mc.vanderbilt.edu/wiki/Main/Hmisc>. Frank has written several books - the most useful of which for 431 students is probably Harrell and Slaughter (2017)
- `mice`, which we'll use (a little) in 431 for multiple imputation to deal with missing data: <http://www.stefvanbuuren.nl/mi/>
- `mosaic`, mostly for its `favstats` summary, but Project MOSAIC is a community of educators you might be interested in: <http://mosaic-web.org/>
- `psych` for its own version of `describe`, but other features are described at <http://personality-project.org/r/psych/>

We also will use a package called `xda` for two functions called `numSummary` and `charSummary`, but that package gets loaded via `devtools` and GitHub by the code in these Notes.

When compiling the Notes from the original code files, these packages will need to be installed (but not loaded) in R, or an error will be thrown when compiling this document. To install all of the packages used within these Notes, type in (or copy and paste) the following commands and run them in the R Console. Again, you only need to install a package once, but you need to reload it every time you start a new session.

```
pkgs <- c("aplpack", "arm", "boot", "car", "devtools", "Epi", "forcats", "GGally",
          "gridExtra", "Hmisc", "knitr", "mice", "mosaic", "NHANES", "pander",
          "psych", "pwr", "tidyverse", "viridis")
install.packages(pkgs)
```

Part A. Exploring Data

Chapter 3

Visualizing Data

Part A of these Notes is designed to ease your transition into working effectively with data, so that you can better understand it. We'll start by visualizing some data from the US National Health and Nutrition Examination Survey, or NHANES. We'll display R code as we go, but we'll return to all of the key coding ideas involved later in the Notes.

3.1 The NHANES data: Collecting a Sample

To begin, we'll gather a random sample of 1,000 subjects participating in NHANES, and then identify several variables of interest about those subjects¹. The motivation for this example came from a Figure in Baumer, Kaplan, and Horton (2017).

```
library(NHANES) # load the NHANES package/library of functions, data

set.seed(431001)
# use set.seed to ensure that we all get the same random sample
# of 1,000 NHANES subjects in our nh_data collection

nh_data <- sample_n(NHANES, size = 1000) %>%
  select(ID, Gender, Age, Height, Weight, BMI, Pulse, Race1, HealthGen, Diabetes)

nh_data
```

```
# A tibble: 1,000 x 10
   ID Gender  Age Height Weight  BMI Pulse  Race1 HealthGen
  <int> <fctr> <int>  <dbl>  <dbl> <dbl> <int>  <fctr>  <fctr>
1 59640 male    54   176  129.0  41.8   74   White    Good
2 59826 female  67   156   50.2  20.5   66   White   Vgood
3 56340 male     9   128   23.3  14.2   86   Black    NA
4 56747 male    33   194  105.1  27.9   68   White   Vgood
5 51754 female  58   167  106.0  37.9   70   White    NA
6 52712 male     6   109   16.9  14.3   NA   White    NA
7 63908 male    55   169   90.6  31.9   62 Mexican Vgood
8 60865 female  25   156   55.0  22.8   58   Other   Vgood
9 66642 male    41   178   89.3  28.2   72   White   Vgood
10 59880 female  45   163   98.3  36.9   80 Hispanic Good
```

¹For more on the NHANES data available in the NHANES package, type ?NHANES in the Console in R Studio.

```
# ... with 990 more rows, and 1 more variables: Diabetes <fctr>
```

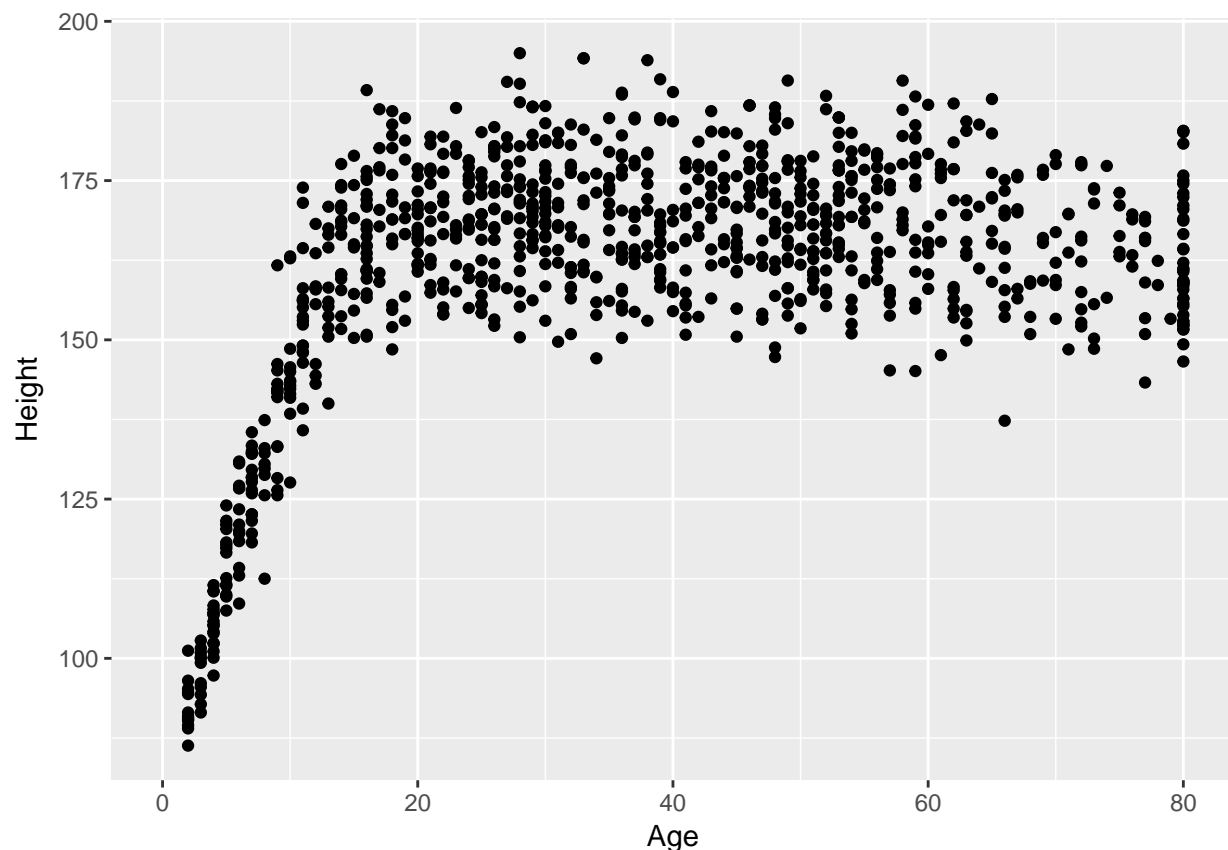
We have 1000 rows (observations) and 10 columns (variables) that describe the subjects listed in the rows.

3.2 Age and Height

Suppose we want to visualize the relationship of Height and Age in our 1,000 NHANES observations. The best choice is likely to be a scatterplot.

```
ggplot(data = nh_data, aes(x = Age, y = Height)) +  
  geom_point()
```

Warning: Removed 25 rows containing missing values (geom_point).



We note several interesting results here.

1. As a warning, R tells us that it has “Removed 25 rows containing missing values (geom_point).” Only 975 subjects plotted here, because the remaining 25 people have missing (NA) values for either Height, Age or both.
2. Unsurprisingly, the measured Heights of subjects grow from Age 0 to Age 20 or so, and we see that a typical Height increases rapidly across these Ages. The middle of the distribution at later Ages is pretty consistent at a Height somewhere between 150 and 175. The units aren’t specified, but we expect they must be centimeters. The Ages are clearly reported in Years.
3. No Age is reported over 80, and it appears that there is a large cluster of Ages at 80. This may be due to a requirement that Ages 80 and above be reported at 80 so as to help mask the identity of those

individuals.²

As in this case, we're going to build most of our visualizations using tools from the `ggplot2` package, which is part of the `tidyverse` series of packages. You'll see similar coding structures throughout this Chapter, most of which are covered as well in Chapter 3 of Golemund and Wickham (2017).

3.3 Subset of Subjects with Known Age and Height

Before we move on, let's manipulate the data set a bit, to focus on only those subjects who have complete data on both Age and Height. This will help us avoid that warning message.

```
nh_dat2 <- nh_data %>%
  filter(complete.cases(Age, Height))

summary(nh_dat2)
```

ID	Gender	Age	Height
Min. :51654	female:498	Min. : 2.0	Min. : 86.3
1st Qu.:56752	male :477	1st Qu.:20.0	1st Qu.:156.4
Median :61453		Median :36.0	Median :165.8
Mean :61602		Mean :37.3	Mean :161.7
3rd Qu.:66484		3rd Qu.:53.0	3rd Qu.:174.1
Max. :71826		Max. :80.0	Max. :195.0

Weight	BMI	Pulse	Race1
Min. : 12.5	Min. :13.2	Min. : 42.0	Black :112
1st Qu.: 57.6	1st Qu.:21.6	1st Qu.: 66.0	Hispanic: 69
Median : 73.4	Median :26.1	Median : 72.0	Mexican :104
Mean : 73.4	Mean :27.0	Mean : 73.7	White :607
3rd Qu.: 90.2	3rd Qu.:31.1	3rd Qu.: 82.0	Other : 83
Max. :198.7	Max. :80.6	Max. :124.0	
NA's :2	NA's :2	NA's :120	

HealthGen	Diabetes
Excellent: 87	No :910
Vgood :276	Yes : 64
Good :276	NA's: 1
Fair :103	
Poor : 15	
NA's :218	

Note that the units and explanations for these variables are contained in the NHANES help file, available via `?NHANES` in the Console of R Studio.

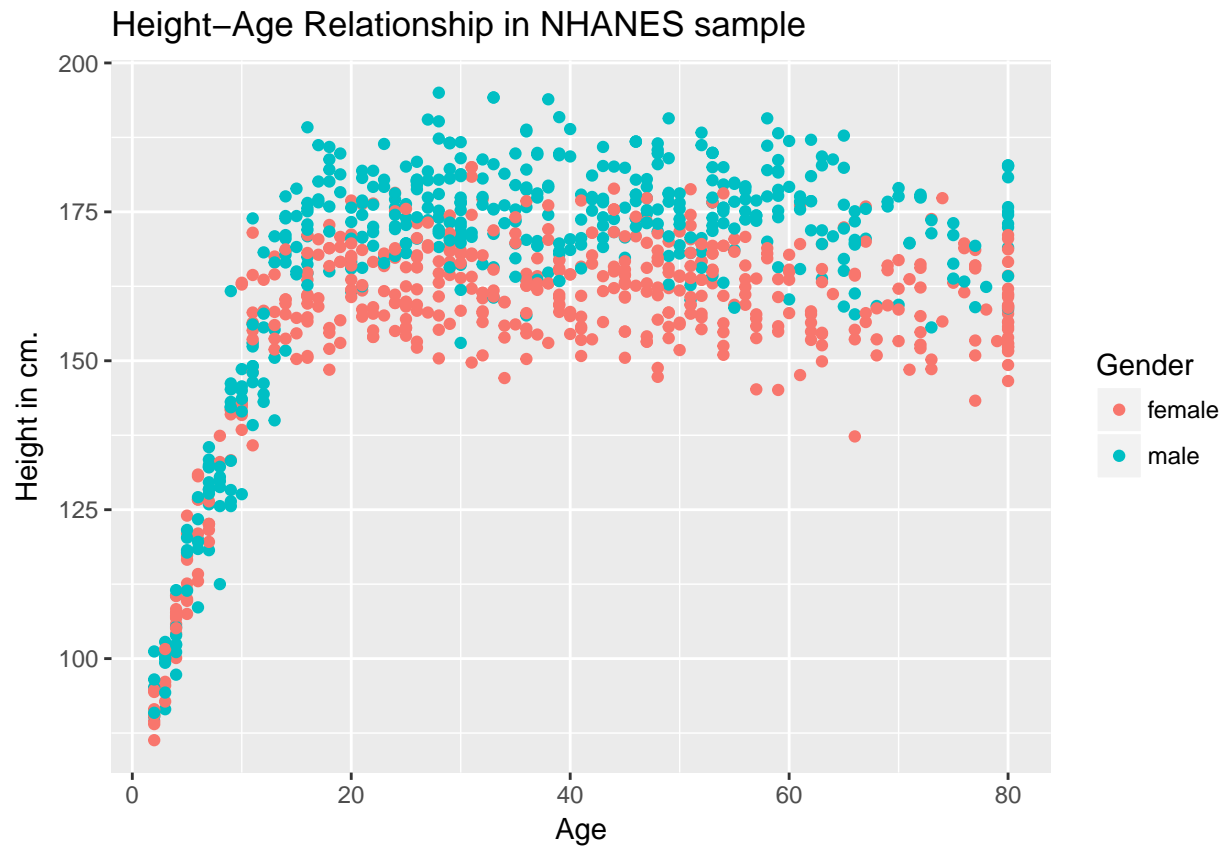
3.4 Age-Height and Gender?

Let's add Gender to the plot using color, and also adjust the y axis label to incorporate the units of measurement.

```
ggplot(data = nh_dat2, aes(x = Age, y = Height, color = Gender)) +
  geom_point() +
```

²If you visit the NHANES help file with `?NHANES`, you will see that subjects 80 years or older were indeed recorded as 80.

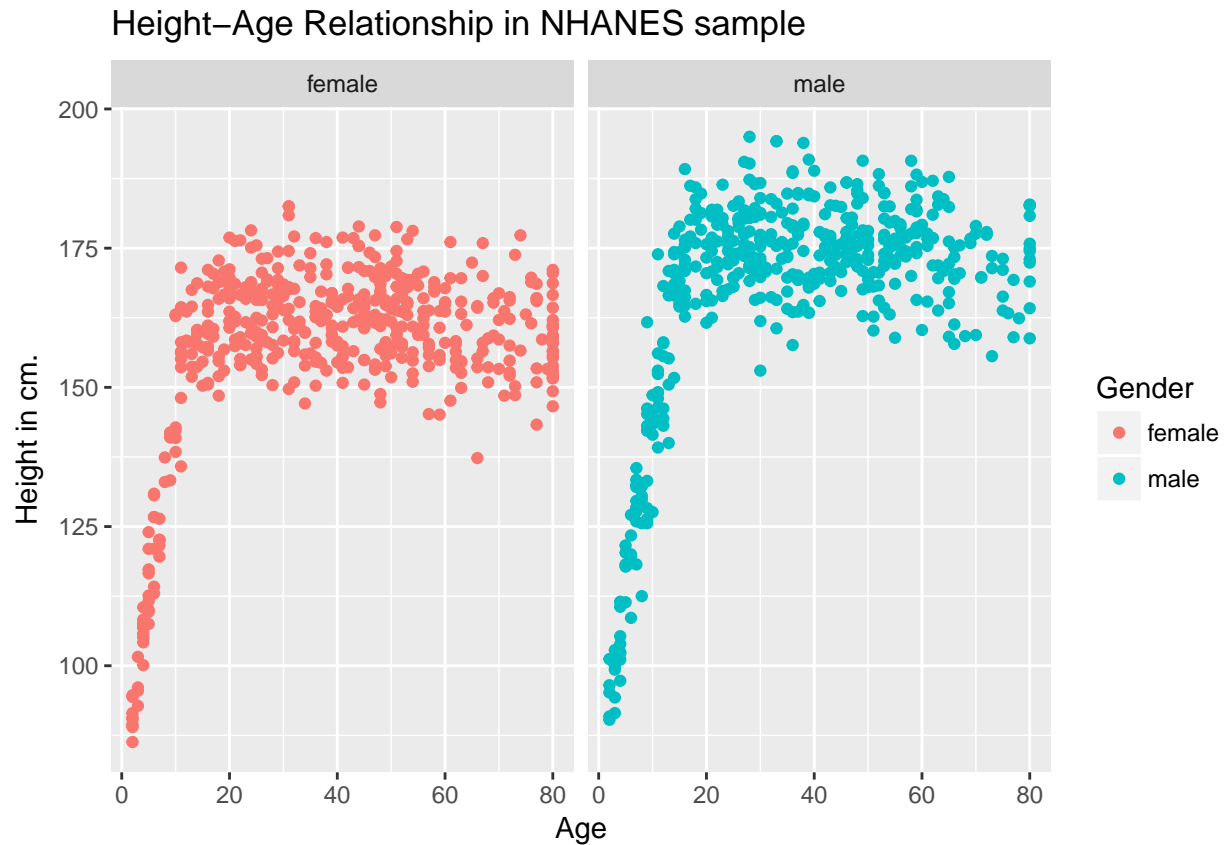
```
labs(title = "Height-Age Relationship in NHANES sample",
     y = "Height in cm.")
```



3.4.1 Can we show the Female and Male relationships in separate panels?

Sure.

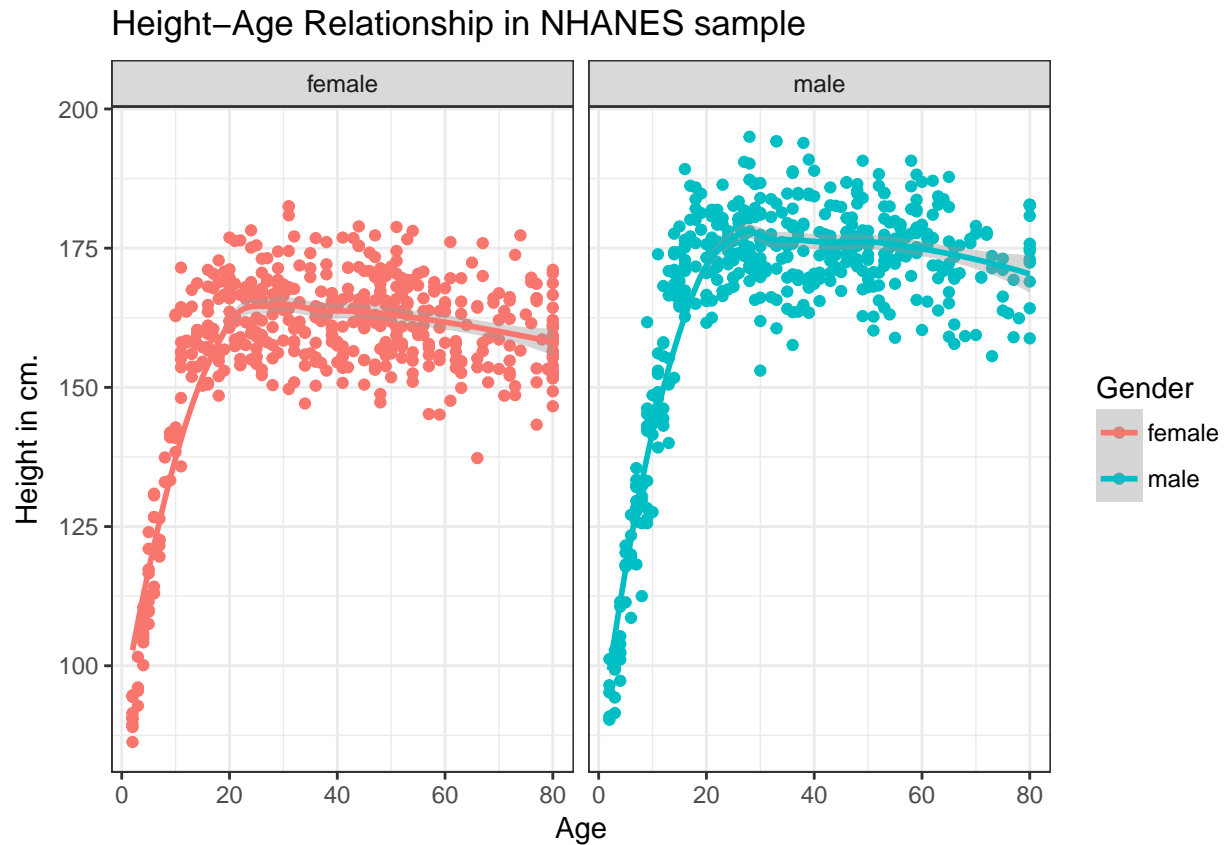
```
ggplot(data = nh_dat2, aes(x = Age, y = Height, color = Gender)) +
  geom_point() +
  labs(title = "Height-Age Relationship in NHANES sample",
       y = "Height in cm.") +
  facet_wrap(~ Gender)
```



3.4.2 Can we add a smooth curve to show the relationship in each plot?

Yep, and let's change the theme of the graph to remove the gray background, too.

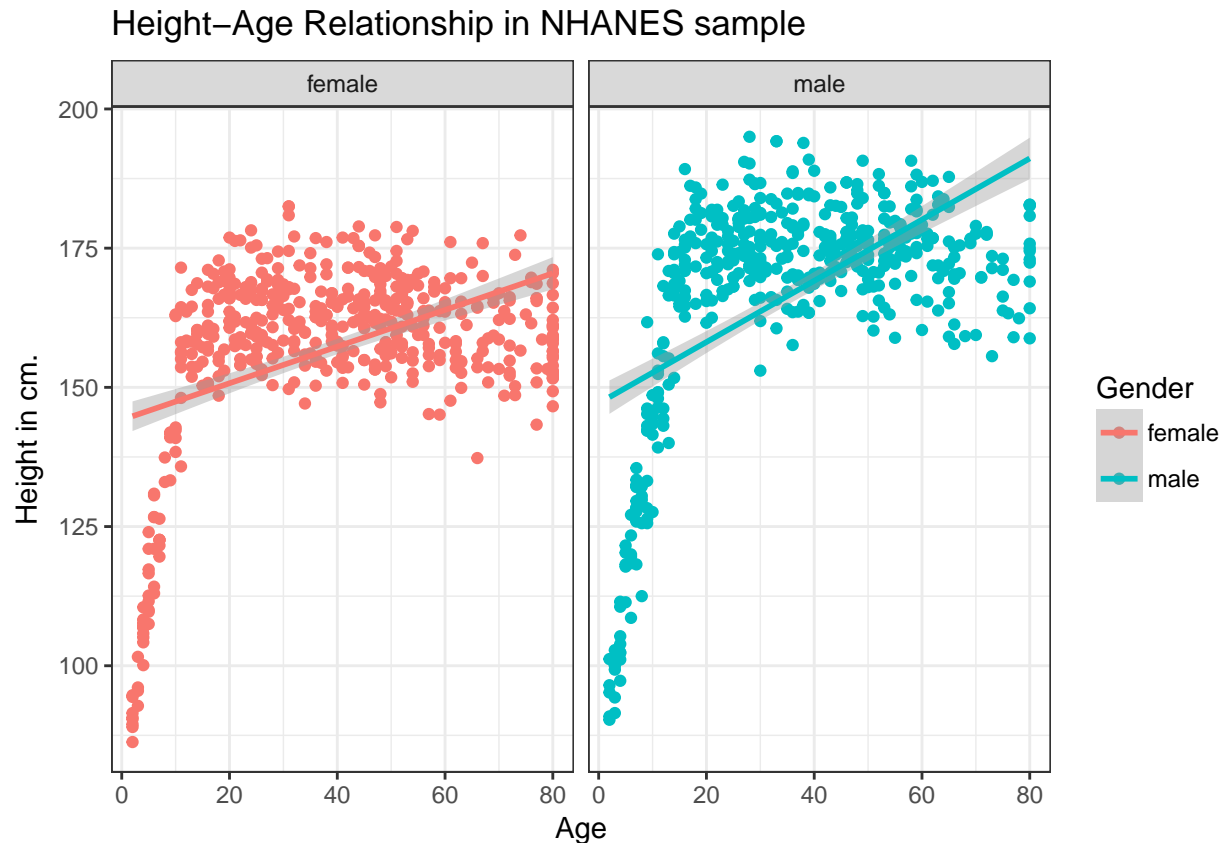
```
ggplot(data = nh_dat2, aes(x = Age, y = Height, color = Gender)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "Height-Age Relationship in NHANES sample",
       y = "Height in cm.") +
  theme_bw() +
  facet_wrap(~ Gender)
```



3.4.3 What if we want to assume straight line relationships?

We could look at a linear model in the plot. Does this make sense here?

```
ggplot(data = nh_dat2, aes(x = Age, y = Height, color = Gender)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Height-Age Relationship in NHANES sample",
       y = "Height in cm.") +
  theme_bw() +
  facet_wrap(~ Gender)
```



3.5 A Subset: Ages 21-79

Suppose we wanted to look at a subset of our sample - those observations (subjects) whose Age is at least 21 and at most 79. We'll create that sample below, and also subset the variables to include nine of particular interest, and remove any observations with any missingness on *any* of the nine variables we're including here.

```
nh_data_2179 <- nh_data %>%
  filter(Age > 20 & Age < 80) %>%
  select(ID, Gender, Age, Height, Weight, BMI, Pulse, Race1, HealthGen, Diabetes) %>%
  na.omit
```

```
nh_data_2179
```

```
# A tibble: 594 x 10
```

	ID	Gender	Age	Height	Weight	BMI	Pulse	Race1	HealthGen
	<int>	<fctr>	<int>	<dbl>	<dbl>	<dbl>	<int>	<fctr>	<fctr>
1	59640	male	54	176	129.0	41.8	74	White	Good
2	59826	female	67	156	50.2	20.5	66	White	Vgood
3	56747	male	33	194	105.1	27.9	68	White	Vgood
4	63908	male	55	169	90.6	31.9	62	Mexican	Vgood
5	60865	female	25	156	55.0	22.8	58	Other	Vgood
6	66642	male	41	178	89.3	28.2	72	White	Vgood
7	59880	female	45	163	98.3	36.9	80	Hispanic	Good
8	71784	female	24	161	50.2	19.3	72	White	Vgood
9	67616	male	63	184	70.0	20.6	82	White	Vgood

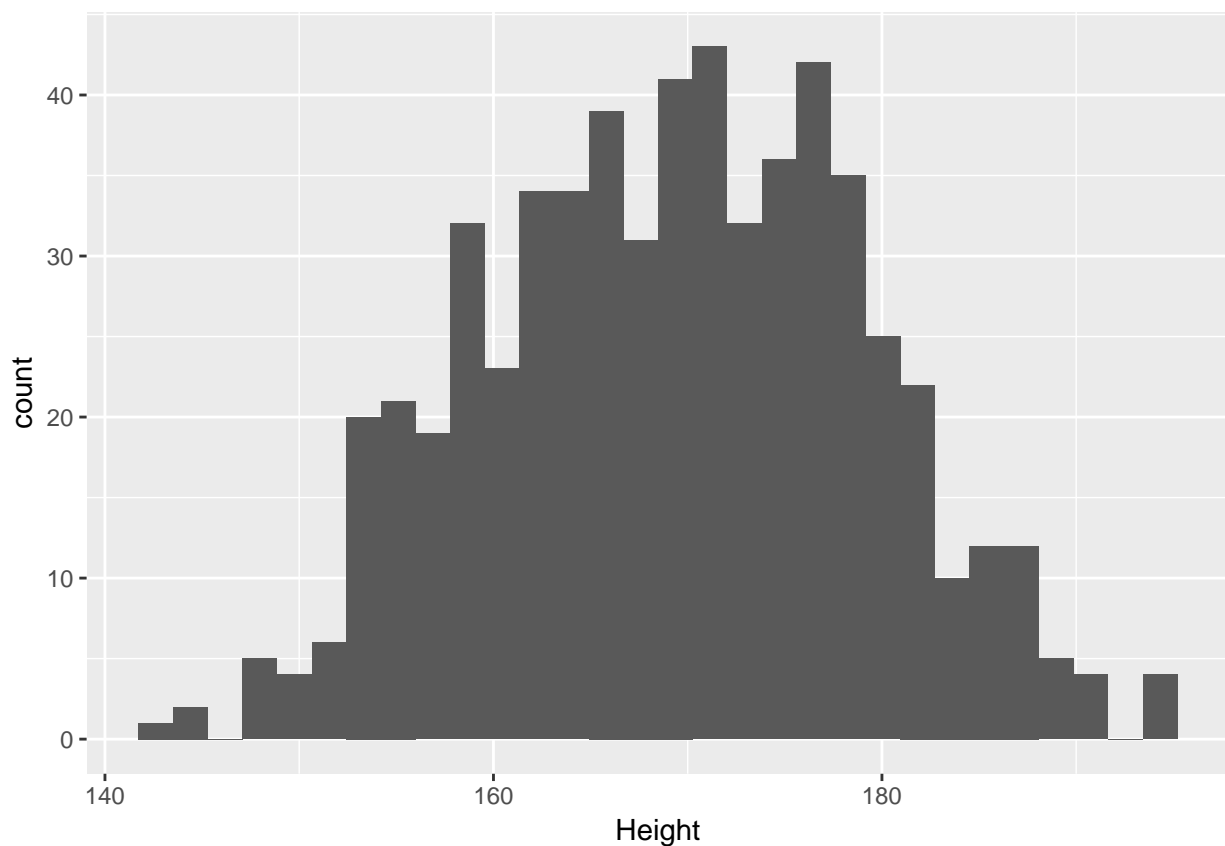
```
10 55391 female    32    161   69.2  26.6   114    Other    Good
# ... with 584 more rows, and 1 more variables: Diabetes <fctr>
```

3.6 Distribution of Heights

What is the distribution of height in this new sample?

```
ggplot(data = nh_data_2179, aes(x = Height)) +
  geom_histogram()
```

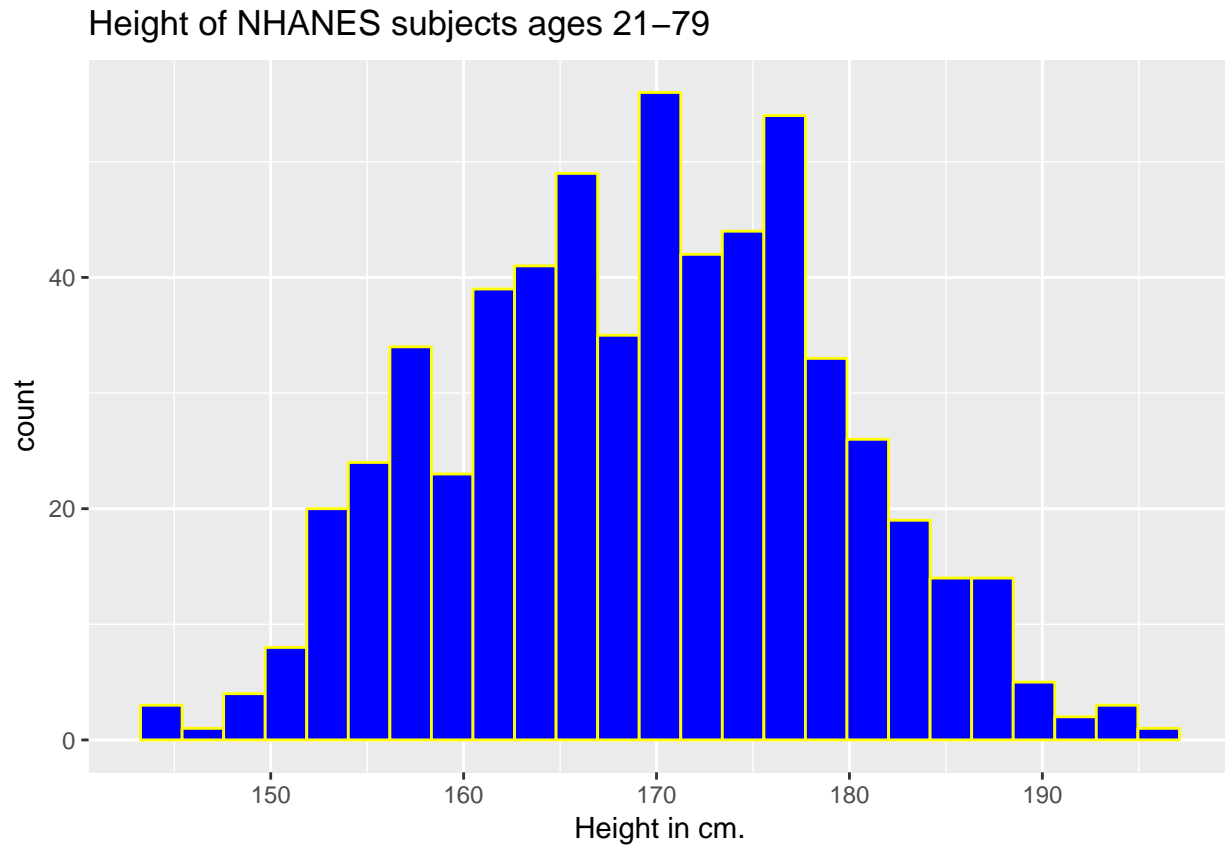
``stat_bin()` using `bins = 30`. Pick better value with `binwidth`.`



We can do several things to clean this up.

1. We'll change the color of the lines for each bar of the histogram.
2. We'll change the fill inside each bar to make them stand out a bit more.
3. We'll add a title and relabel the horizontal (x) axis to include the units of measurement.
4. We'll avoid the warning by selecting a number of bins (we'll use 25 here) into which we'll group the heights before drawing the histogram.

```
ggplot(data = nh_data_2179, aes(x = Height)) +
  geom_histogram(bins = 25, col = "yellow", fill = "blue") +
  labs(title = "Height of NHANES subjects ages 21-79",
       x = "Height in cm.")
```

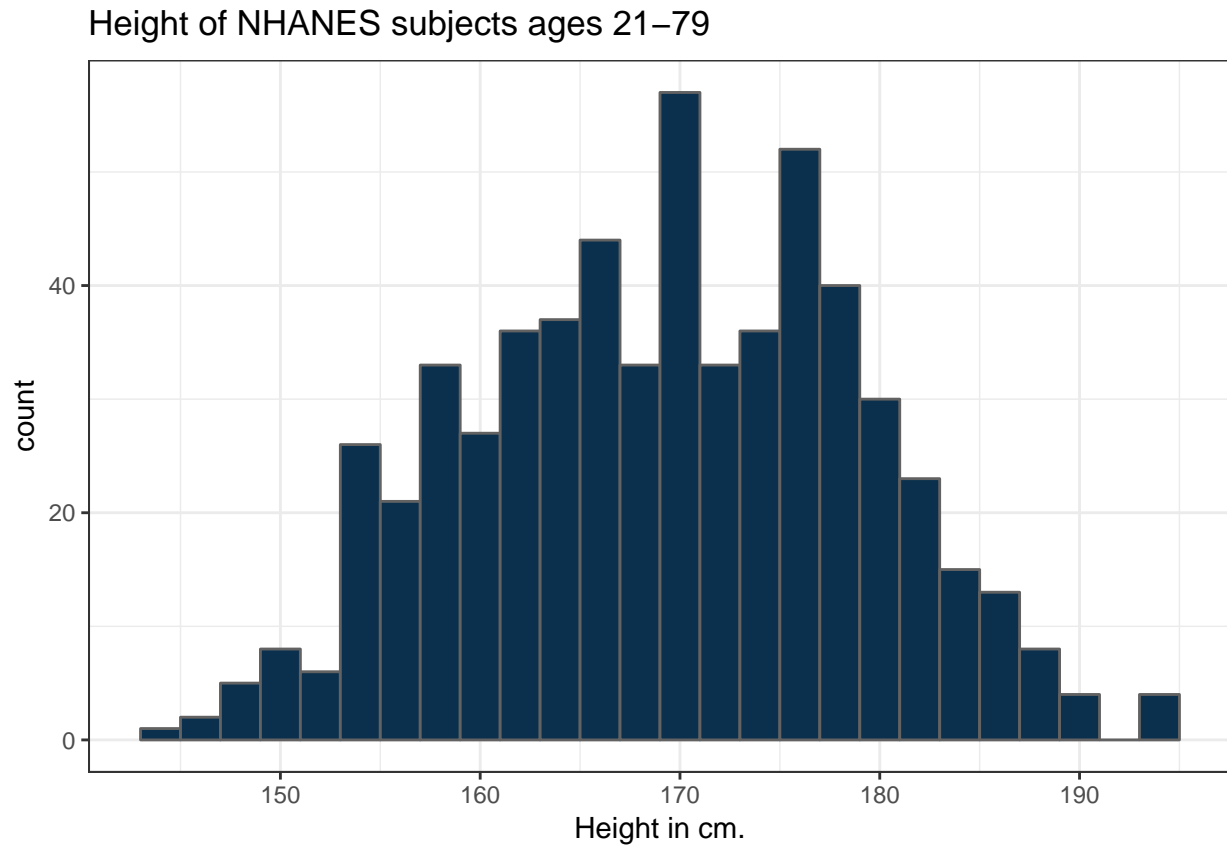


3.6.1 Changing a Histogram's Fill and Color

The CWRU color guide (<https://case.edu/umc/our-brand/visual-guidelines/>) lists the HTML color schemes for CWRU blue and CWRU gray. Let's match that color scheme.

