Statistics Using Technology

Third Edition

Kathryn Kozak

2021-07-24

# Preface

I hope you find this book useful in teaching statistics. When writing this book, I tried to follow the GAISE Standards (GAISE recommendations. (2014, January 05). Retrieved from <http://www.amstat.org/education/gaise/GAISECollege_Recommendations.pdf>

* Teach statistical thinking.
* Focus on conceptual understanding.
* Integrate real data with a context and a purpose.
* Foster active learning.
* Use technology to explore concepts and analyze data.
* Use assessments to improve and evaluate student learning

To this end, I ask students to interpret the results of their calculations. I incorporated the use of technology (R Studio) for most calculations. Because of that you will not find me using any of the computational formulas for standard deviations or correlation and regression since I prefer students understand the concept of these quantities. Also, because I utilize technology you will not find the standard normal table, Student’s t-table, binomial table, chi-square distribution table, and F-distribution table in the book. Another difference between this book and other statistics books is the order of hypothesis testing and confidence intervals. Most books present confidence intervals first and then hypothesis tests. I find that presenting hypothesis testing first and then confidence intervals is more understandable for students. Lastly, I have de-emphasized the use of the z-test. In fact, I only use it to introduce hypothesis testing, and never utilize it again. Two samples should be emphasised over one sample test. Lastly, to aid student understanding and interest, most of the homework and examples utilize real data with multiple variables. The beauty of multiple variables, is that you can ask the students to investigate different analysis with different variables. This way students can work with data and come up with connections of asking questions and using data to answer the questions. Again, I hope you find this book useful for your introductory statistics class.

I want to make a comment about the mathematical knowledge that I assumed the students possess. The course for which I wrote this book has a higher prerequisite than most introductory statistics books. However, I do feel that students can read and understand this book as long as they can read critically. I do not show how to create most of the graphs, but all graphs are created with R Studio. So I hope the mathematical level is appropriate for your course.

The technology that I utilized for creating the graphs and statistical analysis is R Studio. This is a statistical software that are used by statisticians and so using it gives students skills they may need in the future. Please feel free to use any other technology that is more appropriate for your students. Do make sure that you use some technology.

## Acknowledgments:

I would like to thank the following people for taking their valuable time to review the book. Their comments and insights improved this book immensely.

* Daniel Kaplan, Macalester College
* Jane Tanner, Onondaga Community College
* Rob Farinelli, College of Southern Maryland
* Carrie Kinnison, retired engineer
* Sean Simpson, Westchester Community College
* Kim Sonier, Coconino Community College
* Jim Ham, Delta College
* David Straayer, Tacoma Community College
* Kendra Feinstein, Tacoma Community College
* Students of Coconino Community College
* Students of Tacoma Community College

I also want to thank Coconino Community College for granting me a sabbatical so that I would have the time to write the book. On a personal note, I wanted to thank my brother, John Matic, his wife Jenelle, and their children Hannah and Eli for their hospitality when writing the first edition. In addition to allowing my family access to their home, John provided numerous examples and data sets for business applications in this book. I inadvertently left this thank you out of the first edition of the book, and for that I apologize. His help and his family’s hospitality were invaluable to me. Lastly, I want to thank my husband Rich and my son Dylan for supporting me in this project. Without their love and support, I would not have been able to complete the book.

## New to the Third Edition:

The additions to this edition mostly involve adding the commands to create graphs, compute descriptive statistics, finding probabilities, and computing inferential analysis using the open source software R Studio, and the removal of all other technologies. Data Frames with multiple variables and multiple units of measurements were expanded to most of the data. This is to make the course more data-centric. Lastly, minor explanations were made and corrections were made where necessary.

## Packages needed for R Studio:

You will need the following packages installed and loaded in R Studio: arm, mosaic, MASS, Weighted.Desc.Stat.

# Graphical Descriptions of Data

In chapter 1, you were introduced to the concepts of population, which again is a collection of all the measurements from the individuals of interest. Remember, in most cases you can’t collect the entire population, so you have to take a sample. Thus, you collect data either through a sample or a census. Now you have a large number of data values. What can you do with them? No one likes to look at just a set of numbers. One thing is to organize the data into a table or graph. Ultimately though, you want to be able to use that graph to interpret the data, to describe the distribution of the data set, and to explore different characteristics of the data. The characteristics that will be discussed in this chapter and the next chapter are:

1. Center: middle of the data set, also known as the average.
2. Variation: how much the data varies.
3. Distribution: shape of the data (symmetric, uniform, or skewed).
4. Qualitative data: analysis of the data
5. Outliers: data values that are far from the majority of the data.
6. Time: changing characteristics of the data over time.

This chapter will focus mostly on using the graphs to understand aspects of the data, and not as much on how to create the graphs. There is technology that will create most of the graphs, though it is important for you to understand the basics of how to create them.

This textbook uses R Studio to perform all graphical and descriptive statistics, and all statistical inference. When using R Studio, every command is performed the same way. You start off with a goal(explanatory variable ~ response variable, data=data frame\_name,…)

R Studio uses packages to make calculations easier. For this textbook, you will mostly need the package mosaic. There will be others that you will need on occasion, but you will be told that at the time. Most likely, mosaic is already installed in your R Studio. If you wish to install other packages you use the command

install.packages("name of package")

where you replace the name of package with the package you wish to install.

Once the package is installed, then you will need to tell R Studio you want to use it every time you start R Studio. The command to tell R Studio you want to use a package is

library("name of package")

You will need to turn on the package mosaic. The NHANES package contains a data frame that is useful. Both are accessed by doing.

library("mosaic")  
library("NHANES")

Back to the basic command

goal(explanatory variable ~ response variable, data=data frame\_name,…)

The goal depends on what you want to do. If you want to create a graph then you would need

gf\_graphtype(explanatory variable ~   
response variable, data=dataframe\_name, ...)

As an example if you want to create a density plot of cholesterol levels on day 2 from a dataframe called Cholesterol, then your command would be

gf\_density(~day2, data=Cholesterol)

You will see more on what the different commands are that you would use. A word about the … at the end of the command. That means there are other things you can do, but that is up to you if you want to actually do them. They do not need to be used if you don’t want to. The following sections will show you how to create the different graphs that are usually completed in an introductory statistics course.

## Qualitative Data

Remember, qualitative data are words describing a characteristic of the individual. There are several different graphs that are used for qualitative data. These graphs include bar graphs, Pareto charts, and pie charts. Bar graphs can be created using a statistical program like R Studio.

**Bar graphs or charts** consist of the frequencies on one axis and the categories on the other axis. Drawing the bar graph using R is performed using the following command.

gf\_bar(~explanatory variable, data=Dataframe)

### Example: Drawing a Bar Chart\*\*

Data was collected for two semesters in a statistics class. The data frame in is the table #2.1.1. The command

head(data frame)

shows the variables and the first few lines of the data set.

**Table #2.1.1: Statistics class survey**

Class<-read.csv(  
 "https://krkozak.github.io/MAT160/class\_survey.csv")  
head(Class)

## vehicle gender distance\_campus ice\_cream rent  
## 1 None Female 1.5 Cookie Dough 724  
## 2 Mercury Female 14.7 Sherbet 200  
## 3 Ford Female 2.4 Chocolate Brownie. 600  
## 4 Toyota Female 5.2 coffee 0  
## 5 Jeep Male 2.0 Cookie Dough 600  
## 6 Subaru Male 5.0 none 500  
## major height winter  
## 1 Environmental and Sustainability Studies 61 Liked it  
## 2 Administrative Justice 60 Don't like it  
## 3 Bio Chem 68 Liked it  
## 4 66 Loved it  
## 5 Pre-health Careers 71 Loved it  
## 6 Finance 72 No opinion

Every data frame has a code book that describes the data set, the source of the data set, and a listing and description of the variables in the data frame.

**Code book for Data Frame Class**

**Description** Survey results from two semesters of statistics classes at Coconino Community College in the years 2018-2019.

**Format**

This data frame contains the following columns:

vehicle: Type of car a student drives

gender: Self declared gender of a student

distance\_campus: how far a student lives from the Lone Tree Campus of Coconino Community College (miles)

ice\_cream: favorite ice cream flavor

rent: How much a student pays in rent

major: Students declared major

height: height of the student (inches)

winter: Student’s opinion of winter (Love it, Like it, Don’t like, No opinion)

**Source**

Kozak K (2019). Survey results form surveys collected in statistics class at Coconino Community College.

**References**

Kozak, 2019

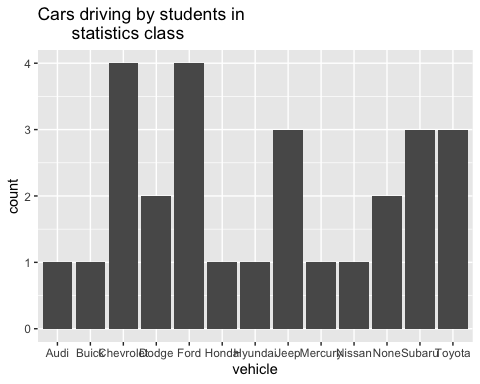
Create a bar graph of vehicle type. To do this in R Studio, use the command

gf\_bar(~variable, data=DataFrame, ...)

where gf\_bar is the goal, vehicle is the name of the response variable (there is no explanatory variable), the dataframe is Class, and a title was added to the graph.

(ref:class-data-bar-cap) Bar Graph for Type of Car Data

gf\_bar(~vehicle, data=Class, title="Cars driving by students in   
 statistics class")



(ref:class-data-bar-cap)

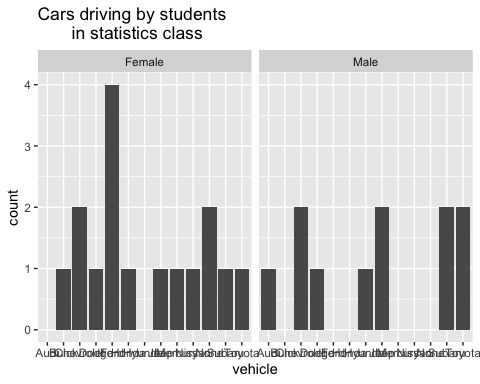
Notice from the graph (Figure @ref(fig:class-data-bar)), you can see that Chevrolet and Ford are the more popular car, with Jeep, Subaru, and Toyota not far behind. Many types seems to be the lesser used, and tied for last place. However, more data would help to figure this out.

* All graphs should have labels on each axis and a title for the graph.\*

The beauty of data frames with multiple variables is that you can answer many questions from the data. Suppose you want to see if gender makes a difference for the type of car a person drives. If you are a car manufacturer, if you knew that certain genders like certain cars, then you would advertise to the different genders. To create a bar graph that separates based on gender, perform the following command in R Studio.

(ref:class-data-bar-gender-cap) Bar Graph for Type of Car Data

gf\_bar(~vehicle|gender, data=Class, title="Cars driving by students   
 in statistics class")



(ref:class-data-bar-gender-cap)

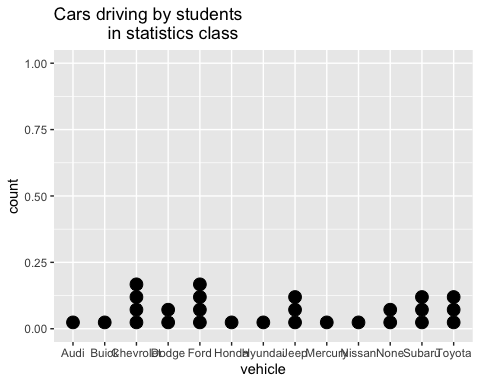
Notice a Ford is driven by females more than any other car, while Chevrolet, Mercury, and Subaru cars are equally driven by males. Obviously a larger sample would be needed to make any conclusions from this data.

There are other types of graphs that can be created for quantitative variables. Another type is known as a dot plot. The command for this graph (Figure @ref(fig:class-data-dot-gender)) is as follows.

(ref:class-data-dot-gender-cap) Dot Plot for Type of Car Data

gf\_dotplot(~vehicle, data=Class, title="Cars driving by students   
 in statistics class")

## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.

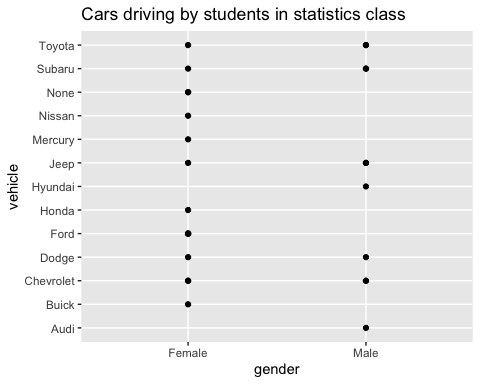


(ref:class-data-dot-gender-cap)

Notice a dot plot is like a bar chart. Both give you the same information. You can also divide a dot plot by gender. Another type of graph that is also useful and similar to the dot plot is a point plot (scatter plot). In this plot (Figure @ref(fig:car-data-gender)) you can graph the explanatory variable versus the response variable. The command for this in R Studio is as follows.

(ref:car-data-gender-cap) Point plot for Type of Car Data versus gender

gf\_point(vehicle~gender, data=Class,   
 title="Cars driving by students in statistics class")

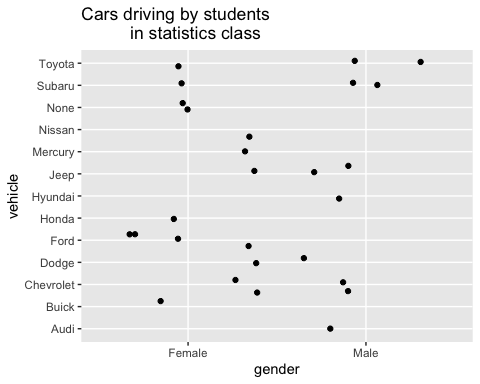


(ref:car-data-gender-cap)

The problem with this graph (Figure @ref(fig:car-data-gender)) is that if there are multiple females who drive a Ford, only one dot is shown. So it is best to spread the dots out using a plot known as a jitter plot. In a jitter plot the dots are randomly moved off the center line. The command for a jitter plot is as follows:

(ref:car-data-gender-jitter-cap) Jitter plot for Type of Car Data versus gender

gf\_jitter(vehicle~gender, data=Class, title="Cars driving by students   
 in statistics class")



(ref:car-data-gender-jitter-cap)

Now you can see (Figure @ref(fig:car-data-gender-jitter)) that there are 4 females who drive a Ford. There is one female who drives a Honda. Other information about other cars and genders can be seen better than in the point plot and the bar graph. Jitter plots are useful to see how many data values are for each qualitative data values.

There are many other types of graphs that can be used on qualitative data. There are spreadsheet software packages that will create most of them, and it is better to look at them to see how to create then. It depends on your data as to which may be useful, but the bar, dot, and jitter plots are really the most useful.

### Homework

1. Eyeglassomatic manufactures eyeglasses for different retailers. The number of lenses for different activities is in table #2.1.2.

**Table #2.1.2: Data for Eyeglassomatic**

Eyeglasses<-read.csv(  
 "https://krkozak.github.io/MAT160/eyglasses.csv")  
head(Eyeglasses)

## activity  
## 1 Grind  
## 2 Grind  
## 3 Grind  
## 4 Grind  
## 5 Grind  
## 6 Grind

**Code book for Data Frame Eyeglasses**

**Description** Activities that an Eyeglass company performs when making eyeglasses, Grind means ground the lenses and put them in frames, multicoat means put tinting or coatings on lenses and then put them in frames, assemble means received frames and lenses from other sources and put them together, make frames means made the frames and put lenses in from other sources, receive finished means received glasses from other source unknown means do not know where the lenses came from.

**Format**

This data frame contains the following columns:

activity: The activity that is completed to make the eyeglasses by Eyeglassomatic

**Source** John Matic provided the data from a company he worked with. The company’s name is fictitious, but the data is from an actual company.

**References** John Matic (2013)

Make a bar chart of this data. State any findings you can see from the graph.

1. Data was collected for two semesters in a statistics class drive. The data frame in is the table #2.1.3.

**Table #2.1.3 Data Frame of Statistics Class Survey**

Class<-read.csv(  
 "https://krkozak.github.io/MAT160/class\_survey.csv")  
head(Class)

## vehicle gender distance\_campus ice\_cream rent  
## 1 None Female 1.5 Cookie Dough 724  
## 2 Mercury Female 14.7 Sherbet 200  
## 3 Ford Female 2.4 Chocolate Brownie. 600  
## 4 Toyota Female 5.2 coffee 0  
## 5 Jeep Male 2.0 Cookie Dough 600  
## 6 Subaru Male 5.0 none 500  
## major height winter  
## 1 Environmental and Sustainability Studies 61 Liked it  
## 2 Administrative Justice 60 Don't like it  
## 3 Bio Chem 68 Liked it  
## 4 66 Loved it  
## 5 Pre-health Careers 71 Loved it  
## 6 Finance 72 No opinion

**Code book for Data Frame Class** see Example #2.1.1

Create a bar graph and dot plot of the variable ice cream. State any findings you can see from the graphs.

1. The number of deaths in the US due to carbon monoxide (CO) poisoning from generators from the years 1999 to 2011 are in table #2.1.4 (Hinatov, 2012). Create a bar chart of this data. State any findings you see from the graph.

**Table #2.1.4: Data of Number of Deaths Due to CO Poisoning**

Area<-read.csv(  
 "https://krkozak.github.io/MAT160/area.csv")  
head(Area)

## deaths  
## 1 Urban  
## 2 Urban  
## 3 Urban  
## 4 Urban  
## 5 Urban  
## 6 Urban

1. Data was collected for two semesters in a statistics class drive. The data frame in is the table #2.1.5. Create a bar graph and dot plot of the variable major. Create a jitter plot of major and gender. State any findings you can see from the graphs.

\*\*Table #2.1.5 Data Frame of Class Survey

Class<-read.csv(  
 "https://krkozak.github.io/MAT160/class\_survey.csv")  
head(Class)

## vehicle gender distance\_campus ice\_cream rent  
## 1 None Female 1.5 Cookie Dough 724  
## 2 Mercury Female 14.7 Sherbet 200  
## 3 Ford Female 2.4 Chocolate Brownie. 600  
## 4 Toyota Female 5.2 coffee 0  
## 5 Jeep Male 2.0 Cookie Dough 600  
## 6 Subaru Male 5.0 none 500  
## major height winter  
## 1 Environmental and Sustainability Studies 61 Liked it  
## 2 Administrative Justice 60 Don't like it  
## 3 Bio Chem 68 Liked it  
## 4 66 Loved it  
## 5 Pre-health Careers 71 Loved it  
## 6 Finance 72 No opinion

**Code book for Data Frame Class** see Example #2.1.1

1. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made during the time period of January 1 to March 31. Table #2.1.6 gives the defect and the number of defects. Create a bar chart of the data and then describe what this tells you about what causes the most defects.

**Table #2.1.6: Data of Defect Type**

Defects<- read.csv(  
 "https://krkozak.github.io/MAT160/defects.csv")   
head(Defects)

## type  
## 1 small  
## 2 small  
## 3 pd  
## 4 flaked  
## 5 scratch  
## 6 spot

**Code book for Data Frame Defects**

**Description** Types of defects that an Eyeglass company sees in the lenses they make into eyeglasses.

**Format**

This data frame contains the following columns:

type: The type of defect that is Seen when making eyeglasses by Eyeglassomatic

**Source** John Matic provided the data from a company he worked with. The company’s name is fictitious, but the data is from an actual company.

**References** John Matic (2013)

1. American National Health and Nutrition Examination (NHANES) surveys is collected every year by the US National Center for Health Statistics (NCHS). The data frame is in table #2.1.7. Create a bar chart of MartialStatus. Create a jitter plot of MaritalStatus versus Education. Describe any findings from the graphs.

