Stat 515: Introduction to Statistics

Chapter 2

Watch This!

- Histograms vs. Bar Chart vs. Line:
  - https://www.youtube.com/watch?v=k8WGdcTt5gc

- Really cool example*:
  - https://www.youtube.com/watch?v=jbkSRLYSojo

Recall: Types of Variables

- **Qualitative(Categorical)**: Observations that belong to a set of categories
  - Examples: gender, hair color, eye color, ethnicity, origin, favorite color, major, etc.

- **Quantitative:** Observations that take on numerical values
  - Examples: Height, weight, age, GPA, etc.

Recall: Types of Variables

- **Quantitative:** Observations that take on numerical values
  - **Discrete:** measured by a whole number
    - Examples: Number of books, children, money, etc
  - **Continuous:** measured on an interval
    - Examples: Height, weight, age, GPA, etc.
    - Note: These are often measured as a discrete variable

Talking about Different Variables

- Knowing what type of variable what we are interested in is important because it tells us what statistics and charts are appropriate for summarizing
- We spend a little time talking about qualitative(categorical) data and much more time talking about quantitative data

Summarizing Qualitative Data

- **Qualitative(Categorical)**: Observations that belong to a set of categories
  - Qualitative variables can be broken up into **classes,** the possible categories that make up the variable

Gender
Male
Female
Rather not say

Hair Color
Red
Blonde
Black
Brown
Other

Eye Color
Blue
Green
Brown
Hazel
Other

Ethnicity
White
Hispanic
Black
Native American
Asian
Other

Color
Red
Blue
Green
Purple
Pink
Yellow
Other

Major
Engineering
Fine Art
Business
Journalism
Chemistry
Medical
Other

Summarizing Qualitative Data: Frequencies

- A **Frequency Distribution** lists each category of the variable and the number or proportion of occurrences for each category of data.

Summarizing Qualitative Data: Frequencies

- **Class Frequency** is the number of occurrences for each class of variable of interest

- **Relative Frequency** is the proportion of observations of a class among all observations of the variable of interest

NOTE!

- **Relative Frequency** is the proportion of observations within a category and is found using the following formula

$$Relative\ Freq. = \frac{frequency}{sum\ of\ all\ frequencies}$$

Relative Frequency is also referred to as a **proportion, $\boldsymbol{\widehat{p}}$ or $\boldsymbol{\rho}$.** This will be really important later in the semester!

Example

- The 2012 South Carolina Republican Primary was held on January 21st. Newt Gingrich, Mitt Romney, Rick Santorum, Ron Paul, Herman Cain, Rick Perry, Jon Huntsman, Michele Bachmann and Gary Johnson were on the ballet for voters to choose from.

Example

Candidate Chosen	Class Frequency - the number of times candidate 'x' was voted for	Relative Frequency- the proportion of times candidate 'x' was voted for
Class = X = Bachmann	491	
Class = X = Cain	6,338	
Class = X = Gingrich	244,065	
Class = X = Huntsman	1,173	
Class = X = Johnson	211	
Class = X = Paul	78,360	
Class = X = Perry	2,534	
Class = X = Romney	168,123	
Class = X = Santorum	102,475	
TOTAL	603,770	

Example

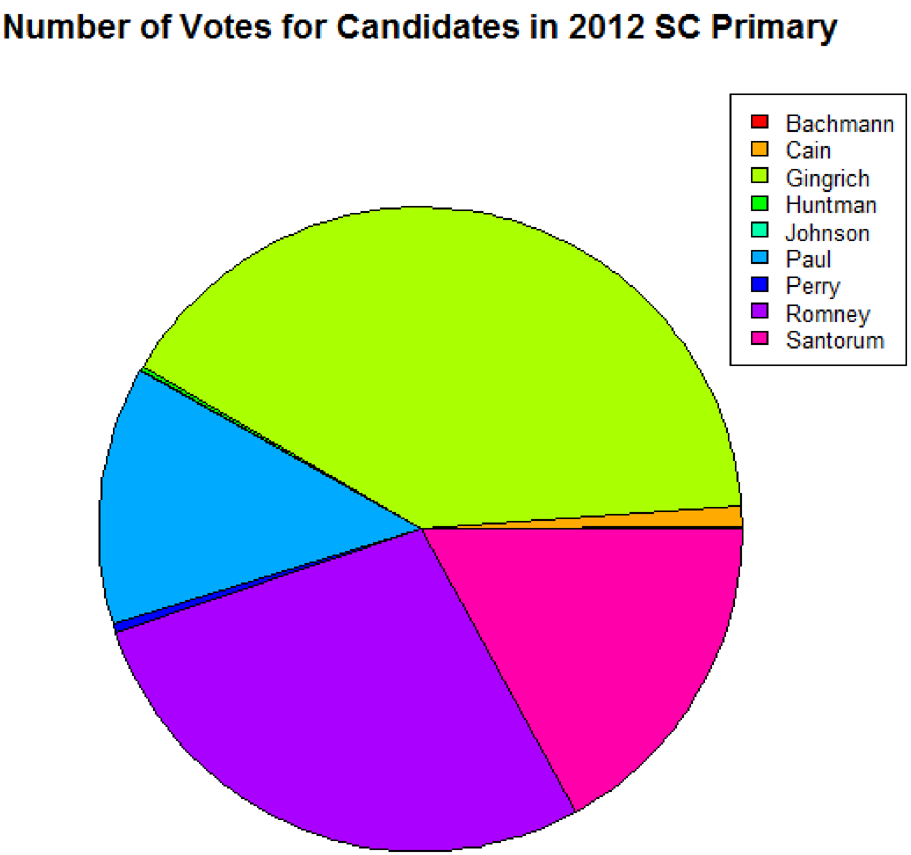
Candidate Chosen	Class Frequency - the number of times candidate ‘x’ was voted for	Relative Frequency- the proportion of times candidate ‘x’ was voted for
Class = X = Bachmann	491	491/603,770 = .0008
Class = X = Cain	6,338	6,338/603,770 = .0105
Class = X = Gingrich	244,065	244,065/603,770 = .4042
Class = X = Huntsman	1,173	1,173/603,770 = .0019
Class = X = Johnson	211	211/603,770 = .0003
Class = X = Paul	78,360	78,360/603,770 = .1298
Class = X = Perry	2,534	2,534/603,770 = .0042
Class = X = Romney	168,123	168,123/603,770 = .2785
Class = X = Santorum	102,475	102,475/603,770 = .1697
TOTAL	603,770	~1