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

ggplot(data = nh_data_2179, aes(x = Height)) +
  geom_histogram(binwidth = 2, col = cwrु.gray, fill = cwrु.blue) +
  labs(title = "Height of NHANES subjects ages 21-79",
       x = "Height in cm.") +
  theme_bw()
```

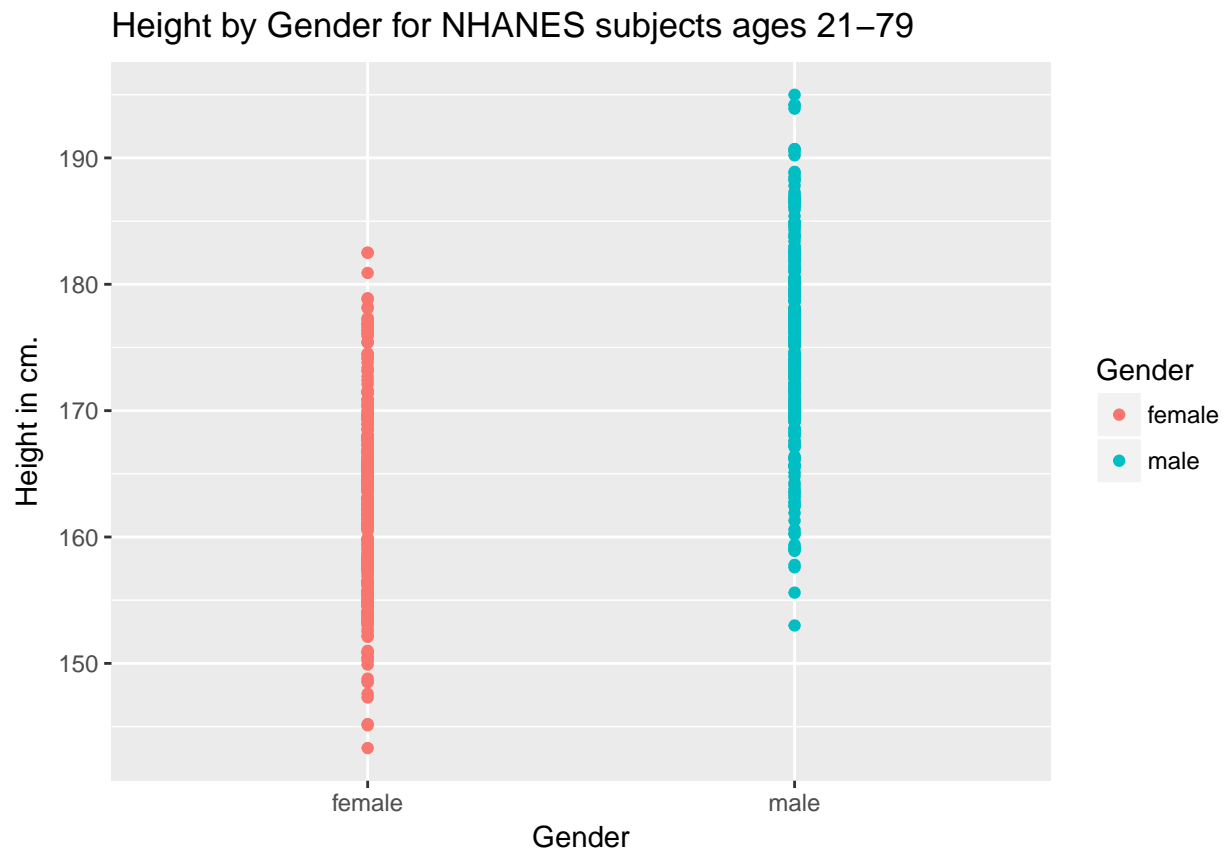


Note the other changes to the graph above.

1. We changed the theme to replace the gray background.
2. We changed the bins for the histogram, to gather observations into groups of 2 cm. each.

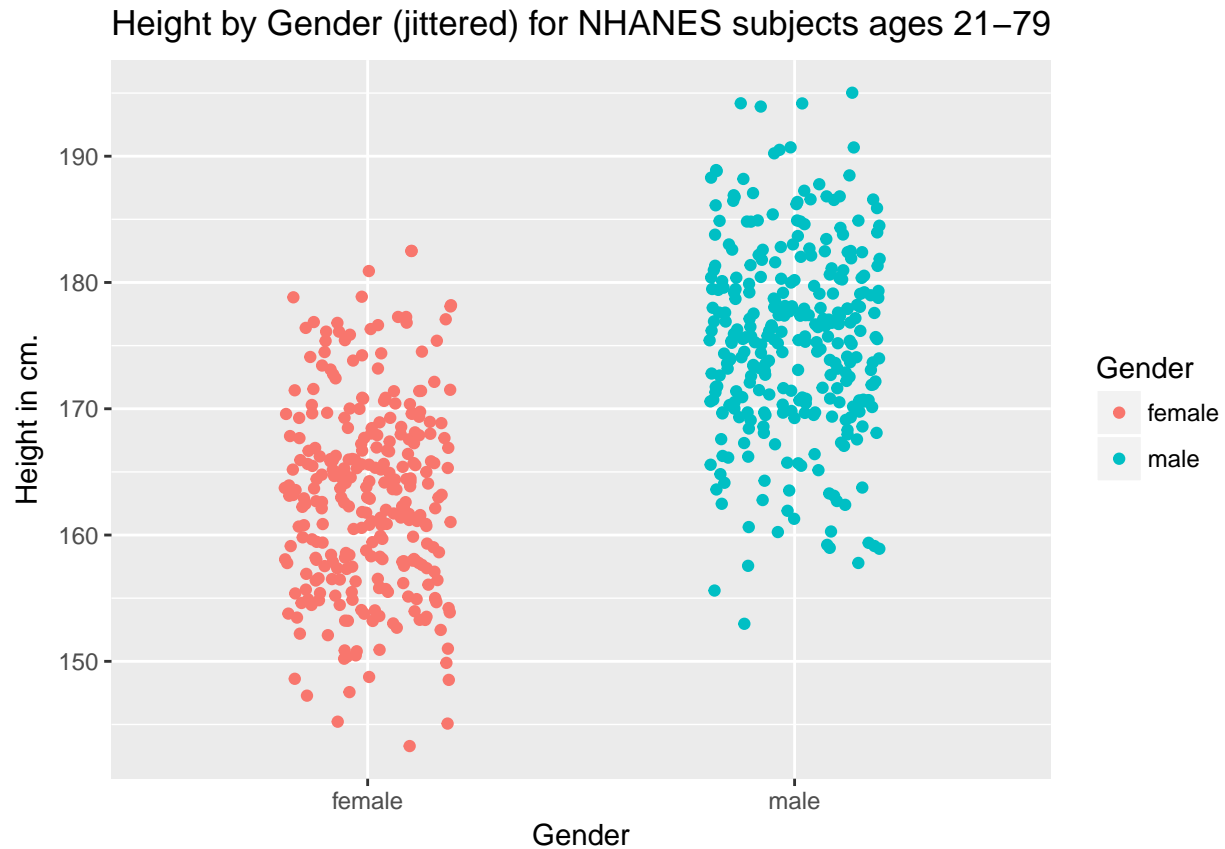
3.7 Height and Gender

```
ggplot(data = nh_data_2179, aes(x = Gender, y = Height, color = Gender)) +  
  geom_point() +  
  labs(title = "Height by Gender for NHANES subjects ages 21-79",  
        y = "Height in cm.")
```

This plot isn't so useful. We can improve things a little by jittering the points horizontally, so that the overlap is reduced.

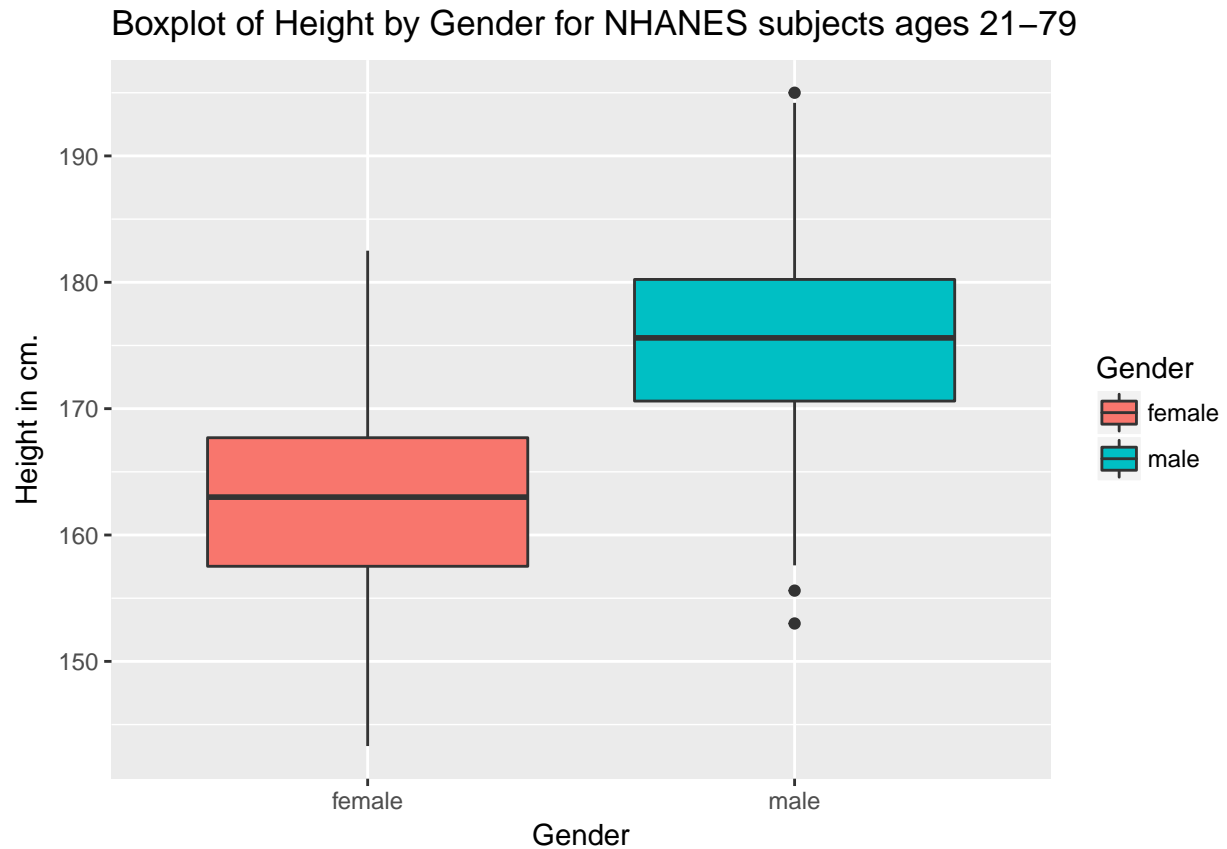
```
ggplot(data = nh_data_2179, aes(x = Gender, y = Height, color = Gender)) +  
  geom_jitter(width = 0.2) +  
  labs(title = "Height by Gender (jittered) for NHANES subjects ages 21-79",  
        y = "Height in cm.")
```



Perhaps it might be better to summarize the distribution in a different way. We might consider a boxplot of the data.

3.7.1 A Boxplot of Height by Gender

```
ggplot(data = nh_data_2179, aes(x = Gender, y = Height, fill = Gender)) +  
  geom_boxplot() +  
  labs(title = "Boxplot of Height by Gender for NHANES subjects ages 21-79",  
        y = "Height in cm.")
```

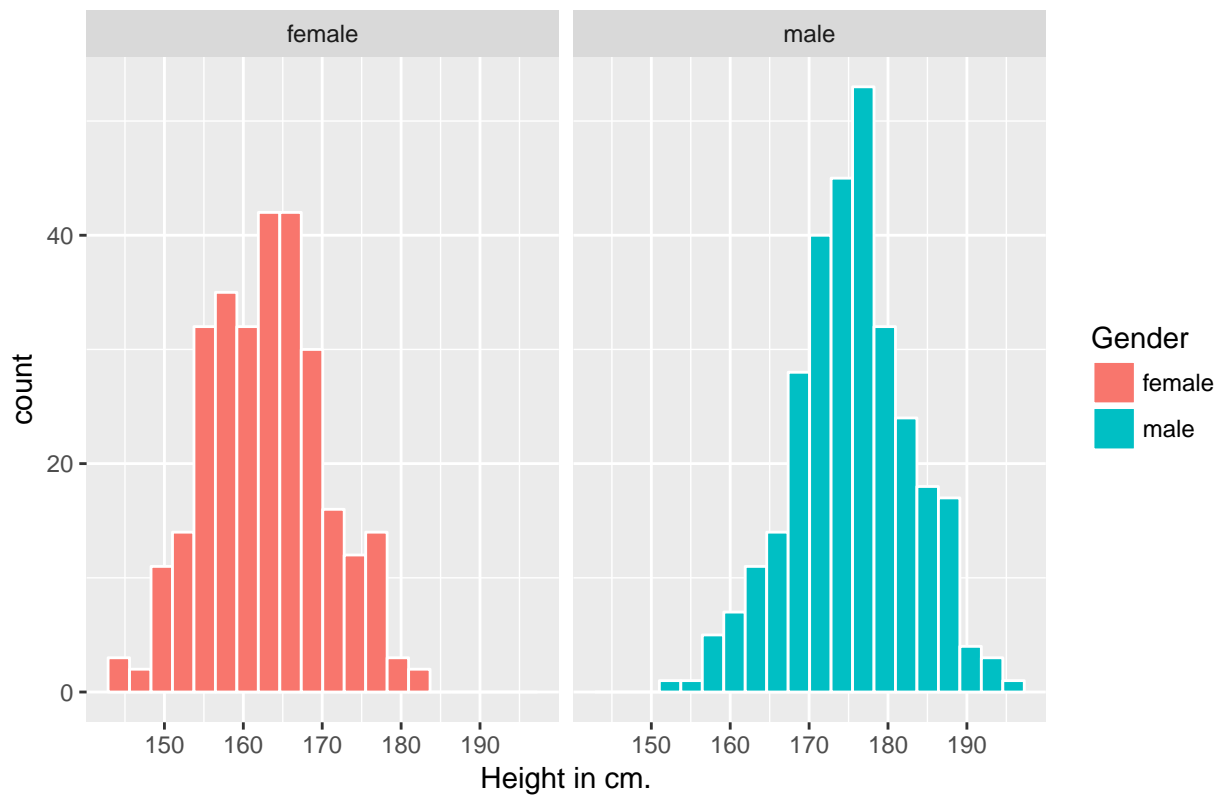


Or perhaps we'd like to see a pair of histograms?

3.7.2 Histograms of Height by Gender

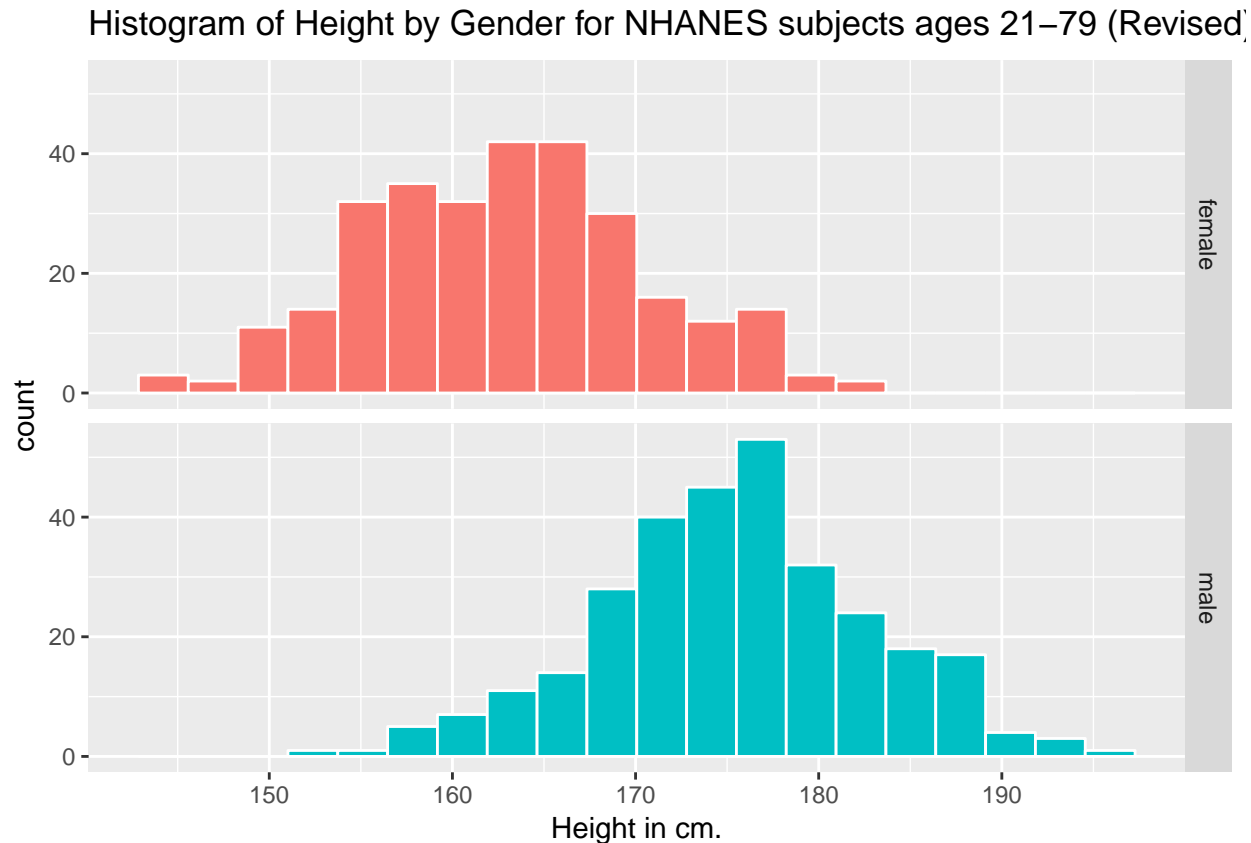
```
ggplot(data = nh_data_2179, aes(x = Height, fill = Gender)) +  
  geom_histogram(color = "white", bins = 20) +  
  labs(title = "Histogram of Height by Gender for NHANES subjects ages 21-79",  
        x = "Height in cm.") +  
  facet_wrap(~ Gender)
```

Histogram of Height by Gender for NHANES subjects ages 21–79



Can we redraw these histograms so that they are a little more comparable, and to get rid of the unnecessary legend?

```
ggplot(data = nh_data_2179, aes(x = Height, fill = Gender)) +
  geom_histogram(color = "white", bins = 20) +
  labs(title = "Histogram of Height by Gender for NHANES subjects ages 21-79 (Revised)",
       x = "Height in cm.") +
  guides(fill = FALSE) +
  facet_grid(Gender ~ .)
```



3.8 A Look at Body-Mass Index

Let's look at a different outcome, the *body-mass index*, or BMI. The definition of BMI for adult subjects (which is expressed in units of kg/m^2) is:

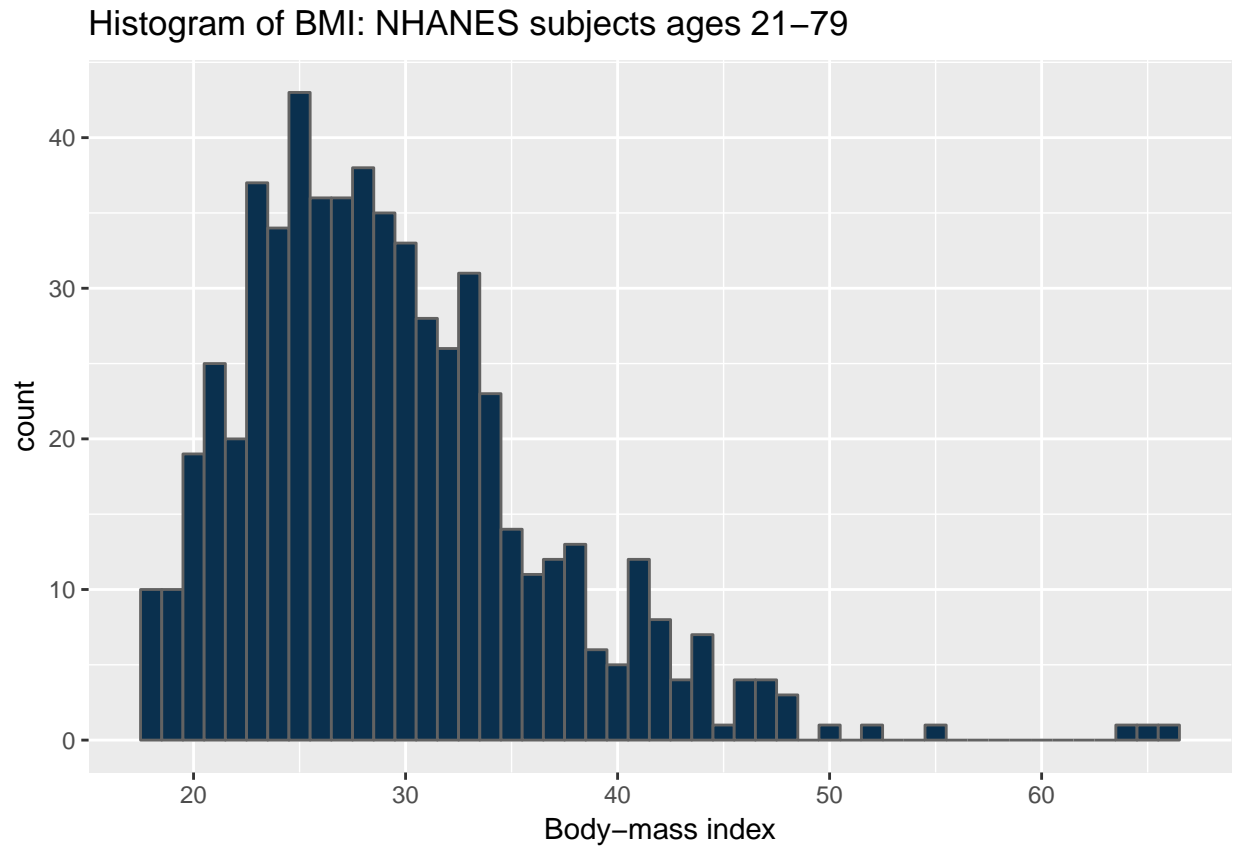
$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in meters})^2} = 703 \times \frac{\text{weight in pounds}}{(\text{height in inches})^2}$$

[BMI is essentially] ... a measure of a person's *thinness* or *thickness*... BMI was designed for use as a simple means of classifying average sedentary (physically inactive) populations, with an average body composition. For these individuals, the current value recommendations are as follow: a BMI from 18.5 up to 25 may indicate optimal weight, a BMI lower than 18.5 suggests the person is underweight, a number from 25 up to 30 may indicate the person is overweight, and a number from 30 upwards suggests the person is obese.

Wikipedia, https://en.wikipedia.org/wiki/Body_mass_index

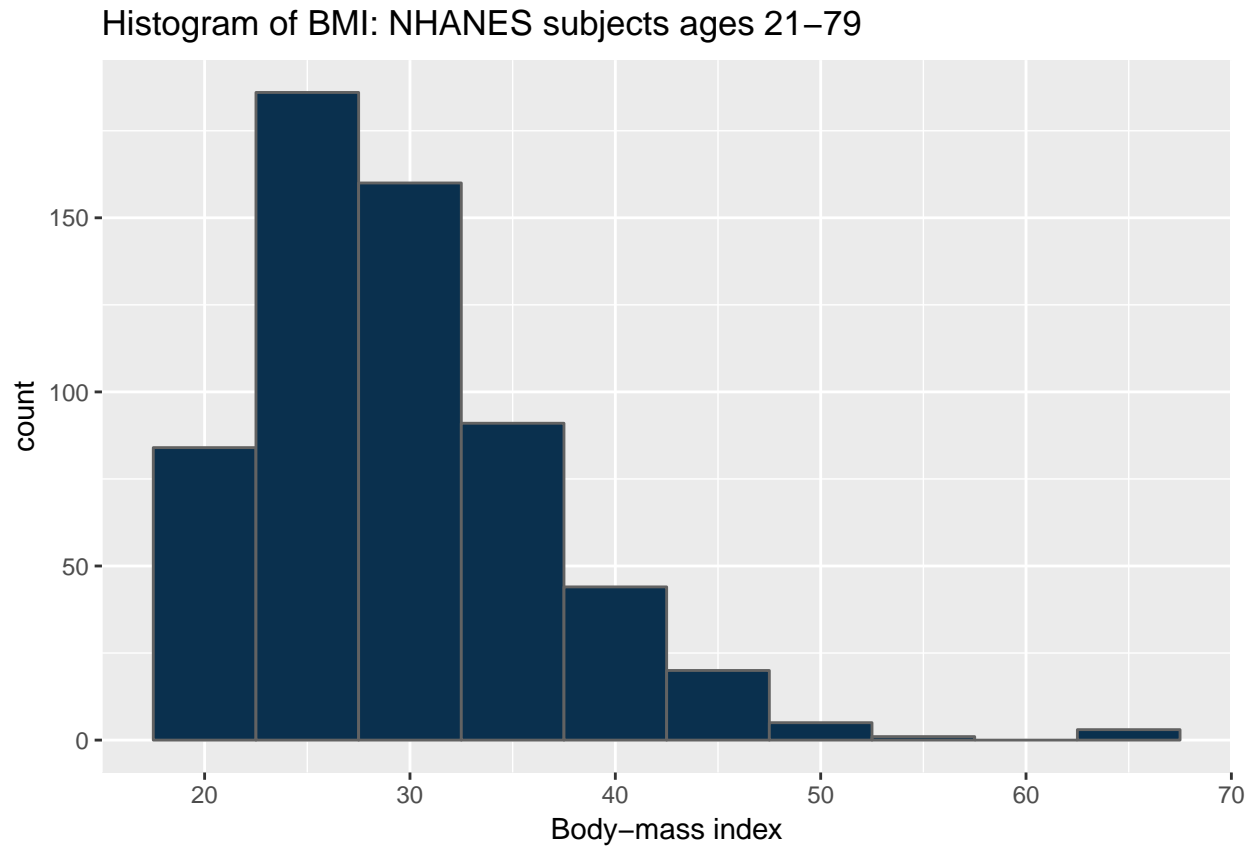
Here's a histogram, again with CWRU colors, for the BMI data.

```
ggplot(data = nh_data_2179, aes(x = BMI)) +
  geom_histogram(binwidth = 1, fill = cwr.blue, col = cwr.gray) +
  labs(title = "Histogram of BMI: NHANES subjects ages 21-79",
       x = "Body-mass index")
```



Note how different this picture looks if instead we bin up groups of 5 kg/m² at a time. Which is the more useful representation will depend a lot on what questions you're trying to answer.

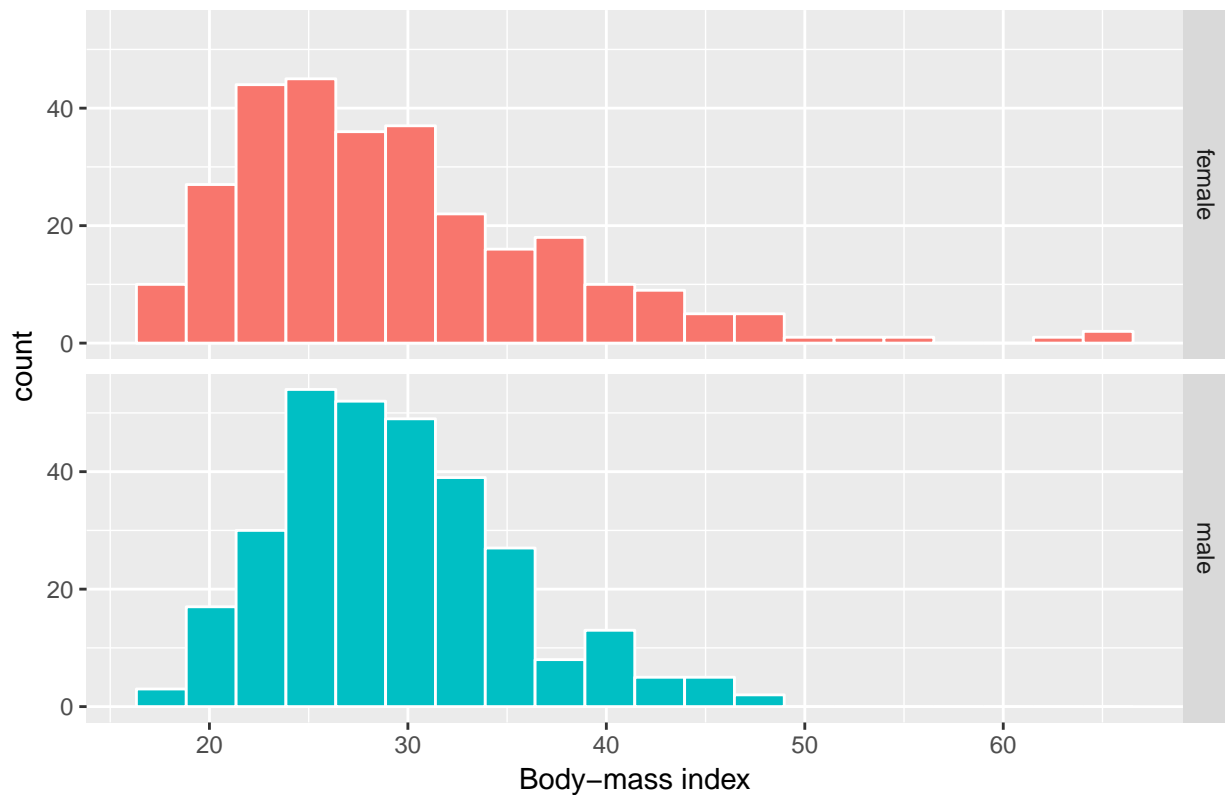
```
ggplot(data = nh_data_2179, aes(x = BMI)) +  
  geom_histogram(binwidth = 5, fill = cwrp.blue, col = cwrp.gray) +  
  labs(title = "Histogram of BMI: NHANES subjects ages 21-79",  
       x = "Body-mass index")
```



3.8.1 BMI by Gender

```
ggplot(data = nh_data_2179, aes(x = BMI, fill = Gender)) +  
  geom_histogram(color = "white", bins = 20) +  
  labs(title = "Histogram of BMI by Gender for NHANES subjects ages 21-79",  
        x = "Body-mass index") +  
  guides(fill = FALSE) +  
  facet_grid(Gender ~ .)
```

Histogram of BMI by Gender for NHANES subjects ages 21–79



As an accompanying numerical summary, we might ask how many people fall into each of these Gender categories, and what is their “average” BMI.

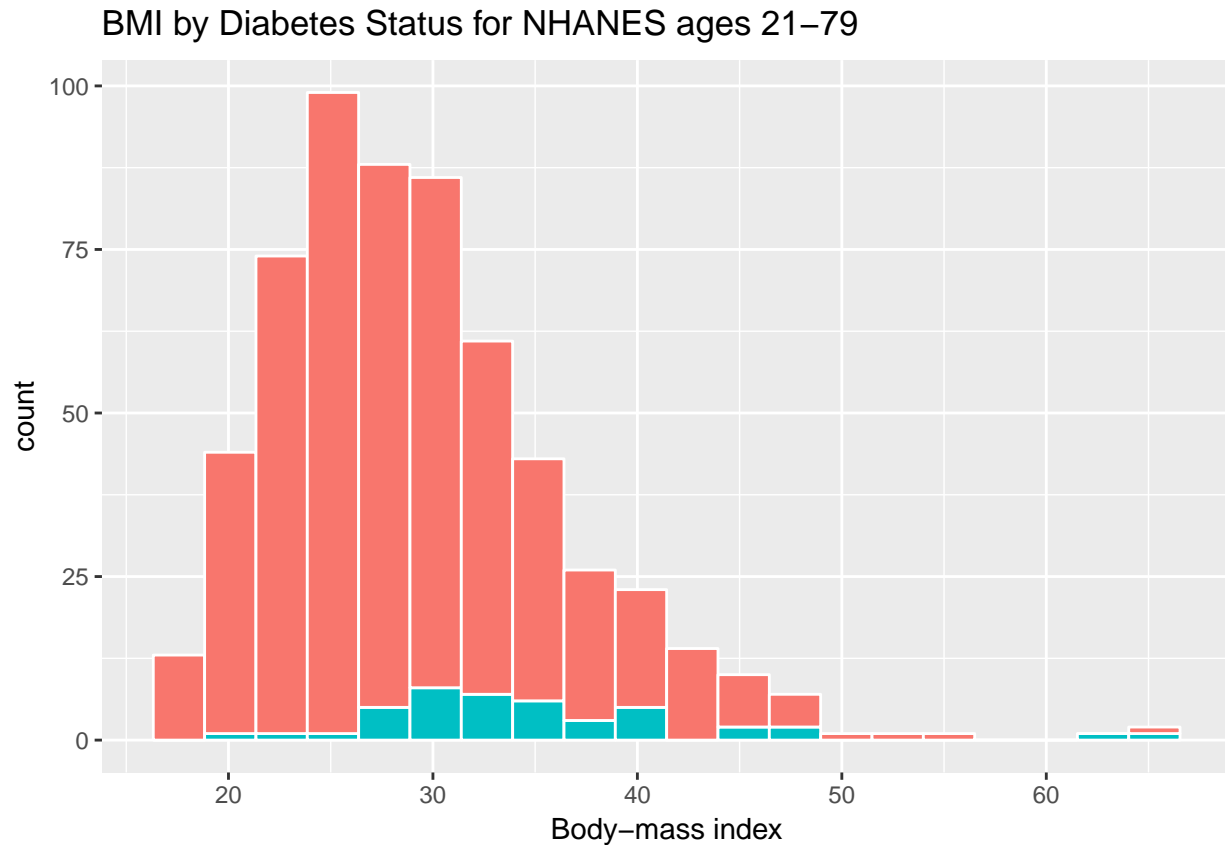
```
nh_data_2179 %>%
  group_by(Gender) %>%
  summarize(count = n(), mean(BMI), median(BMI)) %>%
  knitr::kable()
```

Gender	count	mean(BMI)	median(BMI)
female	290	29.4	27.4
male	304	29.4	28.7

3.8.2 BMI and Diabetes

We can split up our histogram into groups based on whether the subjects have been told they have diabetes.

```
ggplot(data = nh_data_2179, aes(x = BMI, fill = Diabetes)) +
  geom_histogram(color = "white", bins = 20) +
  labs(title = "BMI by Diabetes Status for NHANES ages 21-79",
       x = "Body-mass index") +
  guides(fill = FALSE)
```

How many people fall into each of these Diabetes categories, and what is their “average” BMI?

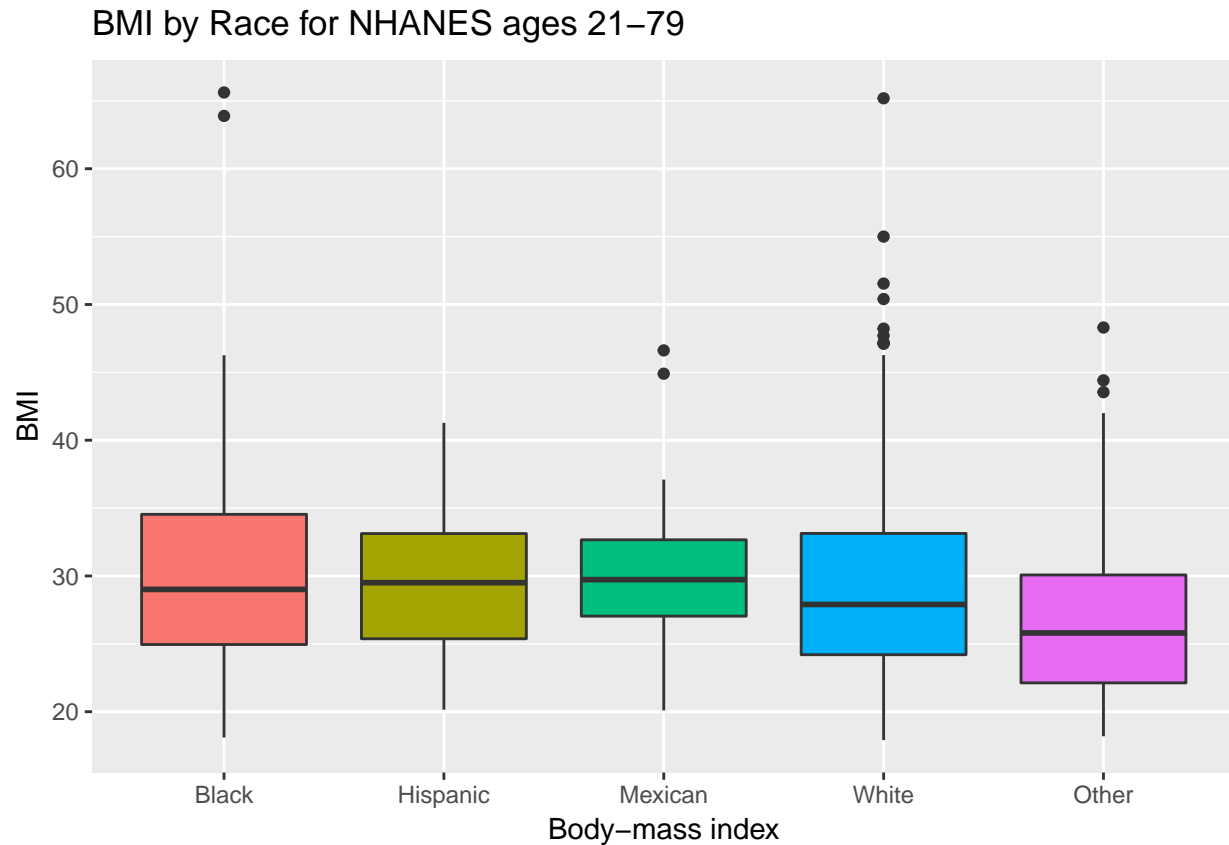
```
nh_data_2179 %>%
  group_by(Diabetes) %>%
  summarize(count = n(), mean(BMI), median(BMI)) %>%
  knitr::kable()
```

Diabetes	count	mean(BMI)	median(BMI)
No	551	28.9	27.9
Yes	43	35.3	33.4

3.8.3 BMI and Race

We can compare the distribution of BMI across Race groups, as well.