**Table #2.1.7: Data Frame NHANES**

head(NHANES)

## # A tibble: 6 × 76  
## ID SurveyYr Gender Age AgeDecade AgeMonths Race1 Race3 Education   
## <int> <fct> <fct> <int> <fct> <int> <fct> <fct> <fct>   
## 1 51624 2009\_10 male 34 " 30-39" 409 White <NA> High School   
## 2 51624 2009\_10 male 34 " 30-39" 409 White <NA> High School   
## 3 51624 2009\_10 male 34 " 30-39" 409 White <NA> High School   
## 4 51625 2009\_10 male 4 " 0-9" 49 Other <NA> <NA>   
## 5 51630 2009\_10 female 49 " 40-49" 596 White <NA> Some College  
## 6 51638 2009\_10 male 9 " 0-9" 115 White <NA> <NA>   
## # … with 67 more variables: MaritalStatus <fct>, HHIncome <fct>,  
## # HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>, HomeOwn <fct>,  
## # Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>, Height <dbl>,  
## # BMI <dbl>, BMICatUnder20yrs <fct>, BMI\_WHO <fct>, Pulse <int>,  
## # BPSysAve <int>, BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>, BPSys2 <int>,  
## # BPDia2 <int>, BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>,  
## # DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>,  
## # UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>, DiabetesAge <int>,  
## # HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,  
## # LittleInterest <fct>, Depressed <fct>, nPregnancies <int>, nBabies <int>,  
## # Age1stBaby <int>, SleepHrsNight <int>, SleepTrouble <fct>,  
## # PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,  
## # TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,  
## # AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,  
## # Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,  
## # RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,  
## # SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>, SameSex <fct>,  
## # SexOrientation <fct>, PregnantNow <fct>

To view the code book for NHANES, type help(“NHANES”) in R Studio after you load the NHANES packages using library(“NHANES”)

## Quantitative Data

There are several different graphs for quantitative data. With quantitative data, you can talk about how the data is distributed, called a distribution. The shape of the distribution can be described from the graphs.

**Histogram**: a graph of frequencies (counts) on the vertical axis and classes on the horizontal axis. The height of the rectangles is the frequency and the width is the class width. The width depends on how many classes (bins) are in the histogram. The shape of a histogram is dependent on the number of bins. In R Studio the command to create a histogram is

gf\_histogram(~response variable, data=Data Frame, title="title   
of the graph")

The last part of the command puts a title on the graph. You type in what ever you want for the title in the quotes.

**Density Plot**: Similar to a histogram, except smoothing is created to smooth out the graph. The shape is not dependent on the number of bins so the distribution is easier to determine from the density plot. In R Studio the command to create a density plot is

gf\_density(~response variable, data=Data Frame, title="title of the graph")

The last part of the command puts a title on the graph. You type in what every you want for the title in the quotes.

**Dot Plot**: Dot plots can be created for both quantitative and qualitative variables. For smaller data frames, a dot plot can be useful to determine the shape of the distribution. The command in R Studio is

gf\_dotplot(~response variable, data=Data Frame, title="title   
of the graph")

The last part of the command puts a title on the graph. You type in what ever you want for the title in the quotes.

### Example: Drawing a Histogram and Density plot

Data was collected for two semesters in a statistics class drive.

**Table #2.2.1: Statistis class survey**

Class<-read.csv(  
 "https://krkozak.github.io/MAT160/class\_survey.csv")  
head(Class)

## vehicle gender distance\_campus ice\_cream rent  
## 1 None Female 1.5 Cookie Dough 724  
## 2 Mercury Female 14.7 Sherbet 200  
## 3 Ford Female 2.4 Chocolate Brownie. 600  
## 4 Toyota Female 5.2 coffee 0  
## 5 Jeep Male 2.0 Cookie Dough 600  
## 6 Subaru Male 5.0 none 500  
## major height winter  
## 1 Environmental and Sustainability Studies 61 Liked it  
## 2 Administrative Justice 60 Don't like it  
## 3 Bio Chem 68 Liked it  
## 4 66 Loved it  
## 5 Pre-health Careers 71 Loved it  
## 6 Finance 72 No opinion

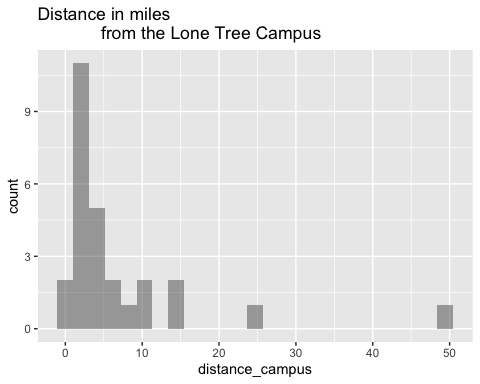
**Code book for Data Frame Class** See Example #2.1.1.

Draw a histogram, density plot, and a dot plot for the variable the distance a student lives from the Lone Tree Campus of Coconino Community College. Describe the story the graphs tell.

**Solution:**

(ref:distance-campus-hist-cap) Histogram of Distance a Student Lives from the Lone Tree Campus

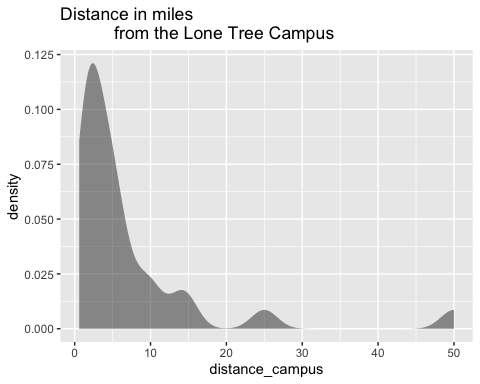
gf\_histogram(~distance\_campus, data=Class, title="Distance in miles   
 from the Lone Tree Campus")



(ref:distance-campus-hist-cap)

(ref:distance-campus-density-cap) Density plot of Distance a Student Lives from the Lone Tree Campus

gf\_density(~distance\_campus, data=Class, title="Distance in miles   
 from the Lone Tree Campus")

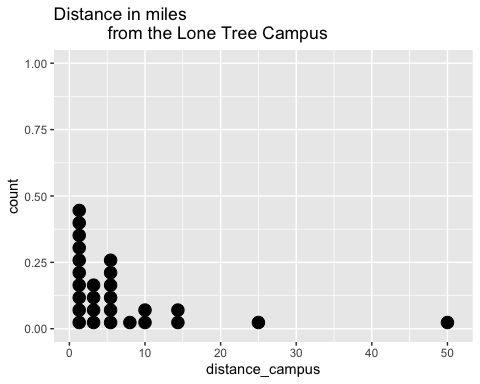


(ref:distance-campus-density-cap)

(ref:distance-campus-dot-cap) Dot Plot of Distance a Student Lives from the Lone Tree Campus

gf\_dotplot(~distance\_campus, data=Class, title="Distance in miles   
 from the Lone Tree Campus")

## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.



(ref:distance-campus-dot-cap)

Notice the histogram, density plot, and dot plot are all very similar, but the density plot is smother. They all tell you similar ideas of the shape of the distribution. Reviewing the graphs you can see that most of the students live within 10 miles of the Lone Tree Campus, in fact most live within 5 miles from the campus. However, there is a student who lives around 50 miles from the Lone Tree Campus. This is a great deal farther from the rest of the data. This value could be considered an outlier. An **outlier** is a data value that is far from the rest of the values. It may be an unusual value or a mistake. It is a data value that should be investigated. In this case, the student lived really far from campus, thus the value is not a mistake, and is just very unusual. The density plot is probably the best plot for most data frames.

There are other aspects that can be discussed, but first some other concepts need to be introduced.

\*\* Shapes of the distribution:\*\*

When you look at a distribution, look at the basic shape. There are some basic shapes that are seen in histograms. Realize though that some distributions have no shape. The common shapes are symmetric, skewed, and uniform. Another interest is how many peaks a graph may have. This is known as modal.

Symmetric means that you can fold the graph in half down the middle and the two sides will line up. You can think of the two sides as being mirror images of each other. Skewed means one “tail” of the graph is longer than the other. The graph is skewed in the direction of the longer tail (backwards from what you would expect). A uniform graph has all the bars the same height.

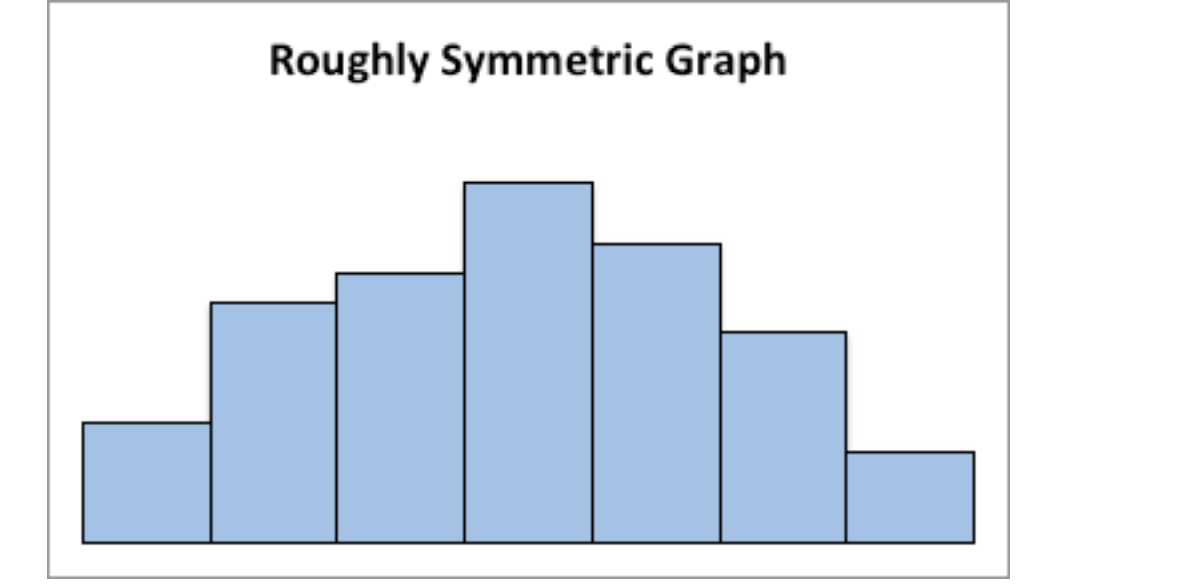
Modal refers to the number of peaks. Unimodal has one peak and bimodal has two peaks. Usually if a graph has more than two peaks, the modal information is not longer of interest.

Other important features to consider are gaps between bars, a repetitive pattern, how spread out is the data, and where the center of the graph is.

**Examples of graphs:**

This graph is roughly symmetric and unimodal:

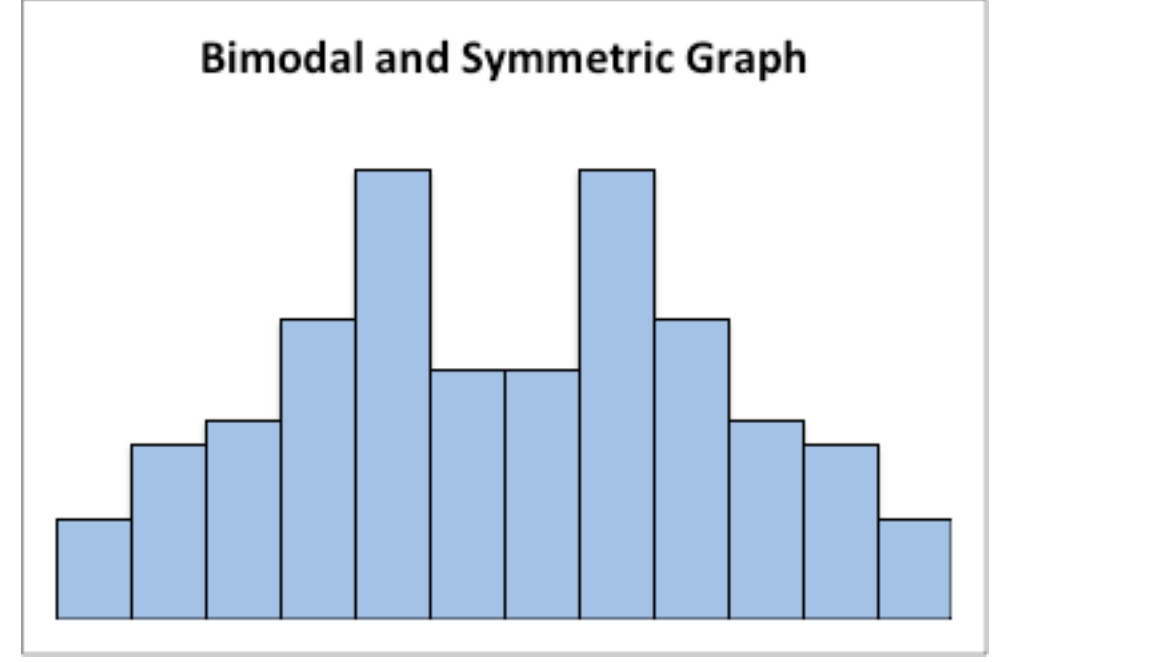
**Graph #.2.1: Symmetric Distribution**



Graph of roughly symmetric graph

This graph is symmetric and bimodal:

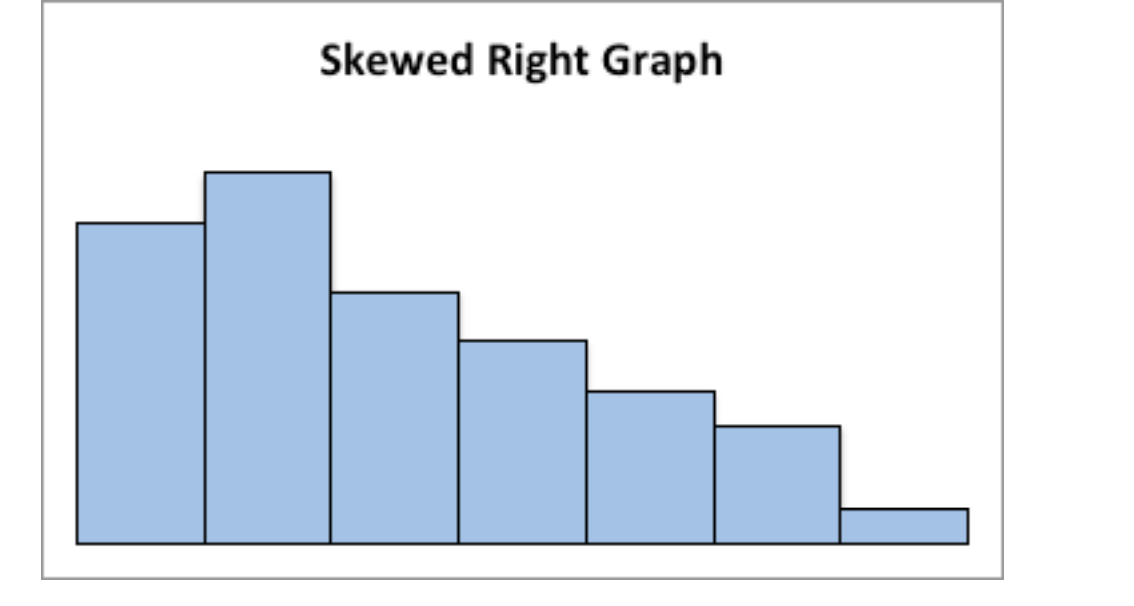
**Graph #2.2.2: Symmetric and Bimodal Distribution**



Graph of symmetric and bimodal graph

This graph is skewed to the right:

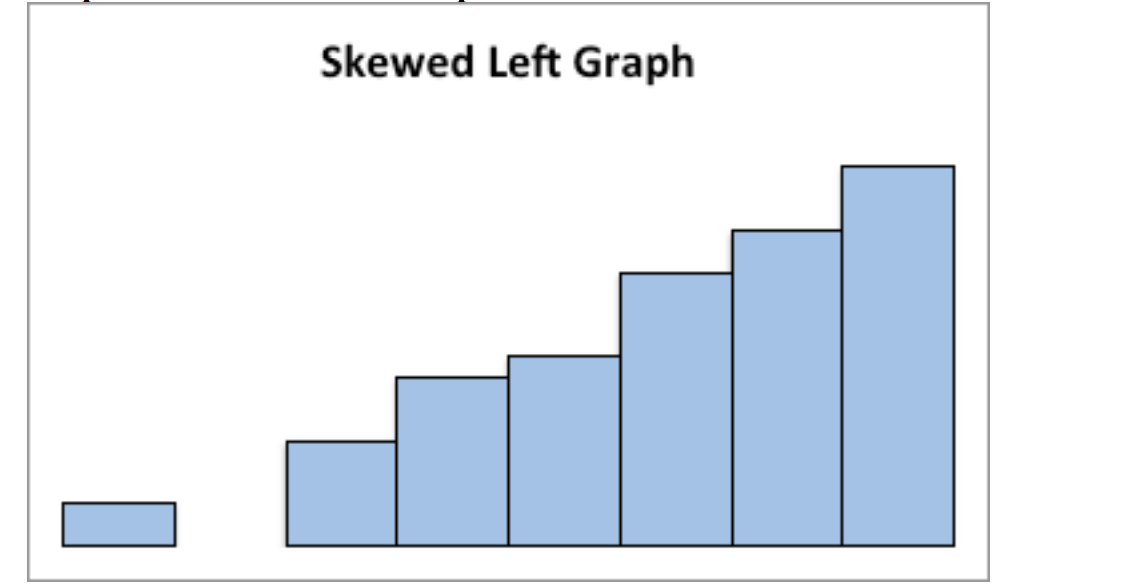
**Graph #2.2.3: Skewed Right Distribution**



Graph of skewed right graph

This graph is skewed to the left and has a gap:

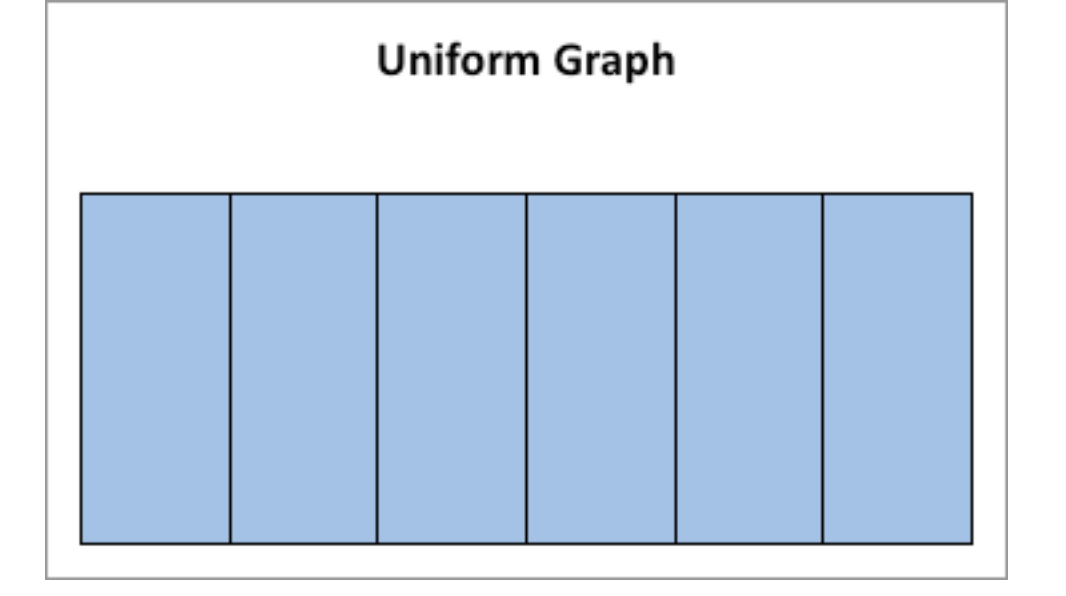
**Graph #2.2.4: Skewed Left Distribution**



Graph of Skewed Left graph

This graph is uniform since all the bars are the same height:

**Graph #2.2.5: Uniform Distribution**



Graph of uniform graph

### Example: Drawing a Histogram and Density plot

Data was collected from the Chronicle of Higher Education for tuition from public four year colleges, private four year colleges, and for profit four year colleges. The data frame is in table #2.2.2. Draw a density plot of instate tuition levels for all four year institutions, and then separate the density plot for instate tuition based on type of institution. Describe any findings from the graph.

table #2.2.2: Tuition of Four Year Colleges

Tuition<-read.csv(  
 "https://krkozak.github.io/MAT160/Tuition\_4\_year.csv")  
head(Tuition)

## INSTITUTION TYPE  
## 1 University of Alaska AnchoragePublic 4-year Public\_4 year  
## 2 University of Alaska FairbanksPublic 4-year Public\_4 year  
## 3 University of Alaska SoutheastPublic 4-year Public\_4 year  
## 4 Alaska Bible CollegePrivate 4-year Private\_4\_year  
## 5 Alaska Pacific UniversityPrivate 4-year Private\_4\_year  
## 6 Alabama Agricultural and Mechanical UniversityPublic 4-year Public\_4 year  
## STATE ROOM\_BOARD INSTATE\_TUITION INSTATE\_TOTAL OUTOFSTATE\_TUITION  
## 1 AK 12200 7688 19888 23858  
## 2 AK 8930 8087 17017 24257  
## 3 AK 9200 7092 16292 19404  
## 4 AK 5700 9300 15000 9300  
## 5 AK 7300 20830 28130 20830  
## 6 AL 8379 9698 18077 17918  
## OUTOFSTATE\_TOTAL  
## 1 36058  
## 2 33187  
## 3 28604  
## 4 15000  
## 5 28130  
## 6 26297

**Code book for Data Frame Tuition**

**Description** Cost of four year institutions.

**Format**

This data frame contains the following columns:

INSTITUTION: Name of four year institution

TYPE: Type of four year institution, Public\_4\_year, Private\_4\_year, For\_profit\_4\_year.