Example

Candidate Chosen	Class Frequency - the number of times candidate ‘x’ was voted for	Relative Frequency- the proportion of times candidate ‘x’ was voted for
Class = X = Bachmann	491	491/603,770 = .0008 = .08%
Class = X = Cain	6,338	6,338/603,770 = .0105 = 1.05%
Class = X = Gingrich	244,065	244,065/603,770 = .4042 = 40.42%
Class = X = Huntsman	1,173	1,173/603,770 = .0019 = .19%
Class = X = Johnson	211	211/603,770 = .0003 = .03%
Class = X = Paul	78,360	78,360/603,770 = .1298 = 12.98%
Class = X = Perry	2,534	2,534/603,770 = .0042 = .42%
Class = X = Romney	168,123	168,123/603,770 = .2785 = 27.85%
Class = X = Santorum	102,475	102,475/603,770 = .1697 = 16.97%
TOTAL	603,770	~100%

Summarizing Qualitative Data: Pie Chart

- Useful when there are a small number of categories



Summarizing Qualitative Data: Pie Chart

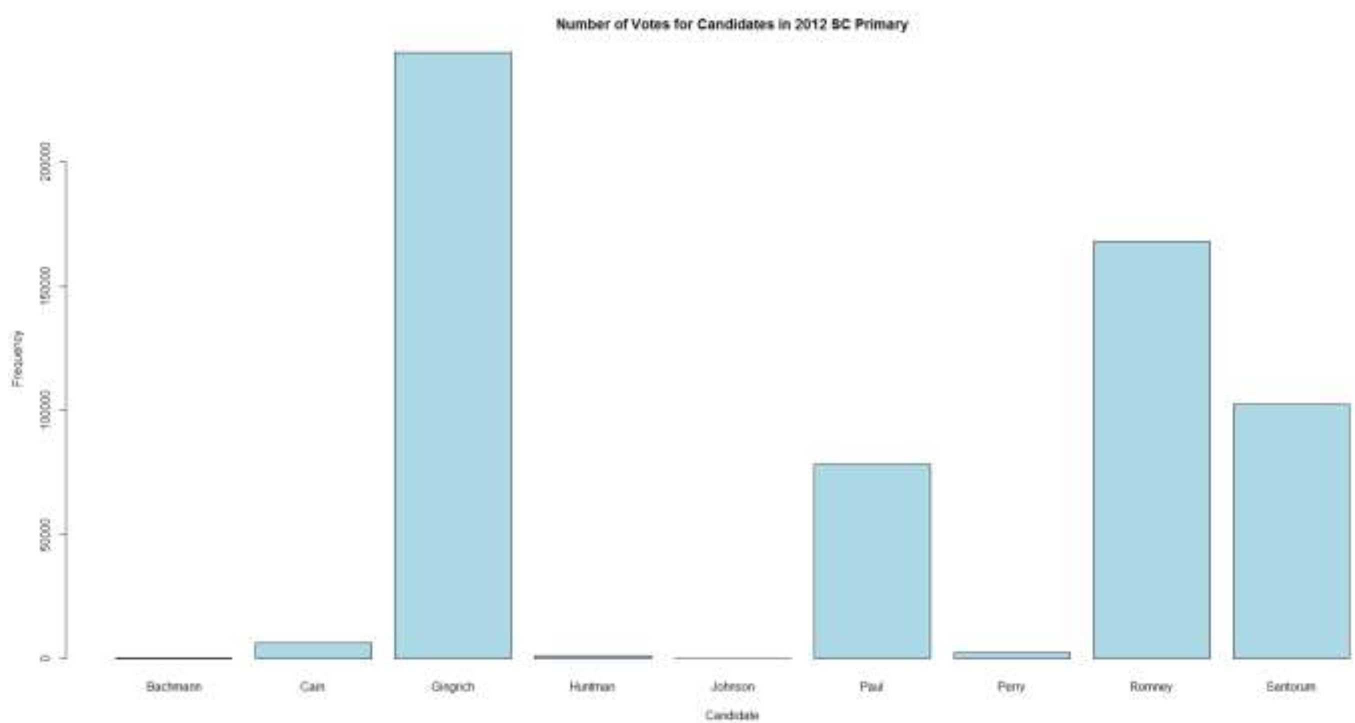
- **R Commands:**

```
####################################################################
#######################Creating a Pie Chart in R#########################
####################################################################
PrimaryVotes<-c(491,6338,244065,1173,211,78360,2534,168123,102475)
PrimaryNames<-c("Bachmann", "Cain", "Gingrich", "Huntman", "Johnson", "Paul", "Perry", "Romney", "Santorum")
#BASIC
pie(PrimaryVotes)
#Add Title and fix labels
pie(PrimaryVotes,labels=PrimaryNames,main="Number of Votes for Candidates in 2012 SC Primary")
#Add colors
colors=rainbow(9)#because we have ten classes
pie(PrimaryVotes, labels=PrimaryNames, col=colors,main="Number of Votes for Candidates in 2012 SC Primary")
#add legend instead of names on the graph
pie(PrimaryVotes, labels=rep("",9), col=colors,main="Number of Votes for Candidates in 2012 SC Primary")
legend("topright", PrimaryNames, cex=0.8, fill=colors)
```