```
ggplot(data = nh_data_2179, aes(x = Race1, y = BMI, fill = Race1)) +
  geom_boxplot() +
  labs(title = "BMI by Race for NHANES ages 21-79",
       x = "Body-mass index") +
  guides(fill = FALSE)
```



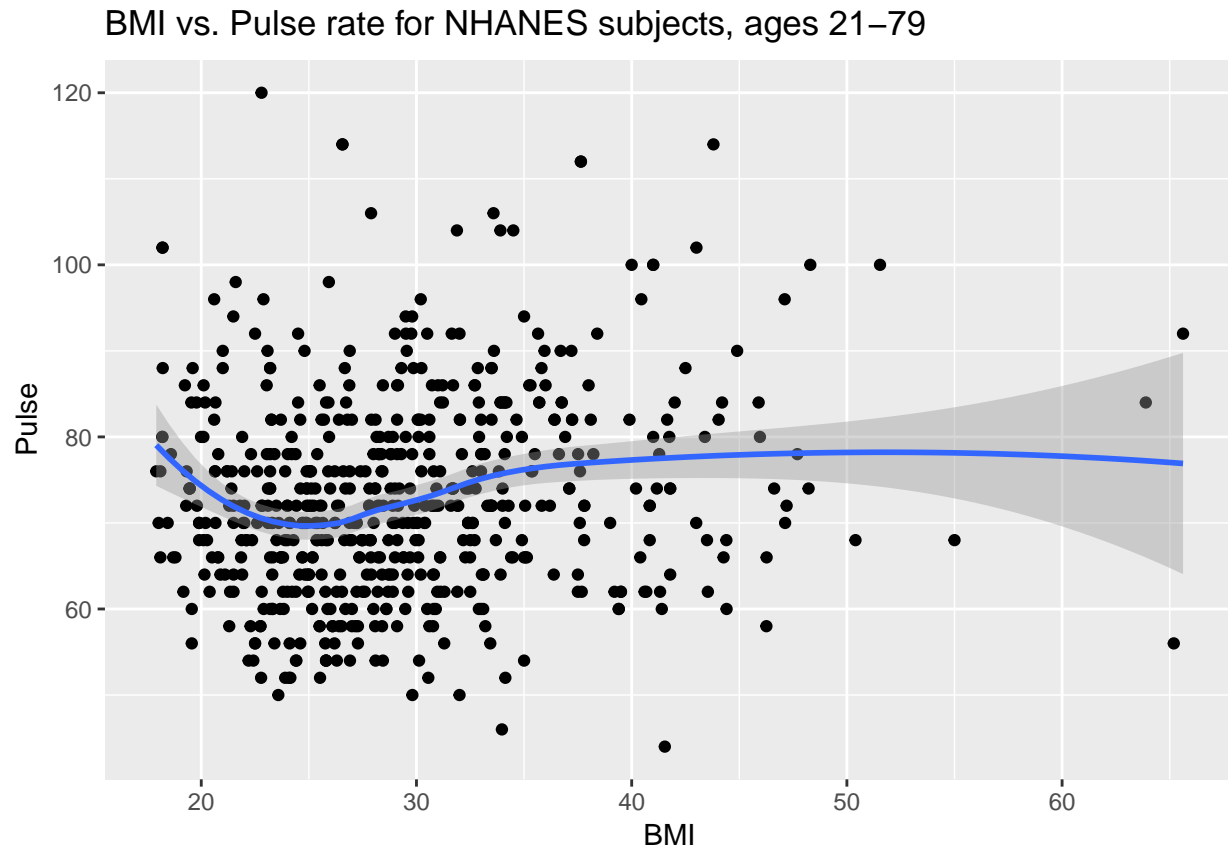
How many people fall into each of these Race1 categories, and what is their “average” BMI?

```
nh_data_2179 %>%
  group_by(Race1) %>%
  summarize(count = n(), mean(BMI), median(BMI)) %>%
  knitr::kable()
```

Race1	count	mean(BMI)	median(BMI)
Black	63	31.0	29.0
Hispanic	44	29.4	29.5
Mexican	50	30.0	29.7
White	387	29.3	27.9
Other	50	27.3	25.8

3.8.4 BMI and Pulse Rate

```
ggplot(data = nh_data_2179, aes(x = BMI, y = Pulse)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "BMI vs. Pulse rate for NHANES subjects, ages 21–79")
```

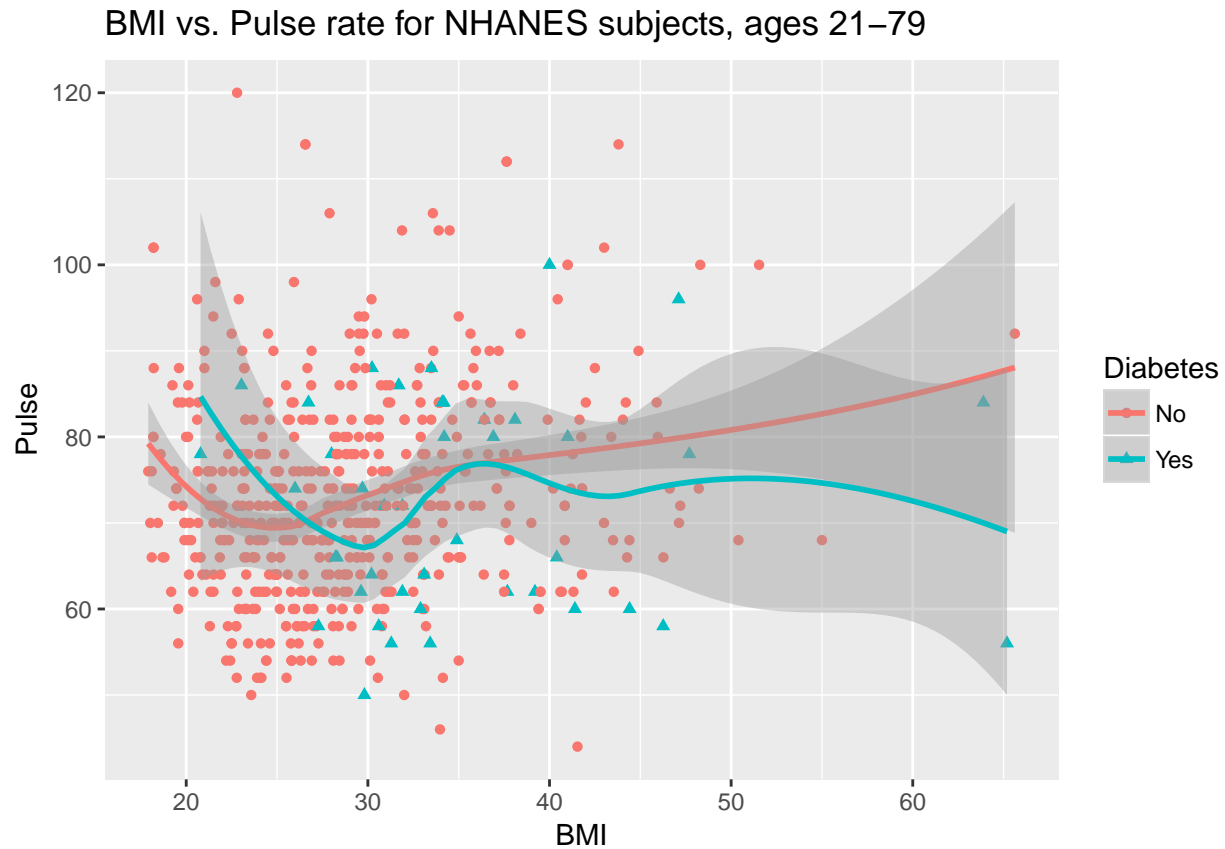


3.8.5 Diabetes vs. No Diabetes

Could we see whether subjects who have been told they have diabetes show different BMI-pulse rate patterns than the subjects who haven't?

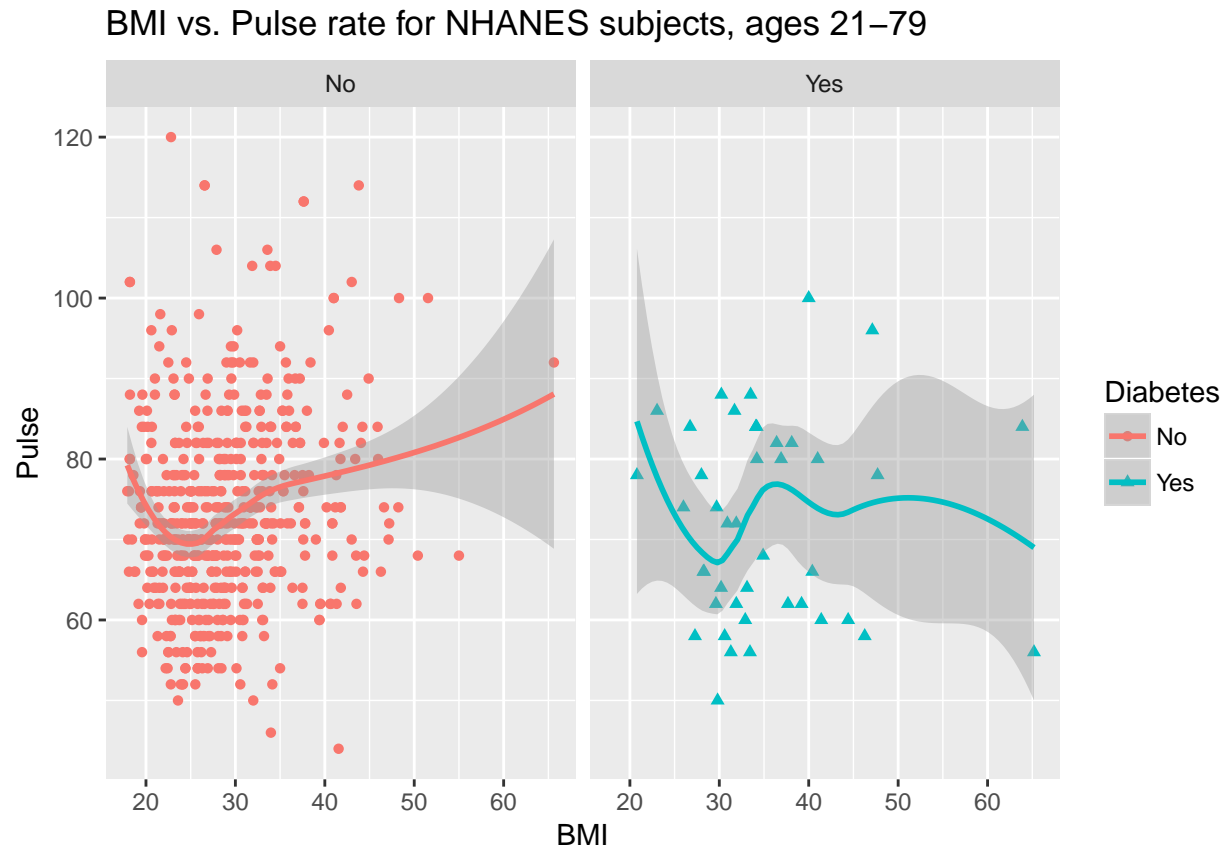
- Let's try doing this by changing the shape *and* the color of the points based on diabetes status.

```
ggplot(data = nh_data_2179,
  aes(x = BMI, y = Pulse,
    color = Diabetes, shape = Diabetes)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "BMI vs. Pulse rate for NHANES subjects, ages 21-79")
```



This plot might be easier to interpret if we faceted by Diabetes status, as well.

```
ggplot(data = nh_data_2179,
  aes(x = BMI, y = Pulse,
    color = Diabetes, shape = Diabetes)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "BMI vs. Pulse rate for NHANES subjects, ages 21–79") +
  facet_wrap(~ Diabetes)
```



3.9 General Health Status

Here's a Table of the General Health Status results. This is a self-reported rating of each subject's health on a five point scale (Excellent, Very Good, Good, Fair, Poor.)

```
nh_data_2179 %>%
  select(HealthGen) %>%
  table()
```

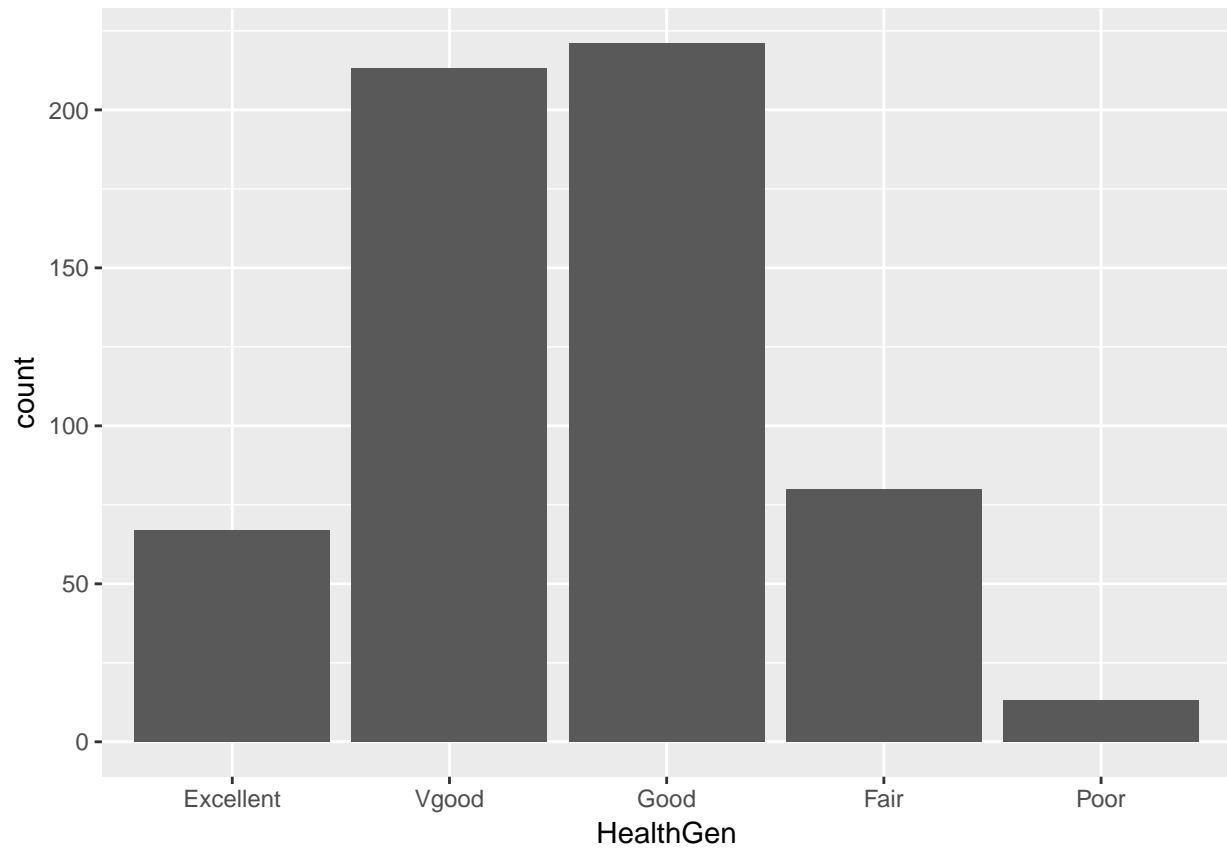
Excellent	Vgood	Good	Fair	Poor
67	213	221	80	13

The HealthGen data are categorical, which means that summarizing them with averages isn't as appealing as looking at percentages, proportions and rates.

3.9.1 Bar Chart for Categorical Data

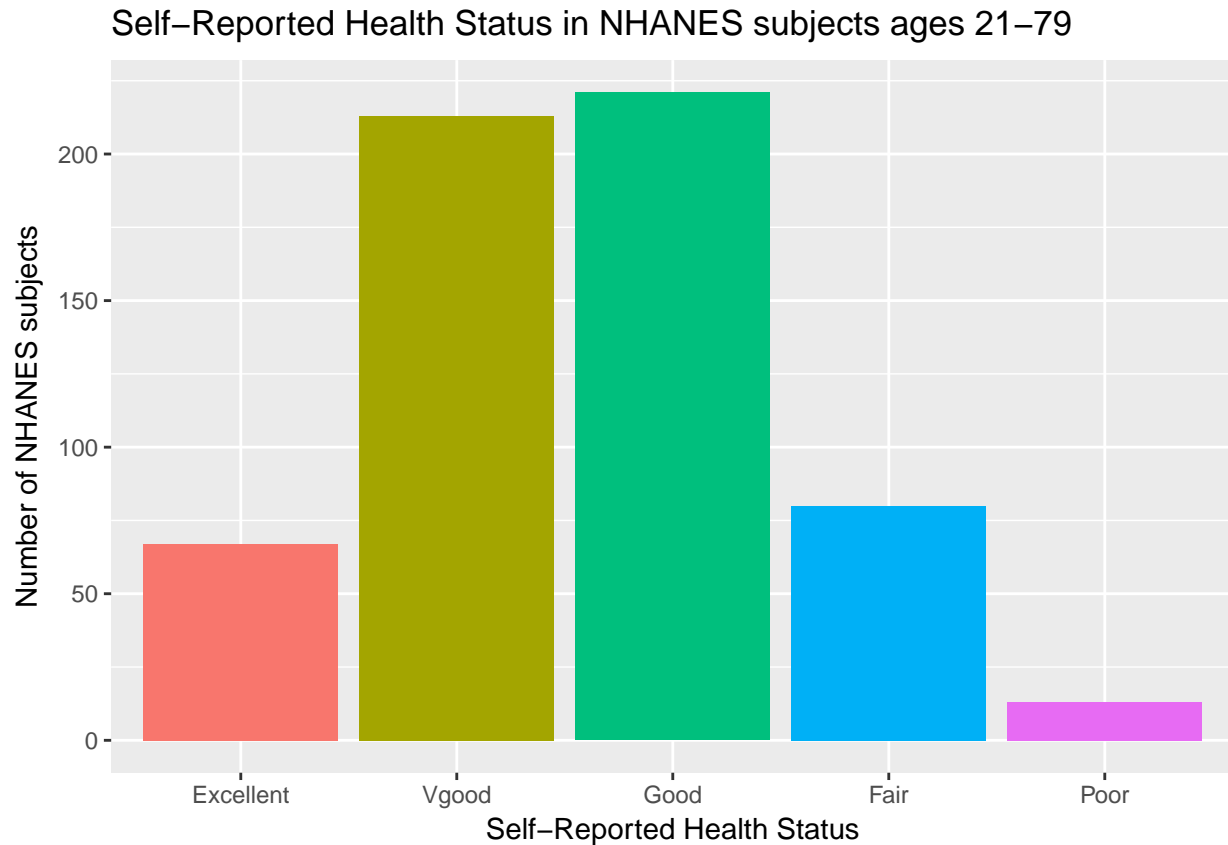
Usually, a **bar chart** is the best choice for a graphing a variable made up of categories.

```
ggplot(data = nh_data_2179, aes(x = HealthGen)) +
  geom_bar()
```



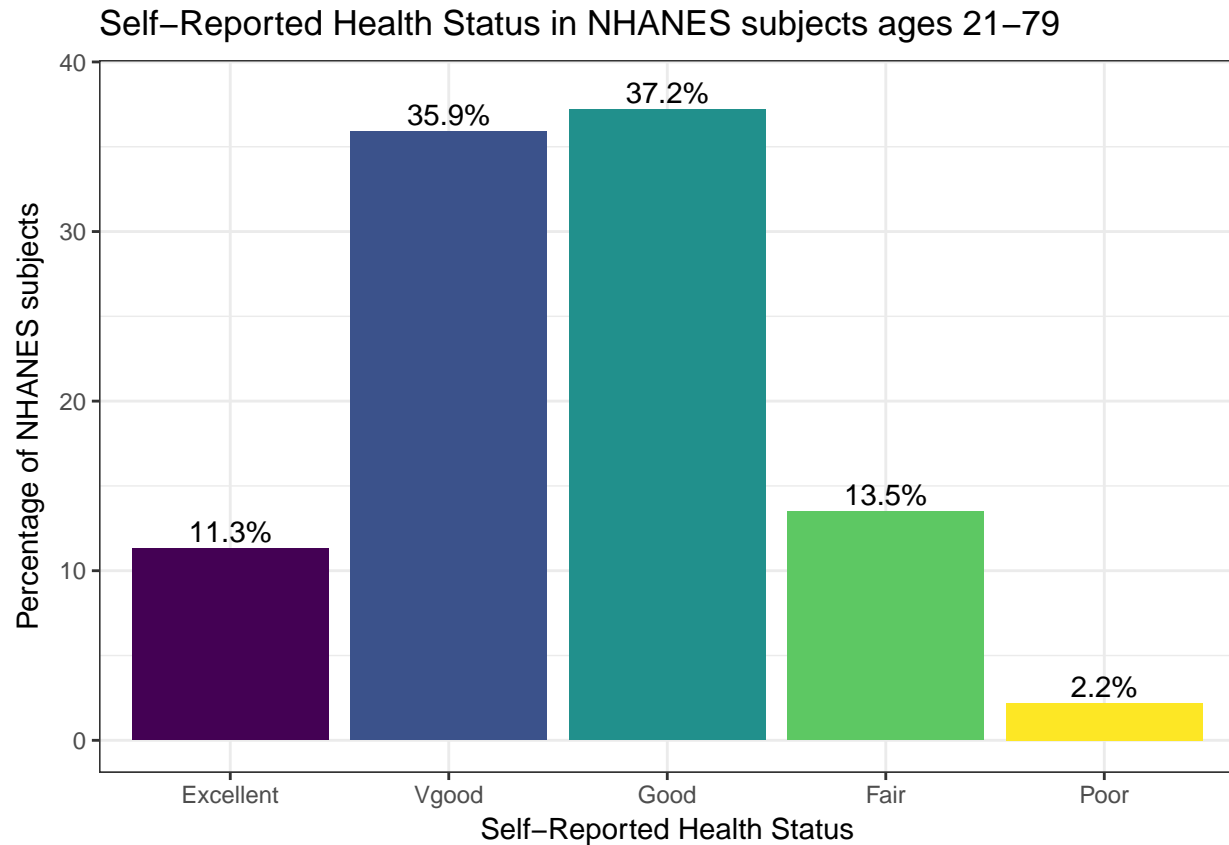
There are lots of things we can do to make this plot fancier.

```
ggplot(data = nh_data_2179, aes(x = HealthGen, fill = HealthGen)) +  
  geom_bar() +  
  guides(fill = FALSE) +  
  labs(x = "Self-Reported Health Status",  
       y = "Number of NHANES subjects",  
       title = "Self-Reported Health Status in NHANES subjects ages 21-79")
```



Or, we can really go crazy...

```
nh_data_2179 %>%
  count(HealthGen) %>%
  ungroup() %>%
  mutate(pct = round(prop.table(n) * 100, 1)) %>%
  ggplot(aes(x = HealthGen, y = pct, fill = HealthGen)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_viridis(discrete = TRUE) +
  guides(fill = FALSE) +
  geom_text(aes(y = pct + 1,      # nudge above top of bar
                label = paste0(pct, '%'), # prettify
                position = position_dodge(width = .9),
                size = 4) +
  labs(x = "Self-Reported Health Status",
       y = "Percentage of NHANES subjects",
       title = "Self-Reported Health Status in NHANES subjects ages 21-79") +
  theme_bw()
```



3.9.2 Working with Tables

We can add a marginal total, and compare subjects by Gender, as follows...

```
nh_data_2179 %>%
  select(Gender, HealthGen) %>%
  table() %>%
  addmargins()
```

	HealthGen					
Gender	Excellent	Vgood	Good	Fair	Poor	Sum
female	34	107	107	34	8	290
male	33	106	114	46	5	304
Sum	67	213	221	80	13	594

If we like, we can make this look a little more polished with the `knitr::kable` function...

```
nh_data_2179 %>%
  select(Gender, HealthGen) %>%
  table() %>%
  addmargins() %>%
  knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor	Sum
female	34	107	107	34	8	290
male	33	106	114	46	5	304
Sum	67	213	221	80	13	594

If we want the proportions of patients within each Gender that fall in each HealthGen category (the row percentages), we can get them, too.

```
nh_data_2179 %>%
  select(Gender, HealthGen) %>%
  table() %>%
  prop.table(.,1) %>%
  knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor
female	0.117	0.369	0.369	0.117	0.028
male	0.109	0.349	0.375	0.151	0.016

To make this a little easier to use, we might consider rounding.

```
nh_data_2179 %>%
  select(Gender, HealthGen) %>%
  table() %>%
  prop.table(.,1) %>%
  round(.,2) %>%
  knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor
female	0.12	0.37	0.37	0.12	0.03
male	0.11	0.35	0.38	0.15	0.02

Another possibility would be to show the percentages, rather than the proportions (which requires multiplying the proportion by 100.) Note the strange "*" function, which is needed to convince R to multiply each entry by 100 here.

```
nh_data_2179 %>%
  select(Gender, HealthGen) %>%
  table() %>%
  prop.table(.,1) %>%
  "*" (100) %>%
  round(.,2) %>%
  knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor
female	11.7	36.9	36.9	11.7	2.76
male	10.9	34.9	37.5	15.1	1.64

And, if we wanted the column percentages, to determine which gender had the higher rate of each HealthGen status level, we can get that by changing the prop.table to calculate 2 (column) proportions, rather than 1 (rows.)

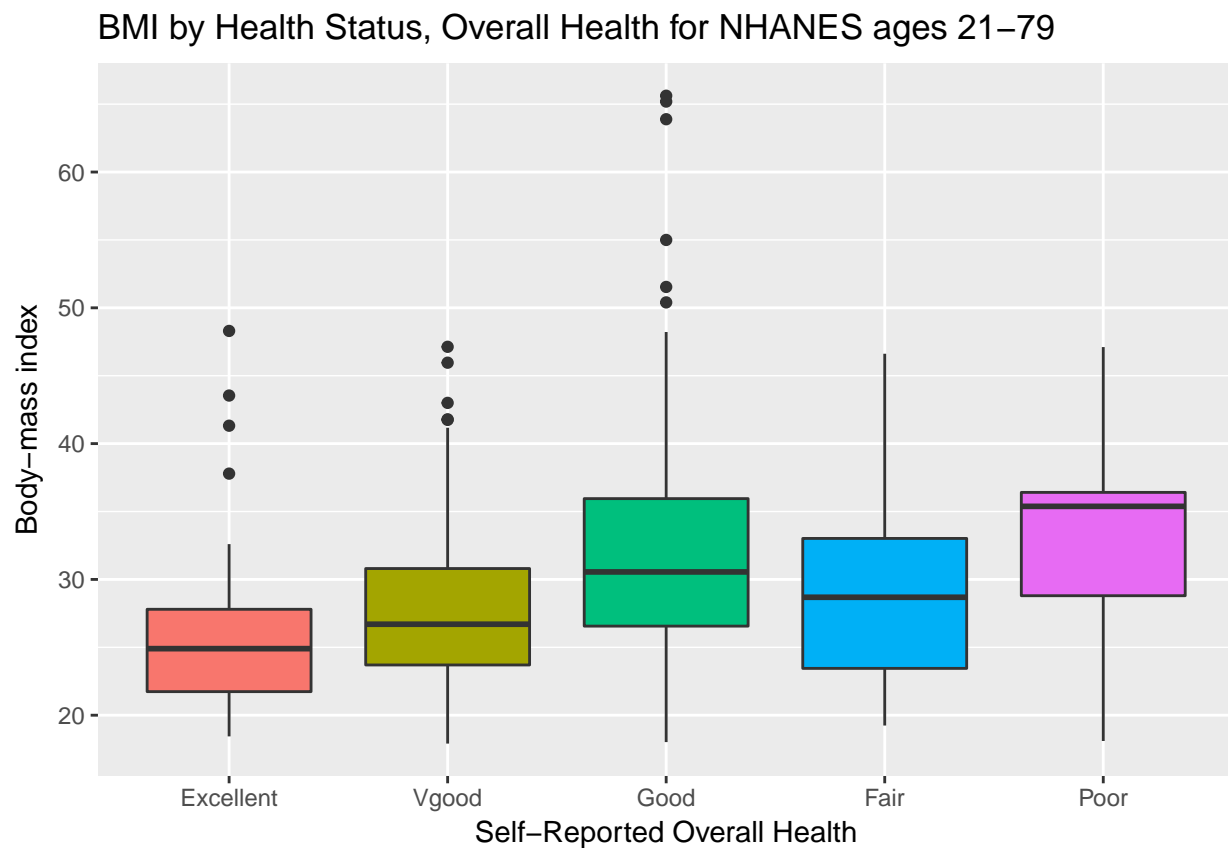
```
nh_data_2179 %>%
  select(Gender, HealthGen) %>%
  table() %>%
  prop.table(.,2) %>%
  "*" (100) %>%
  round(.,2) %>%
  knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor
female	50.8	50.2	48.4	42.5	61.5
male	49.2	49.8	51.6	57.5	38.5

3.9.3 BMI by General Health Status

Let's consider now the relationship between self-reported overall health and body-mass index.

```
ggplot(data = nh_data_2179, aes(x = HealthGen, y = BMI, fill = HealthGen)) +
  geom_boxplot() +
  labs(title = "BMI by Health Status, Overall Health for NHANES ages 21-79",
       y = "Body-mass index", x = "Self-Reported Overall Health") +
  guides(fill = FALSE)
```



We can see that not too many people self-identify with the “Poor” health category.

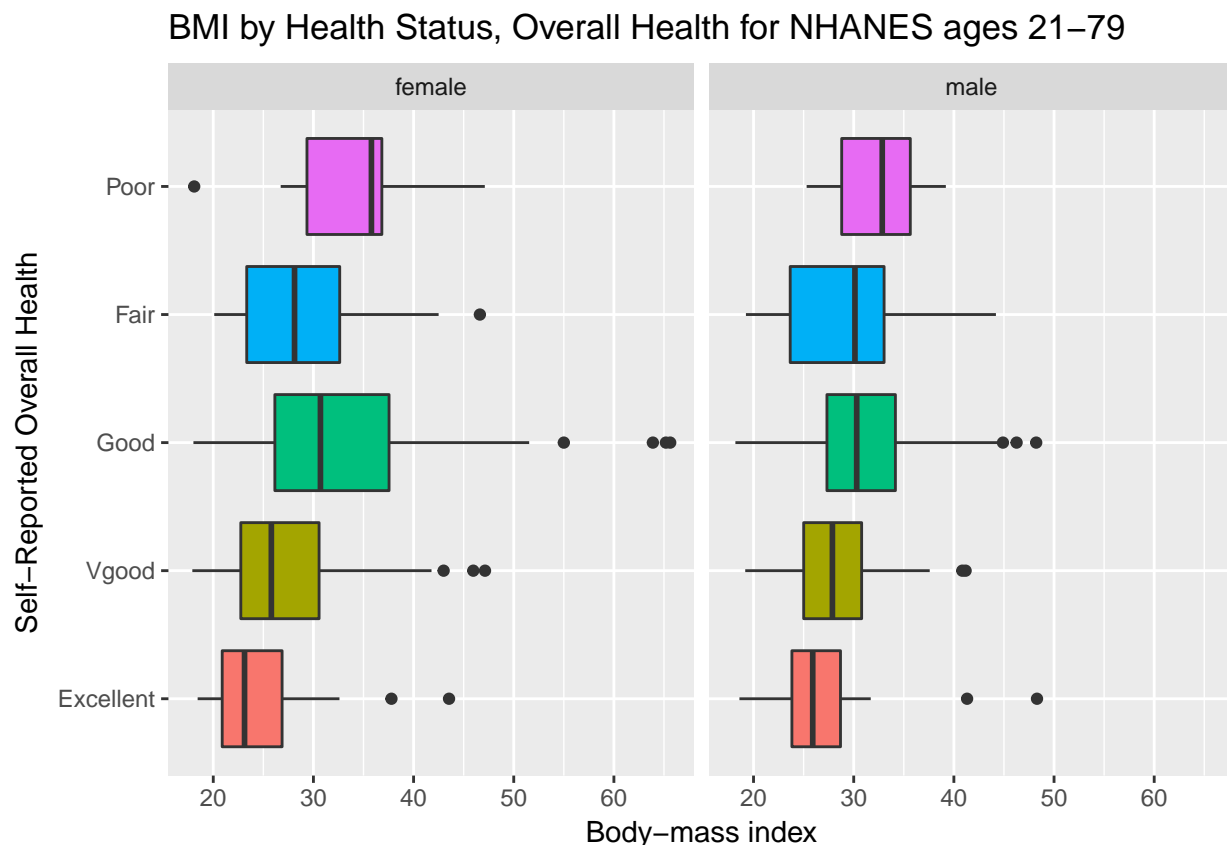
```
nh_data_2179 %>%
  group_by(HealthGen) %>%
  summarize(count = n(), mean(BMI), median(BMI)) %>%
  knitr::kable()
```

HealthGen	count	mean(BMI)	median(BMI)
Excellent	67	25.7	24.9
Vgood	213	27.6	26.7
Good	221	32.0	30.6
Fair	80	29.3	28.7
Poor	13	33.1	35.4

3.9.4 BMI by Gender and General Health Status

We'll start with two panels of boxplots to try to understand the relationships between BMI, General Health Status and Gender. Note the use of `coord_flip` to rotate the graph 90 degrees.

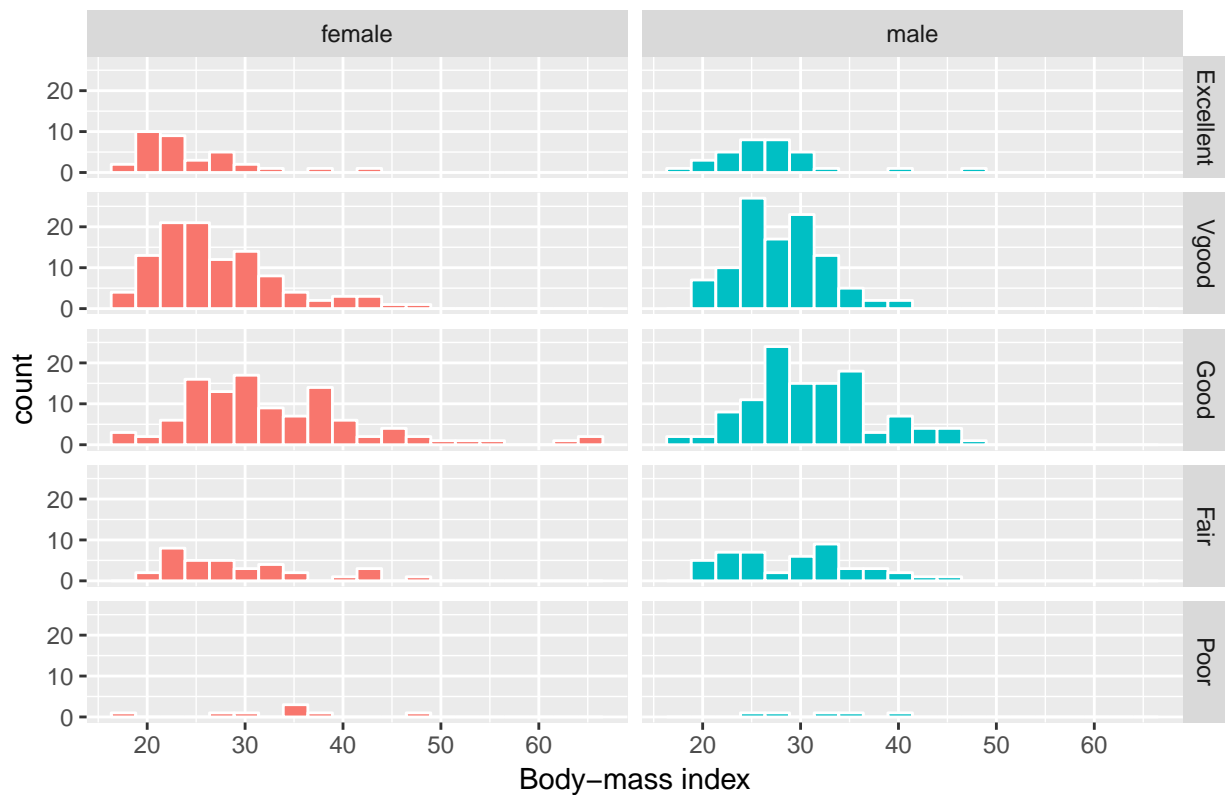
```
ggplot(data = nh_data_2179, aes(x = HealthGen, y = BMI, fill = HealthGen)) +
  geom_boxplot() +
  labs(title = "BMI by Health Status, Overall Health for NHANES ages 21-79",
       y = "Body-mass index", x = "Self-Reported Overall Health") +
  guides(fill = FALSE) +
  facet_wrap(~ Gender) +
  coord_flip()
```



Here's a plot of faceted histograms, which might be used to address similar questions.

```
ggplot(data = nh_data_2179, aes(x = BMI, fill = Gender)) +
  geom_histogram(color = "white", bins = 20) +
  labs(title = "BMI by Gender, Overall Health for NHANES ages 21-79",
       x = "Body-mass index") +
  guides(fill = FALSE) +
  facet_grid(HealthGen ~ Gender)
```

BMI by Gender, Overall Health for NHANES ages 21–79



3.10 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.