STATE: What state the institution resides

ROOM\_BOARD: The cost of room and board at the institution ($)

INSTATE\_TUTION: The cost of instate tuition ($)

INSTATE\_TOTAL: The cost of room and board and instate tuition ($ per year)

OUTOFSTATE\_TUTION: The cost of out of state tuition ($ per year)

OUTOFSTATE\_TOTAL: The cost of room and board and out of state tuition ($ per year)

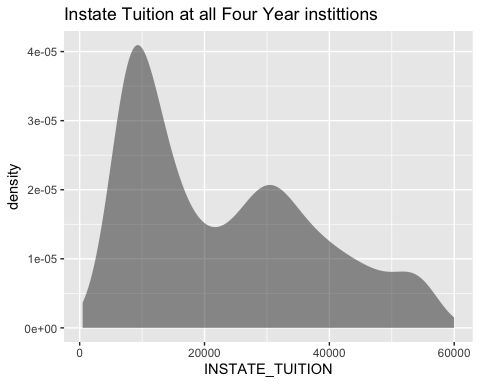
**Source** Tuition and Fees, 1998-99 Through 2018-19. (2018, December 31). Retrieved from <https://www.chronicle.com/interactives/tuition-and-fees>

**References** Chronicle of Higher Education \*, December 31, 2018.

\*\* Soultion \*\*

(ref:tuition-instate-cap) Density Plot for Instate Tuition Levels at all Four-Year Colleges\*\*

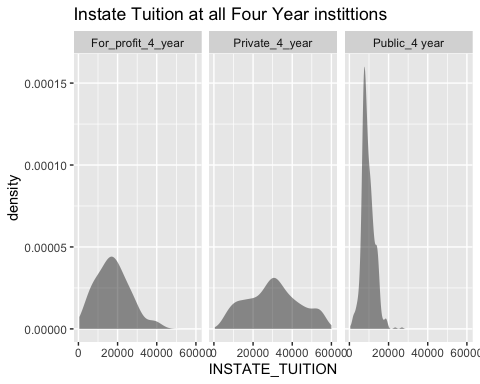
gf\_density(~INSTATE\_TUITION, data=Tuition,   
 title="Instate Tuition at all Four Year instittions")



(ref:tuition-instate-cap)

(ref:tuition-instate-type-cap) Density Plot for Instate Tuition Levels at all Four-Year Colleges\*\*

gf\_density(~INSTATE\_TUITION|TYPE, data=Tuition,   
 title="Instate Tuition at all Four Year instittions")



(ref:tuition-instate-type-cap)

The distribution is skewed right, with no gaps. Most institutions in state is less than $ 20,000 per year though some go as high as $ 60,00 per year. When separated by public versus private and for profit, most public are much less than $ 20,000 per year while private four year cost around $ 30,000 per year, and for profit are around $ 20,000 per year.

There are other types of graphs for quantitative data. They will be explored in the next section.

### Homework

1. The weekly median incomes of males and females for specific occupations, are given in table #2.2.3 (CPS News Releases. (n.d.). Retrieved July 8, 2019, from <https://www.bls.gov/cps/>). Create a density plot for males and females. Discuss any findings from the graph. Note: to put two graphs on the same axis, type %>% at the end of the first command and then type the command for the second graph on the next line. Also, use fill=“pick a color” in the command to plot the graphs with different colors so the two graphs can be easier to distinguish.

table #2.2.3: Weekly median wages for certain occupations

Wages<- read.csv(  
 "https://krkozak.github.io/MAT160/wages.csv")  
head(Wages)

## Occupation Numworkers  
## 1 Management, professional, and related occupations 48808  
## 2 Management, business, and financial operations occupations 19863  
## 3 Management occupations 13477  
## 4 Chief executives 1098  
## 5 General and operations managers 939  
## 6 Legislators 14  
## median\_wage male\_worker male\_wage female\_worker female\_wage  
## 1 1246 23685 1468 25123 1078  
## 2 1355 10668 1537 9195 1168  
## 3 1429 7754 1585 5724 1236  
## 4 2291 790 2488 307 1736  
## 5 1338 656 1427 283 1139  
## 6 NA 10 NA 4 NA

**Code book for Data Frame Wages**

**Description** Median weekly earnings of full-time wage and salary workers by detailed occupation and sex. The Current Population Survey (CPS) is a monthly survey of households conducted by the Bureau of Census for the Bureau of Labor Statistics. It provides a comprehensive body of data on the labor force, employment, unemployment, persons not in the labor force, hours of work, earnings, and other demographic and labor force characteristics.

**Format**

This data frame contains the following columns:

Occupation: Occupations of workers.

Numworkers: The number of workers in each occupation (in thousands of workers)

median\_wage: Median weekly wage ($)

male\_worker: number of male workers (in thousands of workers)

male\_wage: Median weekly wage of male workers ($)

female\_worker: number of female workers (in thousands of workers)

female\_wage: Median weekly wage of female workers ($)

**Source** CPS News Releases. (n.d.). Retrieved July 8, 2019, from <https://www.bls.gov/cps/>

**References** Current Population Survey (CPS) retrieved July 8, 2019.

1. The density of people per square kilometer for certain countries is in table #2.2.4 (World Bank, 2019). Create density plot of density in 2018 for just Sub-Saharan Africa. Describe what story the graph tells.

**Table #2.2.4: Data of Density of People per Square Kilometer**

Density<- read.csv(  
 "https://krkozak.github.io/MAT160/density.csv")   
head(Density)

## Country\_Name Country\_Code Region IncomeGroup  
## 1 Aruba ABW Latin America & Caribbean High income  
## 2 Afghanistan AFG South Asia Low income  
## 3 Angola AGO Sub-Saharan Africa Lower middle income  
## 4 Albania ALB Europe & Central Asia Upper middle income  
## 5 Andorra AND Europe & Central Asia High income  
## 6 Arab World ARB   
## y1961 y1962 y1963 y1964 y1965 y1966 y1967  
## 1 307.988889 312.361111 314.972222 316.844444 318.666667 320.638889 322.527778  
## 2 14.044987 14.323808 14.617537 14.926295 15.250314 15.585020 15.929795  
## 3 4.436891 4.498708 4.555593 4.600180 4.628676 4.637213 4.631622  
## 4 60.576642 62.456898 64.329234 66.209307 68.058066 69.874927 71.737153  
## 5 30.585106 32.702128 34.919149 37.168085 39.465957 41.802128 44.165957  
## 6 8.430860 8.663154 8.903441 9.152526 9.410965 9.679951 9.959490  
## y1968 y1969 y1970 y1971 y1972 y1973 y1974  
## 1 324.366667 326.255556 328.127778 330.222222 332.444444 334.683333 336.266667  
## 2 16.293023 16.686236 17.114913 17.577191 18.060863 18.547565 19.013188  
## 3 4.629544 4.654892 4.724765 4.845413 5.012073 5.211328 5.423422  
## 4 73.805547 75.974270 77.937190 79.848650 81.865912 83.823066 85.770949  
## 5 46.574468 49.059574 51.651064 54.380851 57.217021 60.068085 62.808511  
## 6 10.247580 10.541383 10.839409 11.140162 11.445801 11.762925 12.100336  
## y1975 y1976 y1977 y1978 y1979 y1980 y1981  
## 1 336.983333 336.588889 335.366667 333.905556 333.222222 333.866667 336.483333  
## 2 19.436265 19.825220 20.174779 20.435006 20.542009 20.458461 20.175341  
## 3 5.634074 5.839022 6.042941 6.249063 6.463517 6.690695 6.930654  
## 4 87.767555 89.727226 91.735255 93.659343 95.541314 97.518139 99.491095  
## 5 65.329787 67.610638 69.725532 71.780851 74.080851 76.738298 79.787234  
## 6 12.464221 12.856964 13.276051 13.716559 14.171137 14.634158 15.103942  
## y1982 y1983 y1984 y1985 y1986 y1987 y1988  
## 1 340.805556 345.561111 349.088889 350.144444 348.022222 343.516667 339.327778  
## 2 19.732451 19.204316 18.693582 18.286015 17.976563 17.774920 17.795553  
## 3 7.181319 7.442124 7.712163 7.990693 8.277943 8.574035 8.877878  
## 4 101.615985 103.794161 106.001058 108.202993 110.315146 112.540329 114.683796  
## 5 83.221277 86.951064 90.863830 94.893617 98.972340 103.095745 107.306383  
## 6 15.581254 16.065812 16.557944 17.057705 17.563945 18.075438 18.592082  
## y1989 y1990 y1991 y1992 y1993 y1994 y1995  
## 1 339.066667 345.272222 359.011111 379.08333 402.80000 426.11111 446.24444  
## 2 18.179820 19.012205 20.370396 22.18783 24.22664 26.15527 27.74049  
## 3 9.188078 9.503799 9.825059 10.15270 10.48773 10.83159 11.18570  
## 4 117.808139 119.946788 119.225912 118.50507 117.78420 117.06336 116.34248  
## 5 111.591489 115.976596 120.576596 125.29362 129.72553 133.35532 135.85106  
## 6 19.114029 19.817110 20.358106 20.73408 21.29364 21.84602 22.52760  
## y1996 y1997 y1998 y1999 y2000 y2001 y2002  
## 1 462.22222 474.72778 484.87222 494.47222 504.73889 516.10000 527.73333  
## 2 28.87822 29.64974 30.23277 30.89612 31.82911 33.09590 34.61810  
## 3 11.55107 11.92875 12.32021 12.72709 13.15110 13.59249 14.05263  
## 4 115.62164 114.90077 114.17993 113.45905 112.73821 111.68515 111.35073  
## 5 136.93617 136.86596 136.47234 136.95745 139.12766 143.27872 149.04043  
## 6 23.05216 23.57027 24.08237 24.60020 25.12980 25.67166 26.22642  
## y2003 y2004 y2005 y2006 y2007 y2008 y2009  
## 1 538.98333 548.53889 555.72778 560.18889 562.34444 563.10000 563.63889  
## 2 36.27251 37.87440 39.29522 40.48808 41.51049 42.46282 43.49296  
## 3 14.53556 15.04624 15.58803 16.16259 16.76856 17.40245 18.05910  
## 4 110.93489 110.47223 109.90828 109.21704 108.39478 107.56620 106.84376  
## 5 155.70638 162.22128 167.80213 172.32553 175.92340 178.42979 179.70851  
## 6 26.80081 27.40153 28.03371 28.69994 29.39751 30.11889 30.85858  
## y2010 y2011 y2012 y2013 y2014 y2015 y2016  
## 1 564.82778 566.92222 569.77778 573.10556 576.52222 579.67222 582.62222  
## 2 44.70408 46.13150 47.73056 49.42804 51.11478 52.71207 54.19711  
## 3 18.73446 19.42782 20.13951 20.86771 21.61047 22.36655 23.13506  
## 4 106.31463 106.02901 105.85405 105.66029 105.44175 105.13515 104.96719  
## 5 179.67872 178.18511 175.37660 171.85957 168.53830 165.98085 164.46170  
## 6 31.59402 32.33012 33.06767 33.80379 34.53398 35.25690 35.96876  
## y2017 y2018  
## 1 585.36667 588.02778  
## 2 55.59599 56.93776  
## 3 23.91654 24.71305  
## 4 104.87069 104.61226  
## 5 163.83191 163.84255  
## 6 36.66980 37.37237

**Code book for Data Frame Density**

**Description** Population density of all countries in the world

**Format**

This data frame contains the following columns:

Country\_Name: The name of countries or regions around the world

Country\_Code: The 3 letter code for a country or region

Region: World Banks classification of where the country is in the world

Incomegroup: World Banks classification of what income level the country is considered to be

y1961-y2018: population density for the years 1961 through 2018, people per sq. km of land area, population density is midyear population divided by land area in square kilometers. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship–except for refugees not permanently settled in the country of asylum, who are generally considered part of the population of their country of origin. Land area is a country’s total area, excluding area under inland water bodies, national claims to continental shelf, and exclusive economic zones. In most cases the definition of inland water bodies includes major rivers and lakes.

**Source** Population density (people per sq. km of land area). (n.d.). Retrieved July 9, 2019, from <https://data.worldbank.org/indicator/EN.POP.DNST>

**References** Food and Agriculture Organization and World Bank population estimates.

Since the Density data frame is for all countries, a new data frame must be created with just Su-Saharan Africa. This is created by using the following command

Africa<-  
 Density%>%  
filter(Region=="Sub-Saharan Africa")  
head(Africa)

## Country\_Name Country\_Code Region IncomeGroup  
## 1 Angola AGO Sub-Saharan Africa Lower middle income  
## 2 Burundi BDI Sub-Saharan Africa Low income  
## 3 Benin BEN Sub-Saharan Africa Low income  
## 4 Burkina Faso BFA Sub-Saharan Africa Low income  
## 5 Botswana BWA Sub-Saharan Africa Upper middle income  
## 6 Central African Republic CAF Sub-Saharan Africa Low income  
## y1961 y1962 y1963 y1964 y1965 y1966  
## 1 4.4368910 4.4987078 4.5555932 4.6001797 4.6286757 4.637213  
## 2 111.0762461 113.2134346 115.4371885 117.8461838 120.4976246 123.461449  
## 3 21.8682778 22.1966655 22.5510731 22.9333540 23.3447677 23.786440  
## 4 17.8895468 18.1298465 18.3765387 18.6362939 18.9139985 19.211853  
## 5 0.9046371 0.9242108 0.9452208 0.9667267 0.9881143 1.009235  
## 6 2.4496228 2.4911073 2.5351857 2.5821310 2.6320363 2.685510  
## y1967 y1968 y1969 y1970 y1971 y1972 y1973  
## 1 4.631622 4.629544 4.654892 4.724765 4.845413 5.012073 5.211328  
## 2 126.682944 129.942640 132.940187 135.477959 137.460942 139.005685 140.386527  
## 3 24.257778 24.756917 25.280782 25.827776 26.397410 26.991548 27.613294  
## 4 19.528578 19.861261 20.205314 20.557749 20.918790 21.290837 21.675742  
## 5 1.030635 1.053318 1.078644 1.107609 1.140485 1.177090 1.217356  
## 6 2.742146 2.799759 2.855406 2.907227 2.954377 2.998141 3.041595  
## y1974 y1975 y1976 y1977 y1978 y1979 y1980  
## 1 5.423422 5.634074 5.839022 6.042941 6.249063 6.463517 6.690695  
## 2 141.994977 144.115265 146.840771 150.095210 153.787617 157.758333 161.888551  
## 3 28.267222 28.956767 29.684046 30.449087 31.251667 32.090511 32.965280  
## 4 22.076173 22.494682 22.931422 23.387920 23.869952 24.384708 24.937292  
## 5 1.261116 1.308127 1.358635 1.412540 1.468895 1.526432 1.584296  
## 6 3.089004 3.143547 3.205583 3.274453 3.351092 3.436349 3.530380  
## y1981 y1982 y1983 y1984 y1985 y1986 y1987  
## 1 6.930654 7.181319 7.442124 7.712163 7.990693 8.277943 8.574035  
## 2 166.141744 170.550000 175.137578 179.949494 185.001441 190.293731 195.760826  
## 3 33.878397 34.832512 35.827856 36.864305 37.943429 39.060890 40.220495  
## 4 25.530556 26.163213 26.830793 27.526469 28.245274 28.986455 29.751729  
## 5 1.641713 1.699001 1.757680 1.819983 1.887287 1.960269 2.037842  
## 6 3.634855 3.748648 3.865801 3.978269 4.080659 4.169895 4.248676  
## y1988 y1989 y1990 y1991 y1992 y1993 y1994  
## 1 8.877878 9.188078 9.503799 9.825059 10.152696 10.487727 10.831593  
## 2 201.273287 206.661565 211.797391 216.702726 221.400506 225.780880 229.710553  
## 3 41.440688 42.745796 44.151259 45.667781 47.284525 48.969165 50.675949  
## 4 30.542050 31.359002 32.204072 33.077792 33.980676 34.914020 35.879342  
## 5 2.117529 2.195903 2.270492 2.340307 2.406003 2.468742 2.530410  
## 6 4.324333 4.407419 4.505336 4.620548 4.750130 4.889642 5.032288  
## y1995 y1996 y1997 y1998 y1999 y2000 y2001  
## 1 11.185695 11.551070 11.928748 12.320206 12.727095 13.151097 13.592487  
## 2 233.140304 235.985631 238.400701 240.870794 244.046885 248.398403 254.110008  
## 3 52.372810 54.046284 55.708044 57.380853 59.099840 60.889952 62.759250  
## 4 36.878209 37.912080 38.982259 40.090365 41.237942 42.426689 43.657116  
## 5 2.592370 2.655109 2.718093 2.780555 2.841325 2.899677 2.954984  
## 6 5.172969 5.310336 5.445497 5.578818 5.711281 5.843570 5.974539  
## y2002 y2003 y2004 y2005 y2006 y2007 y2008  
## 1 14.052633 14.535557 15.046238 15.588034 16.162590 16.768559 17.402450  
## 2 261.063590 269.048053 277.713902 286.793692 296.255802 306.160981 316.436994  
## 3 64.698421 66.695238 68.730082 70.789509 72.870672 74.980427 77.127714  
## 4 44.930921 46.252270 47.626349 49.056762 50.545234 52.090720 53.690515  
## 5 3.007856 3.060360 3.115288 3.174489 3.239476 3.309264 3.380162  
## 6 6.103130 6.230025 6.356344 6.482362 6.610275 6.738595 6.859556  
## y2009 y2010 y2011 y2012 y2013 y2014 y2015  
## 1 18.059101 18.734456 19.427818 20.139513 20.867715 21.610475 22.366553  
## 2 327.011994 337.834969 348.847586 360.046262 371.506581 383.344899 395.639797  
## 3 79.325186 81.582645 83.902359 86.282795 88.724619 91.227758 93.791699  
## 4 55.340270 57.036612 58.778914 60.567420 62.400493 64.276378 66.193801  
## 5 3.446964 3.506264 3.556194 3.598805 3.639363 3.685378 3.742022  
## 6 6.962703 7.041587 7.092741 7.121280 7.139783 7.165840 7.212382  
## y2016 y2017 y2018  
## 1 23.135064 23.916538 24.713052  
## 2 408.411137 421.613084 435.178271  
## 3 96.417763 99.106101 101.853920  
## 4 68.151966 70.150892 72.191283  
## 5 3.811240 3.890967 3.977425  
## 6 7.283841 7.377489 7.490412

1. The Affordable Care Act created a market place for individuals to purchase health care plans. In 2014, the premiums for a 27 year old for the different levels health insurance are given in table #2.2.5 ("Health insurance marketplace," 2013). Create a density plot of bronze\_lowest, then silver\_lowest, and gold\_lowest all on the same aces. Use %>% at the end of each command. Describe the story the graphs tells.