Chapter 4

Data Structures and Types of Variables

4.1 Data require structure and context

Descriptive statistics are concerned with the presentation, organization and summary of data, as suggested in Norman and Streiner (2014). This includes various methods of organizing and graphing data to get an idea of what those data can tell us.

As Vittinghoff et al. (2012) suggest, the nature of the measurement determines how best to describe it statistically, and the main distinction is between **numerical** and **categorical** variables. Even this is a little tricky - plenty of data can have values that look like numerical values, but are just numerals serving as labels.

As Bock, Velleman, and De Veaux (2004) point out, the truly critical notion, of course, is that data values, no matter what kind, are useless without their contexts. The Five W's (Who, What [and in what units], When, Where, Why, and often How) are just as useful for establishing the context of data as they are in journalism. If you can't answer Who and What, in particular, you don't have any useful information.

In general, each row of a data frame corresponds to an individual (respondent, experimental unit, record, or observation) about whom some characteristics are gathered in columns (and these characteristics may be called variables, factors or data elements.) Every column / variable should have a name that indicates *what* it is measuring, and every row / observation should have a name that indicates *who* is being measured.

4.2 A New NHANES Adult Sample

In previous work, we spent some time with a sample from the National Health and Nutrition Examination. Now, by changing the value of the `set.seed` function which determines the starting place for the random sampling, and changing some other specifications, we'll generate a new sample describing 500 adult subjects who completed the 2011-12 version of the survey when they were between the ages of 21 and 64.

Note also that what is listed in the NHANES data frame as **Gender** should be more correctly referred to as **sex**. **Sex** is a biological feature of an individual, while **Gender** is a social construct. This is an important distinction, so I'll change the name of the variable. I'm also changing the names of three other variables, to create **Race**, **SBP** and **DBP**.

```
library(NHANES) # load the NHANES package/library of functions, data

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)

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

nh_adults <- sample_n(nh_temp, size = 500)

nh_adults

```

```

# A tibble: 500 x 13
      ID   Sex  Age  Race  Education  BMI  SBP  DBP  Pulse
  <int> <fctr> <int> <fctr>      <fctr> <dbl> <int> <int> <int>
1 64427  male   37  White College Grad 36.5  111   72   56
2 63788  female  40  White High School 18.2  115   74  102
3 66874  female  31  White Some College 27.2   95   52   98
4 69734  male   26  White College Grad 20.6  137   75   74
5 70409  male   44  White High School 29.2  112   71   62
6 68961  female  64  White College Grad 24.2  123   70   80
7 62616  female  37  Asian  8th Grade 19.3  109   73   82
8 70130  male   42  Black High School 31.2  119   71   62
9 71218  male   33  White College Grad 27.7  110   67   68
10 69181  female  37  White  8th Grade 25.0  114   74   82
# ... with 490 more rows, and 4 more variables: PhysActive <fctr>,
#   Smoke100 <fctr>, SleepTrouble <fctr>, HealthGen <fctr>

```

The data consists of 500 rows (observations) on 13 variables (columns). Essentially, we have 13 pieces of information on each of 500 adult NHANES subjects who were included in the 2011-12 panel.

4.2.1 Summarizing the Data's Structure

We can identify the number of rows and columns in a data frame or tibble with the `dim` function.

```
dim(nh_adults)
```

```
[1] 500 13
```

The `str` function provides a lot of information about the structure of a data frame or tibble.

```
str(nh_adults)
```

```

Classes 'tbl_df', 'tbl' and 'data.frame': 500 obs. of 13 variables:
 $ ID      : int  64427 63788 66874 69734 70409 68961 62616 70130 71218 69181 ...
 $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 2 2 1 1 2 2 1 ...
 $ Age     : int   37 40 31 26 44 64 37 42 33 37 ...
 $ Race    : Factor w/ 6 levels "Asian","Black",...: 5 5 5 5 5 5 1 2 5 5 ...
 $ Education : Factor w/ 5 levels "8th Grade","9 - 11th Grade",...: 5 3 4 5 3 5 1 3 5 1 ...
 $ BMI     : num   36.5 18.2 27.2 20.6 29.2 24.2 19.3 31.2 27.7 25 ...
 $ SBP     : int   111 115 95 137 112 123 109 119 110 114 ...
 $ DBP     : int    72 74 52 75 71 70 73 71 67 74 ...
 $ Pulse   : int    56 102 98 74 62 80 82 62 68 82 ...
 $ PhysActive : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1 1 2 2 ...
 $ Smoke100 : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 2 1 1 1 2 ...

```

```
$ SleepTrouble: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 2 ...
$ HealthGen    : Factor w/ 5 levels "Excellent","Vgood",...: 2 3 3 1 3 2 3 3 2 ...
```

To see the first few observations, use `head`, and to see the last few, try `tail`...

```
tail(nh_adults, 5) # shows the last five observations in the data set
```

```
# A tibble: 5 x 13
  ID    Sex Age Race Education BMI SBP DBP Pulse
<int> <fctr> <int> <fctr>      <fctr> <dbl> <int> <int> <int>
1 69692 male   50 Black 9 - 11th Grade 22.7 132 82 60
2 66472 male   61 White Some College 41.3 141 77 62
3 71456 male   21 Mexican 9 - 11th Grade 26.7 113 66 78
4 71420 female 54 Mexican 9 - 11th Grade 32.5 126 69 68
5 63617 male   29 White College Grad 23.2 105 72 76
# ... with 4 more variables: PhysActive <fctr>, Smoke100 <fctr>,
# SleepTrouble <fctr>, HealthGen <fctr>
```

4.2.2 What are the variables?

The variables we have collected are described in the brief table below¹.

Variable	Description	Sample Values
ID	a numerical code identifying the subject	64427, 63788
Sex	sex of subject (2 levels)	male, female
Age	age (years) at screening of subject	37, 40
Race	reported race of subject (6 levels)	White, Asian
Education	educational level of subject (5 levels)	College Grad, High School
BMI	body-mass index, in kg/m ²	36.5, 18.2
SBP	systolic blood pressure in mm Hg	111, 115
DBP	diastolic blood pressure in mm Hg	72, 74
Pulse	60 second pulse rate in beats per minute	56, 102
PhysActive	Moderate or vigorous-intensity sports?	Yes, No
Smoke100	Smoked at least 100 cigarettes lifetime?	Yes, No
SleepTrouble	Told a doctor they have trouble sleeping?	Yes, No
HealthGen	Self-report general health rating (5 lev.)	Vgood, Good

The levels for the multi-categorical variables are:

- **Race:** Mexican, Hispanic, White, Black, Asian, or Other.
- **Education:** 8th Grade, 9 - 11th Grade, High School, Some College, or College Grad.
- **HealthGen:** Excellent, Vgood, Good, Fair or Poor.

4.3 Types of Variables

4.3.1 Quantitative Variables

Variables recorded in numbers that we use as numbers are called **quantitative**. Familiar examples include incomes, heights, weights, ages, distances, times, and counts. All quantitative variables have measurement

¹Descriptions are adapted from the ?NHANES help file. Remember that what NHANES lists as Gender is captured here as Sex, and similarly Race3, BPSysAve and BPDiaAve from NHANES are here listed as Race, SBP and DBP.

units, which tell you how the quantitative variable was measured. Without units (like miles per hour, angstroms, yen or degrees Celsius) the values of a quantitative variable have no meaning.

- It does little good to be promised a salary of 80,000 a year if you don't know whether it will be paid in Euros, dollars, yen or Estonian kroon.
- You might be surprised to see someone whose age is 72 listed in a database on childhood diseases until you find out that age is measured in months.
- Often just seeking the units can reveal a variable whose definition is challenging - just how do we measure “friendliness”, or “success,” for example.
- Quantitative variables may also be classified by whether they are **continuous** or can only take on a **discrete** set of values. Continuous data may take on any value, within a defined range. Suppose we are measuring height. While height is really continuous, our measuring stick usually only lets us measure with a certain degree of precision. If our measurements are only trustworthy to the nearest centimeter with the ruler we have, we might describe them as discrete measures. But we could always get a more precise ruler. The measurement divisions we make in moving from a continuous concept to a discrete measurement are usually fairly arbitrary. Another way to think of this, if you enjoy music, is that, as suggested in Norman and Streiner (2014), a piano is a *discrete* instrument, but a violin is a *continuous* one, enabling finer distinctions between notes than the piano is capable of making. Sometimes the distinction between continuous and discrete is important, but usually, it's not.
 - The `nh_adults` data includes several quantitative variables, specifically Age, BMI, SBP, DBP and Pulse.
 - We know these are quantitative because they have units: Age in years, BMI in kg/m^2 , the BP measurements in mm Hg, and Pulse in beats per minute.
 - Depending on the context, we would likely treat most of these as *discrete* given that are measurements are fairly crude (this is certainly true for Age, measured in years) although BMI is probably *continuous* in most settings, even though it is a function of two other measures (Height and Weight) which are rounded off to integer numbers of centimeters and kilograms, respectively.
- It is also possible to separate out quantitative variables into **ratio** variables or **interval** variables. An interval variable has equal distances between values, but the zero point is arbitrary. A ratio variable has equal intervals between values, and a meaningful zero point. For example, weight is an example of a ratio variable, while IQ is an example of an interval variable. We all know what zero weight is. An intelligence score like IQ is a different matter. We say that the average IQ is 100, but that's only by convention. We could just as easily have decided to add 400 to every IQ value and make the average 500 instead. Because IQ's intervals are equal, the difference between an IQ of 70 and an IQ of 80 is the same as the difference between 120 and 130. However, an IQ of 100 is not twice as high as an IQ of 50. The point is that if the zero point is artificial and moveable, then the differences between numbers are meaningful but the ratios between them are not. On the other hand, most lab test values are ratio variables, as are physical characteristics like height and weight. A person who weighs 100 kg is twice as heavy as one who weighs 50 kg; even when we convert kg to pounds, this is still true. For the most part, we can treat and analyze interval or ratio variables the same way.
 - Each of the quantitative variables in our `nh_adults` data can be thought of as ratio variables.
- Quantitative variables lend themselves to many of the summaries we will discuss, like means, quantiles, and our various measures of spread, like the standard deviation or inter-quartile range. They also have at least a chance to follow the Normal distribution.

4.3.2 Qualitative (Categorical) Variables

Qualitative or categorical variables consist of names of categories. These names may be numerical, but the numbers (or names) are simply codes to identify the groups or categories into which the individuals are divided. Categorical variables with two categories, like yes or no, up or down, or, more generally, 1 and 0,

are called **binary** variables. Those with more than two-categories are sometimes called **multi-categorical** variables.

- When the categories included in a variable are merely names, and come in no particular order, we sometimes call them **nominal** variables. The most important summary of such a variable is usually a table of frequencies, and the mode becomes an important single summary, while the mean and median are essentially useless.
 - In the `nh_adults` data, `Race` is clearly a nominal variable with multiple unordered categories.
- The alternative categorical variable (where order matters) is called **ordinal**, and includes variables that are sometimes thought of as falling right in between quantitative and qualitative variables.
 - Examples of ordinal multi-categorical variables in the `nh_adults` data include the `Education` and `HealthGen` variables.
 - Answers to questions like “How is your overall physical health?” with available responses Excellent, Very Good, Good, Fair or Poor, which are often coded as 1-5, certainly provide a perceived *order*, but a group of people with average health status 4 (Very Good) is not necessarily twice as healthy as a group with average health status of 2 (Fair).
- Sometimes we treat the values from ordinal variables as sufficiently scaled to permit us to use quantitative approaches like means, quantiles, and standard deviations to summarize and model the results, and at other times, we’ll treat ordinal variables as if they were nominal, with tables and percentages our primary tools.
- Note that all binary variables may be treated as ordinal, or nominal.
 - Binary variables in the `nh_adults` data include `Sex`, `PhysActive`, `Smoke100`, `SleepTrouble`. Each can be thought of as either ordinal or nominal.

Lots of variables may be treated as either quantitative or qualitative, depending on how we use them. For instance, we usually think of `age` as a quantitative variable, but if we simply use `age` to make the distinction between “child” and “adult” then we are using it to describe categorical information. Just because your variable’s values are numbers, don’t assume that the information provided is quantitative.

Chapter 5

Summarizing Quantitative Variables

Most numerical summaries that might be new to you are applied most appropriately to quantitative variables. The measures that will interest us relate to:

- the **center** of our distribution,
- the **spread** of our distribution, and
- the **shape** of our distribution.

5.1 The `summary` function for Quantitative data

R provides a small sampling of numerical summaries with the `summary` function, for instance.

```
nh_adults %>%  
  select(Age, BMI, SBP, DBP, Pulse) %>%  
  summary()
```

Age		BMI		SBP		DBP		Pulse	
Min.	:21.0	Min.	:17.8	Min.	: 84	Min.	: 19.0	Min.	: 46
1st Qu.	:31.0	1st Qu.	:24.2	1st Qu.	:109	1st Qu.	: 65.0	1st Qu.	: 64
Median	:42.0	Median	:27.7	Median	:118	Median	: 72.0	Median	: 72
Mean	:42.1	Mean	:28.7	Mean	:119	Mean	: 72.2	Mean	: 73
3rd Qu.	:53.0	3rd Qu.	:32.1	3rd Qu.	:127	3rd Qu.	: 79.0	3rd Qu.	: 80
Max.	:64.0	Max.	:69.0	Max.	:202	Max.	:105.0	Max.	:120
		NA's	:3	NA's	:15	NA's	:15	NA's	:15

This basic summary includes a set of five **quantiles**¹, plus the sample's **mean**.

- **Min.** = the **minimum** value for each variable, so, for example, the youngest subject's Age was 21.
- **1st Qu.** = the **first quartile** (25th percentile) for each variable - for example, 25% of the subjects were Age 31 or younger.
- **Median** = the **median** (50th percentile) - half of the subjects were Age 42 or younger.
- **Mean** = the **mean**, usually what one means by an *average* - the sum of the Ages divided by 500 is 42.1.
- **3rd Qu.** = the **third quartile** (75th percentile) - 25% of the subjects were Age 53 or older.
- **Max.** = the **maximum** value for each variable, so the oldest subject was Age 64.

The summary also specifies the number of missing values for each variable. Here, we are missing 3 of the BMI values, for example.

¹The quantiles (sometimes referred to as percentiles) can also be summarized with a boxplot.


```
[162] 34 34 34 34 35 35 35 35 36 36 36 36 36 36 36 36 37 37 37 37 37 37
[185] 37 37 37 37 37 37 37 37 37 37 38 38 38 38 38 38 38 38 38 38 39 39 39
[208] 39 39 39 39 39 39 39 39 39 39 40 40 40 40 40 40 40 40 40 40 41 41 41
[231] 41 41 41 41 41 41 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 43
[254] 43 43 43 43 43 43 43 43 43 43 44 44 44 44 44 44 44 44 44 44 44 44 45
[277] 45 45 45 45 45 45 46 46 46 46 46 46 46 46 46 46 46 46 46 46 47 47 47
[300] 47 47 47 47 47 47 48 48 48 48 48 48 48 48 48 48 48 48 48 48 49 49 49
[323] 49 49 49 49 49 49 49 49 49 49 50 50 50 50 50 50 50 50 50 50 50 50 50
[346] 50 50 50 50 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 52 52 52
[369] 52 52 52 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 54 54 54
[392] 54 54 54 54 55 55 55 55 55 55 56 56 56 56 56 56 56 56 56 56 56 56 56
[415] 56 56 56 56 56 56 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 58
[438] 58 58 58 58 58 58 59 59 59 59 59 59 59 59 59 59 59 59 59 59 60 60 60
[461] 60 60 60 60 60 60 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 62
[484] 62 62 62 63 63 63 63 63 63 64 64 64 64 64 64 64 64 64 64 64 64 64 64
```

Again, to find the median, we would take the mean of the middle two observations in this sorted data set. That would be the 250th and 251st largest Ages.

```
sort(nh_adults$Age)[250:251]
```

```
[1] 42 42
```

5.2.2 Dealing with Missingness

When calculating a mean, you may be tempted to try something like this...

```
nh_adults %>%
  summarize(mean(Pulse), median(Pulse))
```

```
# A tibble: 1 x 2
  `mean(Pulse)` `median(Pulse)`
      <dbl>         <int>
1         NA           NA
```

This fails because we have some missing values in the Pulse data. We can address this by either omitting the data with missing values before we run the summarize function, or tell the mean and median summary functions to remove missing values².

```
nh_adults %>%
  filter(complete.cases(Pulse)) %>%
  summarize(count = n(), mean(Pulse), median(Pulse))
```

```
# A tibble: 1 x 3
  count `mean(Pulse)` `median(Pulse)`
  <int>      <dbl>         <int>
1   485         73           72
```

Or, we could tell the summary functions themselves to remove NA values.

```
nh_adults %>%
  summarize(mean(Pulse, na.rm=TRUE), median(Pulse, na.rm=TRUE))
```

```
# A tibble: 1 x 2
  `mean(Pulse, na.rm = TRUE)` `median(Pulse, na.rm = TRUE)`
      <dbl>         <int>
1         73           72
```

²We could also use `!is.na` in place of `complete.cases` to accomplish the same thing.

1

73

72

While we eventually discuss the importance of **imputation** when dealing with missing data, this doesn't apply to providing descriptive summaries of actual, observed values.

5.2.3 The Mode of a Quantitative Variable

One other less common measure of the center of a quantitative variable's distribution is its most frequently observed value, referred to as the **mode**. This measure is only appropriate for discrete variables, be they quantitative or categorical. To find the mode, we usually tabulate the data, and then sort by the counts of the numbers of observations.

```
nh_adults %>%
  group_by(Age) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
```

```
# A tibble: 44 x 2
   Age count
<int> <int>
1    56    19
2    50    18
3    28    16
4    37    16
5    42    16
6    49    15
7    24    13
8    27    13
9    39    13
10   46    13
# ... with 34 more rows
```

Note the use of three different “verbs” in our function there - for more explanation of this strategy, visit Grolemund and Wickham (2017).

As an alternative, the **modeest** package's **mfv** function calculates the sample mode (or most frequent value).³

5.3 Measuring the Spread of a Distribution

Statistics is all about variation, so spread or dispersion is an important fundamental concept in statistics. Measures of spread like the inter-quartile range and range (maximum - minimum) can help us understand and compare data sets. If the values in the data are close to the center, the spread will be small. If many of the values in the data are scattered far away from the center, the spread will be large.

5.3.1 The Range and the Interquartile Range (IQR)

The **range** of a quantitative variable is sometimes interpreted as the difference between the maximum and the minimum, even though R presents the actual minimum and maximum values when you ask for a range...

```
nh_adults %>%
  select(Age) %>%
  range()
```

³See the documentation for the **modeest** package's **mlv** function to look at other definitions of the mode.

```
[1] 21 64
```

And, for a variable with missing values, we can use...

```
nh_adults %>%
  select(BMI) %>%
  range(., na.rm=TRUE)
```

```
[1] 17.8 69.0
```

A more interesting and useful statistic is the **inter-quartile range**, or IQR, which is the range of the middle half of the distribution, calculated by subtracting the 25th percentile value from the 75th percentile value.

```
nh_adults %>%
  summarize(IQR(Age), quantile(Age, 0.25), quantile(Age, 0.75))
```

```
# A tibble: 1 x 3
  `IQR(Age)` `quantile(Age, 0.25)` `quantile(Age, 0.75)`
    <dbl>         <dbl>         <dbl>
1      22          31          53
```

We can calculate the range and IQR nicely from the summary information on quantiles, of course:

```
nh_adults %>%
  select(Age, BMI, SBP, DBP, Pulse) %>%
  summary()
```

Age	BMI	SBP	DBP	Pulse
Min. :21.0	Min. :17.8	Min. : 84	Min. : 19.0	Min. : 46
1st Qu.:31.0	1st Qu.:24.2	1st Qu.:109	1st Qu.: 65.0	1st Qu.: 64
Median :42.0	Median :27.7	Median :118	Median : 72.0	Median : 72
Mean :42.1	Mean :28.7	Mean :119	Mean : 72.2	Mean : 73
3rd Qu.:53.0	3rd Qu.:32.1	3rd Qu.:127	3rd Qu.: 79.0	3rd Qu.: 80
Max. :64.0	Max. :69.0	Max. :202	Max. :105.0	Max. :120
	NA's :3	NA's :15	NA's :15	NA's :15

5.3.2 The Variance and the Standard Deviation

The IQR is always a reasonable summary of spread, just as the median is always a reasonable summary of the center of a distribution. Yet, most people are inclined to summarize a batch of data using two numbers: the **mean** and the **standard deviation**. This is really only a sensible thing to do if you are willing to assume the data follow a Normal distribution: a bell-shaped, symmetric distribution without substantial outliers.

But **most data do not (even approximately) follow a Normal distribution**. Summarizing by the median and quartiles (25th and 75th percentiles) is much more robust, explaining R's emphasis on them.

5.3.3 Obtaining the Variance and Standard Deviation in R

Here are the variances of the quantitative variables in the `nh_adults` data. Note the need to include `na.rm = TRUE` to deal with the missing values in some variables.

```
nh_adults %>%
  select(Age, BMI, SBP, DBP, Pulse) %>%
  summarize_all(var, na.rm = TRUE)
```

```
# A tibble: 1 x 5
  Age BMI SBP DBP Pulse
```

```
<dbl> <dbl> <dbl> <dbl> <dbl>
1  157  42.1  234  117  132
```

And here are the standard deviations of those same variables.

```
nh_adults %>%
  select(Age, BMI, SBP, DBP, Pulse) %>%
  summarize_all(sd, na.rm = TRUE)
```

```
# A tibble: 1 x 5
   Age  BMI  SBP  DBP Pulse
<dbl> <dbl> <dbl> <dbl> <dbl>
1  12.5  6.49  15.3  10.8  11.5
```

5.3.4 Defining the Variance and Standard Deviation

Bock, Velleman, and De Veaux (2004) have lots of useful thoughts here, which are lightly edited here.

In thinking about spread, we might consider how far each data value is from the mean. Such a difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel out, leaving an average deviation of zero, so that's not helpful. Instead, we *square* each deviation to obtain non-negative values, and to emphasize larger differences. When we add up these squared deviations and find their mean (almost), this yields the **variance**.

$$\text{Variance} = s^2 = \frac{\Sigma(y - \bar{y})^2}{n - 1}$$

Why almost? It would be the mean of the squared deviations only if we divided the sum by n , but instead we divide by $n - 1$ because doing so produces an estimate of the true (population) variance that is *unbiased*⁴. If you're looking for a more intuitive explanation, this Stack Exchange link awaits your attention.

- To return to the original units of measurement, we take the square root of s^2 , and instead work with s , the **standard deviation**.

$$\text{Standard Deviation} = s = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n - 1}}$$

5.3.5 Empirical Rule Interpretation of the Standard Deviation

For a set of measurements that follow a Normal distribution, the interval:

- Mean \pm Standard Deviation contains approximately 68% of the measurements;
- Mean \pm 2(Standard Deviation) contains approximately 95% of the measurements;
- Mean \pm 3(Standard Deviation) contains approximately all (99.7%) of the measurements.

We often refer to the population or process mean of a distribution with μ and the standard deviation with σ , leading to the Figure below.

But if the data are not from an approximately Normal distribution, then this Empirical Rule is less helpful.

⁴When we divide by $n-1$ as we calculate the sample variance, the average of the sample variances for all possible samples is equal to the population variance. If we instead divided by n , the average sample variance across all possible samples would be a little smaller than the population variance.

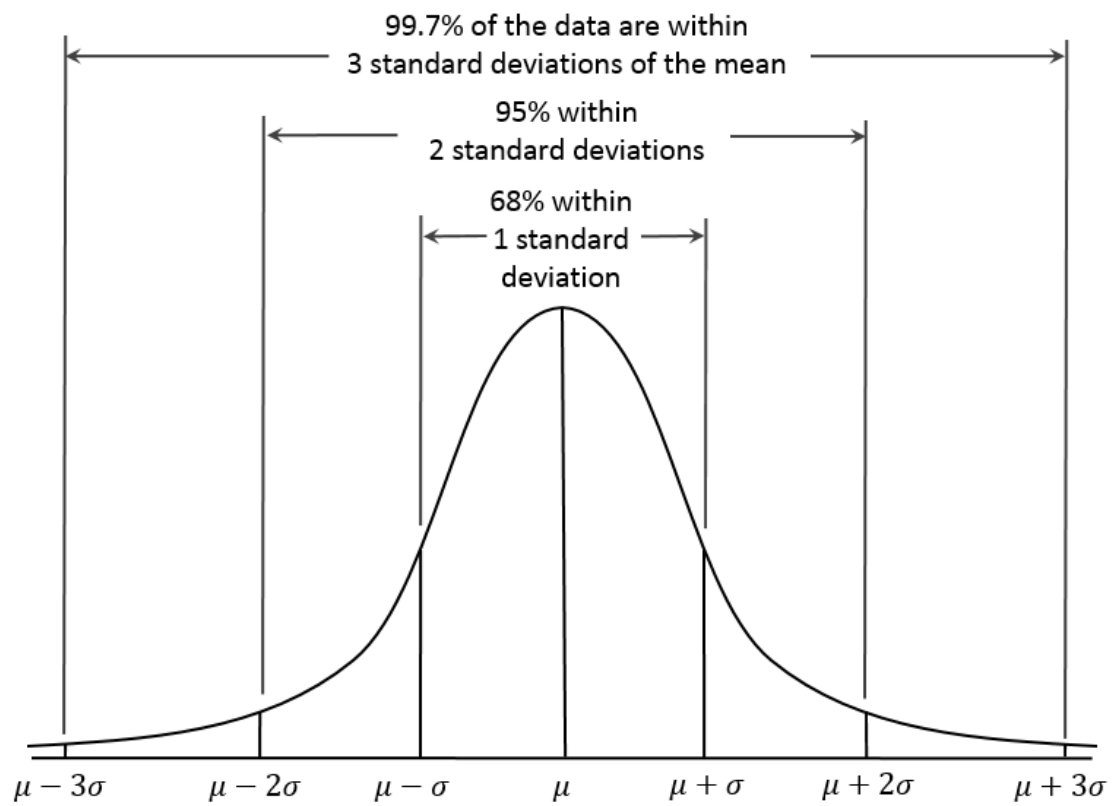


Figure 5.1: The Normal Distribution and the Empirical Rule

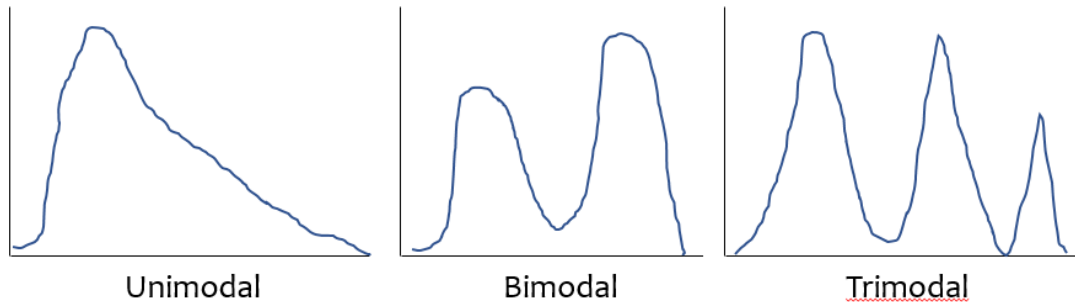


Figure 5.2: Unimodal and Multimodal Sketches

5.3.6 Chebyshev's Inequality: One Interpretation of the Standard Deviation

Chebyshev's Inequality tells us that for any distribution, regardless of its relationship to a Normal distribution, no more than $1/k^2$ of the distribution's values can lie more than k standard deviations from the mean. This implies, for instance, that for **any** distribution, at least 75% of the values must lie within two standard deviations of the mean, and at least 89% must lie within three standard deviations of the mean.

Again, most data sets do not follow a Normal distribution. We'll return to this notion soon. But first, let's try to draw some pictures that let us get a better understanding of the distribution of our data.

5.4 Measuring 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).

5.4.1 Multimodal vs. Unimodal distributions

A unimodal distribution, on some level, is straightforward. It is a distribution with a single mode, or "peak" in the distribution. Such a distribution may be skewed or symmetric, light-tailed or heavy-tailed. We usually describe as multimodal distributions like the two on the right below, which have multiple local maxima, even though they have just a single global maximum peak.

Truly multimodal distributions are usually described that way in terms of shape. For unimodal distributions, skewness and kurtosis become useful ideas.

5.4.2 Skew

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:

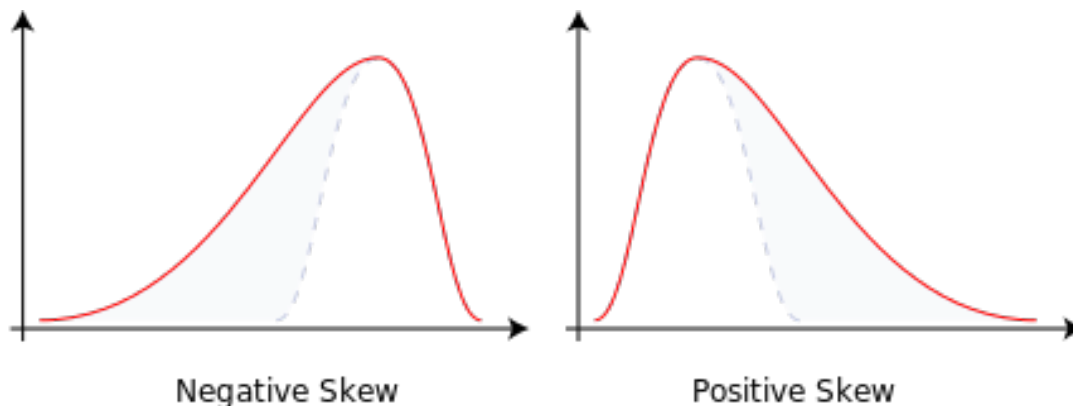


Figure 5.3: Negative (Left) Skew and Positive (Right) Skew

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

- 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.
- 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*.

5.4.3 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. High values of kurtosis measures (and there are several) indicate data which has extreme outliers, or is heavy-tailed.

- A mesokurtic distribution has similar tail behavior to what we would expect from a Normal distribution.
- A leptokurtic distribution is a thinner distribution, with lighter tails (fewer observations far from the center) than we'd expect from a Normal distribution.
- A platykurtic distribution is a flatter distribution, with heavier tails (more observations far from the center) than we'd expect from a Normal distribution.

Graphical tools are in most cases the best way to identify issues related to kurtosis.

5.5 More Detailed Numerical Summaries for Quantitative Variables

5.5.1 favstats in the mosaic package

The `favstats` function adds the standard deviation, and counts of overall and missing observations to our usual `summary` for a continuous variable. Let's look at systolic blood pressure, because we haven't yet.

```
mosaic::favstats(~ SBP, data = nh_adults)
```

```
min  Q1 median  Q3 max mean   sd   n missing
84 109    118 127 202  119 15.3 485      15
```

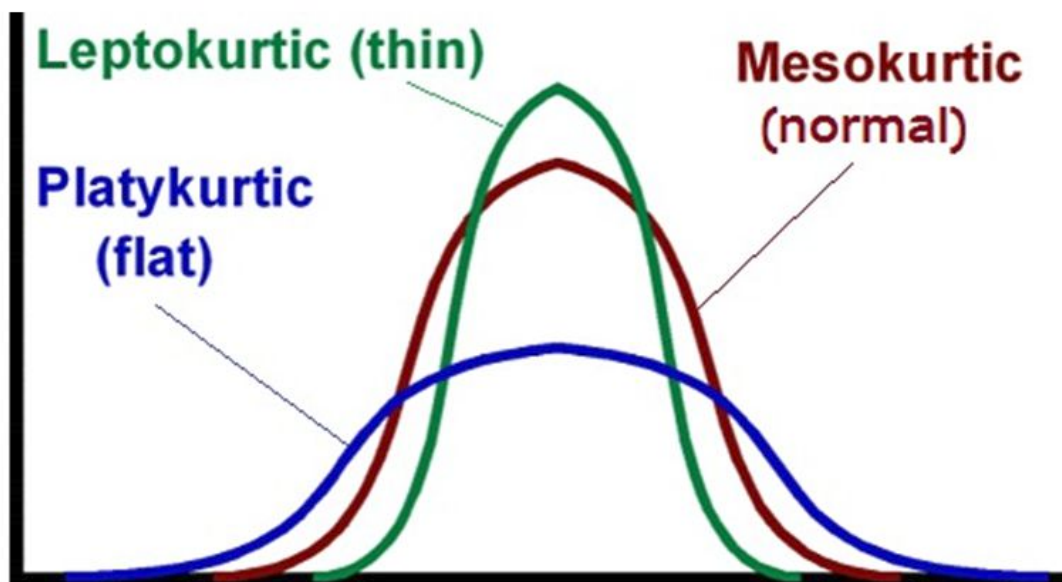


Figure 5.4: The Impact of Kurtosis

We could, of course, duplicate these results with a rather lengthy set of `summarize` pieces...

```
nh_adults %>%
  filter(complete.cases(SBP)) %>%
  summarize(min = min(SBP), Q1 = quantile(SBP, 0.25), median = median(SBP),
            Q3 = quantile(SBP, 0.75), max = max(SBP),
            mean = mean(SBP), sd = sd(SBP), n = n(), missing = sum(is.na(SBP)))
```

```
# A tibble: 1 x 9
  min    Q1 median    Q3  max mean   sd    n missing
<dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1    84   109   118   127   202   119  15.3  485     0
```

The somewhat unusual structure of `favstats` (complete with an easy to forget `~`) is actually helpful. It allows you to look at some interesting grouping approaches, like this:

```
mosaic::favstats(SBP ~ Education, data = nh_adults)
```

```
      Education min    Q1 median    Q3 max mean   sd    n missing
1      8th Grade  95 109   122 126 147   119 14.1  21      3
2 9 - 11th Grade 100 111   115 126 152   118 12.0  57      0
3    High School  89 109   120 129 202   121 19.7  78      3
4   Some College  85 110   118 128 163   119 14.6 149      4
5   College Grad  84 108   116 124 172   117 14.7 180      5
```

Of course, we could accomplish the same comparison with `dplyr` commands, too, but the `favstats` approach has much to offer.

```
nh_adults %>%
  filter(complete.cases(SBP, Education)) %>%
  group_by(Education) %>%
  summarize(min = min(SBP), Q1 = quantile(SBP, 0.25), median = median(SBP),
            Q3 = quantile(SBP, 0.75), max = max(SBP),
```

```
mean = mean(SBP), sd = sd(SBP), n = n(), missing = sum(is.na(SBP)))
```

```
# A tibble: 5 x 10
```

	Education	min	Q1	median	Q3	max	mean	sd	n	missing
	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
1	8th Grade	95	109	122	126	147	119	14.1	21	0
2	9 - 11th Grade	100	111	115	126	152	118	12.0	57	0
3	High School	89	109	120	129	202	121	19.7	78	0
4	Some College	85	110	118	128	163	119	14.6	149	0
5	College Grad	84	108	116	124	172	117	14.7	180	0

5.5.2 describe in the psych package

The `psych` package has a more detailed list of numerical summaries for quantitative variables that looks us look at a group of observations at once.

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

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
Age	1	500	42.1	12.54	42.0	42.1	16.31	21.0	64	43.0	-0.03
BMI	2	497	28.7	6.49	27.7	28.1	5.78	17.8	69	51.2	1.33
SBP	3	485	118.6	15.30	118.0	117.8	13.34	84.0	202	118.0	1.00
DBP	4	485	72.2	10.83	72.0	72.1	10.38	19.0	105	86.0	-0.05
Pulse	5	485	73.0	11.47	72.0	72.5	11.86	46.0	120	74.0	0.46

	kurtosis	se
Age	-1.23	0.56
BMI	4.15	0.29
SBP	3.44	0.69
DBP	1.07	0.49
Pulse	0.45	0.52

The additional statistics presented here are:

- **trimmed** = a trimmed mean (by default in this function, this removes the top and bottom 10% from the data, then computes the mean of the remaining values - the middle 80% of the full data set.)
- **mad** = the median absolute deviation (from the median), which can be used in a manner similar to the standard deviation or IQR to measure spread.
 - If the data are Y_1, Y_2, \dots, Y_n , then the **mad** is defined as $\text{median}(|Y_i - \text{median}(Y_i)|)$.
 - To find the **mad** for a set of numbers, find the median, subtract the median from each value and find the absolute value of that difference, and then find the median of those absolute differences.
 - For non-normal data with a skewed shape but tails well approximated by the Normal, the **mad** is likely to be a better (more robust) estimate of the spread than is the standard deviation.
- a measure of **skew**, which refers to how much asymmetry is present in the shape of the distribution. The measure is not the same as the *nonparametric skew* measure that we will usually prefer. The [Wikipedia page on skewness][https://en.wikipedia.org/wiki/Skewness] is very detailed.
- a measure of **kurtosis**, which refers to how outlier-prone, or heavy-tailed the the shape of the distribution is, mainly as compared to a Normal distribution.
- **se** = the standard error of the sample mean, equal to the sample sd divided by the square root of the sample size.