**Table #2.2.5: Data of Health Insurance Premiums**

Insurance<- read.csv(  
 "https://krkozak.github.io/MAT160/insurance.csv")  
head(Insurance)

## state average\_QHP bronze\_lowest silver\_lowest gold\_lowest catastrophic  
## 1 AK 34 254 312 401 236  
## 2 AL 7 162 200 248 138  
## 3 AR 28 181 231 263 135  
## 4 AZ 106 141 164 187 107  
## 5 DE 19 203 234 282 137  
## 6 FL 102 169 200 229 132  
## second\_silver\_pretax second\_silver\_posttax lowest\_bronze\_posttax  
## 1 312 107 48  
## 2 209 145 98  
## 3 241 145 85  
## 4 166 145 120  
## 5 237 145 111  
## 6 218 145 96  
## silver\_family\_pretax silver\_family\_posttax bronze\_family\_posttax  
## 1 1131 205 0  
## 2 757 282 112  
## 3 873 282 64  
## 4 600 282 192  
## 5 859 282 158  
## 6 789 282 104

**Code book for Data Frame Insurance**

**Description** The Affordable Care Act created a market place for individuals to purchase health care plans.The data is from 2014.

**Format**

This data frame contains the following columns:

state: state of insured.

average\_QHP: The number of qualified health plans

bronze\_lowest: premium for the lowest bronze level of insurance for a single person ($)

silver\_lowest: premium for the lowest silver level of insurance for a single person ($)

gold\_lowest: premium for the lowest gold level of insurance for a single person ($)

catastrophic: premium for the catastrophic level of insurance for a single person ($)

second\_silver\_pretax: premium for the second silver level of insurance for a single person pretax ($)

second\_silver\_posttax: premium for the second silver level of insurance for a single person posttax ($)

second\_bronze\_posttax: premium for the lowest bronze level of insurance for a single person posttax ($)

silver\_family\_pretax: premium for the silver level of insurance for a family pretax ($)

silver\_family\_posttax: premium for the silver level of insurance for a family posttax ($)

bronze\_family\_posttax: premium for the bronze level of insurance for a family posttax ($)

**Source** Health Insurance Market Place Retrieved from website: <http://aspe.hhs.gov/health/reports/2013/marketplacepremiums/ib_premiumslandscape.pdf> premiums for 2014.

**References** Department of Health and Human Services, ASPE. (2013). Health insurance marketplace

1. Students in a statistics class took their first test. The following are the scores they earned. Create a density plot for grades. Describe the shape of the distribution.

**Table #2.2.6: Data of Test 1 Grades**

Firsttest\_1<- read.csv(  
 "https://krkozak.github.io/MAT160/firsttest\_1.csv")  
head(Firsttest\_1)

## grades  
## 1 80  
## 2 79  
## 3 89  
## 4 74  
## 5 73  
## 6 67

1. Students in a statistics class took their first test. The following are the scores they earned. Create a density plot for grades. Describe the shape of the distribution. Compare to the graph in question 4.

**Table #2.2.7: Data of Test 1 Grades**

Firsttest\_2<- read.csv(  
 "https://krkozak.github.io/MAT160/firsttest\_2.csv")  
head(Firsttest\_2)

## grades  
## 1 67  
## 2 67  
## 3 76  
## 4 47  
## 5 85  
## 6 70

## Other Graphical Representations of Data

There are many other types of graphs. Some of the more common ones are the point plot (scatter plot), and a time-series plot. There are also many different graphs that have emerged lately for qualitative data. Many are found in publications and websites. The following is a description of the point plot (scatter plot), and the time-series plot.

**Point Plots or Scatter Plot**

Sometimes you have two different variables and you want to see if they are related in any way. A scatter plot helps you to see what the relationship would look like. A scatter plot is just a plotting of the ordered pairs.

### Example: Scatter Plot

Is there a relationship between systolic blood pressure and weight? To answer this question some data is needed. The data frame NHANES contains this data, but given the size of the data frame, it may be not be very useful to look at the graph of all the data. It makes sense to take a sample from the data frame. A random sample is the better type of sample to take. Once the sample is taken, then a scatter plot can be created. The R studio command for a scatter plot is

gf\_point(response variable ~ explanatory variable, data= Data Frame)

**Solution:**

**Table #2.3.1: Random sample of size 100 from the data frame NHANES**

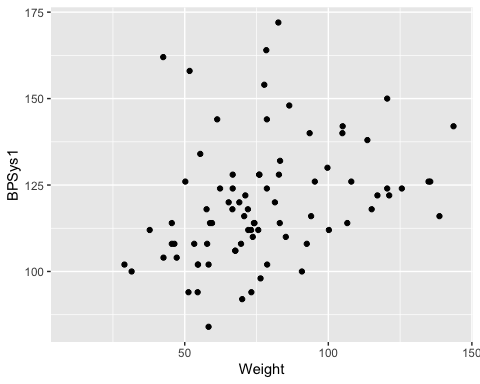
sample\_NHANES <-  
 NHANES%>%  
 sample\_n(size = 100)  
head(sample\_NHANES)

## # A tibble: 6 × 76  
## ID SurveyYr Gender Age AgeDecade AgeMonths Race1 Race3 Education   
## <int> <fct> <fct> <int> <fct> <int> <fct> <fct> <fct>   
## 1 71435 2011\_12 male 4 " 0-9" NA Other Asian <NA>   
## 2 66669 2011\_12 female 63 " 60-69" NA White White College Grad  
## 3 68328 2011\_12 male 5 " 0-9" NA White White <NA>   
## 4 62959 2011\_12 female 13 " 10-19" NA Other Asian <NA>   
## 5 71581 2011\_12 male 36 " 30-39" NA Mexican Mexican High School   
## 6 59005 2009\_10 female 20 " 20-29" 242 Other <NA> Some College  
## # … with 67 more variables: MaritalStatus <fct>, HHIncome <fct>,  
## # HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>, HomeOwn <fct>,  
## # Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>, Height <dbl>,  
## # BMI <dbl>, BMICatUnder20yrs <fct>, BMI\_WHO <fct>, Pulse <int>,  
## # BPSysAve <int>, BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>, BPSys2 <int>,  
## # BPDia2 <int>, BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>,  
## # DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>,  
## # UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>, DiabetesAge <int>,  
## # HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,  
## # LittleInterest <fct>, Depressed <fct>, nPregnancies <int>, nBabies <int>,  
## # Age1stBaby <int>, SleepHrsNight <int>, SleepTrouble <fct>,  
## # PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,  
## # TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,  
## # AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,  
## # Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,  
## # RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,  
## # SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>, SameSex <fct>,  
## # SexOrientation <fct>, PregnantNow <fct>

Preliminary: State the explanatory variable and the response variable Let x=explanatory variable = Weight y=response variable = BPSys1

(ref:weight-bp-sys-data-cap) Scatter Plot of Blood Pressure versus Weight

gf\_point(BPSys1~Weight, data=sample\_NHANES)



(ref:weight-bp-sys-data-cap)

Looking at the graph, it appears that there is a linear relationship between weight and systolic blood pressure though it looks somewhat weak. It also appears to be a positive relationship, thus as weight increases, the systolic blood pressure increases.

**Time-Series**

A time-series plot is a graph showing the data measurements in chronological order, the data being quantitative data. For example, a time-series plot is used to show profits over the last 5 years. To create a time-series plot on R Studio, use the command

gf\_line(response variable ~ explanatory variable, data=Data Frame)

The purpose of a time-series graph is to look for trends over time. Caution, you must realize that the trend may not continue. Just because you see an increase, doesn’t mean the increase will continue forever. As an example, prior to 2007, many people noticed that housing prices were increasing. The belief at the time was that housing prices would continue to increase. However, the housing bubble burst in 2007, and many houses lost value, and haven’t recovered.

### Example: Time-Series Plot\*\*

The bank assets (in billions of Australia dollars (AUD)) of the Reserve Bank of Australia (RBA) and other financial organizations for the time period of September 1 1969, through March 1 2019, are contained in table #2.3.2 (Reserve Bank of Australia, 2019). Create a time-series plot of the total assets of Authorized Deposit-taking Institutions (ADIs) and interpret any findings.

**Table #2.3.2: Data of Date versus RBA Assets**

Australian<- read.csv(  
 "https://krkozak.github.io/MAT160/Australian\_financial.csv")  
head(Australian)

## Date Day Assets\_RBA Assets\_ADIs\_Banks Assets\_ADIs\_Building Assets\_ADIs\_CU  
## 1 Sep-69 0 2.7 NA NA NA  
## 2 Dec-69 90 2.9 NA NA NA  
## 3 Mar-70 180 3.0 NA NA NA  
## 4 Jun-70 270 3.0 NA NA NA  
## 5 Sep-70 360 3.0 NA NA NA  
## 6 Dec-70 450 3.0 NA NA NA  
## Assets\_ADIs\_Total Assets\_RFCs\_MM Assets\_RFCs\_Finance Assets\_RFCs\_Total  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## Assets\_Life.offices Assets\_Life\_funds Assets\_Life\_Total  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## Assets\_Other\_Public\_trusts Assets\_Other\_Cash\_trusts Assets\_Other\_Common\_funds  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## Assets\_Others\_Friendly Assets\_Other\_General\_insurance Assets\_Other\_vehicles  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## Assets\_Unconsolidated  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

**Code book for Data frame Australian**

**Description** The data is a range of economic and financial data produced by the Reserve Bank of Australia and other organizations.

**Format**

This data frame contains the following columns:

Date: quarters from September 1 1969 to March 1, 2019

Day: The number of days since September 1, 1969 using 90 days between starts of a quarter. This column is to make it easier to graph in R Studio, and has no other purpose.

Assets\_RBA: The assets for the Royal Bank of Australia

Assets\_ADIs\_Banks: The assets for Authorized Deposit-taking Institutions (ADIs), Banks

Assets\_ADIs\_Building: The assets for Authorized Deposit-taking Institutions (ADIs), Building societies

Assets\_ADIs\_CU: The assets for Authorized Deposit-taking Institutions (ADIs), Credit Unions

Assets\_ADIs\_Total: The assets for Authorized Deposit-taking Institutions (ADIs), total

Assets\_RFCs\_MM: The assets for Registered Financial Corporations (RFCs), Money Market Corporations

Assets\_RFCs\_Finance: The assets for Registered Financial Corporations (RFCs), Finance companies and general financiers

Assets\_RFCs\_Total: The assets for Registered Financial Corporations (RFCs) total

Assets\_Life offices: The Assets of Life offices and superannuation funds; Life insurance offices

Assets\_Life\_funds: The Assets of Life offices and superannuation funds; Superannuation funds

Assets\_Life\_Total: The Assets of Life offices and superannuation; Total

Assets\_Other\_Public\_trusts: The Assets of Other managed funds; Public unit trusts

Assets\_Other\_Cash\_trusts: The Assets of Other managed funds; Cash management trusts

Assets\_Other\_Common\_funds: The Assets of Other managed funds; Common funds

Assets\_Others\_Friendly: The Assets of Other managed funds; Friendly societies

Assets\_Other\_General\_insurance: The Assets of Other financial institutions; General insurance offices

Assets\_Other\_vehicles: The Assets Other financial institutions; Securitisation vehicles

Assets\_Unconsolidated: The Assets of Unconsolidated; Statutory funds of life insurance offices; Superannuation

**Source** Reserve Bank of Australia. (2019, May 13). Statistical Tables. Retrieved July 10, 2019, from <https://www.rba.gov.au/statistics/tables/>

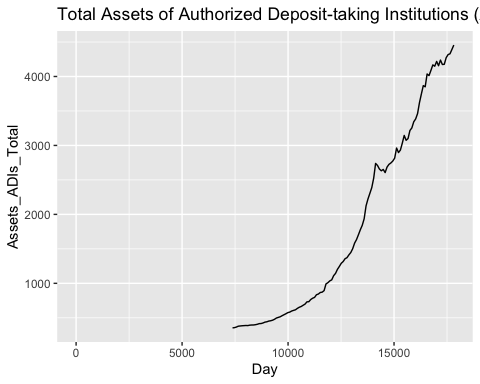
**References** Reserve Bank of Australia and other organizations

**Solution:** variable, x=total assets of Authorized Deposit-taking Institutions (ADIs)

Looking at the code book, one can see that the variable Assets\_ADIs\_Total is the variable in the data frame that is of interest here. With a time series plot, the other variable is time. In this case the variable in the data frame that represents time is Date. The problem with Date is that the units are every quarter. This is not easily interpreted by R Studio, so a column was created called Day. From the code book, this is the number of days since September 1, 1969 using 90 days between starts of a quarter. Even though this isn’t perfect, it will work for determining trends. So create a time series plot of Assets\_ADIs\_Total versus Day. The command is:

(ref:Assets-cap) Time-Series Graph of Total Assets of ADIs versus Time

gf\_line(Assets\_ADIs\_Total~Day, data=Australian, title="Total Assets of Authorized Deposit-taking Institutions (ADIs)")



Time Series Graph

From the graph, total assets of Authorized Deposit-taking Institutions (ADIs) appear to be increasing with a slight dip around 14000 days since September 1, 1969. That would be around the year 2008 (14000 days /360 days per year + 1969).

Be careful when making a graph. If the vertical axis doesn’t start at 0, then the change can look much more dramatic than it really is. For a graph to be useful to the reader, it needs to have a title that explains what the graph contains, the axes should be labeled so the reader knows what each axes represents, each axes should have a scale marked, and it is best if the vertical axis contains 0 to show the relationship.

### Homework

1. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of one of their metacarpal bone (in cm) were collected and are in table #2.3.3 (Prediction of height, 2013). Create a scatter plot of length and height and state if there is a relationship between the height of a person and the length of their metacarpal.

**Table #2.3.3: Data of Metacarpal versus Height**

Metacarpal<- read.csv(  
 "https://krkozak.github.io/MAT160/metacarpal.csv")   
head(Metacarpal)

## length height  
## 1 45 171  
## 2 51 178  
## 3 39 157  
## 4 41 163  
## 5 48 172  
## 6 49 183

**Code book for Data frame Metacarpal**

**Description** When anthropologists analyze human skeletal remains, an important piece of information is living stature. Since skeletons are commonly based on statistical methods that utilize measurements on small bones. The following data was presented in a paper in the American Journal of Physical Anthropology to validate one such method.

**Format**

This data frame contains the following columns:

length: length of Metacarpal I bone in cm

height: stature of skeleton in cm

**Source** Prediction of Height from Metacarpal Bone Length. (n.d.). Retrieved July 9, 2019, from <http://www.statsci.org/data/general/stature.html>

**References** Musgrave, J., and Harneja, N. (1978). The estimation of adult stature from metacarpal bone length. Amer. J. Phys. Anthropology 48, 113-120.

Devore, J., and Peck, R. (1986). Statistics. The Exploration and Analysis of Data. West Publishing, St Paul, Minnesota.

1. Table #2.3.4 contains the value of the house and the amount of rental income in a year that the house brings in (Capital and rental 2013). Create a scatter plot and state if there is a relationship between the value of the house and the annual rental income.

**Table #2.3.4: Data of House Value versus Rental**

House<- read.csv(  
 "https://krkozak.github.io/MAT160/house.csv")  
head(House)

## capital rental  
## 1 61500 6656  
## 2 67500 6864  
## 3 75000 4992  
## 4 75000 7280  
## 5 76000 6656  
## 6 77000 4576

**Code book for Data frame House**

**Description** The data show the capital value and annual rental value of domestic properties in Auckland in 1991.

**Format**

This data frame contains the following columns:

Capital: Selling price of house in Australian dollar (AUD)

rental: rental price of a house in Australian dollar (AUD)

**Source** Capital and rental values of Auckland properties. (2013, September 26). Retrieved from <http://www.statsci.org/data/oz/rentcap.html>

**References** Lee, A. (1994) Data Analysis: An introduction based on R. Auckland: Department of Statistics, University of Auckland. Data courtesy of Sage Consultants Ltd.

1. The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for countries for the year 2011 are in table #2.3.5. Create a scatter plot of the data and state if there appears to be a relationship between life expectancy and the number of births per woman in 2011.

**Table #2.3.5: Data of Life Expectancy versus Fertility Rate**

Fertility<- read.csv(  
 "https://krkozak.github.io/MAT160/fertility.csv")  
head(Fertility)

## country lifexp\_2011 fertilrate\_2011 lifexp\_2000 fertilrate\_2000  
## 1 Macao SAR, China 79.91 1.03 77.62 0.94  
## 2 Hong Kong SAR, China 83.42 1.20 80.88 1.04  
## 3 Singapore 81.89 1.20 78.05 NA  
## 4 Hungary 74.86 1.23 71.25 1.32  
## 5 Korea, Rep. 80.87 1.24 75.86 1.47  
## 6 Romania 74.51 1.25 71.16 1.31  
## lifexp\_1990 fertilrate\_1990  
## 1 75.28 1.69  
## 2 77.38 1.27  
## 3 76.03 1.87  
## 4 69.32 1.84  
## 5 71.29 1.59  
## 6 69.74 1.84

**Code book for Data frame Fertility**

**Description** Data is from the World Bank on the life expectancy of countries and the fertility rates in those countries.

**Format**

This data frame contains the following columns:

Country: Countries in the World

lifexp\_2011: Life expectancy of a person born in 2011

fertilrate\_2011: Fertility rate in the country in 2011

lifexp\_2000: Life expectancy of a person born in 2000

fertilrate\_2000: Fertility rate in the country in 2000

lifexp\_1990: Life expectancy of a person born in 1990

fertilrate\_1990: Fertility rate in the country in 1990

**Source** Life expectancy at birth. (2013, October 14). Retrieved from <http://data.worldbank.org/indicator/SP.DYN.LE00.IN>

**References** Data from World Bank, Life expectancy at birth, total (years)

1. The World Bank collected data on the percentage of gross domestic product (GDP) that a country spends on health expenditures (Current health expenditure (% of GDP), 2019), the fertility rate of the country (Fertility rate, total (births per woman), 2019), and the percentage of woman receiving prenatal care (Pregnant women receiving prenatal care (%), 2019). The data for the countries where this information is available in table #2.3.6. Create a scatter plot of the health expenditure and percentage of woman receiving prenatal care in the year 2014, and state if there appears to be a relationship between percentage spent on health expenditure and the percentage of woman receiving prenatal care.

**Table #2.3.6: Data of Prenatal Care versus Health Expenditure**

Fert\_prenatal<-read.csv(  
 "https://krkozak.github.io/MAT160/fertility\_prenatal.csv")  
head(Fert\_prenatal)

## Country.Name Country.Code Region IncomeGroup  
## 1 Angola AGO Sub-Saharan Africa Lower middle income  
## 2 Armenia ARM Europe & Central Asia Upper middle income  
## 3 Belize BLZ Latin America & Caribbean Upper middle income  
## 4 Cote d'Ivoire CIV Sub-Saharan Africa Lower middle income  
## 5 Ethiopia ETH Sub-Saharan Africa Low income  
## 6 Guinea GIN Sub-Saharan Africa Low income  
## f1960 f1961 f1962 f1963 f1964 f1965 f1966 f1967 f1968 f1969 f1970 f1971 f1972  
## 1 7.478 7.524 7.563 7.592 7.611 7.619 7.618 7.613 7.608 7.604 7.601 7.603 7.606  
## 2 4.786 4.670 4.521 4.345 4.150 3.950 3.758 3.582 3.429 3.302 3.199 3.114 3.035  
## 3 6.500 6.480 6.460 6.440 6.420 6.400 6.379 6.358 6.337 6.316 6.299 6.288 6.284  
## 4 7.691 7.720 7.750 7.781 7.811 7.841 7.868 7.893 7.912 7.927 7.936 7.941 7.942  
## 5 6.880 6.877 6.875 6.872 6.867 6.864 6.867 6.880 6.903 6.937 6.978 7.020 7.060  
## 6 6.114 6.127 6.138 6.147 6.154 6.160 6.168 6.177 6.189 6.205 6.225 6.249 6.277  
## f1973 f1974 f1975 f1976 f1977 f1978 f1979 f1980 f1981 f1982 f1983 f1984 f1985  
## 1 7.611 7.614 7.615 7.609 7.594 7.571 7.540 7.504 7.469 7.438 7.413 7.394 7.380  
## 2 2.956 2.875 2.792 2.712 2.641 2.582 2.538 2.510 2.499 2.503 2.517 2.538 2.559  
## 3 6.285 6.287 6.278 6.250 6.195 6.109 5.992 5.849 5.684 5.510 5.336 5.170 5.019  
## 4 7.939 7.929 7.910 7.877 7.828 7.763 7.682 7.590 7.488 7.383 7.278 7.176 7.078  
## 5 7.094 7.121 7.143 7.167 7.195 7.230 7.271 7.316 7.360 7.397 7.424 7.437 7.435  
## 6 6.306 6.337 6.369 6.402 6.436 6.468 6.500 6.529 6.557 6.581 6.602 6.619 6.631  
## f1986 f1987 f1988 f1989 f1990 f1991 f1992 f1993 f1994 f1995 f1996 f1997 f1998  
## 1 7.366 7.349 7.324 7.291 7.247 7.193 7.130 7.063 6.992 6.922 6.854 6.791 6.734  
## 2 2.578 2.591 2.592 2.578 2.544 2.484 2.400 2.297 2.179 2.056 1.938 1.832 1.747  
## 3 4.886 4.771 4.671 4.584 4.508 4.436 4.363 4.286 4.201 4.109 4.010 3.908 3.805  
## 4 6.984 6.892 6.801 6.710 6.622 6.536 6.454 6.374 6.298 6.224 6.152 6.079 6.006  
## 5 7.418 7.387 7.347 7.298 7.246 7.193 7.143 7.094 7.046 6.995 6.935 6.861 6.769  
## 6 6.637 6.637 6.631 6.618 6.598 6.570 6.535 6.493 6.444 6.391 6.334 6.273 6.211  
## f1999 f2000 f2001 f2002 f2003 f2004 f2005 f2006 f2007 f2008 f2009 f2010 f2011  
## 1 6.683 6.639 6.602 6.568 6.536 6.502 6.465 6.420 6.368 6.307 6.238 6.162 6.082  
## 2 1.685 1.648 1.635 1.637 1.648 1.665 1.681 1.694 1.702 1.706 1.703 1.693 1.680  
## 3 3.703 3.600 3.496 3.390 3.282 3.175 3.072 2.977 2.893 2.821 2.762 2.715 2.676  
## 4 5.932 5.859 5.787 5.717 5.651 5.589 5.531 5.476 5.423 5.372 5.321 5.269 5.216  
## 5 6.659 6.529 6.380 6.216 6.044 5.867 5.690 5.519 5.355 5.201 5.057 4.924 4.798  
## 6 6.147 6.082 6.015 5.947 5.877 5.804 5.729 5.653 5.575 5.496 5.417 5.336 5.256  
## f2012 f2013 f2014 f2015 f2016 f2017 p1986 p1987 p1988 p1989 p1990 p1991 p1992  
## 1 6.000 5.920 5.841 5.766 5.694 5.623 NA NA NA NA NA NA NA  
## 2 1.664 1.648 1.634 1.622 1.612 1.604 NA NA NA NA NA NA NA  
## 3 2.642 2.610 2.578 2.544 2.510 2.475 NA NA NA NA NA 96 NA  
## 4 5.160 5.101 5.039 4.976 4.911 4.846 NA NA NA NA NA NA NA  
## 5 4.677 4.556 4.437 4.317 4.198 4.081 NA NA NA NA NA NA NA  
## 6 5.175 5.094 5.014 4.934 4.855 4.777 NA NA NA NA NA NA 57.6  
## p1993 p1994 p1995 p1996 p1997 p1998 p1999 p2000 p2001 p2002 p2003 p2004 p2005  
## 1 NA NA NA NA NA NA NA NA 65.6 NA NA NA NA  
## 2 NA NA NA NA 82 NA NA 92.4 NA NA NA NA 93.0  
## 3 NA NA NA NA NA 98 95.9 100.0 NA 98 NA NA 94.0  
## 4 NA 83.2 NA NA NA NA 84.3 87.6 NA NA NA NA 87.3  
## 5 NA NA NA NA NA NA NA 26.7 NA NA NA NA 27.6  
## 6 NA NA NA NA NA NA 70.7 NA NA NA 84.3 NA 82.2  
## p2006 p2007 p2008 p2009 p2010 p2011 p2012 p2013 p2014 p2015 p2016 p2017 p2018  
## 1 NA 79.8 NA NA NA NA NA NA NA NA 81.6 NA NA  
## 2 NA NA NA NA 99.1 NA NA NA NA NA 99.6 NA NA  
## 3 94.0 99.2 NA NA NA 96.2 NA NA NA 97.2 97.2 NA NA  
## 4 84.8 NA NA NA NA NA 90.6 NA NA NA 93.2 NA NA  
## 5 NA NA NA NA NA 33.9 NA NA 41.2 NA 62.4 NA NA  
## 6 NA 88.4 NA NA NA NA 85.2 NA NA NA 84.3 NA NA  
## e2000 e2001 e2002 e2003 e2004 e2005 e2006 e2007  
## 1 2.334435 5.483824 4.072288 4.454100 4.757211 3.734836 3.366183 3.211438  
## 2 6.505224 6.536262 5.690812 5.610725 8.227844 7.034880 5.588461 5.445144  
## 3 3.942030 4.228792 3.864327 4.260178 4.091610 4.216728 4.163924 4.568384  
## 4 5.672228 4.850694 4.476869 4.645306 5.213588 5.353556 5.808850 6.259154  
## 5 4.365290 4.713670 4.705820 4.885341 4.304562 4.100981 4.226696 4.801925  
## 6 3.697726 3.884610 4.384152 3.651081 3.365547 2.949490 2.960601 3.013074  
## e2008 e2009 e2010 e2011 e2012 e2013 e2014 e2015  
## 1 3.495036 3.578677 2.736684 2.840603 2.692890 2.990929 2.798719 2.950431  
## 2 4.346749 4.689046 5.264181 3.777260 6.711859 8.269840 10.178299 10.117628  
## 3 4.646109 5.311070 5.764874 5.575126 5.322589 5.727331 5.652458 5.884248  
## 4 6.121604 6.223329 6.146566 5.978840 6.019660 5.074942 5.043462 5.262711  
## 5 4.280639 4.412473 5.466372 4.468978 4.539596 4.075065 4.033651 3.975932  
## 6 2.762090 2.936868 3.067742 3.789550 3.503983 3.461137 4.780977 5.827122  
## e2016  
## 1 2.877825  
## 2 9.927321  
## 3 6.121374  
## 4 4.403621  
## 5 3.974016  
## 6 5.478273

**Code book for Data frame Fert\_prenatal**

**Description** Data is from the World Bank on money spent on expenditure of countries and the percentage of woman receiving prenatal care in those countries.

**Format**

This data frame contains the following columns:

Country.Name: Countries around the world

Country.Code: Three letter country code for countries around the world

Region: Location of a country around the world as classified by the World Bank

IncomeGroup: The income level of a country as classified by the World Bank

f1960-f2017: Fertility rate of a country from 1960-2017

p1986-p2018: Percentage of woman receiving prenatal care in the country in 1986-2018

e200-2016: Expenditure amounts of the countries for medical care in 2000-2016 (% of GDP)

**Source** Fertility rate, total (births per woman). (n.d.). Retrieved July 8, 2019, from <https://data.worldbank.org/indicator/SP.DYN.TFRT.IN> Pregnant women receiving prenatal care (%). (n.d.). Retrieved July 9, 2019, from <https://data.worldbank.org/indicator/SH.STA.ANVC.ZS> Current health expenditure (% of GDP). (n.d.). Retrieved July 9, 2019, from <https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS>

**References** Data from World Bank, fertility rate, expenditure on health, and pregnant woman rate of prenatal care.

1. The Australian Institute of Criminology gathered data on the number of deaths (per 100,000 people) due to firearms during the period 1983 to 1997 ("Deaths from firearms," 2013). The data is in table #2.3.7. Create a time-series plot of the data and state any findings you can from the graph.

**Table #2.3.7: Data of Year versus Number of Deaths due to Firearms**

Firearm<- read.csv(  
 "https://krkozak.github.io/MAT160/rate.csv")  
head(Firearm)

## year rate  
## 1 1983 4.31  
## 2 1984 4.42  
## 3 1985 4.52  
## 4 1986 4.35  
## 5 1987 4.39  
## 6 1988 4.21

**Code book for Data Frame Firearm**

**Description** The data give the number of deaths caused by firearms in Australia from 1983 to 1997, expressed as a rate per 100,000 of population.

**Format**

This data frame contains the following columns:

Year: Years from 1983 to 1997

Rate: Rate of deaths caused by firearms in Australia per 100,000 population

**Source** Deaths from firearms. (2013, September 26). Retrieved from <http://www.statsci.org/data/oz/firearms.html>

**References** Australian Institute of Criminology, 1999.The data was contributed by Rex Boggs, Glenmore State High School, Rockhampton, Queensland, Australia.

1. The economic crisis of 2008 affected many countries, though some more than others. Some people in Australia have claimed that Australia wasn’t hurt that badly from the crisis. The bank assets (in billions of Australia dollars (AUD)) of the Reserve Bank of Australia (RBA) for the time period of September 1 1969 through March 1 2019 are contained in table #2.3.8 (Reserve Bank of Australia, 2019). Create a time-series plot of the assets of the RBA and interpret any findings.

**Table #2.3.8: Data of Date versus RBA Assets**

Australian<- read.csv(  
 "https://krkozak.github.io/MAT160/Australian\_financial.csv")  
head(Australian)

## Date Day Assets\_RBA Assets\_ADIs\_Banks Assets\_ADIs\_Building Assets\_ADIs\_CU  
## 1 Sep-69 0 2.7 NA NA NA  
## 2 Dec-69 90 2.9 NA NA NA  
## 3 Mar-70 180 3.0 NA NA NA  
## 4 Jun-70 270 3.0 NA NA NA  
## 5 Sep-70 360 3.0 NA NA NA  
## 6 Dec-70 450 3.0 NA NA NA  
## Assets\_ADIs\_Total Assets\_RFCs\_MM Assets\_RFCs\_Finance Assets\_RFCs\_Total  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## Assets\_Life.offices Assets\_Life\_funds Assets\_Life\_Total  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## Assets\_Other\_Public\_trusts Assets\_Other\_Cash\_trusts Assets\_Other\_Common\_funds  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## Assets\_Others\_Friendly Assets\_Other\_General\_insurance Assets\_Other\_vehicles  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## Assets\_Unconsolidated  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

**Code book for Data Frame Australian** See Example #2.3.2

1. The consumer price index (CPI) is a measure used by the U.S. government to describe the cost of living. Table #2.3.9 gives the cost of living for the U.S. from the years 1913 through 2019, with the year 1982 being used as the year that all others are compared (Consumer Price Index Data from 1913 to 2019, 2019). Create a time-series plot of the Average Annual CPI and interpret.

**Table #2.3.9: Data of Time versus CPI**

CPI<- read.csv(  
 "https://krkozak.github.io/MAT160/CPI\_US.csv")  
head(CPI)

## Year Jan Feb Mar Apr May June July Aug Sep Oct Nov Dec Annual\_avg  
## 1 1913 9.8 9.8 9.8 9.8 9.7 9.8 9.9 9.9 10.0 10.0 10.1 10.0 9.9  
## 2 1914 10.0 9.9 9.9 9.8 9.9 9.9 10.0 10.2 10.2 10.1 10.2 10.1 10.0  
## 3 1915 10.1 10.0 9.9 10.0 10.1 10.1 10.1 10.1 10.1 10.2 10.3 10.3 10.1  
## 4 1916 10.4 10.4 10.5 10.6 10.7 10.8 10.8 10.9 11.1 11.3 11.5 11.6 10.9  
## 5 1917 11.7 12.0 12.0 12.6 12.8 13.0 12.8 13.0 13.3 13.5 13.5 13.7 12.8  
## 6 1918 14.0 14.1 14.0 14.2 14.5 14.7 15.1 15.4 15.7 16.0 16.3 16.5 15.1  
## PerDec\_Dec Perc\_Avg\_Avg  
## 1 – –  
## 2 1 1  
## 3 2 1  
## 4 12.6 7.9  
## 5 18.1 17.4  
## 6 20.4 18

**Code book for Data frame CPI**

**Description** This table of Consumer Price Index (CPI) data is based upon a 1982 base of 100.

**Format**

This data frame contains the following columns:

Year: Year from 1913 to 2019

Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec: CPI for a particular month

Average\_Avg: The average CPI for a particular year

PerDec\_Dec: Percent change from December to December

Per\_Avg\_Avg: Percent change from Annual Average to Annual Average

**Source** Consumer Price Index Data from 1913 to 2019. (2019, June 12). Retrieved July 10, 2019, from <https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/>

**References** US Inflation Calculator website, 2019.

1. The mean and median incomes income in current dollars is given in Table #2.3.10. Create a time-series plot and interpret.

**Table #2.3.10: Data of US Mean and Median Income**

US\_income<- read.csv(  
 "https://krkozak.github.io/MAT160/US\_income.csv")  
head(US\_income)

## year number med\_income\_current med\_income\_2017 mean\_income\_current  
## 1 2017 127586 61372 61372 86220  
## 2 2016 126224 59039 60309 83143  
## 3 2015 125819 56516 58476 79263  
## 4 2014 124587 53657 55613 75738  
## 5 2013 122952 51939 54744 72641  
## 6 2012 122459 51017 54569 71274  
## mean\_income\_2017  
## 1 86220  
## 2 84931  
## 3 82012  
## 4 78500  
## 5 76565  
## 6 76237

**Code book for Data Frame US\_income**

**Description** This table is of US mean and median incomes in both current dollars and in 2017 dollars.

**Format**

This data frame contains the following columns:

Year: Year from 1975 to 2017

number: Households as of March of the following year. (in thousands)

med\_income\_current: median income of a US household in current dollars

med\_income\_2017: median income of a US household in 2017 CPI-U-RS adjusted dollars

mean\_income\_current: mean income of a US household in current dollars

mean\_income\_2017: mean income of a US household in 2017 CPI-U-RS adjusted dollars

**Source** US Census Bureau. (2018, March 06). Data. Retrieved July 21, 2019, from <https://www.census.gov/programs-surveys/cps/data-detail.html>

**References** U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplements.

Data Sources:

Capital and rental values of Auckland properties. (2013, September 26). Retrieved from <http://www.statsci.org/data/oz/rentcap.html>

Consumer Price Index Data from 1913 to 2019. (2019, June 12). Retrieved July 10, 2019, from <https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/>

CPS News Releases. (n.d.). Retrieved July 8, 2019, from <https://www.bls.gov/cps/>

Current health expenditure (% of GDP). (n.d.). Retrieved July 9, 2019, from <https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS>

Deaths from firearms. (2013, September 26). Retrieved from <http://www.statsci.org/data/oz/firearms.html>

Fertility rate, total (births per woman). (n.d.). Retrieved July 8, 2019, from <https://data.worldbank.org/indicator/SP.DYN.TFRT.IN>

Health Insurance Market Place Retrieved from website: <http://aspe.hhs.gov/health/reports/2013/marketplacepremiums/ib_premiumslandscape.pdf>

John Matic provided the data from a company he worked with. The company’s name is fictitious, but the data is from an actual company.

Kozak K (2019). Survey results form surveys collected in statistics class at Coconino Community College.

Life expectancy at birth. (2013, October 14). Retrieved from <http://data.worldbank.org/indicator/SP.DYN.LE00.IN>

Population density (people per sq. km of land area). (n.d.). Retrieved July 9, 2019, from <https://data.worldbank.org/indicator/EN.POP.DNST>

Prediction of Height from Metacarpal Bone Length. (n.d.). Retrieved July 9, 2019, from <http://www.statsci.org/data/general/stature.html>

Pregnant women receiving prenatal care (%). (n.d.). Retrieved July 9, 2019, from <https://data.worldbank.org/indicator/SH.STA.ANVC.ZS>

Reserve Bank of Australia. (2019, May 13). Statistical Tables. Retrieved July 10, 2019, from <https://www.rba.gov.au/statistics/tables/>

Tuition and Fees, 1998-99 Through 2018-19. (2018, December 31). Retrieved from <https://www.chronicle.com/interactives/tuition-and-fees>

U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplements.

# Numerical Descriptions of Data

Chapter 1 discussed what a population, sample, parameter, and statistic are, and how to take different types of samples. Chapter 2 discussed ways to graphically display data. There was also a discussion of important characteristics: center, variations, distribution, outliers, and changing characteristics of the data over time. Distributions and outliers can be answered using graphical means. Finding the center and variation can be done using numerical methods that will be discussed in this chapter. Both graphical and numerical methods are part of a branch of statistics known as **descriptive statistics**. Later descriptive statistics will be used to make decisions and/or estimate population parameters using methods that are part of the branch called **inferential statistics**.

## Measures of Center

This section focuses on measures of central tendency. Many times you are asking what to expect on average. Such as when you pick a major, you would probably ask how much you expect to earn in that field. If you are thinking of relocating to a new town, you might ask how much you can expect to pay for housing. If you are planting vegetables in the spring, you might want to know how long it will be until you can harvest. These questions, and many more, can be answered by knowing the center of the data set. There are three measures of the “center” of the data. They are the mode, median, and mean. Any of the values can be referred to as the “average.”

The **mode** is the data value that occurs the most frequently in the data. To find it, you count how often each data value occurs, and then determine which data value occurs most often. The mode is not the most useful measure of center. This is because, a data set can have more than one mode. If there is a tie between two values for the most number of times then both values are the mode and the data is called bimodal (two modes). If every data point occurs the same number of times, there is no mode. If there are more than two numbers that appear the most times, then usually there is no mode.

The **median** is the data value in the middle of a sorted list of data. To find it, you put the data in order, and then determine which data value is in the middle of the data set.

The **mean** is the arithmetic average of the numbers. This is the center that most people call the average, though all three – mean, median, and mode – really are averages.

There are no symbols for the mode and the median, but the mean is used a great deal, and statisticians gave it a symbol. There are actually two symbols, one for the population parameter and one for the sample statistic. In most cases you cannot find the population parameter, so you use the sample statistic to estimate the population parameter.

**Population Mean:**

, pronounced mu

*N* is the size of the population.

*x* represents a data value.

means to add up all of the data values.

**Sample Mean**:

, pronounced x bar.

*n* is the size of the sample.

*x* represents a data value.

means to add up all of the data values.

The value for is used to estimate since can’t be calculated in most situations.