5.5.3 describe in the Hmisc package

```
Hmisc::describe(nh_adults %>% select(Age, BMI, SBP, DBP, Pulse))
```

```
nh_adults %>% select(Age, BMI, SBP, DBP, Pulse)
```

5	Variables	500	Observations							

Age										
	n	missing	distinct		Info	Mean	Gmd	.05	.10	
	500	0	44		0.999	42.1	14.48	23	25	
	.25	.50	.75		.90	.95				
	31	42	53		59	61				
lowest : 21 22 23 24 25, highest: 60 61 62 63 64										

BMI										
	n	missing	distinct		Info	Mean	Gmd	.05	.10	
	497	3	203		1	28.73	6.947	19.90	22.00	
	.25	.50	.75		.90	.95				
	24.20	27.70	32.10		36.54	40.82				
lowest : 17.8 18.0 18.1 18.2 18.4, highest: 47.6 48.6 48.8 62.8 69.0										

SBP										
	n	missing	distinct		Info	Mean	Gmd	.05	.10	
	485	15	71		0.999	118.6	16.51	96	101	
	.25	.50	.75		.90	.95				
	109	118	127		137	143				
lowest : 84 85 86 89 91, highest: 163 167 168 172 202										

DBP										
	n	missing	distinct		Info	Mean	Gmd	.05	.10	
	485	15	57		0.999	72.25	12.04	56	59	
	.25	.50	.75		.90	.95				
	65	72	79		86	90				
lowest : 19 41 45 47 49, highest: 100 101 102 103 105										

Pulse										
	n	missing	distinct		Info	Mean	Gmd	.05	.10	
	485	15	31		0.997	72.96	12.81	56	60	
	.25	.50	.75		.90	.95				
	64	72	80		88	92				
lowest : 46 48 50 52 54, highest: 98 100 102 108 120										

The `Hmisc` package's version of `describe` for a distribution of data presents three new ideas, in addition to a more comprehensive list of quartiles (the 5th, 10th, 25th, 50th, 75th, 90th and 95th are shown) and the lowest and highest few observations. These are:

- **distinct** - the number of different values observed in the data.
- **Info** - a measure of how “continuous” the variable is, related to how many “ties” there are in the data, with Info taking a higher value (closer to its maximum of one) if the data are more continuous.
- **Gmd** - the Gini mean difference - a robust measure of spread that is calculated as the mean absolute difference between any pairs of observations. Larger values of Gmd indicate more spread-out distributions.

Chapter 6

Summarizing Categorical Variables

Summarizing categorical variables numerically is mostly about building tables, and calculating percentages or proportions. We'll save our discussion of modeling categorical data for later. Recall that in the `nh_adults` data set we built previously, we had the following categorical variables. The number of levels indicates the number of possible categories for each categorical variable.

Variable	Description	Levels	Type
Sex	sex of subject	2	binary
Race	subject's race	6	nominal
Education	subject's educational level	5	ordinal
PhysActive	Participates in sports?	2	binary
Smoke100	Smoked 100+ cigarettes?	2	binary
SleepTrouble	Trouble sleeping?	2	binary
HealthGen	Self-report health	5	ordinal

6.1 The `summary` function for Categorical data

When R recognizes a variable as categorical, it stores it as a *factor*. Such variables get special treatment from the `summary` function, in particular a table of available values (so long as there aren't too many.)

```
nh_adults %>%  
  select(Sex, Race, Education, PhysActive, Smoke100, SleepTrouble, HealthGen) %>%  
  summary()
```

```
      Sex      Race      Education  PhysActive Smoke100  
female:253 Asian   : 29 8th Grade    : 24 No :225 No :289  
male :247 Black   : 57 9 - 11th Grade: 57 Yes:275 Yes:211  
      Hispanic: 39 High School : 81  
      Mexican : 43 Some College :153  
      White   :322 College Grad :185  
      Other   : 10  
SleepTrouble HealthGen  
No :362 Excellent: 51  
Yes:138 Vgood    :153  
      Good    :172  
      Fair    : 71  
      Poor    : 7
```

NA's : 46

6.2 Tables to describe One Categorical Variable

Suppose we build a table to describe the `HealthGen` distribution.

```
nh_adults %>%
  select(HealthGen) %>%
  table(., useNA = "ifany")
```

```
.
Excellent    Vgood      Good      Fair      Poor      <NA>
      51      153      172      71       7      46
```

What if we want to add a total count?

```
nh_adults %>%
  select(HealthGen) %>%
  table(., useNA = "ifany") %>%
  addmargins()
```

```
.
Excellent    Vgood      Good      Fair      Poor      <NA>      Sum
      51      153      172      71       7      46      500
```

What if we want to leave out the missing responses?

```
nh_adults %>%
  select(HealthGen) %>%
  table(., useNA = "no") %>%
  addmargins()
```

```
.
Excellent    Vgood      Good      Fair      Poor      Sum
      51      153      172      71       7      454
```

Let's put the missing values back in, but now calculate proportions instead. Since the total will just be 1.0, we'll leave that out.

```
nh_adults %>%
  select(HealthGen) %>%
  table(., useNA = "ifany") %>%
  prop.table()
```

```
.
Excellent    Vgood      Good      Fair      Poor      <NA>
    0.102    0.306    0.344    0.142    0.014    0.092
```

Now, we'll calculate percentages by multiplying the proportions by 100.

```
nh_adults %>%
  select(HealthGen) %>%
  table(., useNA = "ifany") %>%
  prop.table() %>%
  "*" (100)
```

```
.
Excellent    Vgood      Good      Fair      Poor      <NA>
```


10.2 30.6 34.4 14.2 1.4 9.2

6.3 The Mode of a Categorical Variable

A common measure applied to a categorical variable is to identify the mode, the most frequently observed value. To find the mode for variables with lots of categories (so that the `summary` may not be sufficient), we usually tabulate the data, and then sort by the counts of the numbers of observations, as we did with discrete quantitative variables.

```
nh_adults %>%
  group_by(HealthGen) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
```

```
# A tibble: 6 x 2
  HealthGen count
  <fctr> <int>
1      Good  172
2     Vgood  153
3     Fair   71
4 Excellent  51
5        NA  46
6     Poor   7
```

6.4 describe in the Hmisc package

```
Hmisc::describe(nh_adults %>%
  select(Sex, Race, Education, PhysActive,
         Smoke100, SleepTrouble, HealthGen))
```

```
nh_adults %>% select(Sex, Race, Education, PhysActive, Smoke100, SleepTrouble, HealthGen)
```

```
7 Variables      500 Observations
-----
Sex
  n missing distinct
500      0         2

Value      female  male
Frequency   253   247
Proportion  0.506  0.494
-----
Race
  n missing distinct
500      0         6

Value      Asian  Black Hispanic  Mexican  White  Other
Frequency   29   57      39      43   322   10
Proportion  0.058  0.114   0.078   0.086  0.644  0.020
-----
Education
```

```

      n missing distinct
500      0          5

Value      8th Grade 9 - 11th Grade      High School      Some College
Frequency      24          57          81          153
Proportion      0.048      0.114      0.162      0.306

Value      College Grad
Frequency      185
Proportion      0.370
-----

PhysActive
      n missing distinct
500      0          2

Value      No  Yes
Frequency  225 275
Proportion 0.45 0.55
-----

Smoke100
      n missing distinct
500      0          2

Value      No  Yes
Frequency  289 211
Proportion 0.578 0.422
-----

SleepTrouble
      n missing distinct
500      0          2

Value      No  Yes
Frequency  362 138
Proportion 0.724 0.276
-----

HealthGen
      n missing distinct
454      46          5

Value      Excellent      Vgood      Good      Fair      Poor
Frequency      51      153      172      71      7
Proportion      0.112      0.337      0.379      0.156      0.015
-----

```

6.5 Cross-Tabulations

It is very common for us to want to describe the association of one categorical variable with another. For instance, is there a relationship between Education and SleepTrouble in these data?

```

nh_adults %>%
  select(Education, SleepTrouble) %>%
  table() %>%
  addmargins()

```

Education	SleepTrouble		
	No	Yes	Sum
8th Grade	15	9	24
9 - 11th Grade	40	17	57
High School	67	14	81
Some College	107	46	153
College Grad	133	52	185
Sum	362	138	500

To get row percentages, we can use:

```
nh_adults %>%
  select(Education, SleepTrouble) %>%
  table() %>%
  prop.table(., 1) %>%
  "*" (100)
```

Education	SleepTrouble	
	No	Yes
8th Grade	62.5	37.5
9 - 11th Grade	70.2	29.8
High School	82.7	17.3
Some College	69.9	30.1
College Grad	71.9	28.1

For column percentages, we use 2 instead of 1 in the `prop.table` function. Here, we'll also round off to two decimal places:

```
nh_adults %>%
  select(Education, SleepTrouble) %>%
  table() %>%
  prop.table(., 2) %>%
  "*" (100) %>%
  round(., 2)
```

Education	SleepTrouble	
	No	Yes
8th Grade	4.14	6.52
9 - 11th Grade	11.05	12.32
High School	18.51	10.14
Some College	29.56	33.33
College Grad	36.74	37.68

Here's another approach, to look at the cross-classification of Race and HealthGen:

```
xtabs(~ Race + HealthGen, data = nh_adults)
```

Race	HealthGen				
	Excellent	Vgood	Good	Fair	Poor
Asian	4	7	9	2	1
Black	7	11	16	11	2
Hispanic	1	9	18	8	0
Mexican	5	6	12	16	1
White	34	115	115	32	3
Other	0	5	2	2	0

6.5.1 Cross-Classifying Three Categorical Variables

Suppose we are interested in `Smoke100` and its relationship to `PhysActive` and `SleepTrouble`.

```
xtabs(~ Smoke100 + PhysActive + SleepTrouble, data = nh_adults)
```

```
, , SleepTrouble = No
```

	PhysActive	
Smoke100	No	Yes
No	99	135
Yes	62	66

```
, , SleepTrouble = Yes
```

	PhysActive	
Smoke100	No	Yes
No	26	29
Yes	38	45

We can also build a **flat** version of this table, as follows:

```
ftable(Smoke100 ~ PhysActive + SleepTrouble, data = nh_adults)
```

		Smoke100	
		No	Yes
PhysActive	SleepTrouble		
	No	99	62
Yes	Yes	26	38
	No	135	66
Yes	Yes	29	45

And we can do this with `dplyr` functions, as well, for example...

```
nh_adults %>%
  select(Smoke100, PhysActive, SleepTrouble) %>%
  table()
```

```
, , SleepTrouble = No
```

	PhysActive	
Smoke100	No	Yes
No	99	135
Yes	62	66

```
, , SleepTrouble = Yes
```

	PhysActive	
Smoke100	No	Yes
No	26	29
Yes	38	45

Chapter 7

The National Youth Fitness Survey (nyfs1)

The `nyfs1.csv` data file comes from the 2012 National Youth Fitness Survey.

The NHANES National Youth Fitness Survey (NNYFS) was conducted in 2012 to collect data on physical activity and fitness levels in order to provide an evaluation of the health and fitness of children in the U.S. ages 3 to 15. The NNYFS collected data on physical activity and fitness levels of our youth through interviews and fitness tests.

In the `nyfs1.csv` data file, I'm only providing a tiny portion of the available information. More on the NNYFS (including information I'm not using) is available at the links below.

- Demographic Information including a complete description of all available variables.
- Body Measures, part of the general Examination data with complete variable descriptions

What I did was merge a few elements from the available demographic information with some elements from the body measures data, reformulated and simplified some variables, and restricted the sample to kids who had a complete set of body measure examinations.

7.1 Looking over the Data Set

To start with, I'll take a look at the `nyfs1` data. One approach is to simply get the size of the set and the names of the available data elements.

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

## next we'll turn that data frame into a more useful tibble
nyfs1 <- tbl_df(nyfs1)

## size of the data frame
dim(nyfs1)
```

```
[1] 1416    7
```

There are 1416 rows (subjects) and 7 columns (variables), by which I mean that there are 1416 kids in the `nyfs1` data frame, and we have 7 pieces of information on each subject.

So, what do we have, exactly?

```
nyfs1 # this is a tibble, has some nice features in a print-out like this
```

```
# A tibble: 1,416 x 7
  subject.id    sex age.exam  bmi      bmi.cat waist.circ
    <int> <fctr>   <int> <dbl>      <fctr>      <dbl>
1     71918 Female      8  22.3      4 Obese      71.9
2     71919 Female     14  19.8 2 Normal weight  79.4
3     71921  Male      3  15.2 2 Normal weight  46.8
4     71922  Male     12  25.9      4 Obese      90.0
5     71923  Male     12  22.5      3 Overweight  72.3
6     71924 Female      8  14.4 2 Normal weight  56.1
7     71925  Male      7  15.9 2 Normal weight  54.5
8     71926  Male      8  17.0 2 Normal weight  59.7
9     71927  Male      3  15.8 2 Normal weight  49.9
10    71928 Female      9  16.0 2 Normal weight  59.9
# ... with 1,406 more rows, and 1 more variables: triceps.skinfold <dbl>
```

Tibbles are a modern reimaging of the main way in which people have stored data in R, called a data frame. Tibbles were developed to keep what time has proven to be effective, and throwing out what is not. We can obtain the structure of the tibble from the `str` function.

```
str(nyfs1)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':  1416 obs. of  7 variables:
 $ subject.id      : int  71918 71919 71921 71922 71923 71924 71925 71926 71927 71928 ...
 $ sex             : Factor w/ 2 levels "Female","Male": 1 1 2 2 2 1 2 2 2 1 ...
 $ 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 17 15.8 16 ...
 $ bmi.cat         : Factor w/ 4 levels "1 Underweight",...: 4 2 2 4 3 2 2 2 2 2 ...
 $ waist.circ      : num   71.9 79.4 46.8 90 72.3 56.1 54.5 59.7 49.9 59.9 ...
 $ triceps.skinfold: num   19.9 15 8.6 22.8 20.5 12.9 6.9 8.8 10.8 13.2 ...
```

7.1.1 subject.id

The first variable, `subject.id` is listed by R as an `int` variable, for integer, which means it consists of whole numbers. However, the information provided by this variable is minimal. This is just an identifying code attributable to a given subject of the survey. This is *nominal* data, which will be of little interest down the line. On some occasions, as in this case, the ID numbers are sequential, in the sense that subject 71919 was included in the data base after subject 71918, but this fact isn't particularly interesting here, because the protocol remained unchanged throughout the study.

7.1.2 sex

The second variable, `sex` is listed as a factor (R uses **factor** to refer to categorical, especially non-numeric information) with two levels, *Female* and *Male*. You'll note that what is stored in the structure is a series of 1 (referring to the first level - Female) and 2 (Male) values. If we want to know how many people fall in each category, we can build a little table.

```
dplyr::select(nyfs1, sex) %>%
  table()
```

```
.
Female  Male
   707    709
```

```
dplyr::select(nyfs1, sex) %>%
  table() %>%
  addmargins() ## add marginal totals
```

```
.
Female   Male    Sum
   707    709   1416
```

```
dplyr::select(nyfs1, sex) %>%
  table() %>%
  prop.table() ## look at the proportions instead
```

```
.
Female   Male
  0.499  0.501
```

Obviously, we don't actually need more than a couple of decimal places for any real purpose.

7.1.3 age.exam

The third variable, `age.exam` is the age of the child at the time of the examination, measured in years. Note that age is a continuous concept, but the measure used here (number of full years alive) is a common discrete approach to measurement. Age, of course, has a meaningful zero point, so this can be thought of as a ratio variable; a child who is 6 is half as old as one who is 12. We can get a table of the observed values.

```
dplyr::select(nyfs1, age.exam) %>%
  table() %>%
  addmargins()
```

```
.
  3    4    5    6    7    8    9   10   11   12   13   14   15   16 Sum
97  111  119  129  123  120   90  109  102  108  113  104   85    6 1416
```

Note that some of the children apparently turned 16 between the time they were initially screened (when they were required to be between 3 and 15 years of age) and the time of the examination. The `sum` listed here is just the total count of all subjects. Since this is a meaningful quantitative variable, we may be interested in a more descriptive summary.

```
dplyr::select(nyfs1, age.exam) %>%
  summary()
```

```
age.exam
Min.   : 3.00
1st Qu.: 6.00
Median : 9.00
Mean    : 8.86
3rd Qu.:12.00
Max.    :16.00
```

These six numbers provide a nice, if incomplete, look at the ages.

- **Min.** = the minimum, or youngest age at the examination was 3 years old.
- **1st Qu.** = the first quartile (25th percentile) of the ages was 6. This means that 25 percent of the subjects were age 6 or less.
- **Median** = the second quartile (50th percentile) of the ages was 9. This is often used to describe the center of the data. Half of the subjects were age 9 or less.
- **3rd Qu.** = the third quartile (75th percentile) of the ages was 12

- Max. = the maximum, or oldest age at the examination was 16 years.

7.1.4 bmi

The fourth variable, `bmi`, is the body-mass index of the child. The BMI is a person's weight in kilograms divided by his or her height in meters squared. Symbolically, $BMI = \text{weight in kg} / (\text{height in m})^2$. This is a continuous concept, measured to as many decimal places as you like, and it has a meaningful zero point, so it's a ratio variable.

```
dplyr::select(nyfs1, bmi) %>%
  summary()
```

```
      bmi
Min.   :11.9
1st Qu.:15.8
Median :17.7
Mean   :18.8
3rd Qu.:20.9
Max.   :38.8
```

Why would a table of these BMI values not be a great idea, for these data? A hint is that R represents this variable as `num` or numeric in its depiction of the data structure, and this implies that R has some decimal values stored.

```
dplyr::select(nyfs1, bmi) %>%
  table()
```

```
.
11.9 12.6 12.7 12.9 13 13.1 13.2 13.3 13.4 13.5 13.6 13.7 13.8 13.9 14
1 1 1 1 2 2 1 1 3 4 5 4 5 11 7
14.1 14.2 14.3 14.4 14.5 14.6 14.7 14.8 14.9 15 15.1 15.2 15.3 15.4 15.5
12 9 11 11 11 9 17 20 23 13 14 18 27 24 32
15.6 15.7 15.8 15.9 16 16.1 16.2 16.3 16.4 16.5 16.6 16.7 16.8 16.9 17
18 20 21 30 27 15 18 30 12 25 20 22 13 21 23
17.1 17.2 17.3 17.4 17.5 17.6 17.7 17.8 17.9 18 18.1 18.2 18.3 18.4 18.5
14 20 14 10 19 13 17 18 14 17 13 10 9 8 15
18.6 18.7 18.8 18.9 19 19.1 19.2 19.3 19.4 19.5 19.6 19.7 19.8 19.9 20
10 17 10 11 4 13 15 12 8 25 6 6 16 8 13
20.1 20.2 20.3 20.4 20.5 20.6 20.7 20.8 20.9 21 21.1 21.2 21.3 21.4 21.5
9 7 12 7 3 9 5 6 11 7 5 6 8 9 8
21.6 21.7 21.8 21.9 22 22.1 22.2 22.3 22.4 22.5 22.6 22.7 22.8 22.9 23
6 7 16 6 13 7 7 8 6 4 4 5 2 10 7
23.1 23.2 23.3 23.4 23.5 23.6 23.7 23.8 23.9 24 24.1 24.2 24.3 24.4 24.5
3 8 3 5 4 3 2 4 4 5 1 4 3 5 5
24.6 24.7 24.8 24.9 25 25.1 25.2 25.3 25.4 25.5 25.6 25.7 25.8 25.9 26
4 3 6 4 3 2 4 2 3 3 4 5 3 3 2
26.1 26.2 26.3 26.4 26.5 26.6 26.7 26.8 27 27.2 27.3 27.4 27.5 27.6 27.7
1 4 2 1 2 1 2 1 2 1 2 2 1 2 1
27.9 28.1 28.2 28.4 28.5 28.6 28.7 28.8 28.9 29 29.2 29.5 29.7 29.8 30.1
2 2 2 1 1 2 2 1 3 1 3 1 2 3 2
30.2 30.4 30.5 30.7 30.8 30.9 31.1 31.3 31.4 31.5 31.7 31.8 32 32.2 32.4
4 1 2 1 1 1 1 1 2 1 1 2 2 1 1
32.6 32.9 33.2 33.5 34 34.4 34.6 34.7 35.9 37 38.8
1 1 1 1 1 1 1 1 1 1 1
```


7.1.5 bmi.cat

Our next variable, `bmi.cat`, is a four-category ordinal variable, which divides the sample according to BMI into four groups. The BMI categories use sex-specific 2000 BMI-for-age (in months) growth charts prepared by the Centers for Disease Control for the US. We can get the breakdown from a table of the variable's values.

```
dplyr::select(nyfs1, bmi.cat) %>%
  table() %>%
  addmargins()
```

1 Underweight	2 Normal weight	3 Overweight	4 Obese
42	926	237	211
Sum			
1416			

In terms of percentiles by age and sex from the growth charts, the meanings of the categories are:

- Underweight (BMI < 5th percentile)
- Normal weight (BMI 5th to < 85th percentile)
- Overweight (BMI 85th to < 95th percentile)
- Obese (BMI ≥ 95th percentile)

Note how I've used labels in the `bmi.cat` variable that include a number at the start so that the table results are sorted in a rational way. R sorts tables alphabetically, in general.

7.1.6 waist.circ

The sixth variable is `waist.circ`, which is the circumference of the child's waist, in centimeters. Again, this is a numeric variable, so perhaps we'll stick to the simple summary, rather than obtaining a table of observed values.

```
dplyr::select(nyfs1, waist.circ) %>%
  summary()
```

```
waist.circ
Min.   : 42.5
1st Qu.: 55.0
Median : 63.0
Mean    : 65.3
3rd Qu.: 72.9
Max.    :112.4
```

7.1.7 triceps.skinfold

The seventh and final variable is `triceps.skinfold`, which is measured in millimeters. This is one of several common locations used for the assessment of body fat using skinfold calipers, and is a frequent part of growth assessments in children. Again, this is a numeric variable according to R.

```
dplyr::select(nyfs1, triceps.skinfold) %>%
  summary()
```

```
triceps.skinfold
Min.   : 4.0
1st Qu.: 9.0
Median :11.8
```

```

Mean    :13.4
3rd Qu.:16.6
Max.    :38.2

```

7.2 Summarizing the Data Set

The `summary` function can be applied to the whole tibble. For numerical and integer variables, this function produces the five number summary, plus the mean. For categorical (factor) variables, it lists the count for each category.

```
summary(nyfs1)
```

```

      subject.id      sex      age.exam      bmi
Min.   :71918  Female:707   Min.    : 3.00   Min.    :11.9
1st Qu.:72313   Male  :709   1st Qu.: 6.00   1st Qu.:15.8
Median :72698                      Median : 9.00   Median :17.7
Mean   :72703                      Mean   : 8.86   Mean   :18.8
3rd Qu.:73096                      3rd Qu.:12.00   3rd Qu.:20.9
Max.   :73492                      Max.   :16.00   Max.   :38.8

      bmi.cat      waist.circ      triceps.skinfold
1 Underweight   : 42   Min.    : 42.5   Min.    : 4.0
2 Normal weight:926   1st Qu.: 55.0   1st Qu.: 9.0
3 Overweight   :237   Median : 63.0   Median :11.8
4 Obese        :211   Mean    : 65.3   Mean    :13.4
                      3rd Qu.: 72.9   3rd Qu.:16.6
                      Max.    :112.4   Max.    :38.2

```

7.2.1 The Five Number Summary, Quantiles and IQR

The **five number summary** is most famous when used to form a box plot - it's the minimum, 25th percentile, median, 75th percentile and maximum. Our usual `summary` adds the mean.

```

nyfs1 %>%
  select(bmi) %>%
  summary()

```

```

      bmi
Min.   :11.9
1st Qu.:15.8
Median :17.7
Mean   :18.8
3rd Qu.:20.9
Max.   :38.8

```

As an alternative, we can use the `$` notation to indicate the variable we wish to study inside a data set, and we can use the `fivenum` function to get the five numbers used in developing a box plot.

```
fivenum(nyfs1$bmi)
```

```
[1] 11.9 15.8 17.7 20.9 38.8
```

- As mentioned in 5.3.1, the **inter-quartile range**, or IQR, is sometimes used as a competitor for the standard deviation. It's the difference between the 75th percentile and the 25th percentile. The 25th percentile, median, and 75th percentile are referred to as the quartiles of the data set, because, together, they split the data into quarters.

```
IQR(nyfs1$bmi)
```

```
[1] 5.1
```

We can obtain **quantiles** (percentiles) as we like - here, I'm asking for the 1st and 99th

```
quantile(nyfs1$bmi, probs=c(0.01, 0.99))
```

```
 1% 99%
13.5 32.0
```

7.3 Additional Summaries from favstats

If we're focusing on a single variable, the `favstats` function in the `mosaic` package can be very helpful. Rather than calling up the entire `mosaic` library here, I'll just specify the function within the library.

```
mosaic::favstats(nyfs1$bmi)
```

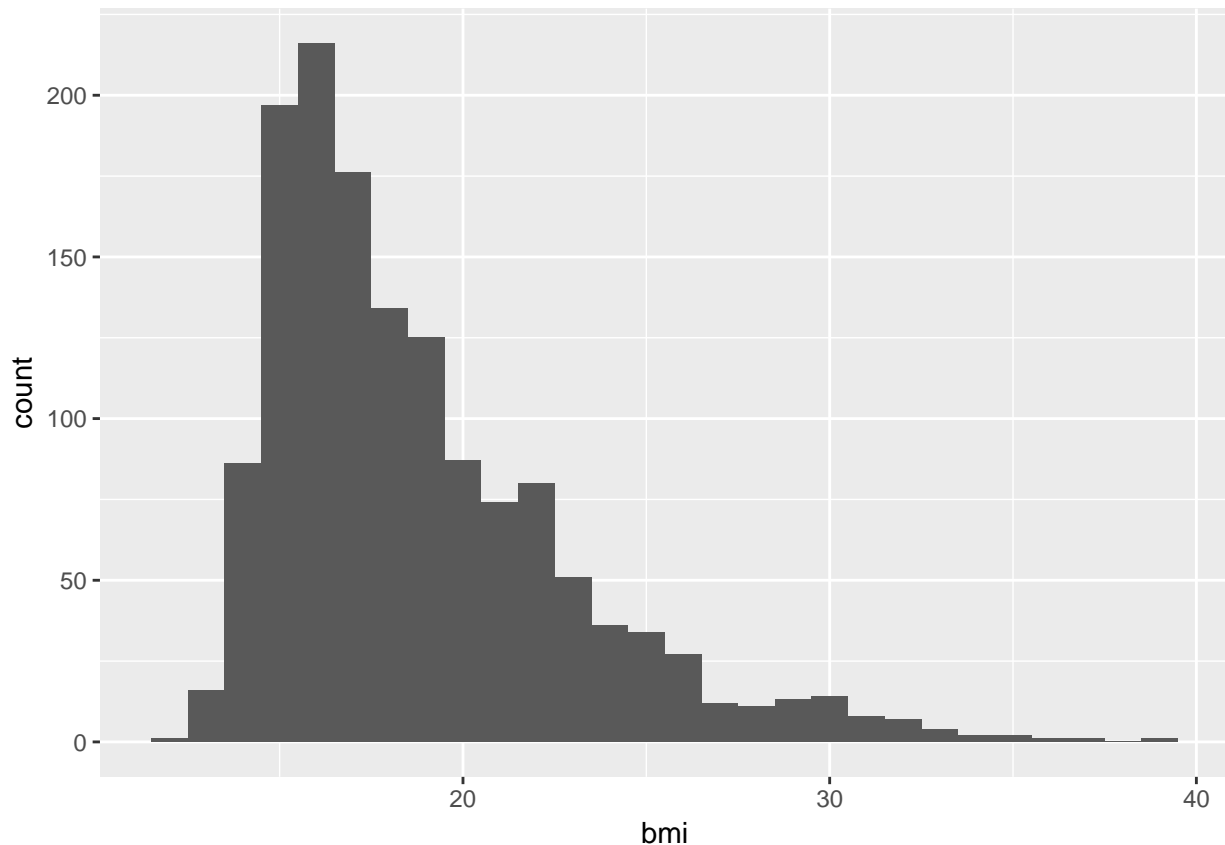
```
   min   Q1 median   Q3  max mean   sd    n missing
11.9 15.8   17.7 20.9 38.8 18.8 4.08 1416         0
```

This adds three useful results to the base summary - the standard deviation, the sample size and the number of missing observations.

7.4 The Histogram

As we saw in 3, obtaining a basic **histogram** of, for example, the BMIs in the `nyfs1` data is pretty straightforward.

```
ggplot(data = nyfs1, aes(x = bmi)) +
  geom_histogram(binwidth = 1)
```



7.4.1 Freedman-Diaconis Rule to select bin width

If we like, we can suggest a particular number of cells for the histogram, instead of accepting the defaults. In this case, we have $n = 1416$ observations. The **Freedman-Diaconis rule** can be helpful here. That rule suggests that we set the bin-width to

$$h = \frac{2 * IQR}{n^{1/3}}$$

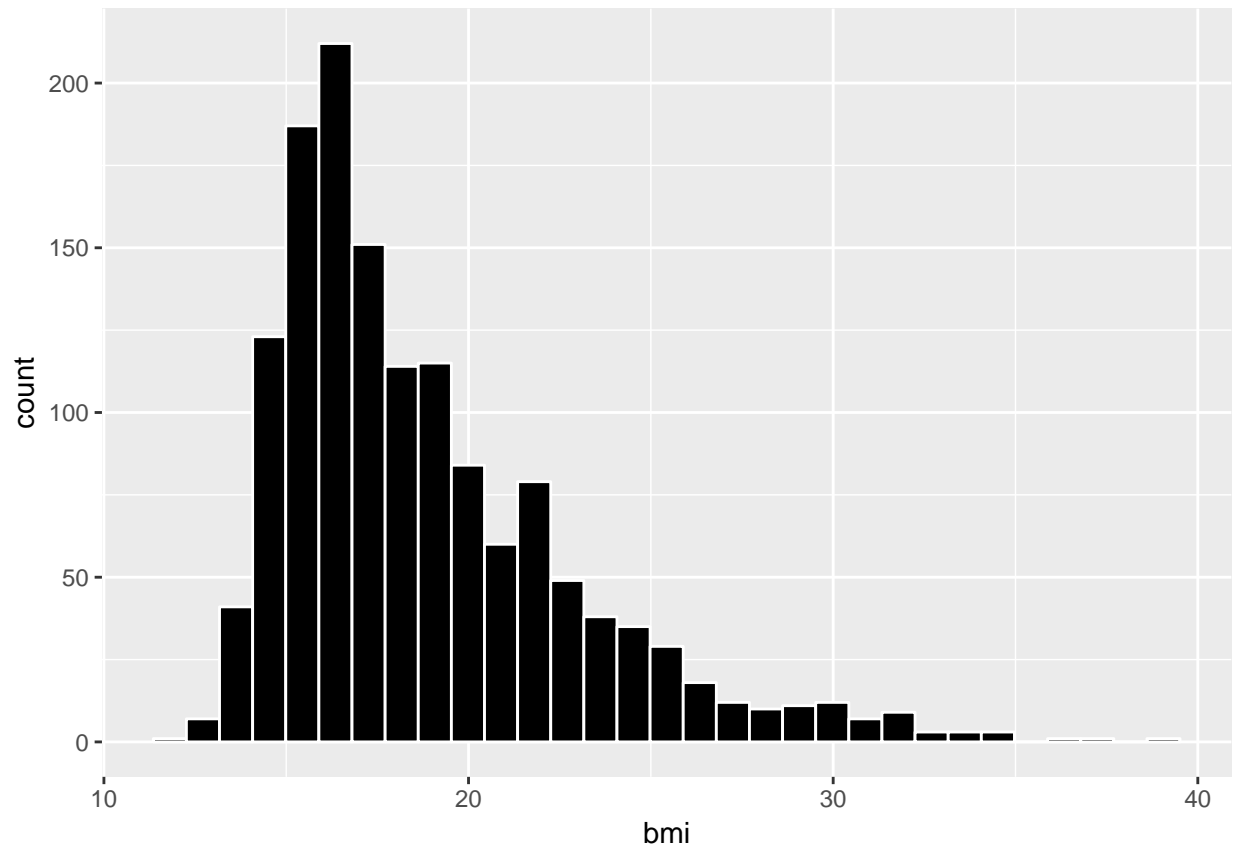
so that the number of bins is equal to the range of the data set (maximum - minimum) divided by h .

For the `bmi` data in the `nyfs1` tibble, we have

- IQR of 5.1, $n = 1416$ and range = 26.9
- Thus, by the Freedman-Diaconis rule, the optimal binwidth h is 0.908, or, realistically, 1.
- And so the number of bins would be 29.615, or, realistically 30.

Here, we'll draw the graph again, using the Freedman-Diaconis rule to identify the number of bins, and also play around a bit with the fill and color of the bars.

```
bw <- 2 * IQR(nyfs1$bmi) / length(nyfs1$bmi)^(1/3)
ggplot(data = nyfs1, aes(x = bmi)) +
  geom_histogram(binwidth=bw, color = "white", fill = "black")
```

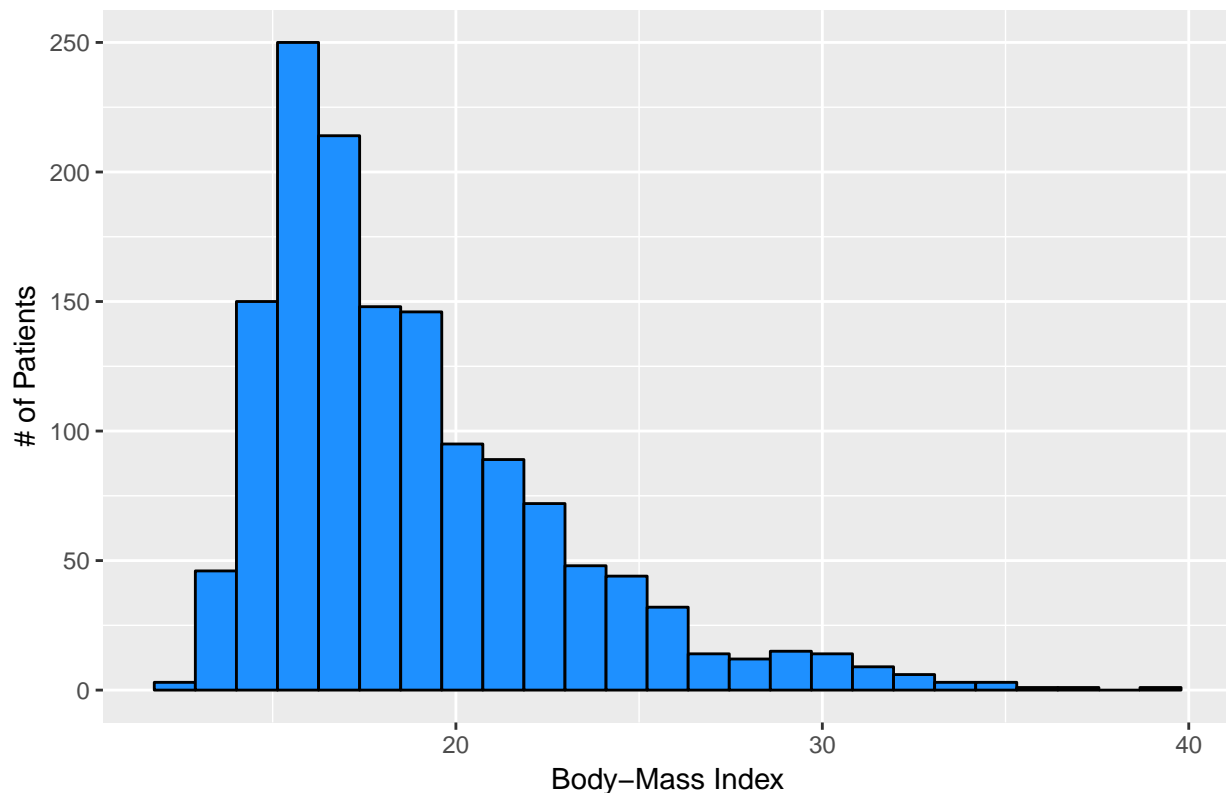


This is a nice start, but it is by no means a finished graph.

Let's improve the axis labels, add a title, and fill in the bars with a distinctive blue and use a black outline around each bar. I'll just use 25 bars, because I like how that looks in this case, and optimizing the number of bins is rarely important.

```
ggplot(data = nyfs1, aes(x = bmi)) +  
  geom_histogram(bins=25, color = "black", fill = "dodgerblue") +  
  labs(title = "Histogram of Body-Mass Index Results in the nyfs1 data",  
        x = "Body-Mass Index", y = "# of Patients")
```

Histogram of Body–Mass Index Results in the nyfs1 data



7.5 A Note on Colors

The simplest way to specify a color is with its name, enclosed in parentheses. My favorite list of R colors is <http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>. In a pinch, you can find it by googling **Colors in R**. You can also type `colors()` in the R console to obtain a list of the names of the same 657 colors.

When using colors to make comparisons, you may be interested in using a scale that has some nice properties. I suggest the `viridis` package to help with this work. The `viridis` package vignette describes four color scales (`viridis`, `magma`, `plasma` and `inferno`) that are designed to be colorful, robust to colorblindness and gray scale printing, and perceptually uniform, which means (as the package authors describe it) that values close to each other have similar-appearing colors and values far away from each other have more different-appearing colors, consistently across the range of values.

7.6 The Stem-and-Leaf

We might consider a **stem-and-leaf display** (a John Tukey invention) to show the actual data values while retaining the shape of a histogram. The `scale` parameter can help expand the size of the diagram, so you can see more of the values. Stem and leaf displays are usually used for relatively small samples, perhaps with 10-200 observations, so we'll first take a sample of 150 of the BMI values from the complete set gathered in the `nyfs1` tibble.

```
set.seed(431) # set a seed for the random sampling so we can replicate the results
```

```
sampleA <- sample_n(nyfs1, 150, replace = FALSE) # draw a sample of 150 unique rows from nyfs1
stem(sampleA$bmi) # build a stem-and-leaf for those 150 sampled BMI values
```

The decimal point is at the |

```
13 | 129
14 | 001224455566778889
15 | 02344455567789999
16 | 0000112233345667779
17 | 001225556677789
18 | 0111346677888899
19 | 111224555578889
20 | 0113334456899
21 | 014568
22 | 11349
23 | 012479
24 | 478
25 | 05669
26 | 03
27 | 05
28 |
29 |
30 | 27
31 |
32 | 4
33 |
34 | 67
```

We can see that the minimum BMI value in this small sample is 13.1 and the maximum BMI value is 34.7.

Here's a summary of all variables for these 150 observations.

```
summary(sampleA)
```

subject.id	sex	age.exam	bmi
Min. :71935	Female:68	Min. : 3.00	Min. :13.1
1st Qu.:72302	Male :82	1st Qu.: 6.00	1st Qu.:15.9
Median :72688		Median :10.00	Median :18.1
Mean :72679		Mean : 9.45	Mean :19.0
3rd Qu.:73080		3rd Qu.:13.00	3rd Qu.:20.6
Max. :73490		Max. :15.00	Max. :34.7

bmi.cat	waist.circ	triceps.skinfold
1 Underweight : 4	Min. : 45.6	Min. : 5.6
2 Normal weight:103	1st Qu.: 55.4	1st Qu.: 9.2
3 Overweight : 21	Median : 64.7	Median :12.2
4 Obese : 22	Mean : 66.5	Mean :13.6
	3rd Qu.: 72.8	3rd Qu.:16.6
	Max. :108.4	Max. :34.8

If we really wanted to, we could obtain a stem-and-leaf of all of the BMI values in the entire `nyfs1` data. The `scale` parameter lets us see some more of the values.

```
stem(nyfs1$bmi, scale = 2)
```



```

28    22 | 11349
23    23 | 012479
17    24 | 478
14    25 | 05669
9     26 | 03
7     27 | 05
HI: 30.2 30.7 32.4 34.6 34.7

```

We can also produce back-to-back stem and leaf plots to compare, for instance, body-mass index by sex.

```

samp.F <- filter(sampleA, sex=="Female")
samp.M <- filter(sampleA, sex=="Male")

aplpack::stem.leaf.backback(samp.F$bmi, samp.M$bmi)

```

```

-----
1 | 2: represents 1.2, leaf unit: 0.1
    samp.F$bmi      samp.M$bmi
-----
3          921| 13 |
16   9876654422100| 14 |55788          5
21          98444| 15 |023555677999    17
33   776653210000| 16 |1233479        24
(2)          91| 17 |0022555667778    37
33          9887410| 18 |113667889    (9)
26          9888555411| 19 |12257        36
16          9954310| 20 |133468        31
9              0| 21 |14568        25
              | 22 |11349        20
8          910| 23 |247        15
5              8| 24 |47        12
4          95| 25 |066        10
              | 26 |03         7
              | 27 |05         5
              | 28 |
-----
HI: 30.2 32.4          HI: 30.7 34.6 34.7
n:          68          82
-----

```

7.7 The Dot Plot to display a distribution

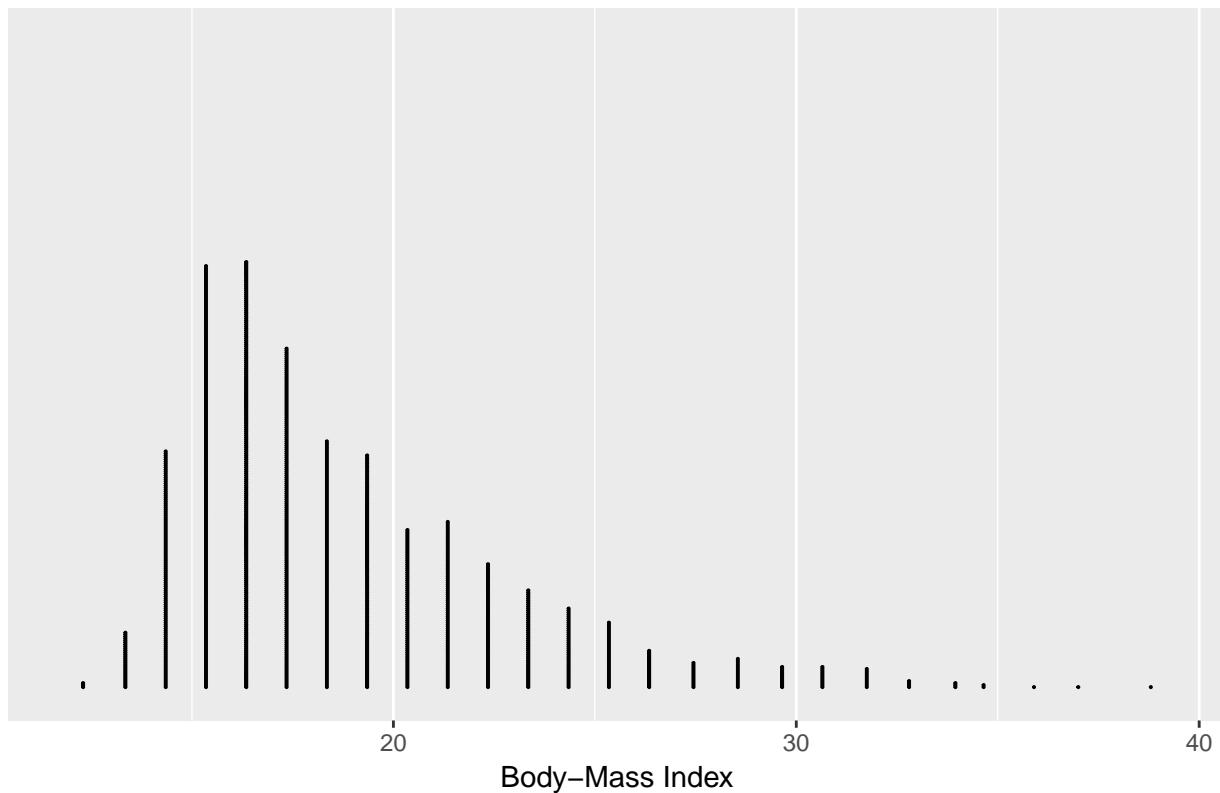
We can plot the distribution of a single continuous variable using the `dotplot` geom:

```

ggplot(data = nyfs1, aes(x = bmi)) +
  geom_dotplot(dotsize = 0.05, binwidth=1) +
  scale_y_continuous(NULL, breaks = NULL) + # hides y-axis since it is meaningless
  labs(title = "Dotplot of nyfs1 Body-Mass Index data",
       x = "Body-Mass Index")

```

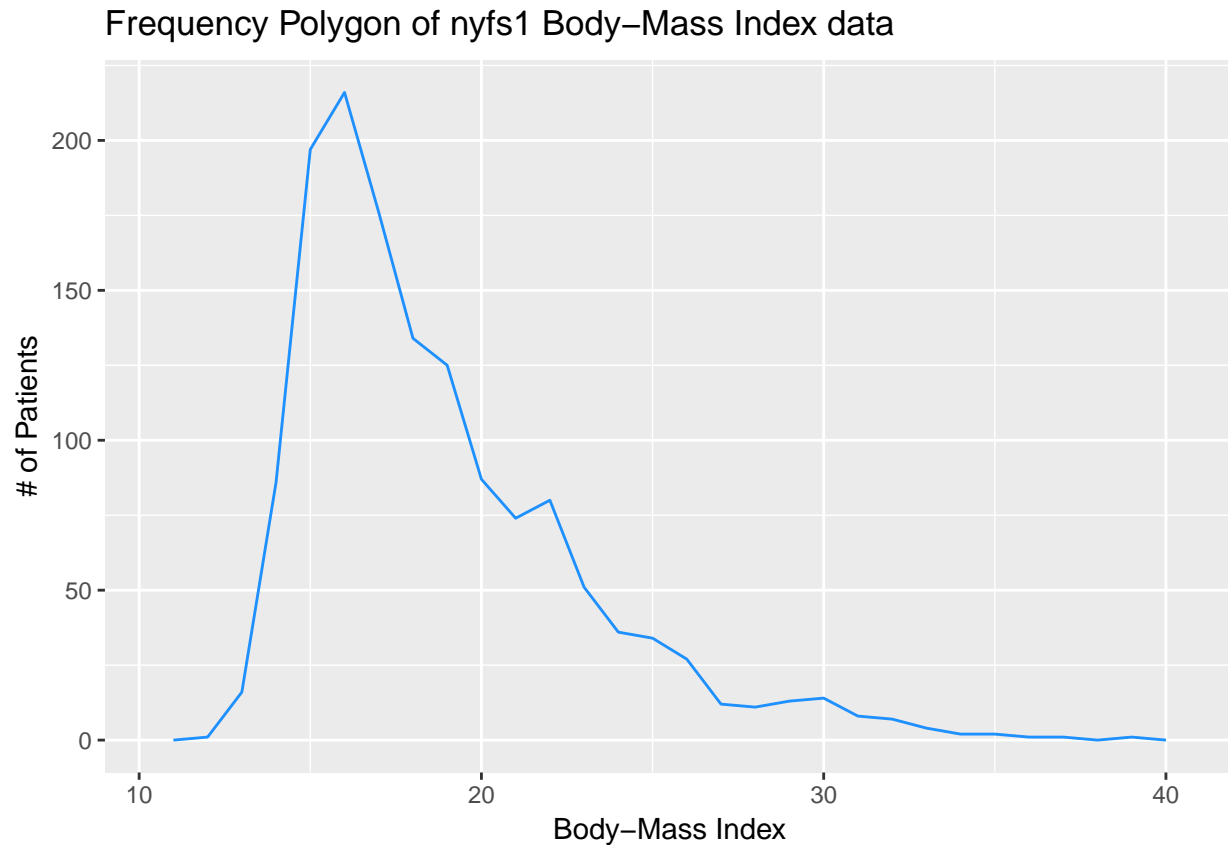
Dotplot of nyfs1 Body–Mass Index data



7.8 The Frequency Polygon

We can plot the distribution of a single continuous variable using the `freqpoly` geom:

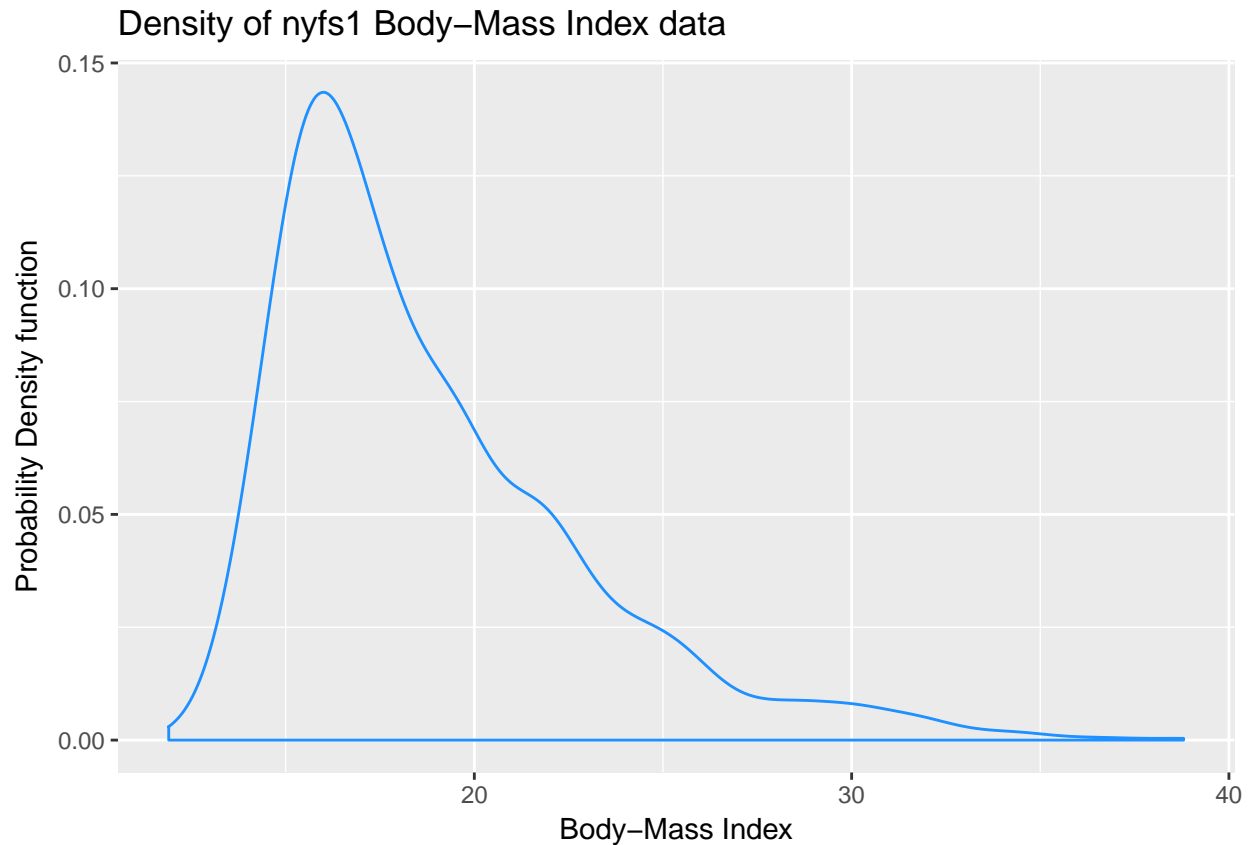
```
ggplot(data = nyfs1, aes(x = bmi)) +  
  geom_freqpoly(binwidth = 1, color = "dodgerblue") +  
  labs(title = "Frequency Polygon of nyfs1 Body-Mass Index data",  
       x = "Body-Mass Index", y = "# of Patients")
```



7.9 Plotting the Probability Density Function

We can also produce a density function, which has the effect of smoothing out the bumps in a histogram or frequency polygon, while also changing what is plotted on the y-axis.

```
ggplot(data = nyfs1, aes(x = bmi)) +  
  geom_density(kernel = "gaussian", color = "dodgerblue") +  
  labs(title = "Density of nyfs1 Body-Mass Index data",  
        x = "Body-Mass Index", y = "Probability Density function")
```



So, what's a density function?

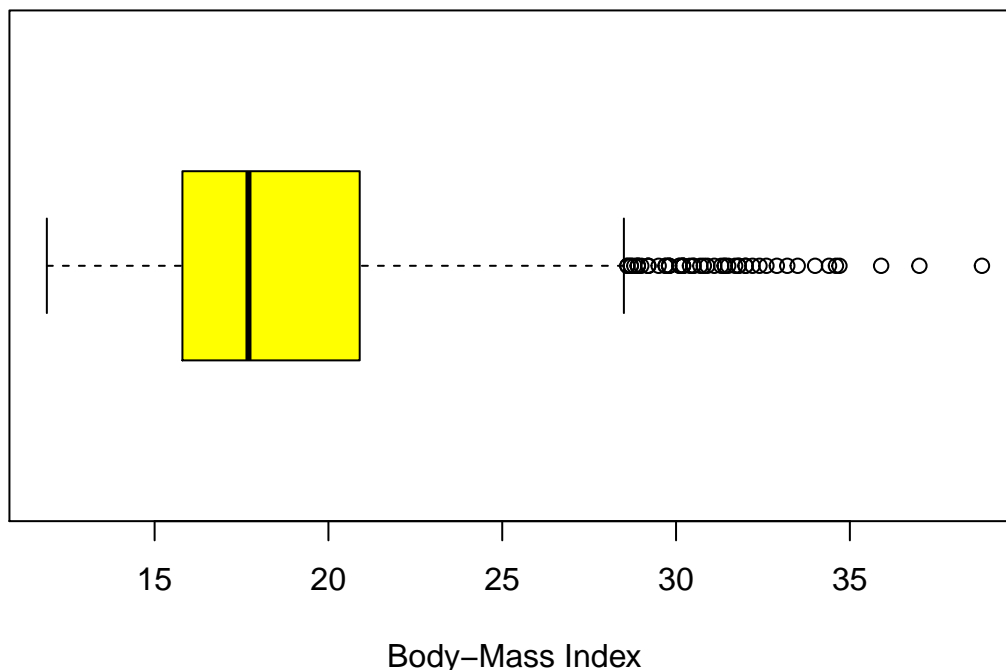
- A probability density function is a function of a continuous variable, x , that represents the probability of x falling within a given range. Specifically, the integral over the interval (a,b) of the density function gives the probability that the value of x is within (a,b) .
- If you're interested in exploring more on the notion of density functions for continuous (and discrete) random variables, some nice elementary material is available at Khan Academy.

7.10 The Boxplot

Sometimes, it's helpful to picture the five-number summary of the data in such a way as to get a general sense of the distribution. One approach is a **boxplot**, sometimes called a box-and-whisker plot.

```
boxplot(nyfs1$bmi, col="yellow", horizontal=T, xlab="Body-Mass Index",
main="BMI for 1416 kids in the NYFS")
```

BMI for 1416 kids in the NYFS



The boxplot is another John Tukey invention.

- R draws the box (here in yellow) so that its edges of the box fall at the 25th and 75th percentiles of the data, and the thick line inside the box falls at the median (50th percentile).
- The whiskers then extend out to the largest and smallest values that are not classified by the plot as candidate *outliers*.
- An outlier is an unusual point, far from the center of a distribution.
- Note that I've used the **horizontal** option to show this boxplot in this direction. Most comparison boxplots, as we'll see below, are oriented vertically.

The boxplot's **whiskers** that are drawn from the first and third quartiles (i.e. the 25th and 75th percentiles) out to the most extreme points in the data that do not meet the standard of "candidate outliers." An outlier is simply a point that is far away from the center of the data - which may be due to any number of reasons, and generally indicates a need for further investigation.

Most software, including R, uses a standard proposed by Tukey which describes a "candidate outlier" as any point above the **upper fence** or below the **lower fence**. The definitions of the fences are based on the inter-quartile range (IQR).

If $IQR = 75\text{th percentile} - 25\text{th percentile}$, then the upper fence is $75\text{th percentile} + 1.5 IQR$, and the lower fence is $25\text{th percentile} - 1.5 IQR$.

So for these BMI data,

- the upper fence is located at $20.9 + 1.5(5.1) = 28.55$
- the lower fence is located at $15.8 - 1.5(5.1) = 8.15$

In this case, we see no points identified as outliers in the low part of the distribution, but quite a few identified that way on the high side. This tends to identify about 5% of the data as a candidate outlier, *if* the data follow a Normal distribution.

- This plot is indicating clearly that there is some asymmetry (skew) in the data, specifically right skew.
- The standard R uses is to indicate as outliers any points that are more than 1.5 inter-quartile ranges away from the edges of the box.

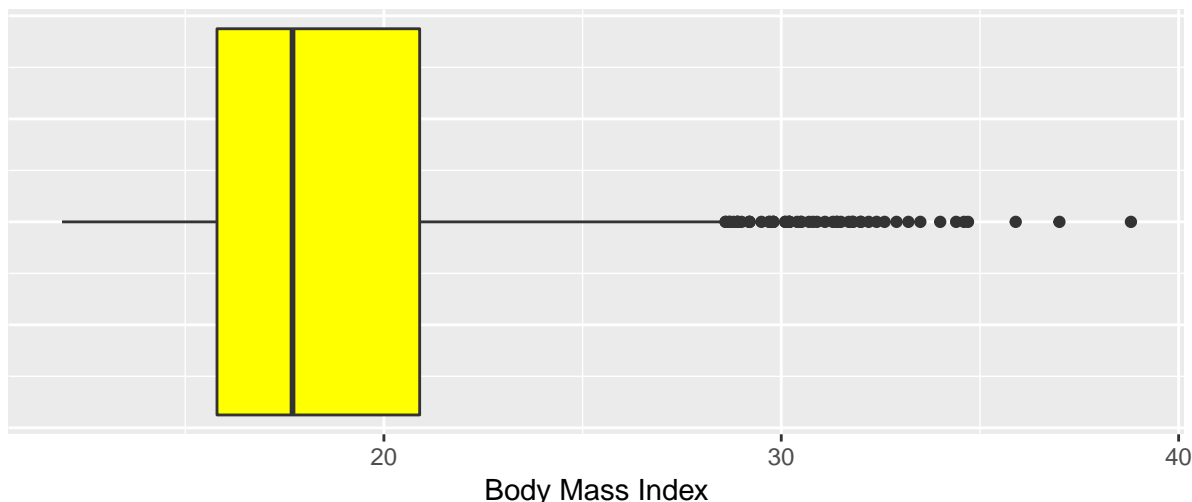
The horizontal orientation I've chosen here clarifies the relationship of direction of skew to the plot. A plot like this, with multiple outliers on the right side is indicative of a long right tail in the distribution, and hence, positive or right skew - with the mean being larger than the median. Other indications of skew include having one side of the box being substantially wider than the other, or one side of the whiskers being substantially longer than the other. More on skew later.

7.10.1 Drawing a Boxplot for One Variable in ggplot2

The `ggplot2` library easily handles comparison boxplots for multiple distributions, as we'll see in a moment. However, building a boxplot for a single distribution requires a little trickiness.

```
ggplot(nyfs1, aes(x = 1, y = bmi)) +
  geom_boxplot(fill = "yellow") +
  coord_flip() +
  labs(title = "Boxplot of BMI for 1416 kids in the NYFS",
       y = "Body Mass Index",
       x = "") +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```

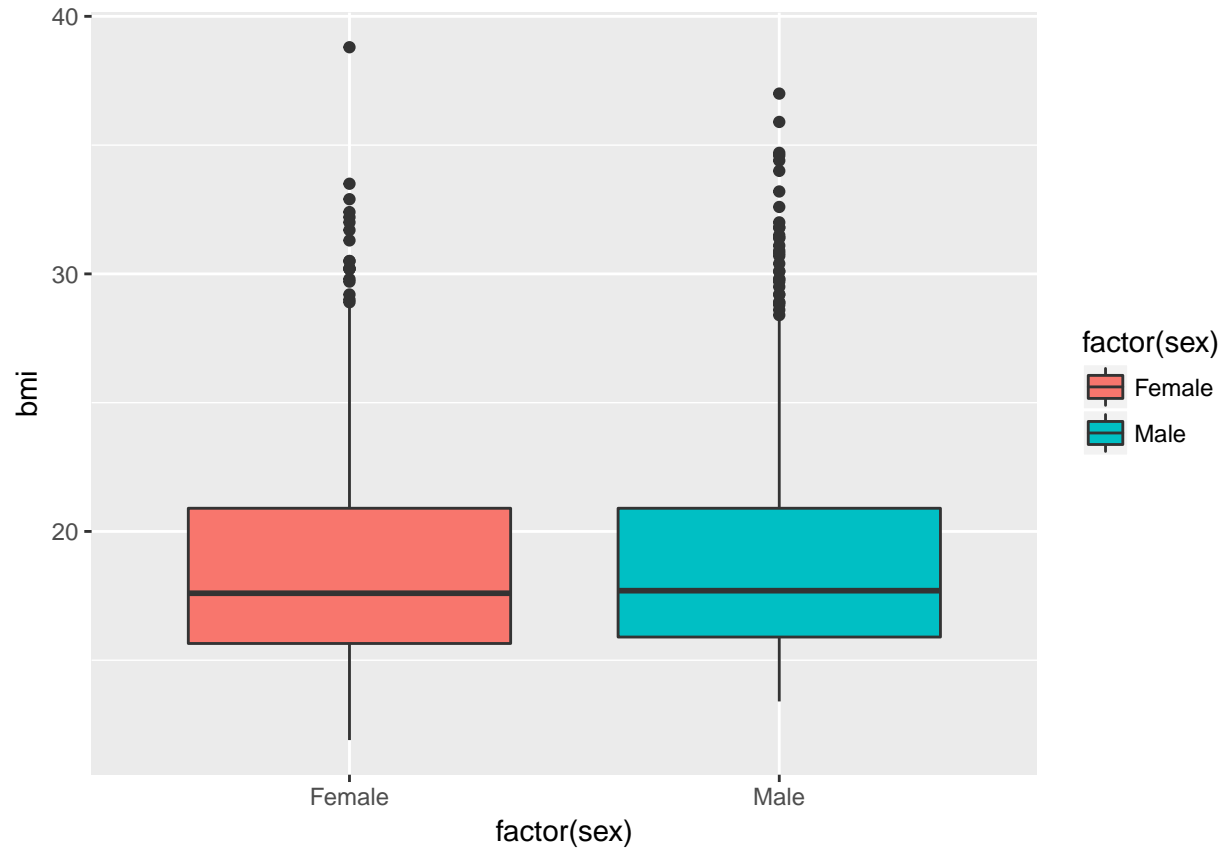
Boxplot of BMI for 1416 kids in the NYFS



7.11 A Simple Comparison Boxplot

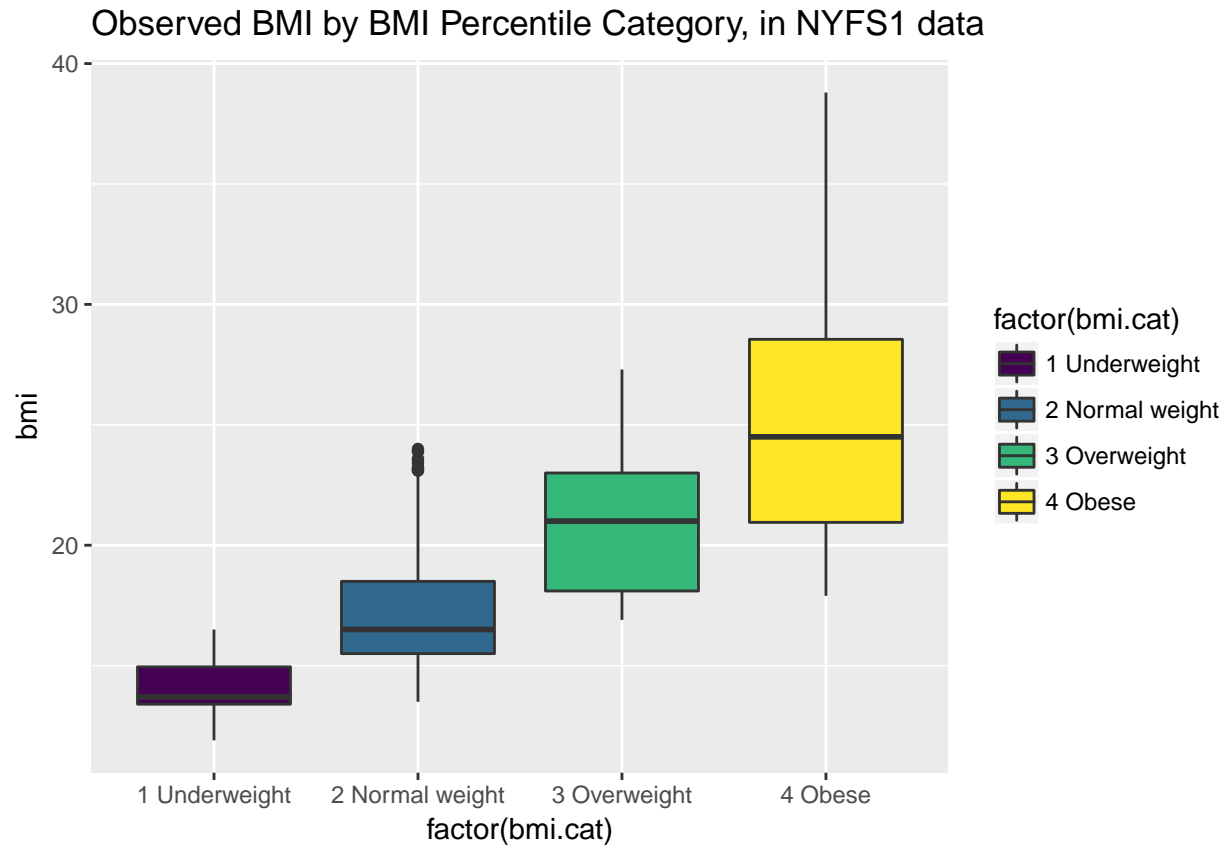
Boxplots are most often used for comparison. We can build boxplots using `ggplot2`, as well, and we'll discuss that in detail later. For now, here's a boxplot built to compare the `bmi` results by the child's sex.

```
ggplot(nyfs1, aes(x = factor(sex), y = bmi, fill=factor(sex))) +
  geom_boxplot()
```



Let's look at the comparison of observed BMI levels across the four categories in our `bmi.cat` variable, now making use of the `viridis` color scheme.

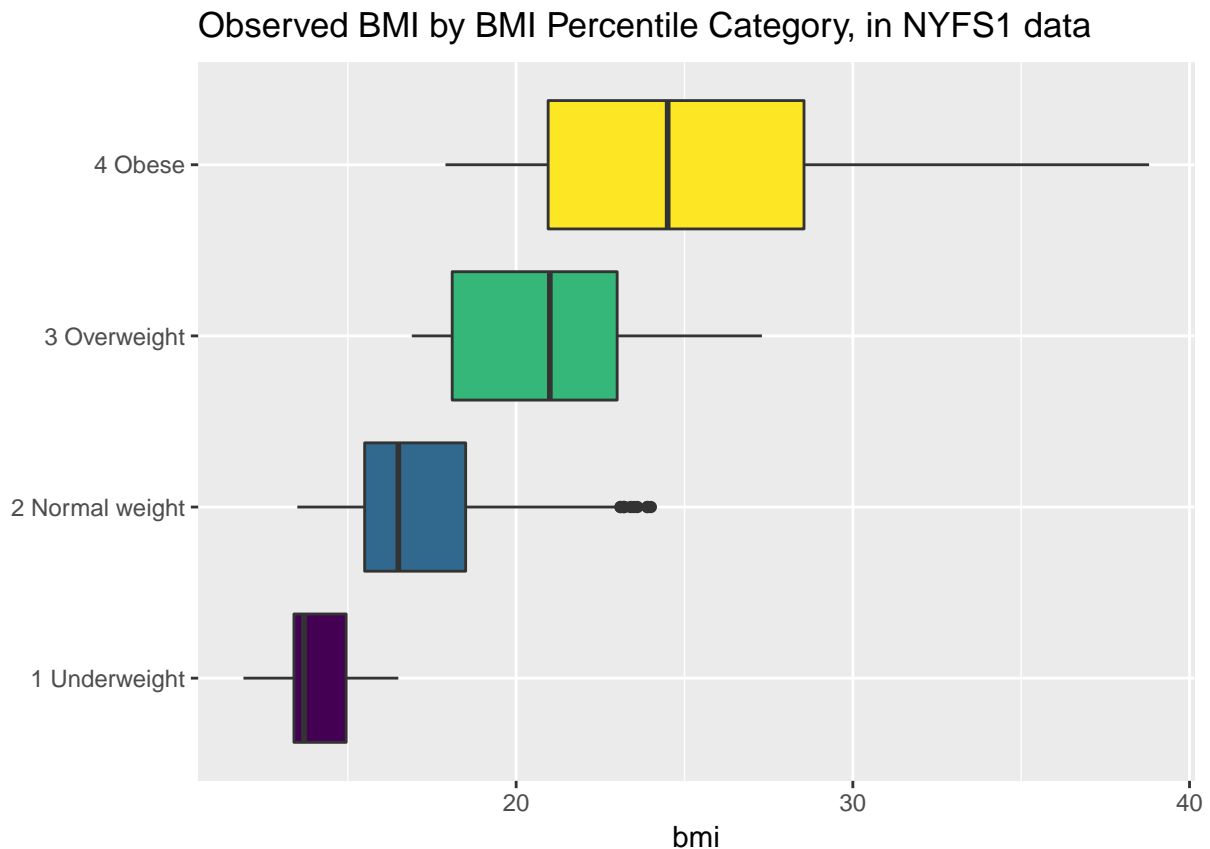
```
ggplot(nyfs1, aes(x = factor(bmi.cat), y = bmi, fill = factor(bmi.cat))) +
  geom_boxplot() +
  scale_fill_viridis(discrete=TRUE) +
  # above line uses viridis palette to identify color choices
  labs(title = "Observed BMI by BMI Percentile Category, in NYFS1 data")
```



Note that the BMI categories incorporate additional information (in particular the age and sex of the child) beyond the observed BMI, and so the observed BMI levels overlap quite a bit across the four categories. As a graph, that's not bad, but what if we want to improve it further?

Let's turn the boxes in the horizontal direction, and get rid of the perhaps unnecessary `bmi.cat` labels.

```
ggplot(nyfs1, aes(x = factor(bmi.cat), y = bmi, fill = factor(bmi.cat))) +
  geom_boxplot() +
  scale_fill_viridis(discrete=TRUE) +
  coord_flip() +
  guides(fill=FALSE) +
  labs(title = "Observed BMI by BMI Percentile Category, in NYFS1 data", x = "")
```

7.12 Using describe in the psych library

For additional numerical summaries, one option would be to consider using the `describe` function from the `psych` library.

```
psych::describe(nyfs1$bmi)
```

```
vars      n mean  sd median trimmed mad  min  max range skew kurtosis
X1       1 1416 18.8 4.08  17.7   18.2 3.26 11.9 38.8  26.9 1.35    1.97
se
X1 0.11
```

This package provides, in order, the following...

- **n** = the sample size
- **mean** = the sample mean
- **sd** = the sample standard deviation
- **median** = the median, or 50th percentile
- **trimmed** = mean of the middle 80% of the data
- **mad** = median absolute deviation
- **min** = minimum value in the sample
- **max** = maximum value in the sample
- **range** = max - min
- **skew** = skewness measure, described below (indicates degree of asymmetry)
- **kurtosis** = kurtosis measure, described below (indicates heaviness of tails, degree of outlier-proneness)
- **se** = standard error of the sample mean = sd / square root of sample size, useful in inference

7.12.1 The Trimmed Mean

The **trimmed mean** trim value in R indicates proportion of observations to be trimmed from each end of the outcome distribution before the mean is calculated. The `trimmed` value provided by the `psych::describe` package describes what this particular package calls a 20% trimmed mean (bottom and top 10% of BMIs are removed before taking the mean - it's the mean of the middle 80% of the data.) I might call that a 10% trimmed mean in some settings, but that's just me.

```
mean(nyfs1$bmi, trim=.1)
```

```
[1] 18.2
```

7.12.2 The Median Absolute Deviation

An alternative to the IQR that is fancier, and a bit more robust, is the **median absolute deviation**, which, in large sample sizes, for data that follow a Normal distribution, will be (in expectation) equal to the standard deviation. The MAD is the median of the absolute deviations from the median, multiplied by a constant (1.4826) to yield asymptotically normal consistency.

```
mad(nyfs1$bmi)
```

```
[1] 3.26
```

7.13 Assessing Skew

A relatively common idea is to assess **skewness**, several measures of which (including the one below, sometimes called type 3 skewness, or Pearson's moment coefficient of skewness) are available. Many models assume a Normal distribution, where, among other things, the data are symmetric around the mean.

Skewness measures asymmetry in the distribution - left skew (mean < median) is indicated by negative skewness values, while right skew (mean > median) is indicated by positive values. The skew value will be near zero for data that follow a Normal distribution.

7.13.1 Non-parametric Skew via skew₁

A simpler measure of skew, sometimes called the **nonparametric skew** and closely related to Pearson's notion of median skewness, falls between -1 and +1 for any distribution. It is just the difference between the mean and the median, divided by the standard deviation.

- Values greater than +0.2 are sometimes taken to indicate fairly substantial right skew, while values below -0.2 indicate fairly substantial left skew.

```
(mean(nyfs1$bmi) - median(nyfs1$bmi))/sd(nyfs1$bmi)
```

```
[1] 0.269
```

There is a function in the `Love-boost.R` script called `skew1` that can be used to do these calculations, so long as the variable has no missing data.

```
skew1(nyfs1$bmi)
```

```
[1] 0.269
```

The Wikipedia page on skewness, from which some of this material is derived, provides definitions for several other skewness measures.

7.14 Assessing Kurtosis (Heavy-Tailedness)

Another measure of a distribution's shape that can be found in the **psych** library is the **kurtosis**. Kurtosis is an indicator of whether the distribution is heavy-tailed or light-tailed as compared to a Normal distribution. Positive kurtosis means more of the variance is due to outliers - unusual points far away from the mean relative to what we might expect from a Normally distributed data set with the same standard deviation.

- A Normal distribution will have a kurtosis value near 0, a distribution with similar tail behavior to what we would expect from a Normal is said to be *mesokurtic*
- Higher kurtosis values (meaningfully higher than 0) indicate that, as compared to a Normal distribution, the observed variance is more the result of extreme outliers (i.e. heavy tails) as opposed to being the result of more modest sized deviations from the mean. These heavy-tailed, or outlier prone, distributions are sometimes called *leptokurtic*.
- Kurtosis values meaningfully lower than 0 indicate light-tailed data, with fewer outliers than we'd expect in a Normal distribution. Such distributions are sometimes referred to as *platykurtic*, and include distributions without outliers, like the Uniform distribution.