### Example: Finding the Mean and Median using R

Suppose a vet wants to find the average weight of cats. The weights (in kg) of cats are in table #3.1.1.

Table #3.1.1: Cats’ Weights

head(cats)

## Sex Bwt Hwt  
## 1 F 2.0 7.0  
## 2 F 2.0 7.4  
## 3 F 2.0 9.5  
## 4 F 2.1 7.2  
## 5 F 2.1 7.3  
## 6 F 2.1 7.6

The head command shows the variable names and the first few unit of observation rows

Find the mean and median of the weight of a cat.

**Solution:**

Before starting any mathematics problem, it is always a good idea to define the unknown in the problem. In statistics, you want to define the variable. The symbol for the variable is *x*.

The variable is *x* = weight of a cat

Mean: To find with R Studio, perform the command:

df\_stats(~Bwt, cats, mean)

## response mean  
## 1 Bwt 2.723611

The mean weight is 2.72 kg

Median: To find with R Studio, perform the command:

df\_stats(~Bwt, cats, median)

## response median  
## 1 Bwt 2.7

The median weight is 2.7 kg also. It appears the average weight is 2.7 kg of all cats.

### Example: Finding Mean and Median with filtering

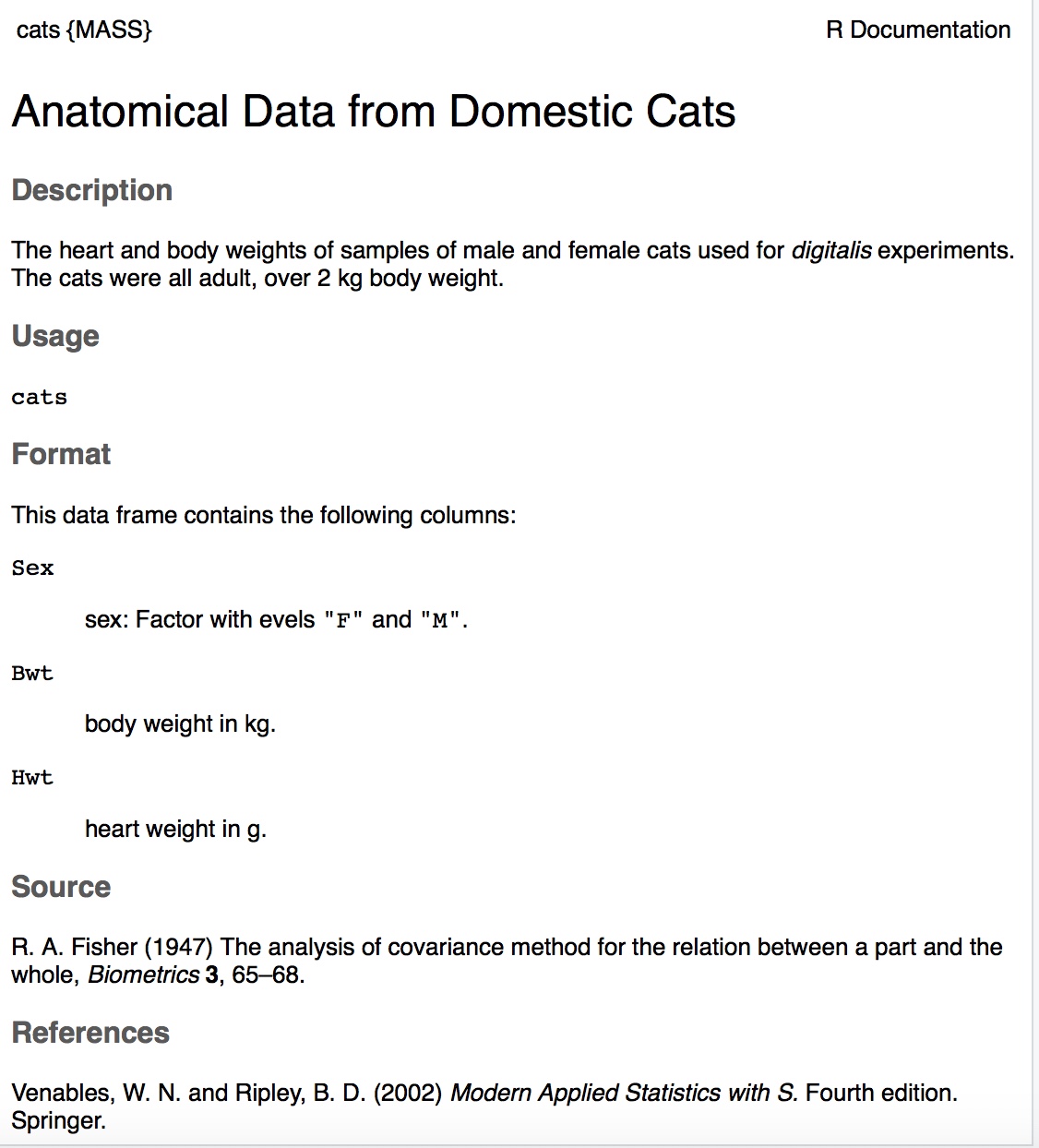
Looking at the data frame for cats weights table(#3.1.1), you see that there are several variables You may want to know what the other variables are. A \* Code Book \* describes the data set, explains what the variables are including the units, and the source of the data frame. To review the code book for a data frame, complete the following command.

??cats

Then click on MASS::cats.

The output looks like:

**Image #3.1.1: Code book for cats data frame**



Code book for cats data

Suppose you want to know if male cats weight more than female cats. Looking at the variables, you notice that there is a variable for the sex of the cat. You can look at the weights of males and females separately. This looks like:

**Solution:** To find the mean and median, separated by sex, use this command in R Studio:

df\_stats(Bwt~Sex, data=cats, mean, median)

## response Sex mean median  
## 1 Bwt F 2.359574 2.3  
## 2 Bwt M 2.900000 2.9

Notice that the female cats’ mean weigh 2.4 kg and the male cats’ mean weigh 2.9 kg The median weight of female cats is 2.3 kg and for males is is 2.9 kg So it does appear that males cats weight a bit more than the female cats.

There are many different summary statistics that can be found. An example is the minimum and maximum value. In this example, you will see how to find the min and max values and then filter them out of a data set to see what effect they have on the mean and median.

### Example: Affect of Extreme Values on Mean and Median\*\*

Find the minimum and maximum values of cats weights.

**Solution** The command in R Studio for finding the minimum and maximum is very similar to how to find the mean and median, In fact all summary statistics start with

df\_stats(~variable, data=Data Frame, desired statistics)

Here is the command in R Studio for the minimum and maximum of cat’s body weight.

df\_stats(~Bwt, data=cats, min, max)

## response min max  
## 1 Bwt 2 3.9

The minimum weight of a cat in this data frame is 2 kg and the maximum weight of a cat is 3.9 kg.

Now create two new data sets. One data set will exclude the maximum value. You can call it anything you want, but it would make sense to call it something like nomax. The command to create the new data set is:

nomax<-filter(cats, Bwt<3.9)

Then create a data set that excludes the minimum value; call it nomin:

nomin<-filter(cats, Bwt>2)

The <- is the way to indicate to R what the data set nomin is equivalent to what follows the symbol. Notice that it doesn’t look like anything happened, but new data sets were created in the background. Now you can find the mean and median of each new data set:

df\_stats(~Bwt, data=nomax, mean, median)

## response mean median  
## 1 Bwt 2.707042 2.7

The mean without the maximum value is 2.70 kg, and the median is 2.7 kg.

df\_stats(~Bwt, data=nomin, mean, median)

## response mean median  
## 1 Bwt 2.74964 2.7

The mean without the minimum value is 2.75 kg, and the median is 2.7 kg.

From Example 3.1.1, the mean of the data set with all the values is 2.72 kg where the median is 2.7 kg. Notice that when the maximum value was excluded from the data set, the mean decreased a little but the median didn’t change, and when the minimum value was excluded from the data set, the mean increased a little but the median didn’t change. The mean is much higher than the median. Why is this? This is because the mean is affected by extreme values, while the median is not. We say the median is a much more resistant measure of center because it isn’t affected by extreme values as much.

An outlier is a data value that is very different from the rest of the data. It can be really high or really low. Extreme values may be an outlier if the extreme value is far enough from the center. If there are extreme values in the data, the median is a better measure of the center than the mean. If there are no extreme values, the mean and the median will be similar so most people use the mean. The mean is not a resistant measure because it is affected by extreme values. The median is a resistant measure because it not affected by extreme values.

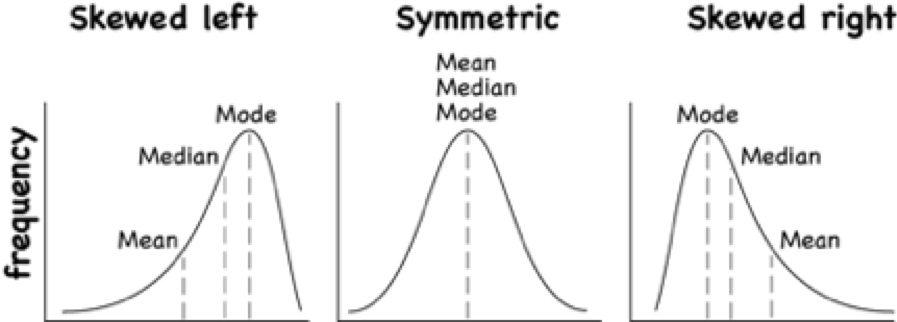
As a consumer you need to be aware that people choose the measure of center that best supports their claim. When you read an article in the newspaper and it talks about the “average” it usually means the mean but sometimes it refers to the median. Some articles will use the word “median” instead of “average” to be more specific. If you need to make an important decision and the information says “average”, it would be wise to ask if the “average” is the mean or the median before you decide.

As an example, suppose that a company wants to use the mean salary as the average salary for the company. This is because the high salaries of the administrators will pull the mean higher. The company can say that the employees are paid well because the average is high. However, the employees want to use the median since it discounts the extreme values of the administration and will give a lower value of the average. This will make the salaries seem lower and that a raise is in order.

Why use the mean instead of the median? The reason is because when multiple samples are taken from the same population, the sample means tend to be more consistent than other measures of the center.

To understand how the different measures of center related to skewed or symmetric distributions, see Graph #3.1.2. As you can see sometimes the mean is smaller than the median, sometimes the mean is larger than the median, and sometimes they are the same values.

**Graph #3.1.2: Mean, Median, Mode as Related to a Distribution**



Graphs of symmetric, skewed right and skewed left

One last type of average is a weighted average. **Weighted averages** are used quite often in different situations. Some teachers use them in calculating a student’s grade in the course, or a grade on a project. Some employers use them in employee evaluations. The idea is that some activities are more important than others. As an example, a full time teacher at a community college may be evaluated on their service to the college, their service to the community, whether their paperwork is turned in on time, and their teaching. However, teaching is much more important than whether their paperwork is turned in on time. When the evaluation is completed, more weight needs to be given to the teaching and less to the paperwork. This is a weighted average.

**Weighted Average**

where *w* is the weight of the data value, *x*.

### Example: Weighted Average

In your biology class, your final grade is based on several things: a lab score, scores on two major tests, and your score on the final exam. There are 100 points available for each score. The lab score is worth 15% of the course, the two exams are worth 25% of the course each, and the final exam is worth 35% of the course. Suppose you earned scores of 95 on the labs, 83 and 76 on the two exams, and 84 on the final exam. Compute your weighted average for the course.

**Solution:**

Variable: *x* = score

A weighted average can be found using technology. The commands for finding the weighted mean using R Studio is as follows:

x<-c(type in the scores with commas in between)  
  
 w<-c(type in the weights as decimals with commas in between  
  
 weighted.mean(x,w)

The *x* and *w* represent the variables, <- means make the variables equivalent to what follows, the c( means combine all the values in the () as one combined variable.

For this example, the commands would be

x<-c(95, 83, 76, 84)  
w<-c(.15, .25, .25, .35)  
weighted.mean(x,w)

## [1] 83.4

Your weighted mean in the biology class is 83.4%. Using the traditional grading scale, you have a B in the class.

### Example: Weighted Average

The faculty evaluation process at John Jingle University rates a faculty member on the following activities: teaching, publishing, committee service, community service, and submitting paperwork in a timely manner. The process involves reviewing student evaluations, peer evaluations, and supervisor evaluation for each teacher and awarding him/her a score on a scale from 1 to 10 (with 10 being the best). The weights for each activity are 20 for teaching, 18 for publishing, 6 for committee service, 4 for community service, and 2 for paperwork.

1. One faculty member had the following ratings: 8 for teaching, 9 for publishing, 2 for committee work, 1 for community service, and 8 for paperwork. Compute the weighted average of the evaluation.

**Solution:**

Variable: *x* = rating, *w* = weight

x<-c(8, 9, 2, 1, 8)  
w<-c(20, 18, 6, 4, 2)  
weighted.mean(x,w)

## [1] 7.08

The weighted average is 7.08.

1. Another faculty member had ratings of 6 for teaching, 8 for publishing, 9 for committee work, 10 for community service, and 10 for paperwork. Compute the weighted average of the evaluation.

**Solution:**

x<-c(6, 8, 9, 10, 10)  
w<-c(20, 18, 6, 4, 2)  
weighted.mean(x,w)

## [1] 7.56

The weighted average for this employee is 7.56.

1. Which faculty member had the higher average evaluation?

**Solution:**

The second faculty member has a higher average evaluation.

The last thing to mention is which average is used on which type of data.

Mode can be found on nominal, ordinal, interval, and ratio data, since the mode is just the data value that occurs most often. You are just counting the data values.

Median can be found on ordinal, interval, and ratio data, since you need to put the data in order. As long as there is order to the data you can find the median.

Mean can be found on interval and ratio data, since you must have numbers to add together.

### Homework

\*\* Use Technology on all problems. State the variable on all problems.\*\*

1. Cholesterol levels were collected from patients certain days after they had a heart attack and are in table #3.1.2. Find the mean and median for cholesterol levels 2 days after the heart attack.

table #3.1.2

Cholesterol<-read.csv(  
 "https://krkozak.github.io/MAT160/cholesterol.csv")  
head(Cholesterol)

## patient day2 day4 day14  
## 1 1 270 218 156  
## 2 2 236 234 NA  
## 3 3 210 214 242  
## 4 4 142 116 NA  
## 5 5 280 200 NA  
## 6 6 272 276 256

**Code book for Data Frame Cholesterol**

**Description** Cholesterol levels were collected from patients certain days after they had a heart attack

This data frame contains the following columns:

Patient: Patient number

day2: Cholesterol level of patient 2 days after heart attack. (mg/dL)

day4: Cholesterol level of patient 4 days after heart attack. (mg/dL)

day14: Cholesterol level of patient 14 days after heart attack. (mg/dL)

**Source** Ryan, B. F., Joiner, B. L., & Ryan, Jr, T. A. (1985). Cholesterol levels after heart attack.Retrieved from <http://www.statsci.org/data/general/cholest.html>

**References** Ryan, Joiner & Ryan, Jr, 1985

1. The lengths (in kilometers) of rivers on the South Island of New Zealand and what body of water they flow into are listed in table #3.1.2 (Lee, 1994). Find the mean and median length of rivers that flow into the Pacific Ocean and the mean and median length of rivers that flow into the Tasman Sea.

\*\* Table #3.1.2: Lengths of Rivers (km) in New Zealand\*\*

Length<-read.csv(  
 "https://krkozak.github.io/MAT160/length.csv")  
head(Length)

## river length flowsto  
## 1 Clarence 209 Pacific  
## 2 Conway 48 Pacific  
## 3 Waiau 169 Pacific  
## 4 Hurunui 138 Pacific  
## 5 Waipara 64 Pacific  
## 6 Ashley 97 Pacific

**Code book for data frame Length**

**Description**

Rivers in New Zealand, the lengths of river and what body of water the river flows into

This data frame contains the following columns:

River: Name of the river

length: how long the river is in kilometers

flowsto: what body of water the river flows into Pacific Ocean is Pacific and the Tasman Sea is Tasman

**Source** Lee, A. (1994). Data analysis: An introduction based on r. Auckland. Retrieved from <http://www.statsci.org/data/oz/nzrivers.html>

**References** Lee, A. (1994). Data analysis: An introduction based on r. Auckland.

1. Print-O-Matic printing company’s employees have salaries that are contained in table #3.1.3.

**Table #3.1.3: Salaries of Print-O-Matic Printing Company Employees**

Pay<-read.csv(  
 "https://krkozak.github.io/MAT160/pay.csv")  
head(Pay)

## employee salary  
## 1 CEO 272500  
## 2 Driver 58456  
## 3 CD74 100702  
## 4 CD65 57380  
## 5 Embellisher 73877  
## 6 Folder 65270

**Code book for data frame Pay**

**Description**

Salaries of Print-O-Matic printing company’s employees

This data frame contains the following columns:

employee:employees position in the company

salary: salary of that employee (Australian dollars (AUD))

**Source** John Matic provided the data from a company he worked with. The company’s name is fictitious, but the data is from an actual company.

**References** John Matic (2013)

1. Find the mean and median.
2. Find the mean and median with the CEO’s salary removed.
3. What happened to the mean and median when the CEO’s salary was removed? Why?
4. If you were the CEO, who is answering concerns from the union that employees are underpaid, which average (mean or median) using the complete data set of the complete data set would you prefer? Why?
5. If you were a platen worker, who believes that the employees need a raise, which average (mean or median) using the complete data set would you prefer? Why?
6. Print-O-Matic printing company spends specific amounts on fixed costs every month. The costs of those fixed costs are in table #3.1.4.

**Table #3.1.4: Fixed Costs for Print-O-Matic Printing Company**

Cost<-read.csv(  
 "https://krkozak.github.io/MAT160/cost.csv")  
head(Cost)

## charges cost  
## 1 Bank charges 482  
## 2 Cleaning 2208  
## 3 Computer expensive 2471  
## 4 Lease payments 2656  
## 5 Postage 2117  
## 6 Uniforms 2600

**Code book for data frame Cost**

**Description** fixed monthly charges for Print-0-Matic printing company

This data frame contains the following columns:

charges: Categories of monthly fixed charges

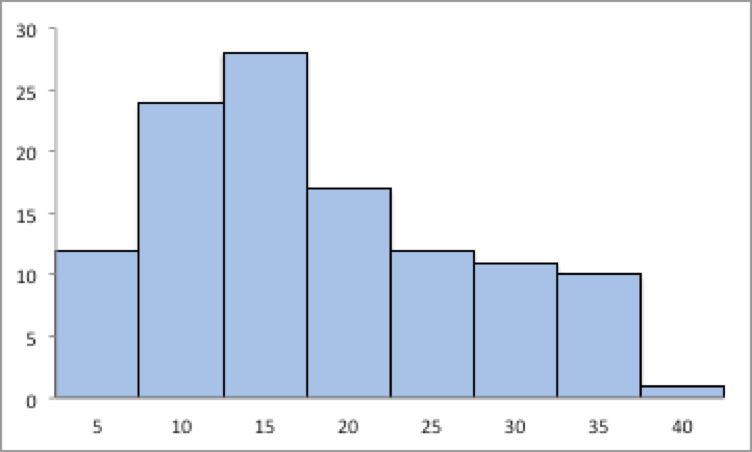
cost: fixed month costs (AUD)

**Source** John Matic provided the data from a company he worked with. The company’s name is fictitious, but the data is from an actual company.

**References** John Matic (2013)

1. Find the mean and median.
2. Find the mean and median with the bank charges removed.
3. What happened to the mean and median when the bank charges was removed? Why?
4. If it is your job to oversee the fixed costs, which average (mean or median) using the complete data set would you prefer to use when submitting a report to administration to show that costs are low? Why?
5. If it is your job to find places in the budget to reduce costs, which average (mean or median) using the complete data set would you prefer to use when submitting a report to administration to show that fixed costs need to be reduced? Why?
6. Looking at graph #3.1.1, state if the graph is skewed left, skewed right, or symmetric and then state which is larger, the mean or the median?