Here's a table:

Fewer outliers than a Normal	Approximately Normal	More outliers than a Normal
Light-tailed <i>platykurtic</i> (kurtosis < 0)	"Normalish" <i>mesokurtic</i> (kurtosis = 0)	Heavy-tailed <i>leptokurtic</i> (kurtosis > 0)

```
psych::kurtosi(nyfs1$bmi)
```

```
[1] 1.97
```

7.14.1 The Standard Error of the Sample Mean

The **standard error** of the sample mean, which is the standard deviation divided by the square root of the sample size:

```
sd(nyfs1$bmi)/sqrt(length(nyfs1$bmi))
```

```
[1] 0.108
```

7.15 The describe function in the Hmisc library

The **Hmisc** library has lots of useful functions. It's named for its main developer, Frank Harrell. The **describe** function in **Hmisc** knows enough to separate numerical from categorical variables, and give you separate (and detailed) summaries for each.

- For a categorical variable, it provides counts of total observations (n), the number of missing values, and the number of unique categories, along with counts and percentages falling in each category.
- For a numerical variable, it provides:
 - counts of total observations (n), the number of missing values, and the number of unique values
 - an Info value for the data, which indicates how continuous the variable is (a score of 1 is generally indicative of a completely continuous variable with no ties, while scores near 0 indicate lots of ties, and very few unique values)
 - the sample Mean
 - many sample percentiles (quantiles) of the data, specifically (5, 10, 25, 50, 75, 90, 95, 99)

- either a complete table of all observed values, with counts and percentages (if there are a modest number of unique values), or
- a table of the five smallest and five largest values in the data set, which is useful for range checking

```
Hmisc::describe(nyfs1)
```

```
nyfs1
```

```
7 Variables      1416 Observations
```

```
subject.id
```

n	missing	distinct	Info	Mean	Gmd	.05	.10
1416	0	1416	1	72703	525.3	71994	72073
.25	.50	.75	.90	.95			
72313	72698	73096	73331	73414			

```
lowest : 71918 71919 71921 71922 71923, highest: 73488 73489 73490 73491 73492
```

```
sex
```

n	missing	distinct
1416	0	2

Value	Female	Male
Frequency	707	709
Proportion	0.499	0.501

```
age.exam
```

n	missing	distinct	Info	Mean	Gmd	.05	.10
1416	0	14	0.994	8.855	4.235	3	4
.25	.50	.75	.90	.95			
6	9	12	14	15			

Value	3	4	5	6	7	8	9	10	11	12
Frequency	97	111	119	129	123	120	90	109	102	108
Proportion	0.069	0.078	0.084	0.091	0.087	0.085	0.064	0.077	0.072	0.076

Value	13	14	15	16
Frequency	113	104	85	6
Proportion	0.080	0.073	0.060	0.004

```
bmi
```

n	missing	distinct	Info	Mean	Gmd	.05	.10
1416	0	191	1	18.8	4.321	14.30	14.80
.25	.50	.75	.90	.95			
15.80	17.70	20.90	24.45	27.00			

```
lowest : 11.9 12.6 12.7 12.9 13.0, highest: 34.6 34.7 35.9 37.0 38.8
```

```
bmi.cat
```

n	missing	distinct
1416	0	4

Value	1 Underweight	2 Normal weight	3 Overweight	4 Obese
Frequency		42	926	237
Proportion		0.030	0.654	0.167

```

-----
waist.circ
      n missing distinct      Info      Mean      Gmd      .05      .10
1416      0      462          1    65.29    14.23    49.30    51.10
.25      .50      .75      .90      .95
55.00    63.00    72.93    82.35    90.40

lowest :  42.5  43.4  44.1  44.4  44.7, highest: 108.4 108.5 110.4 111.0 112.4
-----
triceps.skinfold
      n missing distinct      Info      Mean      Gmd      .05      .10
1416      0      236          1    13.37     6.279    6.775    7.400
.25      .50      .75      .90      .95
9.000    11.800   16.600   21.750   25.600

lowest :  4.0  4.6  4.9  5.0  5.2, highest: 34.3 34.8 36.0 36.2 38.2
-----

```

More on the Info value in Hmisc::describe is available [here](#)

7.16 xda from GitHub for numerical summaries for exploratory data analysis

```
## next two commands needed if xda is not already installed
```

```
library(devtools)
install_github("ujjwalkarn/xda")
```

Skipping install of 'xda' from a github remote, the SHA1 (fb68f0da) has not changed since last install.
Use `force = TRUE` to force installation

```
xda::numSummary(nyfs1)
```

```

      n      mean      sd      max      min range nunique
subject.id  1416 72702.70 454.75 73492.0 71918.0 1574.0   1416
age.exam    1416   8.86   3.68   16.0    3.0   13.0    14
bmi         1416  18.80   4.08   38.8   11.9   26.9   191
waist.circ  1416  65.29  12.85  112.4   42.5   69.9   462
triceps.skinfold 1416  13.37   5.83   38.2    4.0   34.2   236

      nzeros      iqr lowerbound upperbound noutlier kurtosis
subject.id      0 784.0   71136.75   74272.2      0  -1.193
age.exam        0   6.0    -3.00    21.0      0  -1.198
bmi             0   5.1     8.15    28.5     53   1.973
waist.circ      0  17.9    28.15    99.8     22   0.384
triceps.skinfold 0   7.6    -2.40    28.0     31   1.149

      skewness      mode miss miss%      1%      5%      25%
subject.id  0.00815 71918.0   0    0 71933.1 71993.75 72312.8
age.exam    0.08202   6.0    0    0   3.0    3.00   6.0
bmi         1.34804  15.5    0    0  13.5   14.30  15.8
waist.circ  0.85106  55.4    0    0  46.1   49.30  55.0
triceps.skinfold 1.15791   8.0    0    0   5.6    6.77   9.0

      50%      75%      95%      99%
subject.id 72697.5 73096.2 73414.2 73478

```

age.exam	9.0	12.0	15.0	15
bmi	17.7	20.9	27.0	32
waist.circ	63.0	72.9	90.4	102
triceps.skinfold	11.8	16.6	25.6	31

Most of the elements of this `numSummary` should be familiar. Some new pieces include:

- `nunique` = number of unique values
- `nzeroes` = number of zeroes
- `noutlier` = number of outliers (using a standard that isn't entirely transparent to me)
- `miss` = number of rows with missing value
- `miss%` = percentage of total rows with missing values $((\text{miss}/n)*100)$
- `5%` = 5th percentile value of that variable (value below which 5 percent of the observations may be found)

```
xda::charSummary(nyfs1)
```

```

              n miss miss% unique
sex          1416    0     0      2
bmi.cat 1416    0     0      4

                                top5levels:count
sex                                Male:709, Female:707
bmi.cat 2 Normal weight:926, 3 Overweight:237, 4 Obese:211, 1 Underweight:42
```

The `top5levels:count` provides the top 5 unique values for each variable, sorted by their counts.

7.17 What Summaries to Report

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.

Chapter 8

Assessing Normality

Data are well approximated by a Normal distribution if the shape of the data's distribution is a good match for a Normal distribution with mean and standard deviation equal to the sample statistics.

- the data are symmetrically distributed about a single peak, located at the sample mean
- the spread of the distribution is well characterized by a Normal distribution with standard deviation equal to the sample standard deviation
- the data show outlying values (both in number of candidate outliers, and size of the distance between the outliers and the center of the distribution) that are similar to what would be predicted by a Normal model.

We have several tools for assessing Normality of a single batch of data, including:

- a histogram with superimposed Normal distribution
- histogram variants (like the boxplot) which provide information on the center, spread and shape of a distribution
- the Empirical Rule for interpretation of a standard deviation
- a specialized *normal Q-Q plot* (also called a normal probability plot or normal quantile-quantile plot) designed to reveal differences between a sample distribution and what we might expect from a normal distribution of a similar number of values with the same mean and standard deviation

8.1 Empirical Rule Interpretation of the Standard Deviation

For a set of measurements that follows a Normal distribution, the interval:

- Mean \pm Standard Deviation contains approximately 68% of the measurements;
- Mean \pm 2(Standard Deviation) contains approximately 95% of the measurements;
- Mean \pm 3(Standard Deviation) contains approximately all (99.7%) of the measurements.

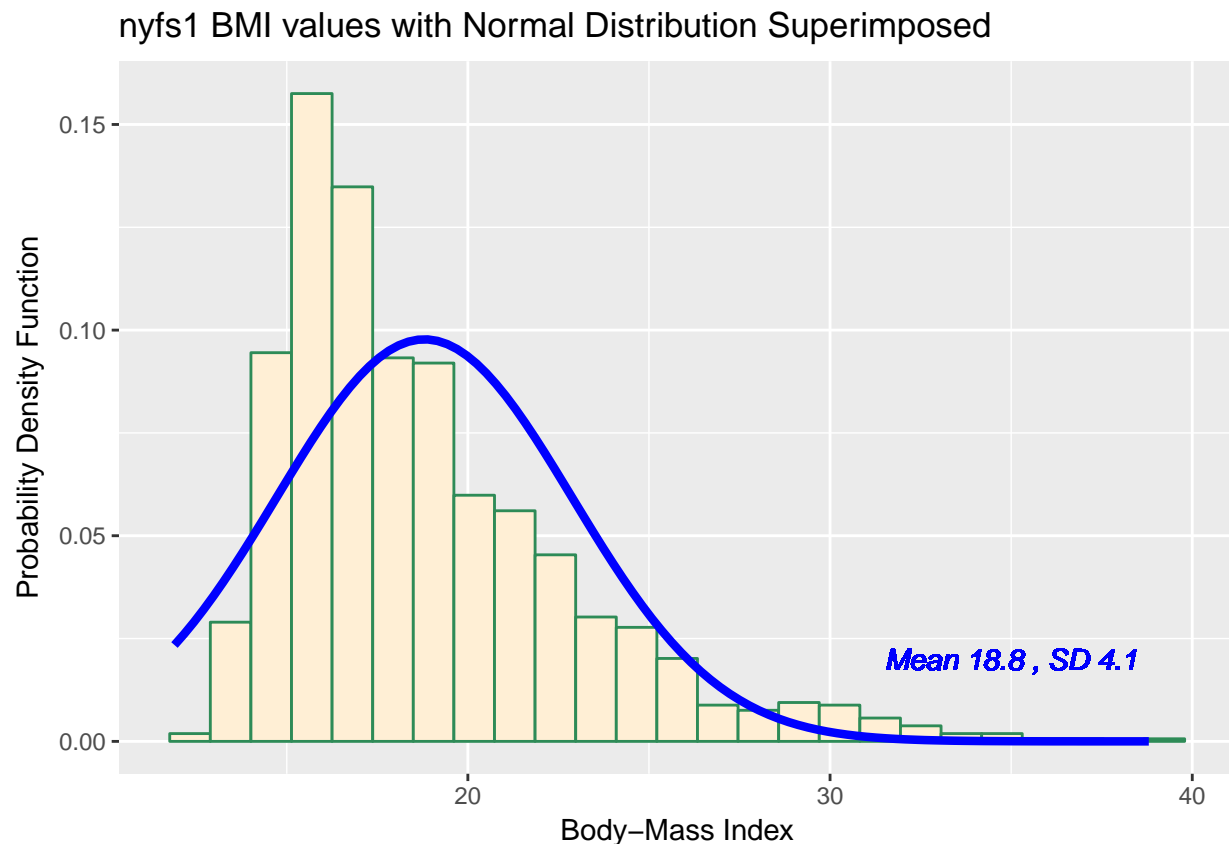
Again, most data sets do not follow a Normal distribution. We will occasionally think about transforming or re-expressing our data to obtain results which are better approximated by a Normal distribution, in part so that a standard deviation can be more meaningful.

For the BMI data we have been studying, here again are some summary statistics...

```
mosaic::favstats(nyfs1$bmi)
```

min	Q1	median	Q3	max	mean	sd	n	missing
11.9	15.8	17.7	20.9	38.8	18.8	4.08	1416	0


```
x = 35, y = 0.02, color="blue", fontface = "italic") +
labs(title = "nyfs1 BMI values with Normal Distribution Superimposed",
     x = "Body-Mass Index", y = "Probability Density Function")
```



Does it seem as though the Normal model (as shown in the blue density curve) is an effective approximation to the observed distribution shown in the bars of the histogram?

We'll return shortly to the questions:

- Does a Normal distribution model fit our data well? *and*
- If the data aren't Normal, but we want to use a Normal model anyway, what should we do?

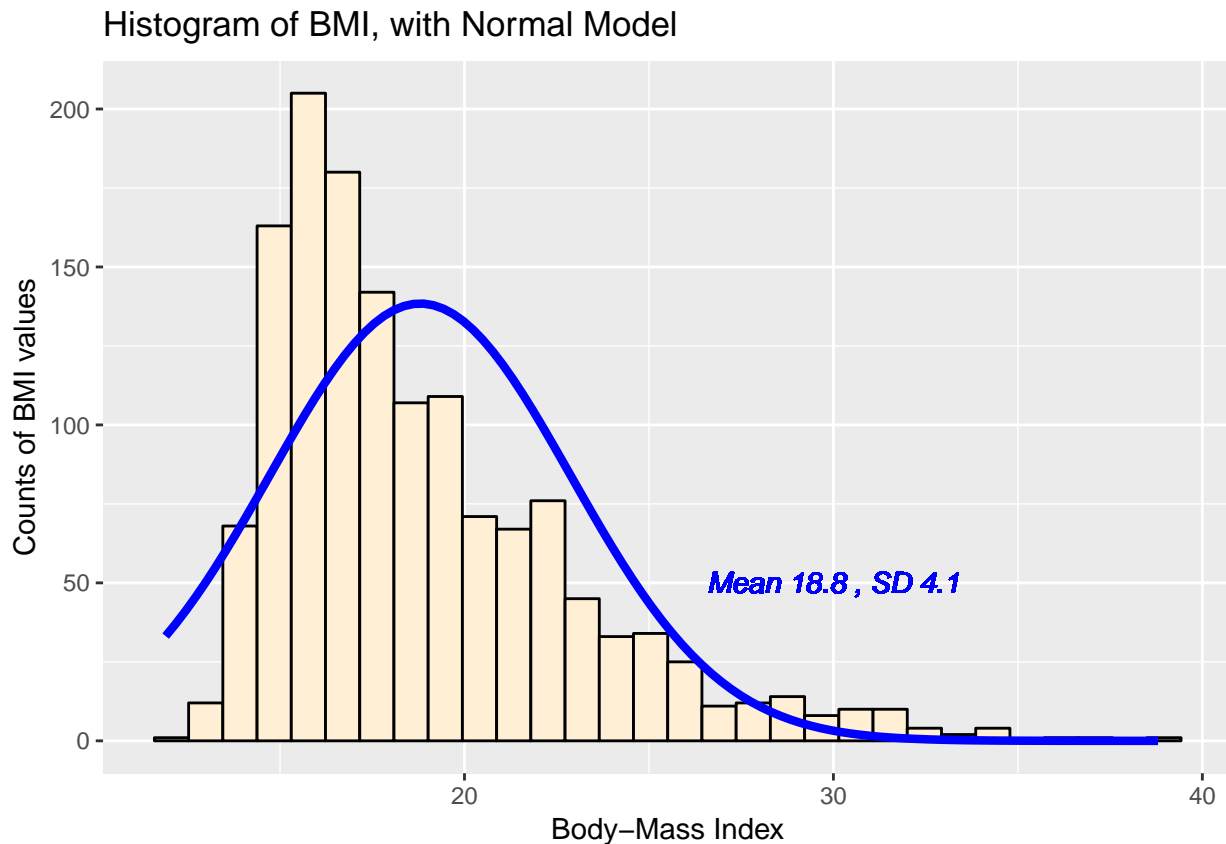
8.3.1 Histogram of BMI with Normal model (with Counts)

But first, we'll demonstrate an approach to building a histogram of counts (rather than a probability density) and then superimposing a Normal model.

```
## ggplot of counts of bmi with Normal model superimposed
## Source: https://stat.ethz.ch/pipermail/r-help/2009-September/403220.html

ggplot(nyfs1, aes(x = bmi)) +
  geom_histogram(bins = 30, fill = "papayawhip", color = "black") +
  stat_function(fun = function(x, mean, sd, n)
    n * dnorm(x = x, mean = mean, sd = sd),
    args = with(nyfs1,
      c(mean = mean(bmi), sd = sd(bmi), n = length(bmi))),
    col = "blue", lwd = 1.5) +
```

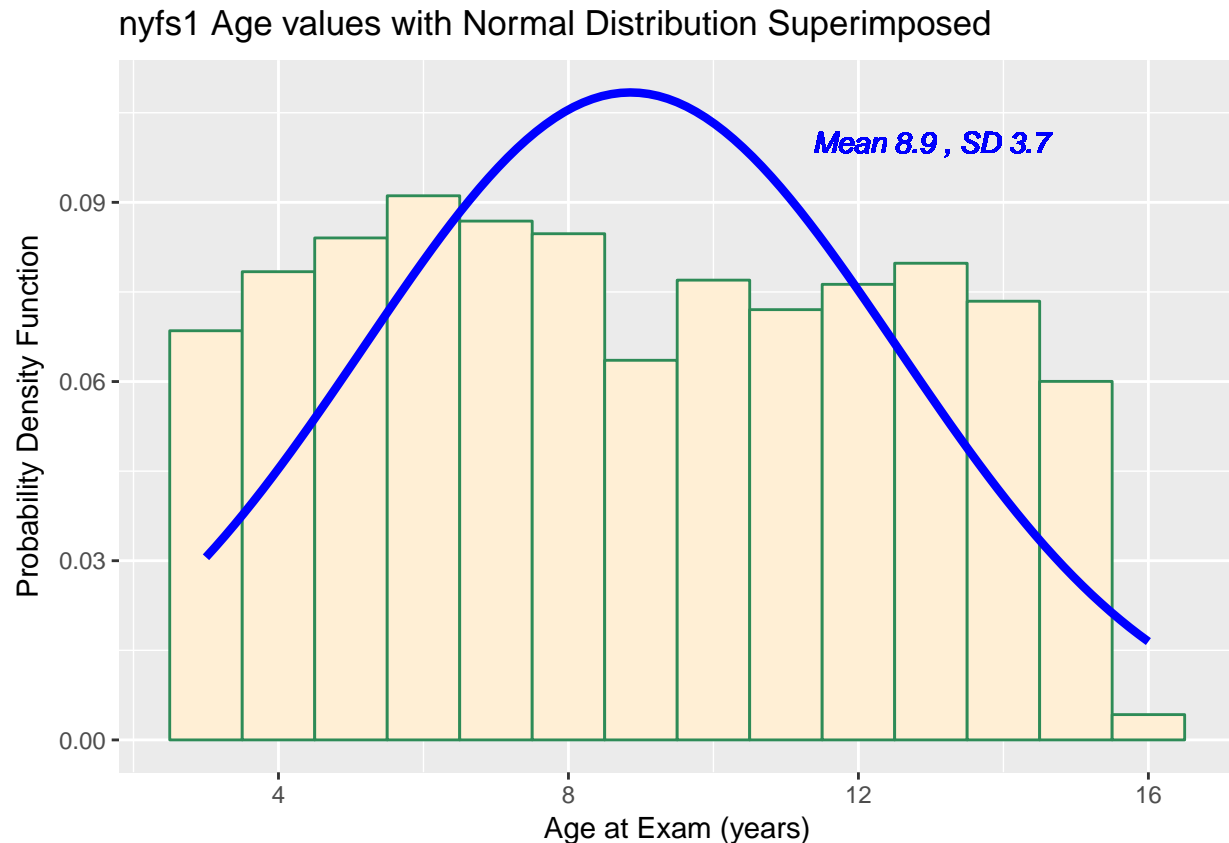
```
geom_text(aes(label = paste("Mean", round(mean(nyfs1$bmi),1),
                             ", SD", round(sd(nyfs1$bmi),1))),
          x = 30, y = 50, color="blue", fontface = "italic") +
labs(title = "Histogram of BMI, with Normal Model",
     x = "Body-Mass Index", y = "Counts of BMI values")
```



8.4 Does a Normal model work well for the Ages?

Now, suppose we instead look at the `age.exam` data. Do these data appear to follow a Normal distribution?

```
ggplot(nyfs1, aes(x=age.exam)) +
  geom_histogram(aes(y = ..density..), binwidth=1,
                 fill = "papayawhip", color = "seagreen") +
  stat_function(fun = dnorm,
               args = list(mean = mean(nyfs1$age.exam),
                           sd = sd(nyfs1$age.exam)),
               lwd = 1.5, col = "blue") +
  geom_text(aes(label = paste("Mean", round(mean(nyfs1$age.exam),1),
                              ", SD", round(sd(nyfs1$age.exam),1))),
            x = 13, y = 0.1, color="blue", fontface = "italic") +
labs(title = "nyfs1 Age values with Normal Distribution Superimposed",
     x = "Age at Exam (years)", y = "Probability Density Function")
```



```
mosaic::favstats(nyfs1$age.exam)
```

```
min Q1 median Q3 max mean sd n missing
3 6 9 12 16 8.86 3.68 1416 0
```

The mean is 8.86 and the standard deviation is 3.68 so if the `age.exam` data really were Normally distributed, we'd expect to see:

- About 68% of the data in the range (5.17, 12.54). In fact, 781 of the 1416 Age values are in this range, or 55.2%.
- About 95% of the data in the range (1.49, 16.22). In fact, 1416 of the 1416 Age values are in this range, or 100%.
- About 99.7% of the data in the range (-2.19, 19.9). In fact, 1416 of the 1416 Age values are in this range, or 100%.

How does the Normal approximation work for age, according to the Empirical Rule?

There is a function in the `Love-boost.R` script called `Emp_Rule` that can be used to do these calculations, so long as the variable has no missing data.

```
Emp_Rule(nyfs1$bmi)
```

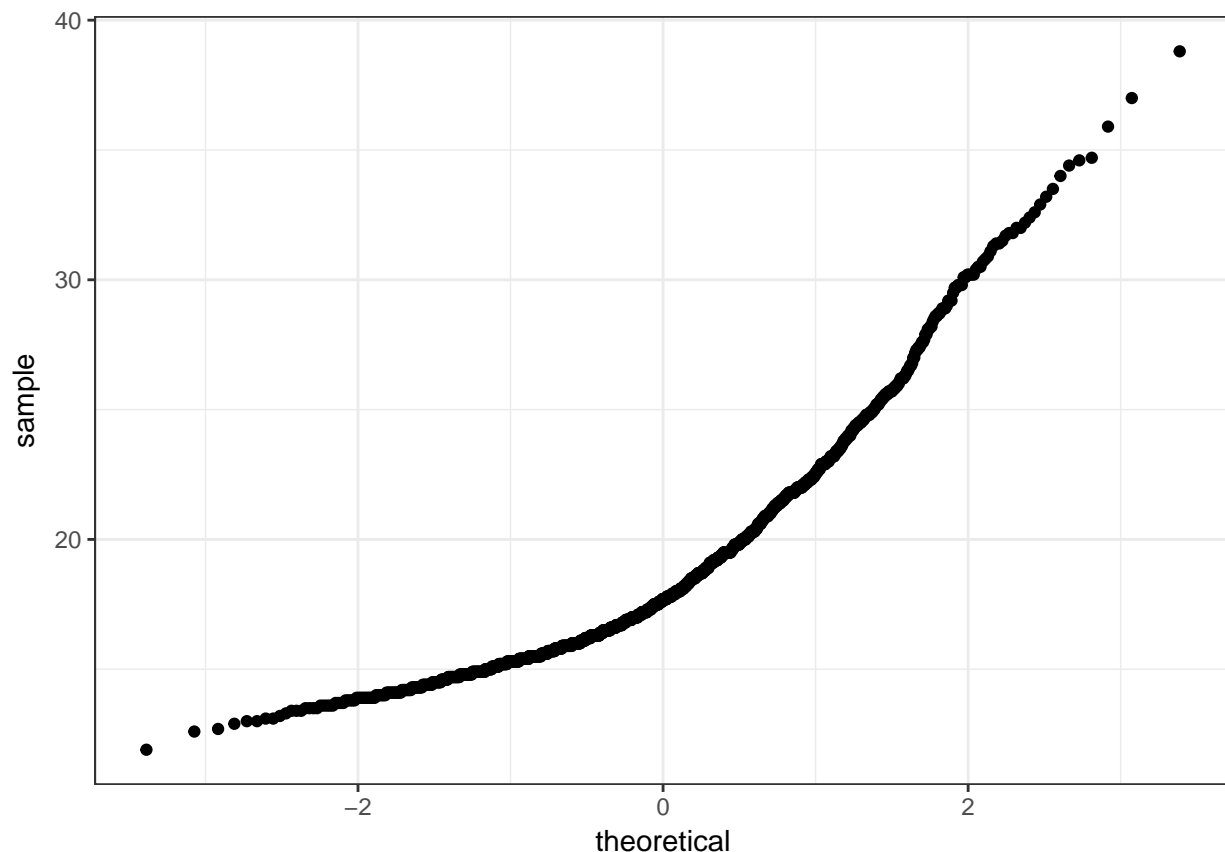
	count	proportion
Mean +/- 1 SD	1074	0.7585
Mean +/- 2 SD	1344	0.9492
Mean +/- 3 SD	1393	0.9838
Entire Data Set	1416	1

8.5 The Normal Q-Q Plot

A normal probability plot (or normal quantile-quantile plot) of the BMI results from the `nyfs1` data, developed using `ggplot2` is shown below. In this case, this is a picture of 1416 BMI results. The idea of a normal Q-Q plot is that it plots the observed sample values (on the vertical axis) and then, on the horizontal, the expected or theoretical quantiles that would be observed in a standard normal distribution (a Normal distribution with mean 0 and standard deviation 1) with the same number of observations.

A Normal Q-Q plot will follow a straight line when the data are (approximately) Normally distributed. When the data have a different shape, the plot will reflect that.

```
ggplot(nyfs1, aes(sample = bmi)) +  
  geom_point(stat="qq") +  
  theme_bw() # eliminate the gray background
```



8.5.1 A Fancy ggplot2 function - `gg_qq`

The lengthy function below (which is also part of the `Love-boost.R` script) produces a normal Q-Q plot with some appealing features, including a Q-Q line, and confidence interval estimates - original source [here](https://gist.github.com/rentrop/d39a8406ad8af2a1066c).

```
## Code from https://gist.github.com/rentrop/d39a8406ad8af2a1066c  
## Slightly modified by T. Love
```

```
gg_qq <- function(x, distribution = "norm", ..., line.estimate = NULL, conf = 0.95,  
                  labels = names(x)){  
  q.function <- eval(parse(text = paste0("q", distribution)))
```

```

d.function <- eval(parse(text = paste0("d", distribution)))
x <- na.omit(x)
ord <- order(x)
n <- length(x)
P <- ppoints(length(x))
df <- data.frame(ord.x = x[ord], z = q.function(P, ...))

if(is.null(line.estimate)){
  Q.x <- quantile(df$ord.x, c(0.25, 0.75))
  Q.z <- q.function(c(0.25, 0.75), ...)
  b <- diff(Q.x)/diff(Q.z)
  coef <- c(Q.x[1] - b * Q.z[1], b)
} else {
  coef <- coef(line.estimate(ord.x ~ z))
}

zz <- qnorm(1 - (1 - conf)/2)
SE <- (coef[2]/d.function(df$z)) * sqrt(P * (1 - P)/n)
fit.value <- coef[1] + coef[2] * df$z
df$upper <- fit.value + zz * SE
df$lower <- fit.value - zz * SE

if(!is.null(labels)){
  df$label <- ifelse(df$ord.x > df$upper | df$ord.x < df$lower, labels[ord], "")
}

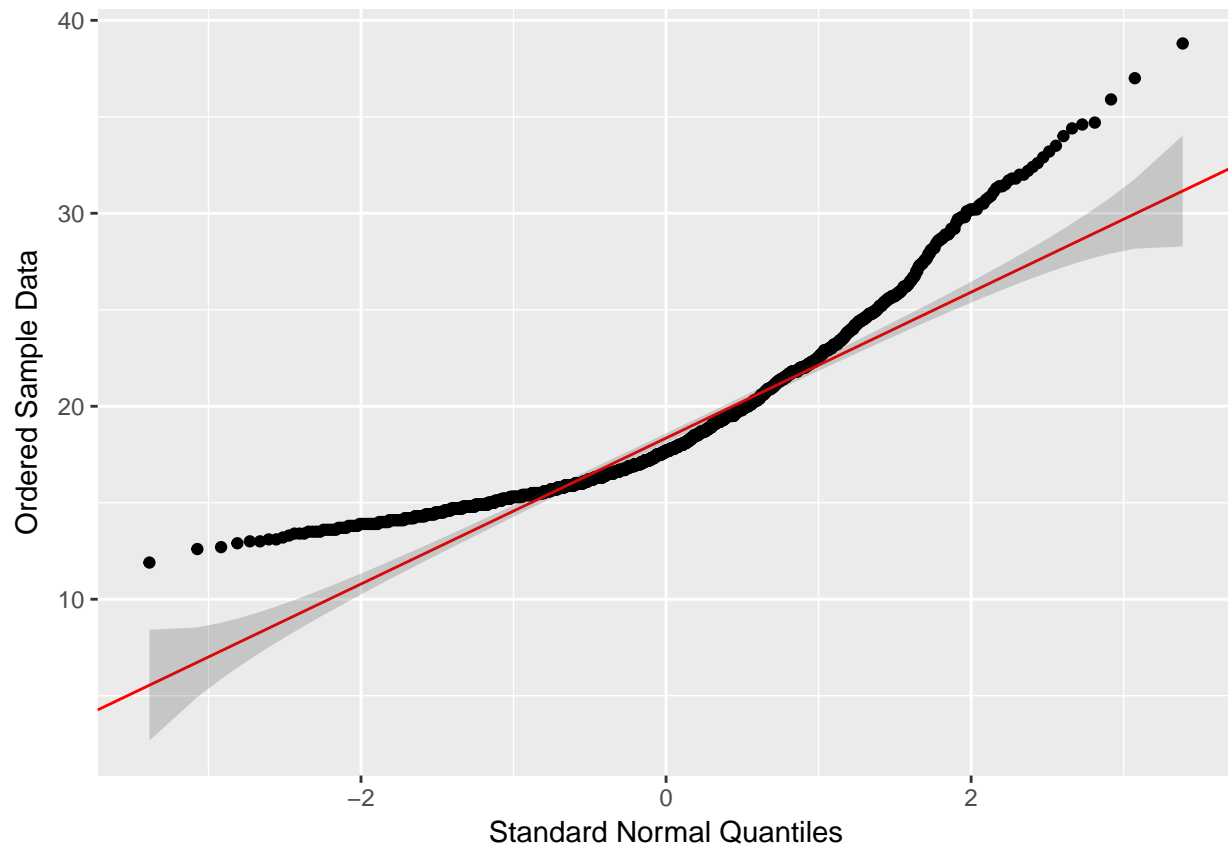
p <- ggplot(df, aes(x=z, y=ord.x)) +
  geom_point() +
  geom_abline(intercept = coef[1], slope = coef[2], color="red") +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha=0.2) +
  labs(y = "Ordered Sample Data", x = "Standard Normal Quantiles")

if(!is.null(labels)) p <- p + geom_text(aes(label = label))
print(p)
## can print 25th and 75th percentiles if desired by removing the
## single # in the following line
# coef
}

```

Having set up the function, we can now use it to look at the BMI data.

```
gg_qq(nyfs1$bmi)
```



8.6 Interpreting the Normal Q-Q Plot

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)

8.6.1 Data from a Normal distribution shows up as a straight line in a Normal Q-Q plot

We'll demonstrate the looks that we can obtain from a Normal Q-Q plot in some simulations. First, here is an example of a Normal Q-Q plot, and its associated histogram, for a sample of 200 observations simulated from a Normal distribution.

```
set.seed(123431) # so the results can be replicated

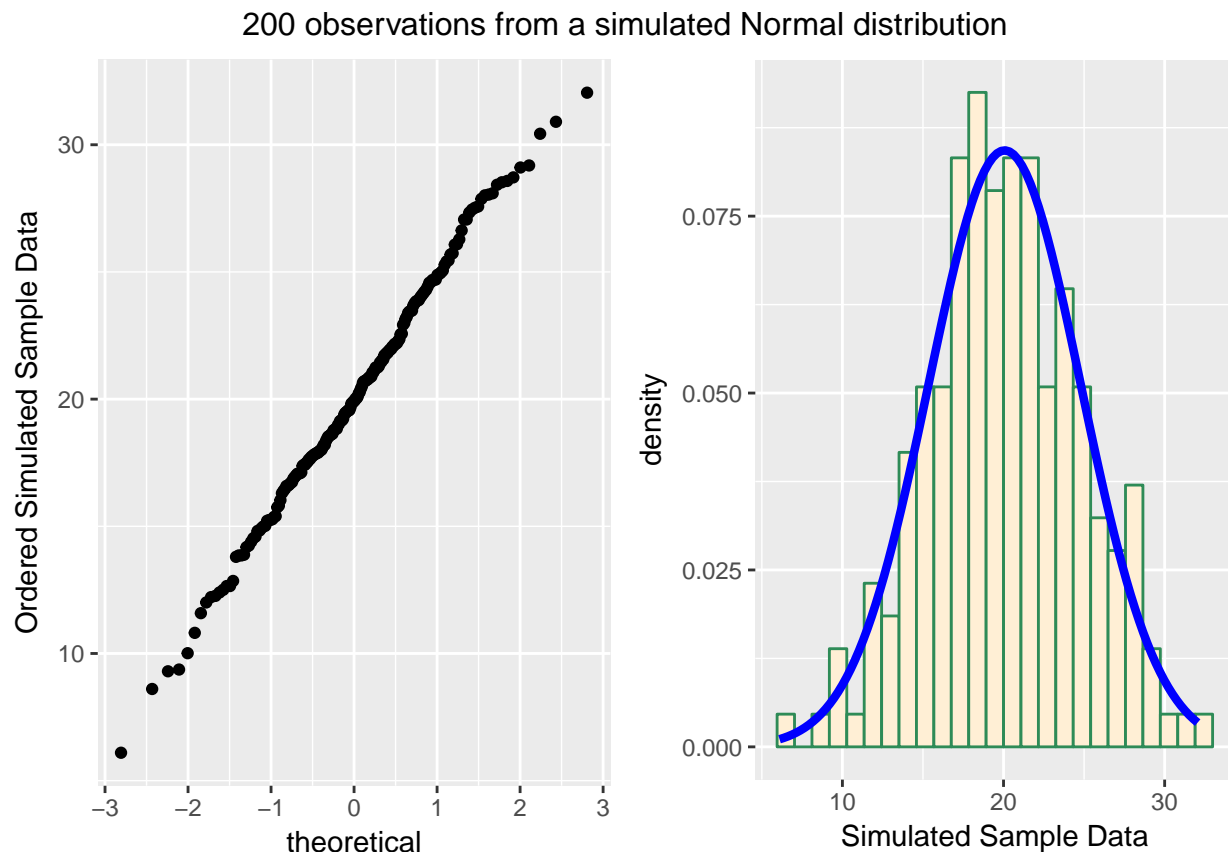
# simulate 200 observations from a Normal(20, 5) distribution and place them
# in the d variable within the temp.1 data frame
temp.1 <- data.frame(d = rnorm(200, mean = 20, sd = 5))

# left plot - basic Normal Q-Q plot of simulated data
p1 <- ggplot(temp.1, aes(sample = d)) +
```

```
geom_point(stat="qq") +
labs(y = "Ordered Simulated Sample Data")

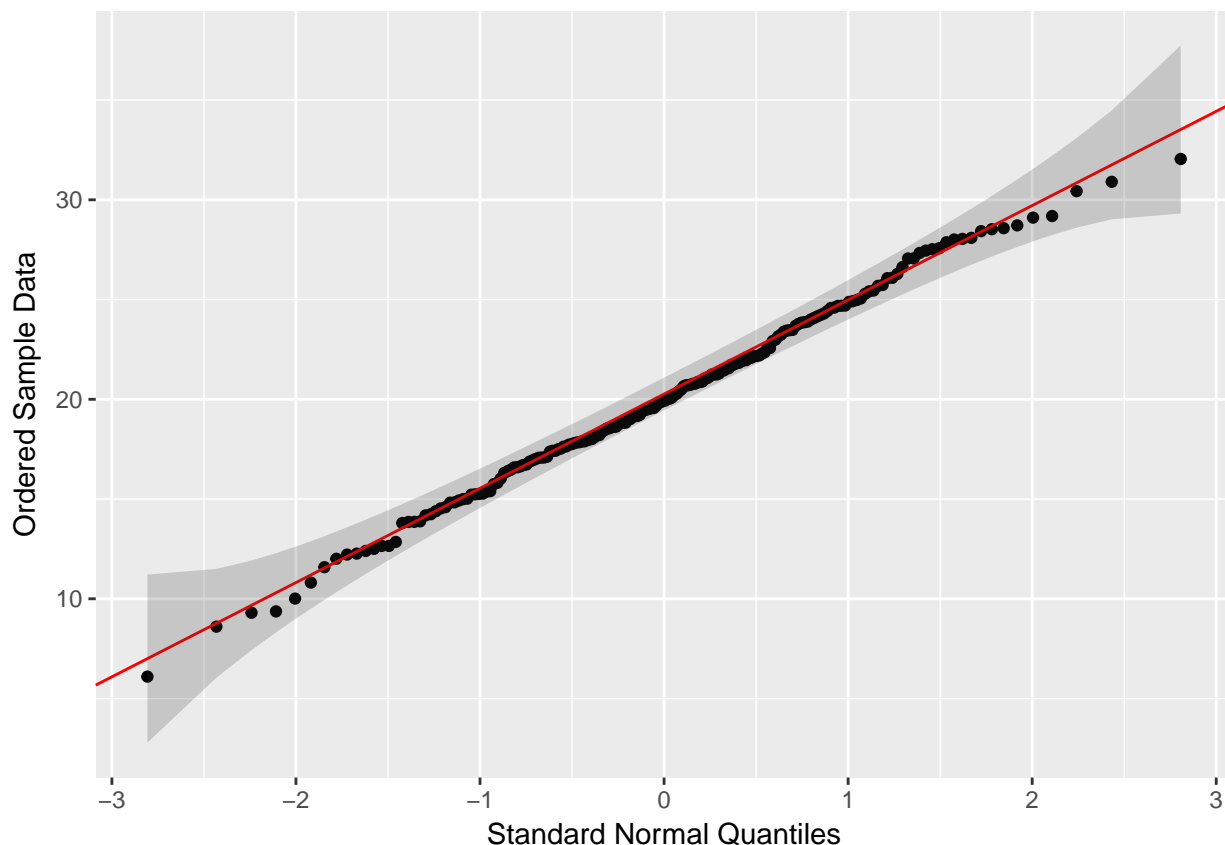
# right plot - histogram with superimposed normal distribution
p2 <- ggplot(temp.1, aes(x = d)) +
  geom_histogram(aes(y = ..density..),
    bins=25, fill = "papayawhip", color = "seagreen") +
  stat_function(fun = dnorm,
    args = list(mean = mean(temp.1$d),
      sd = sd(temp.1$d)),
    lwd = 1.5, col = "blue") +
  labs(x = "Simulated Sample Data")

gridExtra::grid.arrange(p1, p2, ncol=2,
  top = "200 observations from a simulated Normal distribution")
```



And here is another look at that same simulated data...

```
gg_qq(temp.1$d)
```



So, what are the characteristics of this simulation? The data appear to be well-modeled by the Normal distribution, because: - the points on the Normal Q-Q plot follow a straight line, in particular - there is no substantial curve (such as we'd see with data that were skewed) - there is no particularly surprising behavior (curves away from the line) at either tail, so there's no obvious problem with outliers

8.6.2 Skew is indicated by monotonic curves in the Normal Q-Q plot

Data that come from a skewed distribution appear to curve away from a straight line in the Q-Q plot.