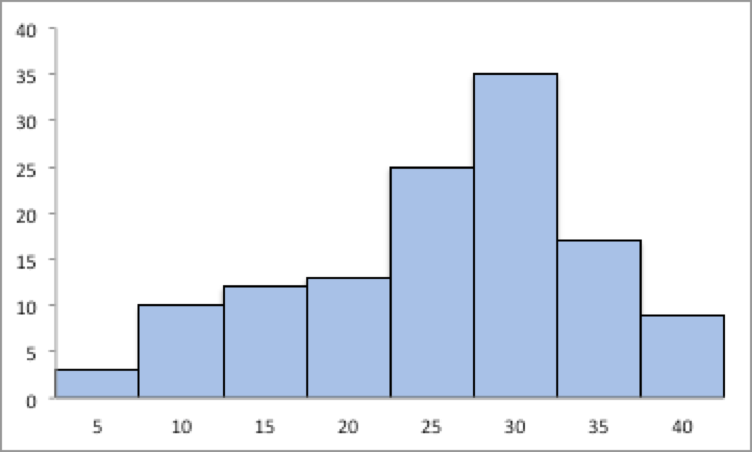
**Graph #3.1.1: Skewed or Symmetric Graph**



Graphs of symmetric, skewed right and skewed left

1. Looking at graph #3.1.2, state if the graph is skewed left, skewed right, or symmetric and then state which is larger, the mean or the median?

**Graph #3.1.2: Skewed or Symmetric Graph**



Graphs of symmetric, skewed right and skewed left

1. An employee at Coconino Community College (CCC) is evaluated based on goal setting and accomplishments toward the goals, job effectiveness, competencies, and CCC core values. Suppose for a specific employee, goal 1 has a weight of 30%, goal 2 has a weight of 20%, job effectiveness has a weight of 25%, competency 1 has a weight of 4%, competency 2 has a weight of 3%, competency 3 has a weight of 3%, competency 4 has a weight of 3%, competency 5 has a weight of 2%, and core values has a weight of 10%. Suppose the employee has scores of 3.0 for goal 1, 3.0 for goal 2, 2.0 for job effectiveness, 3.0 for competency 1, 2.0 for competency 2, 2.0 for competency 3, 3.0 for competency 4, 4.0 for competency 5, and 3.0 for core values. Find the weighted average score for this employee. If an employee has a score less than 2.5, they must have a Performance Enhancement Plan written. Does this employee need a plan?
2. An employee at Coconino Community College (CCC) is evaluated based on goal setting and accomplishments toward goals, job effectiveness, competencies, CCC core values. Suppose for a specific employee, goal 1 has a weight of 20%, goal 2 has a weight of 20%, goal 3 has a weight of 10%, job effectiveness has a weight of 25%, competency 1 has a weight of 4%, competency 2 has a weight of 3%, competency 3 has a weight of 3%, competency 4 has a weight of 5%, and core values has a weight of 10%. Suppose the employee has scores of 2.0 for goal 1, 2.0 for goal 2, 4.0 for goal 3, 3.0 for job effectiveness, 2.0 for competency 1, 3.0 for competency 2, 2.0 for competency 3, 3.0 for competency 4, and 4.0 for core values. Find the weighted average score for this employee. If an employee that has a score less than 2.5, they must have a Performance Enhancement Plan written. Does this employee need a plan?
3. A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives an 85 on test 1, a 76 on test 2, an 83 on test 3, a 74 on the homework, a 65 on the project, and a 79 on the final, what grade did the student earn in the course?
4. A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives a 92 on test 1, an 85 on test 2, a 95 on test 3, a 92 on the homework, a 55 on the project, and an 83 on the final, what grade did the student earn in the course?

## Measures of Spread

Variability is an important idea in statistics. If you were to measure the height of everyone in your classroom, every observation gives you a different value. That means not every student has the same height. Thus there is variability in people’s heights. If you were to take a sample of the income level of people in a town, every sample gives you different information. There is variability between samples too. Variability describes how the data are spread out. If the data are very close to each other, then there is low variability. If the data are very spread out, then there is high variability. How do you measure variability? It would be good to have a number that measures it. This section will describe some of the different measures of variability, also known as variation.

In example #3.1.1, the average weight of a cat was calculated to be 2.72 kg. How much does this tell you about the weight of all cats? Can you tell if most of the weights were close to 2.72 kg or were the weights really spread out? The highest weight and the lowest weight are known, but is there more that you can tell? All you know is that the center of the weights is 2.72 kg.

You need more information.

The **range** of a set of data is the difference between the highest and the lowest data values (or maximum and minimum values). The **interval** is the lowest and highest values. The range is one value while the interval is two.

### Example: Range

From example #3.1.2, the maximum is 3.9 kg and the minimum is 2 kg. So the range is . But what does that tell you? You don’t know if the weights are really spread out, or if they are close together.

Unfortunately, range doesn’t really provide a very accurate picture of the variability. A better way to describe how the data is spread out is needed. Instead of looking at the distance the highest value is from the lowest how about looking at the distance each value is from the mean. This distance is called the **deviation**. You might want to find the average of the deviation. Though the calculation for finding the average deviation is not very straight forward, you end up with a value called the **variance**. The symbol for the population variance is , and it is the average squared distance from the mean. Statisticians like the variance, but many other people who work with statistics use a descriptive statistics which is the square root of the variance. This gives you the average distance from the mean. This is called the standard deviation, and is denoted with the letter .

The standard deviation is the average (mean) distance from a data point to the mean. It can be thought of as how much a typical data point differs from the mean.

The **sample variance** formula: , where is the sample mean, *n* is the sample size, and means to find the sum of the values.The on the bottom has to do with a concept called degrees of freedom. Basically, it makes the sample variance a better approximation of the population variance.

The **sample standard deviation** formula: .

The **population variance** formula: , where is the Greek letter sigma and represents the population variance, is the population mean, and *N* is the size of the population.

The **population standard deviation** formula:

Both the sample variance and sample standard deviation can be found using technology. If using R Studio, you would use

df\_stats(~variable, data=data frame, var, sd)

The next example will demonstrate this command.

### Example: Finding the Standard Deviation

For the data frame *Cats* from example #3.2.1, find the variance and standard derivation for weight of cats. Then find the variance and standard deviation separated by sex of the cat.

**Solution:** The variance and standard deviation for all cats is found by performing the command:

df\_stats(~Bwt, data=cats, var, sd)

## response var sd  
## 1 Bwt 0.2355225 0.4853066

The variance for all cats is 0.24 and the standard deviation is 0.49 kg.

To find out the mean, variance, and standard deviation for each sex of the cats, use the command:

df\_stats(Bwt~Sex, data=cats, mean, var, sd)

## response Sex mean var sd  
## 1 Bwt F 2.359574 0.07506938 0.2739879  
## 2 Bwt M 2.900000 0.21854167 0.4674844

You can see that the mean weight of females cats is 2.36 kg, the variance is 0.075 , and the standard deviation is 0.27 kg. For males cats, the mean is 2.9 kg, the variance is 0.22 , and the standard deviation is 0.47 kg. This means that female cats weigh less than males and since the variance and standard deviations are much less for female cats than males cats, female cats’ weights are more consistent than male cats.

In general a “small” variance and standard deviation means the data is close together (more consistent) and a “large” variance and standard deviation means the data is spread out (less consistent). Sometimes you want consistent data and sometimes you don’t. As an example if you are making bolts, you want the lengths to be very consistent so you want a small standard deviation. If you are administering a test to see who can be a pilot, you want a large standard deviation so you can tell who are the good pilots and who are the not so good pilots.

What do “small” and “large” standard deviation mean? To a bicyclist whose average speed is 20 mph, *s* = 20 mph is huge. To an airplane whose average speed is 500 mph, *s* = 20 mph is nothing. The “size” of the variation depends on the size of the numbers in the problem and the mean. Another situation where you can determine whether a standard deviation is small or large is when you are comparing two different samples such as in example #3.2.2. A sample with a smaller standard deviation is more consistent than a sample with a larger standard deviation.

Many other books and authors stress that there is a computational formula for calculating the standard deviation. However, this formula doesn’t give you an idea of what standard deviation is and what you are doing. It is only good for doing the calculations quickly. It goes back to the days when standard deviations were calculated by hand, and the person needed a quick way to calculate the standard deviation. It is an archaic formula that this author is trying to eradicate. It is not necessary anymore, computers will do the calculations for you with as much meaning as this formula gives. It is suggested that you never use it. If you want to understand what the standard deviation is doing, then you should use the definition formula. If you want an answer quickly, use a computer.

**Use of Standard Deviation**

One of the uses of the standard deviation is to describe how a population is distributed. This describes where much of the data is for most distributions. A general rule is that about 95% of the data is within 2 standard deviations of the mean. This is not perfect, but is works for many distributions. There are rules like the empirical rule and Chebyshev’s theorem that give you more detailed percentages, but 95% in 2 standard deviations is a very good approximation.

### Example: the gneral rule

The U.S. Weather Service has provided the information in table #3.2.1 about the total monthly/annual number of reported tornadoes in Oklahoma for the years 1950 to 2018. (US Department of Commerce & Noaa, 2016)

**Table #3.2.1: Monthyl/Annual Number of tornadoes in Oklahoma**

Tornado<-read.csv(  
 "https://krkozak.github.io/MAT160/Tornado\_OK.csv")  
head(Tornado)

## Year Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Annual  
## 1 1950 0 1 1 5 12 1 0 0 2 1 0 0 23  
## 2 1951 0 2 0 11 11 11 4 2 1 1 0 0 43  
## 3 1952 0 0 0 7 5 5 4 1 0 0 0 0 22  
## 4 1953 0 4 7 9 8 13 4 2 0 0 5 2 54  
## 5 1954 0 0 7 13 19 4 4 2 3 1 0 0 53  
## 6 1955 1 1 0 15 32 22 4 2 0 0 0 0 77

**Code book for data frame Pulse**

**Description** The U.S. Weather Service has collected data on the monthly and annual number of tornadoes in Oklahoma.

This data frame contains the following columns:

Year: Year from 1950-2018

Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec: Tornado numbers in each moth of the year

Annual: Total number of tornadoes for each year

**Source** US Department of Commerce, & Noaa. (2016, November 15). 1950 Oklahoma Tornadoes. Retrieved from <https://www.weather.gov/oun/tornadodata-ok-1950>

**References** The data was supplied by The U.S. Weather Service

Find the general interval that contains about 95% of the data.

**Solution:**

Variable: *x* = number of annual tornadoes in Oklahoma

Find the mean and standard deviation:

df\_stats(~Annual, data=Tornado, mean, sd)

## response mean sd  
## 1 Annual 56.02899 27.56061

The mean is tornadoes and the standard deviation is tornadoes. The interval will be

About 95% of the years have between 0.8 or 1 and 111 tornadoes in Oklahoma.

The general rule says that about 95% of the data is within two standard deviations of the mean. That percentage is fairly high. There isn’t much data outside two standard deviations. A rule that can be followed is that if a data value is within two standard deviations, then that value is a common data value. If the data value is outside two standard deviations of the mean, either above or below, then the number is uncommon. It could even be called unusual. An easy calculation that you can do to figure it out is to find the difference between the data point and the mean, and then divide that answer by the standard deviation. As a formula this would be

If you don’t know the population mean, , and the population standard deviation, , then use the sample mean, , and the sample standard deviation, *s*, to estimate the population parameter values. Realize that using the sample standard deviation may not actually be very accurate.

### Example: Determining If a Value Is Unusual

1. In 1974, there were 45 tornadoes in Oklahoma. Is this value unusual? Why or why not?

**Solution:**

Variable: *x* = number of tornadoes in Oklahoma

To answer this question, first find how many standard deviations 45 is from the mean. From example #3.2.3, we know and . For *x*=45,

Since this value is between -2 and 2, then it is not unusual to have 45 tornadoes in a year in Oklahoma. The z value is negative, so that means that 45 is less than the mean number of tornadoes.

1. In 1999, there were 145 tornadoes in the Oklahoma. Is this value unusual? Why or why not?

**Solution:**

Variable: *x* = number of tornadoes in Oklahoma

For this question the *x* = 145,

Since this value is more than 2, then it is unusual to have only 145 tornadoes in a year in Oklahoma.

### Homework

**Use Technology on all problems. State the variable on all problems.**

1. Cholesterol levels were collected from patients certain days after they had a heart attack and are in table #3.2.2. Find the mean, median, range, variance, and standard deviation for cholesterol levels 2 days after the heart attack.

**table #3.3.2: Cholesterol Levels of Patients After Heart Attack**

Cholesterol<-read.csv(  
 "https://krkozak.github.io/MAT160/cholesterol.csv")  
head(Cholesterol)

## patient day2 day4 day14  
## 1 1 270 218 156  
## 2 2 236 234 NA  
## 3 3 210 214 242  
## 4 4 142 116 NA  
## 5 5 280 200 NA  
## 6 6 272 276 256

**Code book for Data Frame Cholesterol** See problem 3.1.1 in Section 3.1 homework.

1. The lengths (in kilometers) of rivers on the South Island of New Zealand and what body of water they flow into are listed in table #3.1.2 (Lee, 1994). Find the mean, median, range, variance, and standard deviation of the length of rivers that flow into the Pacific Ocean and the mean, median, range, variance, and standard deviation of the length of rivers that flow into the Tasman Sea. Compare and contrast the length of rivers that flow to the Pacific Ocean versus the ones that flow into the Tasman Sea using both measures of spread and measures of variability.

**Table #3.3.3: Lengths of Rivers (km) Flowing to Pacific Ocean**

Length<-read.csv(  
 "https://krkozak.github.io/MAT160/length.csv")  
head(Length)

## river length flowsto  
## 1 Clarence 209 Pacific  
## 2 Conway 48 Pacific  
## 3 Waiau 169 Pacific  
## 4 Hurunui 138 Pacific  
## 5 Waipara 64 Pacific  
## 6 Ashley 97 Pacific

**Code book for data frame Length** See problem 3.1.2 in Section 3.1 homework.

1. Print-O-Matic printing company’s employees have salaries that are contained in table #3.2.4. Find the mean, median, range, variance, and standard deviation for the salaries of all employees.

**Table #3.2.4: Salaries of Print-O-Matic Printing Company Employees**

Pay<-read.csv(  
 "https://krkozak.github.io/MAT160/pay.csv")  
head(Pay)

## employee salary  
## 1 CEO 272500  
## 2 Driver 58456  
## 3 CD74 100702  
## 4 CD65 57380  
## 5 Embellisher 73877  
## 6 Folder 65270

**Code book for data frame Pay** See problem 3.1.3 in Section 3.1 homework.

1. Print-O-Matic printing company spends specific amounts on fixed costs every month. The costs of those fixed costs are in table #3.2.5. Find the mean, median, range, variance, and standard deviation for the fixed costs.

**Table #3.2.5: Fixed Costs for the Print-O-Matic Printing Company**

Cost<-read.csv(  
 "https://krkozak.github.io/MAT160/cost.csv")  
head(Cost)

## charges cost  
## 1 Bank charges 482  
## 2 Cleaning 2208  
## 3 Computer expensive 2471  
## 4 Lease payments 2656  
## 5 Postage 2117  
## 6 Uniforms 2600

**Code book for Data frame Cost** See problem 3.1.4 in Section 3.1 homework.

1. The data frame Pulse (Table 3.2.6) contains various variables about a person including their pulse rates before the subject exercised and after the subject ran in place for one minute.

**Table #3.2.6: Pulse Rates of people Before and After Exercise**

Pulse<-read.csv(  
 "https://krkozak.github.io/MAT160/pulse.csv")  
options(width = 60)  
head(Pulse)

## height weight age gender smokes alcohol exercise ran  
## 1 170 68 22 male yes yes moderate sat  
## 2 182 75 26 male yes yes moderate sat  
## 3 180 85 19 male yes yes moderate ran  
## 4 182 85 20 male yes yes low sat  
## 5 167 70 22 male yes yes low sat  
## 6 178 86 21 male yes yes low sat  
## pulse\_before pulse\_after year  
## 1 70 71 93  
## 2 80 76 93  
## 3 68 125 95  
## 4 70 68 95  
## 5 92 84 96  
## 6 76 80 98

**Code book for data frame Pulse**

**Description** Students in an introductory statistics class (MS212 taught by Professor John Eccleston and Dr Richard Wilson at The University of Queensland) participated in a simple experiment. The students took their own pulse rate. They were then asked to flip a coin. If the coin came up heads, they were to run in place for one minute. Otherwise they sat for one minute. Then everyone took their pulse again. The pulse rates and other physiological and lifestyle data are given in the data.

Five class groups between 1993 and 1998 participated in the experiment. The lecturer, Richard Wilson, was concerned that some students would choose the less strenuous option of sitting rather than running even if their coin came up heads, In the years 1995-1998 a different method of random assignment was used. In these years, data forms were handed out to the class before the experiment. The forms were pre-assigned to either running or non-running and there were an equal number of each. In 1995 and 1998 not all of the forms were returned so the numbers running and sitting was still not entirely controlled.

This data frame contains the following columns:

height: height of subject in cm

weight: weight of subject in kg

age: age of subject in years

gender: sex of subject, male, female

Smokes: whether a subject regularly smokes, yes means does smoke, no means does not smoke

alcohol: whether a subject regularly drinks alcohol, yes means the person does, no means the person does not

exercise: whether a subject exercises, low, moderate, high

ran: whether a subject ran one minute between pulse measurements (ran) or sat between pulse measurement (sat)

pulse\_before: the pulse rate before a subject either ran or sat (bpm)

pulse\_after: the pulse rate after a subject either ran or sat (bpm)

year: what year the data was collected (93-98)

**Source** Pulse rates before and after exercise. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/ms212.html>

**References** The data was supplied by Dr Richard J. Wilson, Department of Mathematics, University of Queensland.

Create a data frame that contains only males, who drink alcohol, but do not smoke. Then compare the pulse before and the pulse after using the mean and standard deviation. Discuss whether pulse before or pulse after has a higher mean and larger spread. To create a new data frame with just males, who drink alcohol, but do not smoke, use the following command, where the new name is Males:

Males<-  
Pulse%>%  
 filter(gender=="male", smokes == "no", alcohol == "yes")

1. The data frame Pulse (Table 3.2.6) contains various variables about a person including their pulse rates before the subject exercised and after after the subject ran in place for one minute. Create a data frame that contains females, who do not smoke but do drink alcohol. Compare the pulse rate before and after exercise using the mean and standard deviation. Discuss whether pulse before or pulse after has a higher mean and larger spread.
2. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) and a likert scale immediately before and after the Reiki treatment (Olson & Hanson, 1997) and the data is in table #3.2.7.

**Table #3.2.7: Pain Measurements Before and After Reiki Treatment**

Reiki<- read.csv(  
 "https://krkozak.github.io/MAT160/reki.csv")  
head(Reiki)

## vas.before vas.after likert\_before likert\_after  
## 1 6 3 2 1  
## 2 2 1 2 1  
## 3 2 0 3 0  
## 4 9 1 3 1  
## 5 3 0 2 0  
## 6 3 2 2 2

**Code book for data frame Reiki**

**Description** The purpose of this study was to explore the usefulness of Reiki as an adjuvant to opioid therapy in the management of pain. Since no studies in this area could be found, a pilot study was carried out involving 20 volunteers experiencing pain at 55 sites for a variety of reasons, including cancer. All Reiki treatments were provided by a certified second-degree Reiki therapist. Pain was measured using both a visual analogue scale (VAS) and a Likert scale immediately before and after the Reiki treatment. Both instruments showed a highly significant (p < 0.0001) reduction in pain following the Reiki treatment.

This data frame contains the following columns:

vas.before: pain measured using a visual analogue scale (VAS) before Reiki treatment

vas.after: pain measured using a visual analogue scale (VAS) after Reiki treatment

likert\_before: pain measured using a likert before Reiki treatment

likert\_after: pain measured using a likert after Reiki treatment

**Source** Olson, K., & Hanson, J. (1997). Using reiki to manage pain: a preliminary report. Cancer Prev Control, 1(2), 108-13. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9765732>

**References** Using Reiki to manage pain: a preliminary report. Olson K1, Hanson J., Cancer Prev Control 1997, Jun; 1(2): 108-13.

Since the data was collected both before and after the treatment for all of the units of observations, you want to look at the effect size of the treatment. You want to find the difference between before and after for the pain scale. First you must create a new data frame that adds a column for the difference in before and after. This data is known as paired data. To create the new column in a new data frame called Newreiki use the following commands

Newreiki<-Reiki%>%  
 mutate(vas.diff=vas.before-vas.after)  
head(Newreiki)

## vas.before vas.after likert\_before likert\_after vas.diff  
## 1 6 3 2 1 3  
## 2 2 1 2 1 1  
## 3 2 0 3 0 2  
## 4 9 1 3 1 8  
## 5 3 0 2 0 3  
## 6 3 2 2 2 1

Now find the mean and standard deviation of the vas.diff variable in Newreiki. Perform similar commands to create the likert.diff variable. Then find the mean and standard deviation for likert.diff, and compare and contrast the vas and likert methods for describing pain.

8.Yearly rainfall amounts (in millimeters) in Sydney, Australia, are in table #3.2.8 (Annual maximums of, 2013). a. Calculate the mean and standard deviation. b. Suppose Sydney, Australia received 300 mm of rainfall in a year. Would this be unusual?

\*\* Table #3.2.8: Yewarly rainfall amounts in Sydney, Australia\*\*

Rainfall<-read.csv(  
 "https://krkozak.github.io/MAT160/rainfall.csv")  
head(Rainfall)

## amount  
## 1 146.8  
## 2 383.0  
## 3 90.9  
## 4 178.1  
## 5 267.5  
## 6 95.5

**Code book for data frame Rainfall**

**Description** Daily rainfall (in millimeters) was recorded over a 47-year period in Turramurra, Sydney, Australia. For each year, the wettest day was identified (that having the greatest rainfall). The data show the rainfall recorded for the 47 annual maxima.

This data frame contains the following columns:

amount: daily rainfall (mm)

**Source** Annual maximums of daily rainfall in Sydney. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/sydrain.html>

**References** Rayner J.C.W. and Best D.J. (1989) Smooth tests of goodness of fit. Oxford: Oxford University Press. Hand D.J., Daly F., Lunn A.D., McConway K.J., Ostrowski E. (1994). A Handbook of Small Data Sets. London: Chapman & Hall. Data set 157. Thanks to Jim Irish of the University of Technology, Sydney, for assistance in identifying the correct units for this data.

## Ranking

Along with the center and the variability, another useful numerical measure is the ranking of a number. A **percentile** is a measure of ranking. It represents a location measurement of a data value to the rest of the values. Many standardized tests give the results as a percentile. Doctors also use percentiles to track a child’s growth.

The **kth percentile** is the data value that has k% of the data at or below that value.

### Example: Interpreting Percentile

1. What does a score of the 90th percentile mean?

**Solution:**

This means that 90% of the scores were at or below this score. (A person did the same as or better than 90% of the test takers.)

1. What does a score of the 70th percentile mean?

**Solution:**

This means that 70% of the scores were at or below this score.

### Example: Percentile Versus Score

If the test was out of 100 points and you scored at the 80th percentile, what was your score on the test?

**Solution:**

You don’t know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were really low, you could have still failed the test. On the other hand, if many of the scores were high you could have gotten a 95% or more.

There are special percentiles called **quartiles**. Quartiles are numbers that divide the data into fourths. One fourth (or a quarter) of the data falls between consecutive quartiles.

**To find the quartiles:**

The command in R Studio is

df\_stats(~variable, data=data frame, summary)

If you record the quartiles together with the maximum and minimum you have five numbers. This is known as the five-number summary. The five-number summary consists of the minimum, the first quartile (*Q1*), the median, the third quartile (*Q3*), and the maximum (in that order).

The interquartile range, *IQR*, is the difference between the first and third quartiles, *Q1* and *Q3*. Half of the data (50%) falls in the interquartile range. If the *IQR* is “large” the data is spread out and if the *IQR* is “small” the data is closer together.

Interquartile Range (*IQR*)

Determining probable outliers from IQR: **fences**

A value that is less than (this value is often referred to as a **low** **fence**) is considered an outlier.

Similarly, a value that is more than (the **high** **fence**) is considered an outlier.

A boxplot (or box-and-whisker plot) is a graphical display of the five-number summary. It can be drawn vertically or horizontally. The basic format is a box from *Q1* to *Q3*, a vertical line across the box for the median and horizontal lines as whiskers extending out each end to the minimum and maximum. The minimum and maximum can be represented with dots. Don’t forget to label the tick marks on the number line and give the graph a title.

An alternate form of a Boxplot, known as a modified box plot, only extends the left line to the smallest value greater than the *low fence*, and extends the left line to the largest value less than the *high fence*, and displays markers (dots, circles or asterisks) for each outlier.

If the data are *symmetrical*, then the box plot will be visibly symmetrical. If the data distribution has a left skew or a right skew, the line on that side of the box plot will be visibly long. If the plot is symmetrical, and the four quartiles are all about the same length, then the data are likely a near *uniform* distribution. If a box plot is symmetrical, and both outside lines are noticeably longer than the Q1 to median and median to Q3 distance, the distribution is then probably *bell-shaped.*

### Example: Five-number Summary and Boxplot

Find the five-number summary, the interquartile range (*IQR*), and draw a box-and-whiskers plot for the weight of cats.

head(cats)

## Sex Bwt Hwt  
## 1 F 2.0 7.0  
## 2 F 2.0 7.4  
## 3 F 2.0 9.5  
## 4 F 2.1 7.2  
## 5 F 2.1 7.3  
## 6 F 2.1 7.6

**Solution:**

Variable: *x* = weight of cats To compute the five-number summary on R Studio, use the command:

df\_stats(~Bwt, data=cats, summary)

## response Min. 1st Qu. Median Mean 3rd Qu. Max.  
## 1 Bwt 2 2.3 2.7 2.723611 3.025 3.9

Not R Studo also calculates the mea as part of the summary command, but the five-number summary is jus the five numbers:

Minimum: 2 kg *Q1*: 2.3 kg Median: 2.7 kg *Q3*: 3.025 kg Maximum: 3.9 kg

To find the interquartile range, *IQR*, find , so

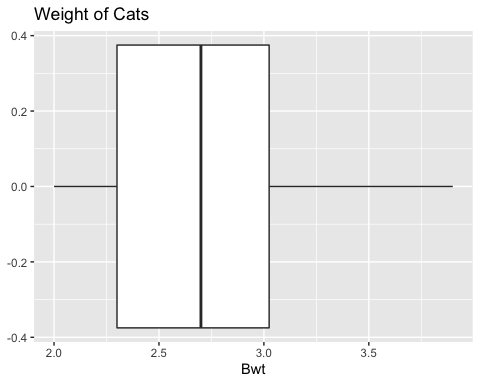
To create a boxplot use the command

gf\_boxplot(~variable, data=data frame)

This is a modified boxplot which shows the outliers in the data.

(ref:cats1-data-box-cap) Box Plot of Weights of Cats

gf\_boxplot(~Bwt, data=cats, title="Weight of Cats")



(ref:cats1-data-box-cap)

There are no outliers since there are no dots outside of the fences.

### Example: Separating based on a factor

Find the five-number summary of the weights of cats separated by the sex of the cat. Then create a box plot of the weights of cats for each sex of the cat.

**Solution:**

Variable: *x\_1* = weight of female cat

Variable: *x\_2* = weight of male cat

To find the five-number summary separated based on gender use the following command:

df\_stats(~Bwt|Sex, data=cats, summary)

## response Sex Min. 1st Qu. Median Mean 3rd Qu. Max.  
## 1 Bwt F 2 2.15 2.3 2.359574 2.5 3.0  
## 2 Bwt M 2 2.50 2.9 2.900000 3.2 3.9

The five-number summary for female cats is (in kg)

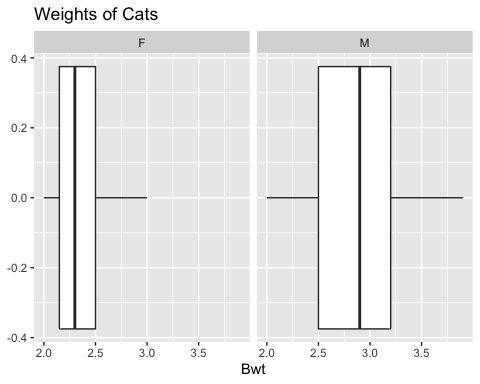
Minimum: 2 Q1: 2.15 Median: 2.3 Q3: 2.5 Maximum: 3.0

The five-number summary for male cats is (in kg)

Minimum: 2 Q1: 2.50 Median: 2.9 Q3: 3.2 Maximum: 3.9

(ref:cats2-data-box-cap) Box Plot of Cats Weights Separated by Sex

gf\_boxplot(~Bwt|Sex, data=cats, title="Weights of Cats")



(ref:cats2-data-box-cap)

Notice that the weights of female cats has a median less than male cats, and in fact it can be seen that the Q1 to Q3 of the female cats is less than the Q1 to Q3 of the male cats.

### Example: Putting it all together

The time (in 1/50 seconds) between successive pulses along a nerve fiber (“Time between nerve,” 2013) are given in table #3.3.1.

**Table #3.3.1: Successive pulses along a nerve fiber**

Nerve<-read.csv(  
 "https://krkozak.github.io/MAT160/Nerve\_pulse.csv")  
head(Nerve)

## time  
## 1 10.5  
## 2 1.5  
## 3 2.5  
## 4 5.5  
## 5 29.5  
## 6 3.0

**Code book for data frame Nerve**

**Description** The data gives the time between 800 successive pulses along a nerve fiber. There are 799 observations rounded to the nearest half in units of 1/50 second.

This data frame contains the following columns:

time: time between successive Pulses along a nerve fiber, 1/50 second.

**Source** *Time between nerve pulses*. (2019, July 3). Retrieved from <<http://www.statsci.org/data/general/nerve.html>

**References** Fatt, P., and Katz, B. (1952). Spontaneous subthreshold activity at motor nerve endings. Journal of Physiology 117, 109-128.

Cox, D. R., and Lewis, P. A. W. (1966). The Statistical Analysis of Series of Events. Methuen, London.

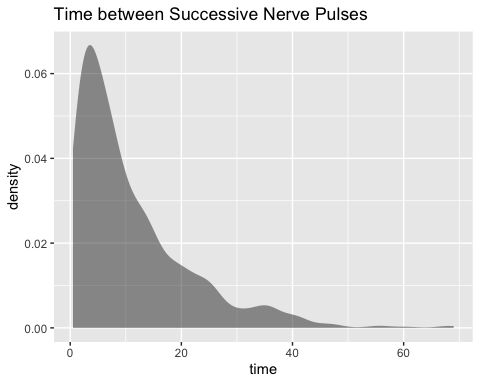
Jorgensen, B. (1982). The Generalized Inverse-Gaussian Distribution. Springer-Verlag.

**Solution:**

First, it might be useful to look at a visualization of the data, so create a density plot

(ref:density-data-density-cap) Density Plot of Nerve Pulses

gf\_density(~time, data=Nerve, title="Time between Successive Nerve Pulses")



(ref:density-data-density-cap)

From the graph (Figure @ref(fig:density-data-density)), the data appears to be skewed right. Most of the time between successive nerve pulses appear to be around 5 or 10 1/50 second, but there are some times that are 60 1/50 second.

df\_stats(~time, data=Nerve, mean, median, sd, summary)

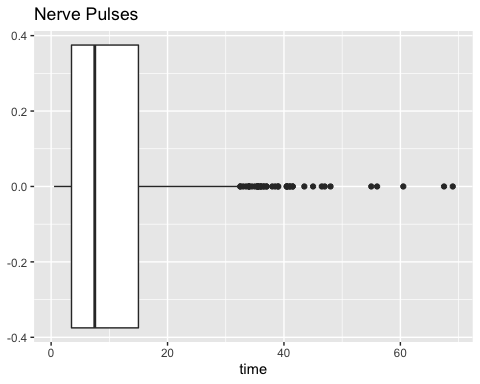
## response mean median sd Min. 1st Qu. Median  
## 1 time 10.95119 7.5 10.45956 0.5 3.5 7.5  
## Mean 3rd Qu. Max.  
## 1 10.95119 15 69

Numerical descriptions might also be useful. Using technology, the mean is 11 1/50 second,the median is 7.5 1/50 second, the standard deviation is 10.5 1/50 second, and the five-number summary is minimum = 3.5, Q1 = 3.5, median = 7.5, Q3 = 15, and maximum = 69 1/50 second.

To visualize the five-number summary, create a box plot.

(ref:nerve-data-box-cap) Box Plot of Health Expenditure

gf\_boxplot(~time, data=Nerve, title="Nerve Pulses")



(ref:nerve-data-box-cap)

Since there are many dots outside the upper fence the data has many outliers. From all of this information, one could say that nerve pulses between successive pulses is around 11 1/50 second, with a spread of 19.5 1/50 second. Most of the values are round 11 1/50 second, but they are not very consistent. The density plot and boxplot show that there is a great deal of spread of the data and it is skewed to the right. This means mostly the speed is around 11 1/50 second, but there is a great deal of variability in the values.

### Homework

**Use Technology on all problems. State the variable on all problems.**

1. Suppose you take a standardized test and you are in the 10th percentile. What does this percentile mean? Can you say that you failed the test? Explain.
2. Suppose your child takes a standardized test in mathematics and scores in the 96th percentile. What does this percentile mean? Can you say your child passed the test? Explain.
3. Suppose your child is in the 83rd percentile in height and 24th percentile in weight. Describe what this tells you about your child’s stature.
4. Suppose your work evaluates the employees and places them on a percentile ranking. If your evaluation is in the 65th percentile, do you think you are working hard enough? Explain.
5. Cholesterol levels were collected from patients certain days after they had a heart attack and are in table #3.3.2.

**table #3.3.2: Cholesterol Levels of Patients After Heart Attack**

Cholesterol<-read.csv(  
 "https://krkozak.github.io/MAT160/cholesterol.csv")  
head(Cholesterol)

## patient day2 day4 day14  
## 1 1 270 218 156  
## 2 2 236 234 NA  
## 3 3 210 214 242  
## 4 4 142 116 NA  
## 5 5 280 200 NA  
## 6 6 272 276 256

**Code book for Data Frame Cholesterol** See problem 3.1.1 in Section 3.1 homework.

Find the five-number summary and interquartile range (IQR) for the cholesterol level on day 2, and draw a boxplot

1. The lengths (in kilometers) of rivers on the South Island of New Zealand and what body of water they flow into are listed in table #3.3.3 (Lee, 1994).

**Table #3.3.3: Lengths of Rivers (km) Flowing to Pacific Ocean**

Length<-read.csv(  
 "https://krkozak.github.io/MAT160/length.csv")  
head(Length)

## river length flowsto  
## 1 Clarence 209 Pacific  
## 2 Conway 48 Pacific  
## 3 Waiau 169 Pacific  
## 4 Hurunui 138 Pacific  
## 5 Waipara 64 Pacific  
## 6 Ashley 97 Pacific

**Code book for data frame Length** See problem 3.1.2 in Section 3.1 homework.

Find the five-number summary and interquartile range (IQR) for the lengths of rivers that go to the Pacific Ocean and ones that go to the Tasman Sea, and draw a boxplot of both.

1. Print-O-Matic printing company’s employees have salaries that are contained in table #3.3.4. Find the five number summary and draw a boxplot for the salaries of all employees.

**Table #3.3.4: Salaries of Print-O-Matic Printing Company Employees**

Pay<-read.csv(  
 "https://krkozak.github.io/MAT160/pay.csv")  
head(Pay)

## employee salary  
## 1 CEO 272500  
## 2 Driver 58456  
## 3 CD74 100702  
## 4 CD65 57380  
## 5 Embellisher 73877  
## 6 Folder 65270

**Code book for data frame Pay** See problem 3.1.3 in Section 3.1 homework.

1. The data frame Pulse (Table 3.3.5) contains various variables about a person including their pulse rates before the subject exercised and after after the subject ran in place for one minute.

**Table #3.3.5: Pulse Rates of Males Before and After Exercise**

Pulse<-read.csv(  
 "https://krkozak.github.io/MAT160/pulse.csv")  
head(Pulse)

## height weight age gender smokes alcohol exercise ran  
## 1 170 68 22 male yes yes moderate sat  
## 2 182 75 26 male yes yes moderate sat  
## 3 180 85 19 male yes yes moderate ran  
## 4 182 85 20 male yes yes low sat  
## 5 167 70 22 male yes yes low sat  
## 6 178 86 21 male yes yes low sat  
## pulse\_before pulse\_after year  
## 1 70 71 93  
## 2 80 76 93  
## 3 68 125 95  
## 4 70 68 95  
## 5 92 84 96  
## 6 76 80 98

**Code book for data frame Pulse** See Problem 3.2.5 in Section 3.2 Homework

Create a data frame that contains only people who drink alcohol, but do not smoke. Then find the five number summary and draw a boxplot for both males and females separately.

1. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) and a likert scale immediately before and after the Reiki treatment (Olson & Hanson, 1997) and the data is in table #3.3.6.

**Table #3.2.7: Pain Measurements Before and After Reiki Treatment**

Reiki<- read.csv(  
 "https://krkozak.github.io/MAT160/reki.csv")  
head(Reiki)

## vas.before vas.after likert\_before likert\_after  
## 1 6 3 2 1  
## 2 2 1 2 1  
## 3 2 0 3 0  
## 4 9 1 3 1  
## 5 3 0 2 0  
## 6 3 2 2 2

**Code book for data frame Reiki** see problem 3.2.7 in Section 3.2 Homework

Find the five number summary for both the before and after VAS scores and draw boxplots of before and after VAS scores. To draw two boxplots at the same time, after the command to create the first box plot type %>% before pressing enter. Then type the command for the second boxplot after the + symbol. Then press enter. You may want to graph each boxplot as a different color. To do this, the command would be

gf\_boxplot(~variable, data=data frame, color="red")

You can pick any color you want. Just replace the word red with the color you want to use. Now compare and contrast the before and after VAS scores.

Data Sources:

*Annual maximums of daily rainfall in Sydney*. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/sydrain.html>

Lee, A. (1994). *Data analysis: An introduction based on r. Auckland*. Retrieved from <http://www.statsci.org/data/oz/nzrivers.html>

*Life expectancy in southeast Asia*. (2013, September 23). Retrieved from <http://apps.who.int/gho/data/node.main.688>

Olson, K., & Hanson, J. (1997). Using reiki to manage pain: a preliminary report. *Cancer Prev Control*, *1*(2), 108-13. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9765732>

*Pulse rates before and after exercise*. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/ms212.html>

Ryan, B. F., Joiner, B. L., & Ryan, Jr, T. A. (1985). *Cholesterol levels after heart attack*. Retrieved from <http://www.statsci.org/data/general/cholest.html>

*Time between nerve pulses*. (2019, July 3). Retrieved from <http://www.statsci.org/data/general/nerve.html>

*Time of passages of play in rugby*. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/rugby.html>

US Department of Commerce, & Noaa. (2016, November 15). 1950 Oklahoma Tornadoes. Retrieved from <https://www.weather.gov/oun/tornadodata-ok-1950>

*UV radiation: Burden of disease by country*. (2013, September 4). Retrieved from <http://apps.who.int/gho/data/node.main.165?lang=en>