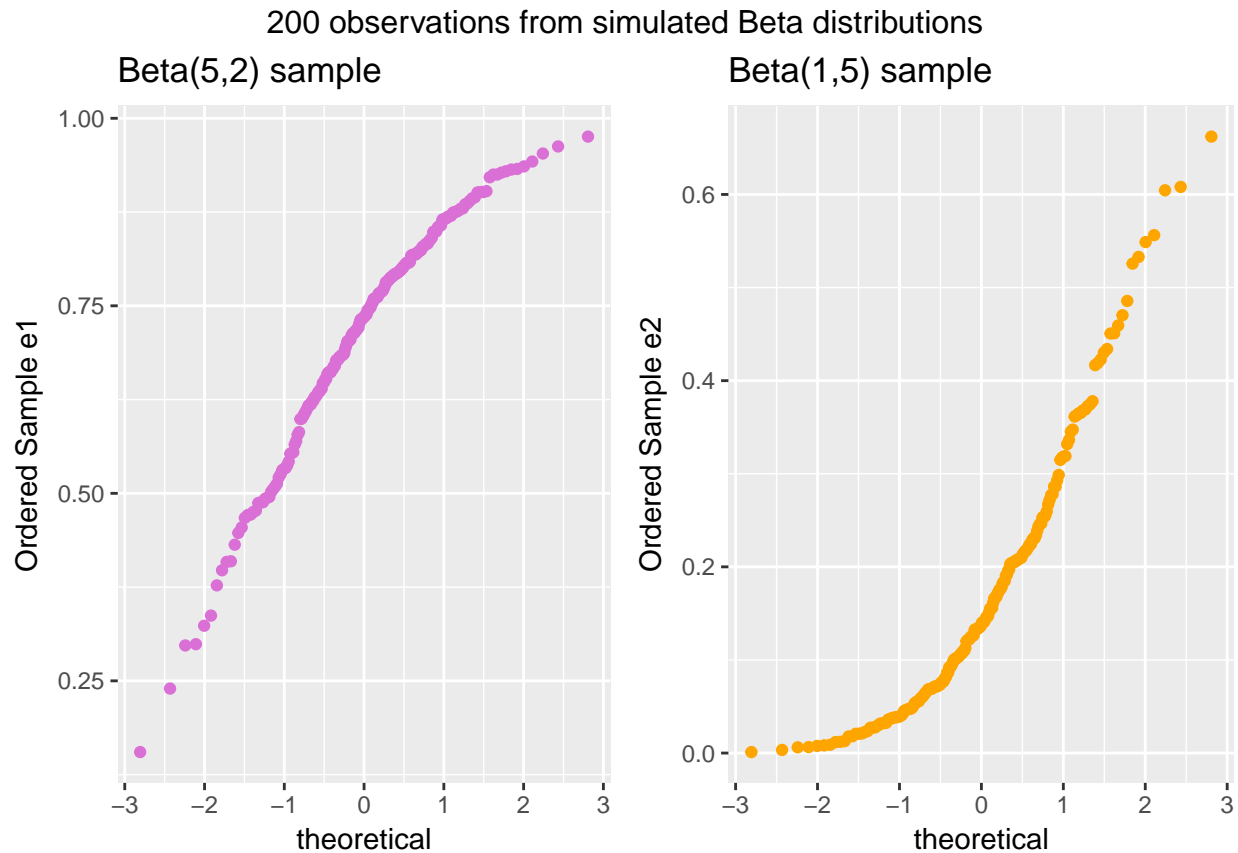
```
set.seed(123431) # so the results can be replicated

# simulate 200 observations from a beta(5, 2) distribution into the e1 variable
# simulate 200 observations from a beta(1, 5) distribution into the e2 variable
temp.2 <- data.frame(e1 = rbeta(200, 5, 2), e2 = rbeta(200, 1, 5))

p1 <- ggplot(temp.2, aes(sample = e1)) +
  geom_point(stat="qq", color = "orchid") +
  labs(y = "Ordered Sample e1", title = "Beta(5,2) sample")

p2 <- ggplot(temp.2, aes(sample = e2)) +
  geom_point(stat="qq", color = "orange") +
  labs(y = "Ordered Sample e2", title = "Beta(1,5) sample")

gridExtra::grid.arrange(p1, p2, ncol=2, top = "200 observations from simulated Beta distributions")
```

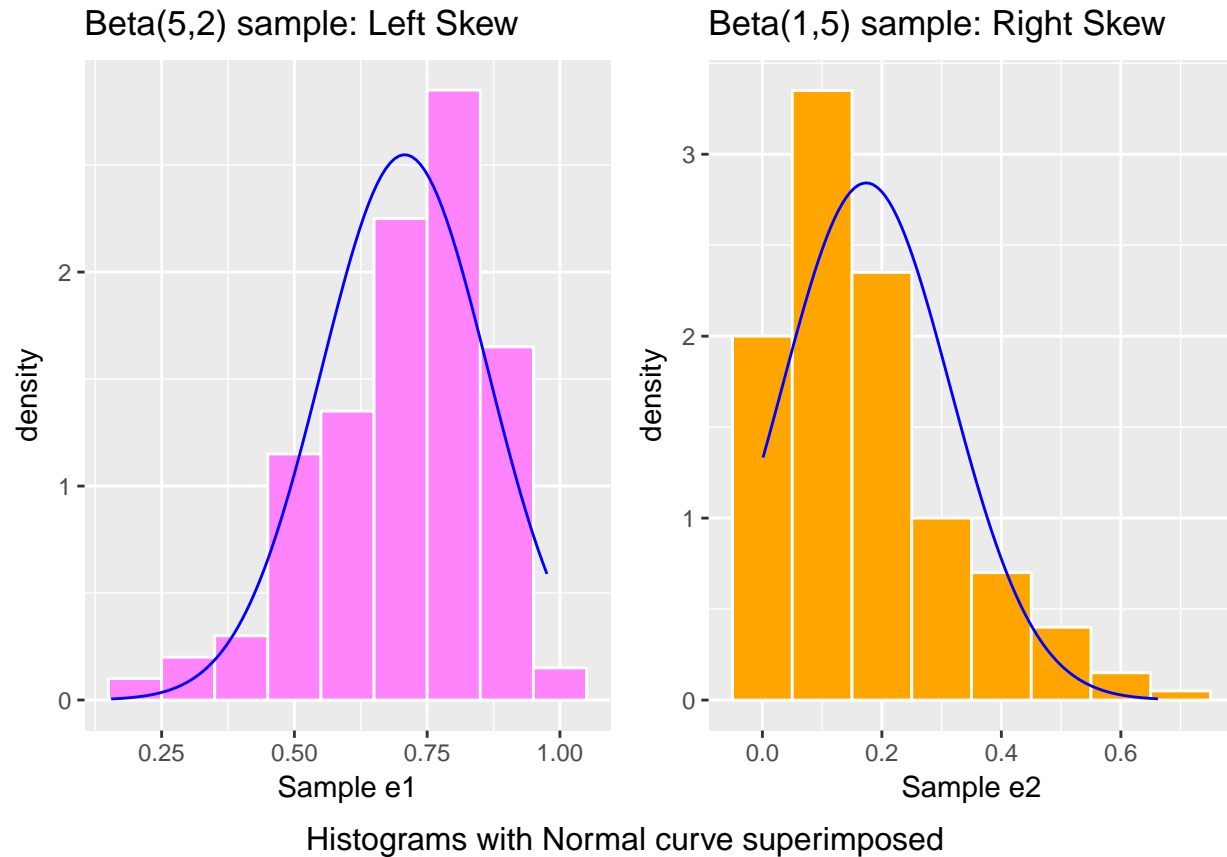



Note the bends away from a straight line in each sample. The non-Normality may be easier to see in a histogram.

```
p1 <- ggplot(temp.2, aes(x = e1)) +
  geom_histogram(aes(y = ..density..),
    binwidth=0.1, fill = "orchid1", color = "white") +
  stat_function(fun = dnorm,
    args = list(mean = mean(temp.2$e1),
      sd = sd(temp.2$e1)),
    col = "blue") +
  labs(x = "Sample e1", title = "Beta(5,2) sample: Left Skew")

p2 <- ggplot(temp.2, aes(x = e2)) +
  geom_histogram(aes(y = ..density..),
    binwidth=0.1, fill = "orange1", color = "white") +
  stat_function(fun = dnorm,
    args = list(mean = mean(temp.2$e2),
      sd = sd(temp.2$e2)),
    col = "blue") +
  labs(x = "Sample e2", title = "Beta(1,5) sample: Right Skew")

gridExtra::grid.arrange(p1, p2, ncol=2,
  bottom = "Histograms with Normal curve superimposed")
```



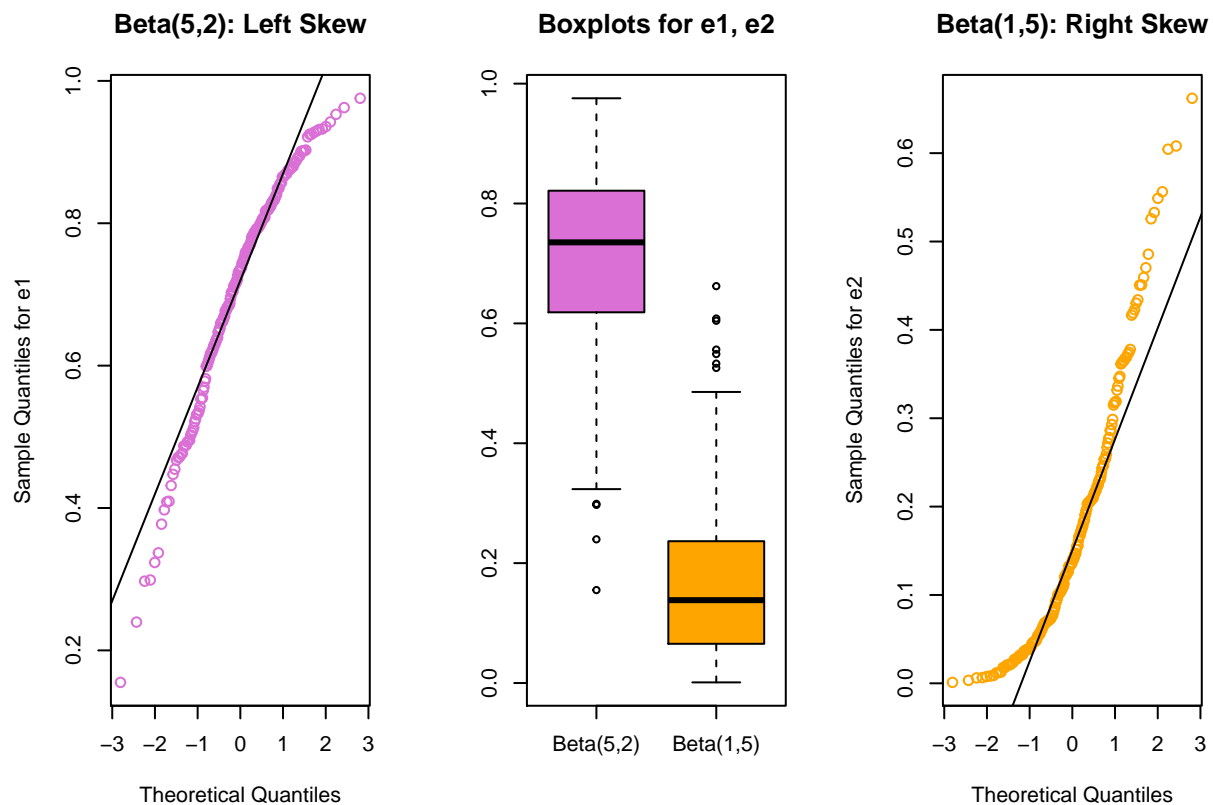
8.6.3 Direction of Skew

In each of these pairs of plots, we see the same basic result.

- The left plot (for data e1) shows left skew, with a longer tail on the left hand side and more clustered data at the right end of the distribution.
- The right plot (for data e2) shows right skew, with a longer tail on the right hand side, the mean larger than the median, and more clustered data at the left end of the distribution.

You may want to see the lines to help you see what's happening in the Q-Q plots. You can do this with our fancy approach, or with the qqnorm-qqline combination from base R.

```
par(mfrow=c(1,3))
qqnorm(temp.2$e1, col="orchid", main="Beta(5,2): Left Skew",
       ylab="Sample Quantiles for e1")
qqline(temp.2$e1)
boxplot(temp.2$e1, temp.2$e2, names=c("Beta(5,2)", "Beta(1,5)"),
       col=c("orchid", "orange"), main="Boxplots for e1, e2")
qqnorm(temp.2$e2, col="orange", main="Beta(1,5): Right Skew",
       ylab="Sample Quantiles for e2")
qqline(temp.2$e2)
```



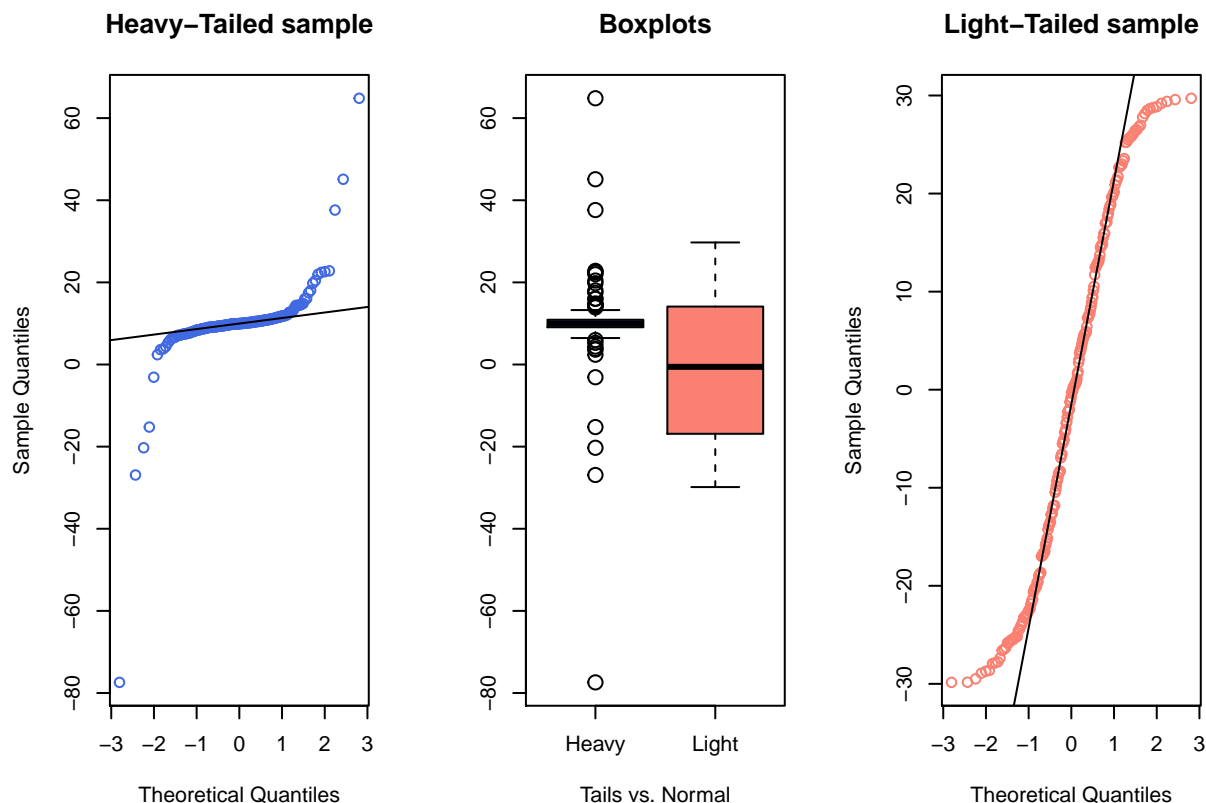
```
par(mfrow=c(1,1))
```

8.6.4 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, as shown on the left below.
- Light-tailed but symmetric distributions are indicated by “S” shapes in the plot, as shown on the right below.

```
set.seed(4311)
# sample 200 observations from each of two probability distributions
samp1 <- rcauchy(200, location=10, scale = 1) # use a Cauchy distribution
samp2 <- runif(200, -30, 30) # a uniform distribution on (-30, 30)

par(mfrow=c(1,3)) ## set up plot window for one row, three columns
qqnorm(samp1, col="royalblue", main="Heavy-Tailed sample")
qqline(samp1)
boxplot(samp1, samp2, names=c("Heavy", "Light"), cex=1.5,
        col=c("royalblue", "salmon"), main="Boxplots",
        xlab="Tails vs. Normal")
qqnorm(samp2, col="salmon", main="Light-Tailed sample")
qqline(samp2)
```



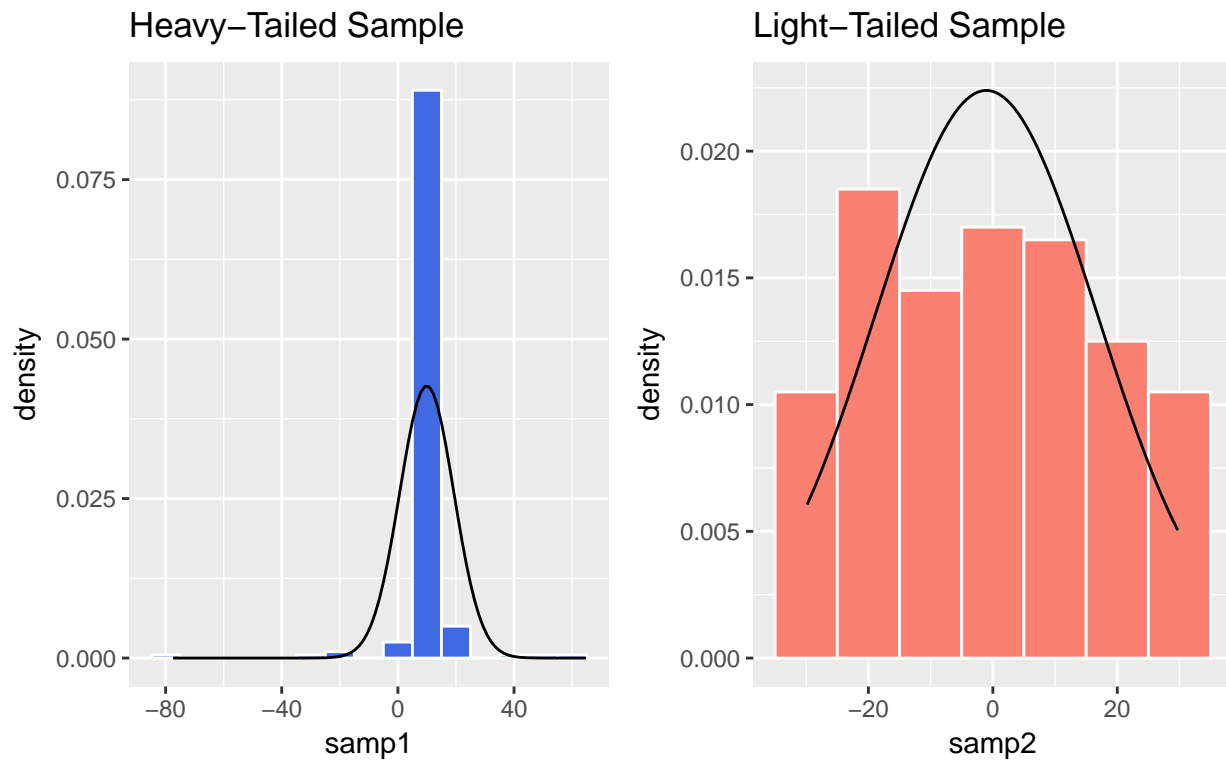
```
par(mfrow=c(1,1)) ## return to usual plot window
```

And, we can verify these initial conclusions with histograms.

```
temp.3 <- data.frame(samp1, samp2)
p1 <- ggplot(temp.3, aes(x = samp1)) +
  geom_histogram(aes(y = ..density..),
    binwidth=10, fill = "royalblue", color = "white") +
  stat_function(fun = dnorm, args = list(mean = mean(temp.3$samp1),
    sd = sd(temp.3$samp1))) +
  labs(x = "samp1", title = "Heavy-Tailed Sample")

p2 <- ggplot(temp.3, aes(x = samp2)) +
  geom_histogram(aes(y = ..density..),
    binwidth=10, fill = "salmon", color = "white") +
  stat_function(fun = dnorm, args = list(mean = mean(temp.3$samp2),
    sd = sd(temp.3$samp2))) +
  labs(x = "samp2", title = "Light-Tailed Sample")

gridExtra::grid.arrange(p1, p2, ncol=2,
  bottom = "Heavy- and light-tailed distributions
  with superimposed Normal models")
```



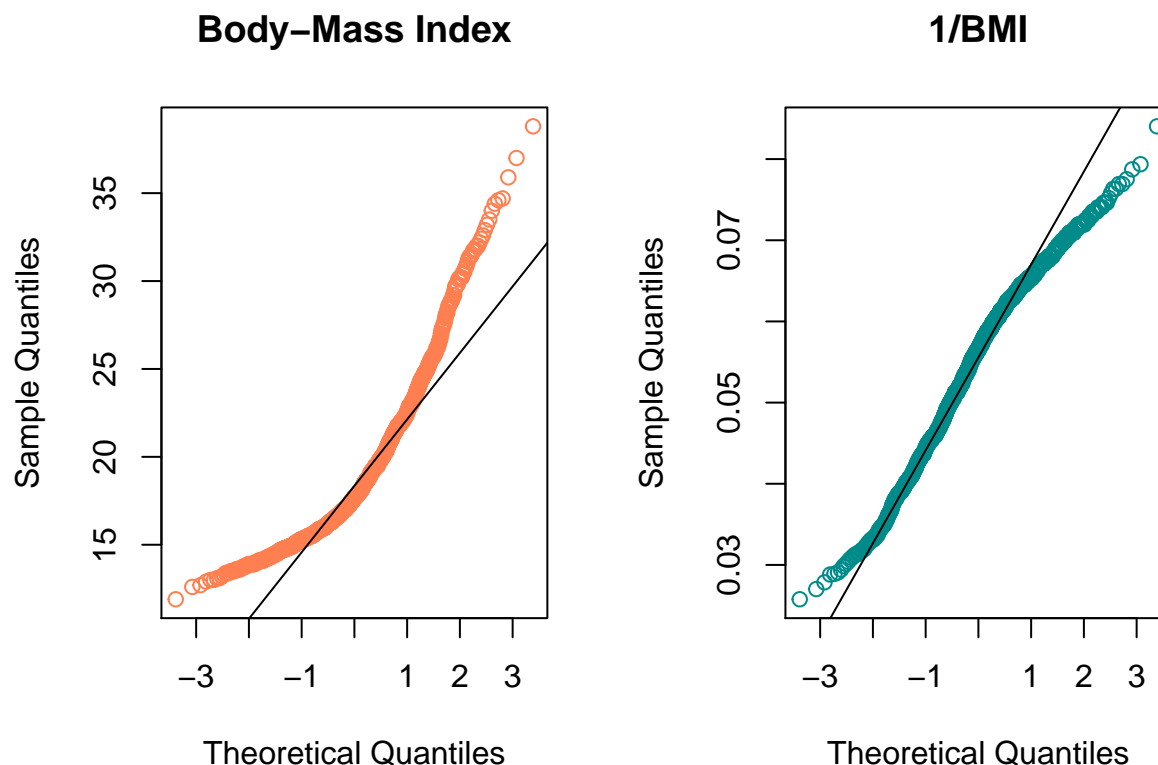
Heavy- and light-tailed distributions
with superimposed Normal models

```
rm(samp1, samp2, temp.1, temp.2, temp.3, p1, p2) # cleaning up
```

8.7 Does a Normal Distribution Fit the nyfs1 Data Well?

- Skewness is indicated by curves in the Normal Q-Q plot. Compare these two plots - the left is the original BMI data from the NYFS data frame, and the right plot shows the inverse of those values.

```
par(mfrow=c(1,2)) ## set up plot window for one row, two columns
qqnorm(nyfs1$bmi, main="Body-Mass Index", col="coral")
qqline(nyfs1$bmi)
qqnorm(1/(nyfs1$bmi), main="1/BMI", col="darkcyan")
qqline(1/nyfs1$bmi)
```



```
par(mfrow=c(1,1)) ## return to usual plot window
```

- The left plot shows fairly substantial **right** or *positive* skew
- The right plot shows there's much less skew after the inverse has been taken.
- Our conclusion is that a Normal model is a far better fit to 1/BMI than it is to BMI.

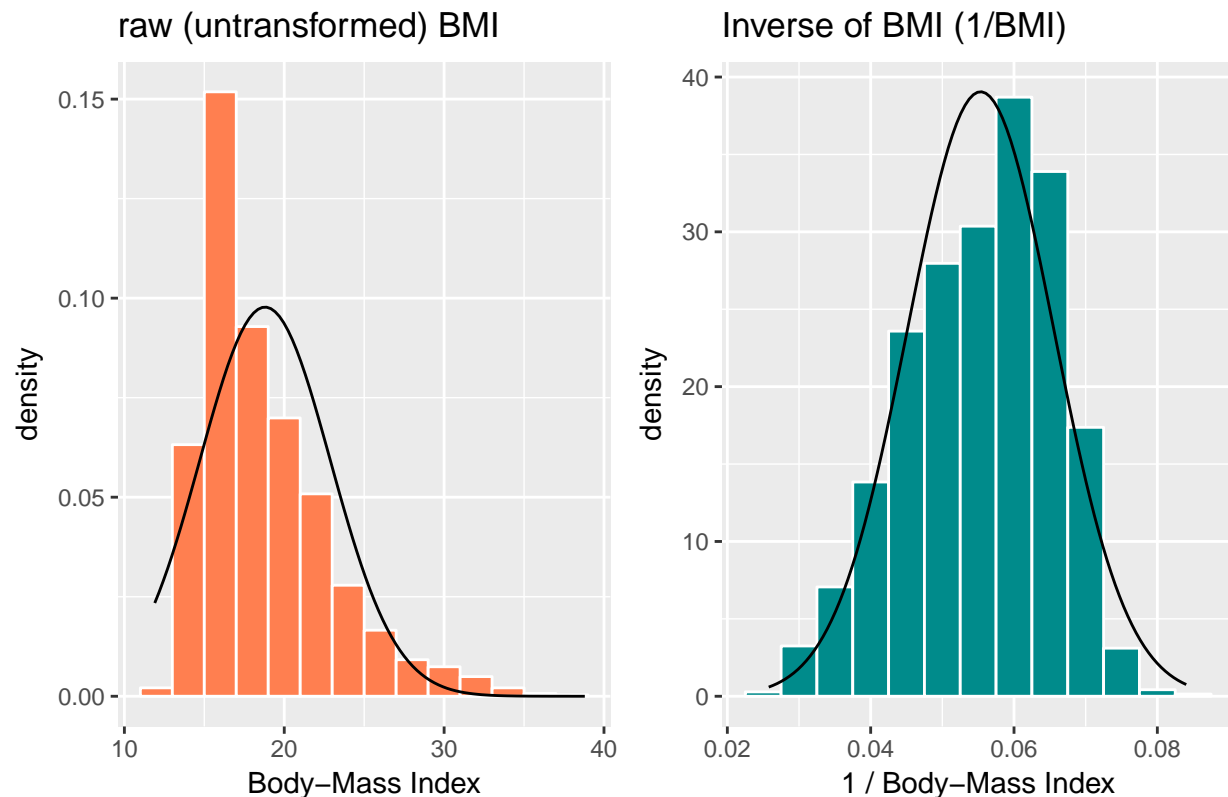
The effect of taking the inverse here may be clearer from the histograms below, with Normal density functions superimposed.

```
p1 <- ggplot(nyfs1, aes(x = bmi)) +
  geom_histogram(aes(y = ..density..),
    binwidth=2, fill = "coral", color = "white") +
  stat_function(fun = dnorm,
    args = list(mean = mean(nyfs1$bmi), sd = sd(nyfs1$bmi))) +
  labs(x = "Body-Mass Index", title = "raw (untransformed) BMI")

p2 <- ggplot(nyfs1, aes(x = 1/bmi)) +
  geom_histogram(aes(y = ..density..),
    binwidth=0.005, fill = "darkcyan", color = "white") +
  stat_function(fun = dnorm,
    args = list(mean = mean(1/nyfs1$bmi),
      sd = sd(1/nyfs1$bmi))) +
  labs(x = "1 / Body-Mass Index",
    title = "Inverse of BMI (1/BMI)")

gridExtra::grid.arrange(p1, p2, ncol=2,
  top = textGrob("Comparing BMI to 1/BMI", gp=gpar(fontsize=15)))
```

Comparing BMI to 1/BMI



```
# this approach to top label lets us adjust the size of type used
# in the main title
# note that you'll need to have called library(grid) or
# require(grid) for this to work properly
```

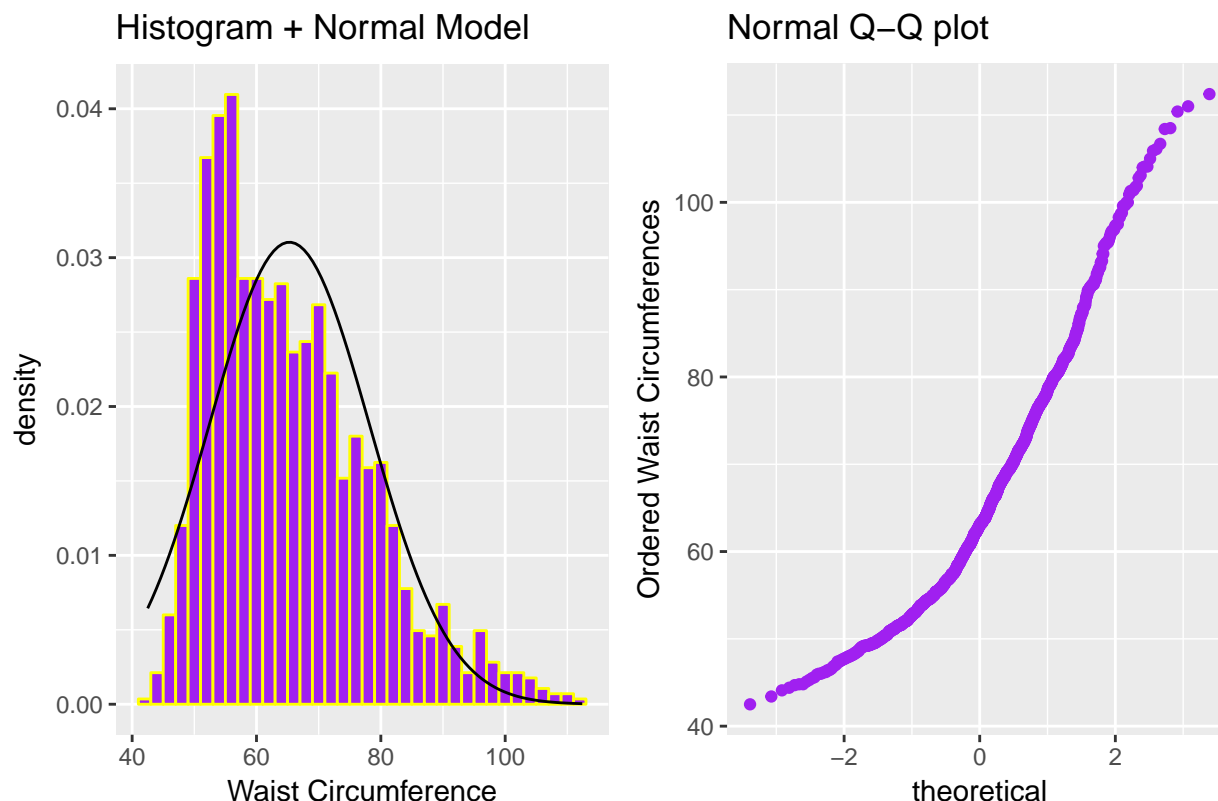
8.8 Are the Waist Circumferences Normally Distributed?

```
p1 <- ggplot(nyfs1, aes(x = waist.circ)) +
  geom_histogram(aes(y = ..density..),
    binwidth=2, fill = "purple", color = "yellow") +
  stat_function(fun = dnorm, args = list(mean = mean(nyfs1$waist.circ),
    sd = sd(nyfs1$waist.circ))) +
  labs(x = "Waist Circumference", title="Histogram + Normal Model")

p2 <- ggplot(nyfs1, aes(sample = waist.circ)) +
  geom_point(stat="qq", color = "purple") +
  labs(y = "Ordered Waist Circumferences", title="Normal Q-Q plot")

# this approach to top label lets us adjust
# the size and font (here bold) used in the main title
gridExtra::grid.arrange(p1, p2, ncol=2,
  top = textGrob("Waist Circumference in the nyfs1 data",
    gp=gpar(fontsize=15,font=2)))
```

Waist Circumference in the nyfs1 data



```
## clean up
rm(p1, p2)
```

Baumer, Benjamin S., Daniel T. Kaplan, and Nicholas J. Horton. 2017. *Modern Data Science with R*. Boca Raton, FL: CRC Press. <https://mdsr-book.github.io/>.

Bock, David E., Paul F. Velleman, and Richard D. De Veaux. 2004. *Stats: Modelling the World*. Boston MA: Pearson Addison-Wesley.

Cetinkaya-Rundel, Mine. 2017. "Teaching Data Science to New useRs." bit.ly/user2017.

Fox, John, and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Second. Thousand Oaks CA: Sage. <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel-Hierarchical Models*. New York: Cambridge University Press. <http://www.stat.columbia.edu/~gelman/arm/>.

Gelman, Andrew, and Deborah Nolan. 2017. *Teaching Statistics: A Bag of Tricks*. Second. Oxford, UK: Oxford University Press.

Grolemund, Garrett, and Hadley Wickham. 2017. *R for Data Science*. O'Reilly. <http://r4ds.had.co.nz/>.

Harrell, Frank E., and James C. Slaughter. 2017. *Biostatistics for Biomedical Research*. Vanderbilt University School of Medicine. biostat.mc.vanderbilt.edu/ClinStat.

Ismay, Chester, and Albert Y. Kim. 2017. *ModernDive: An Introduction to Statistical and Data Sciences via R*. <http://moderndive.com/>.

Norman, Geoffrey R., and David L. Streiner. 2014. *Biostatistics: The Bare Essentials*. Fourth. People's

Medical Publishing House.

Vittinghoff, Eric, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Second. Springer-Verlag, Inc. <http://www.biostat.ucsf.edu/vgsm/>.