More Examples: http://www.harding.edu/fmccown/r/

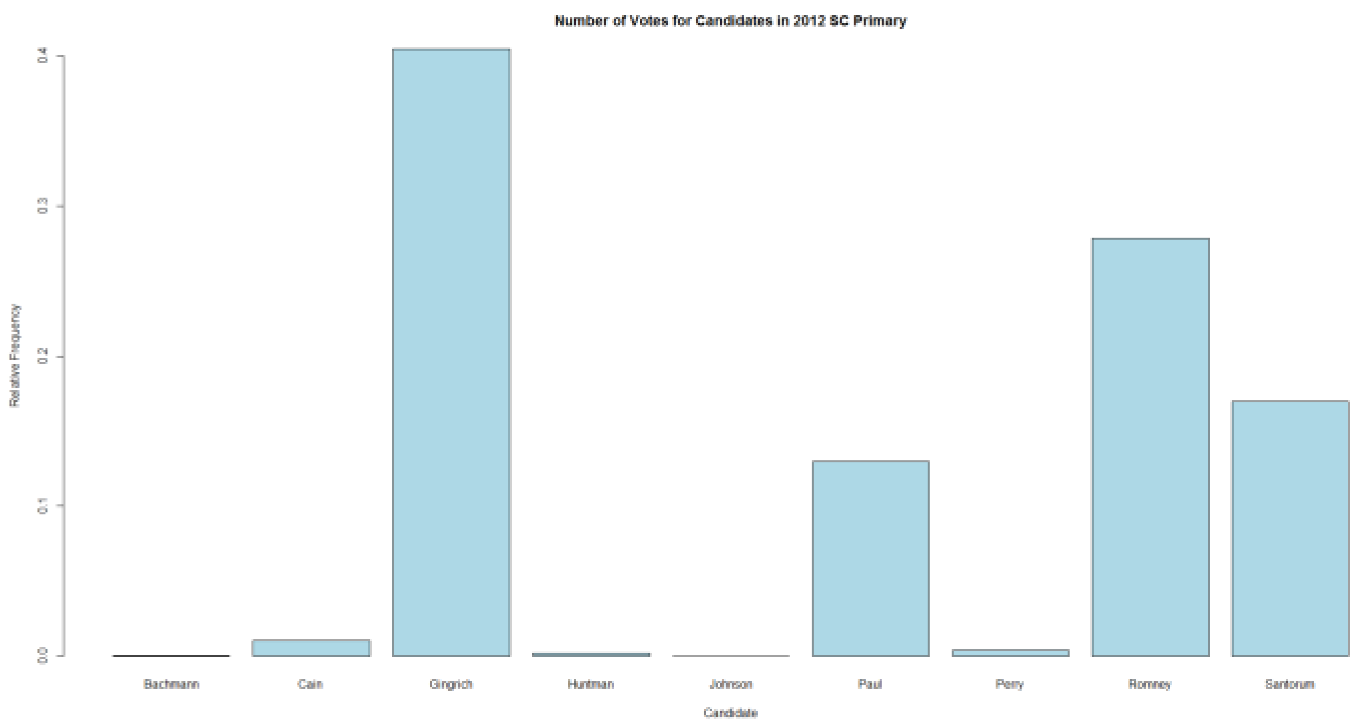
Summarizing Qualitative Data: Bar Graph

- Useful when there are many categories of the variable
- Useful to compare groups



Summarizing Qualitative Data: Bar Graph

- **Note:** the relative frequency chart has the same shape but a different y-axis



Summarizing Qualitative Data: Bar Graph

- **R Commands:**

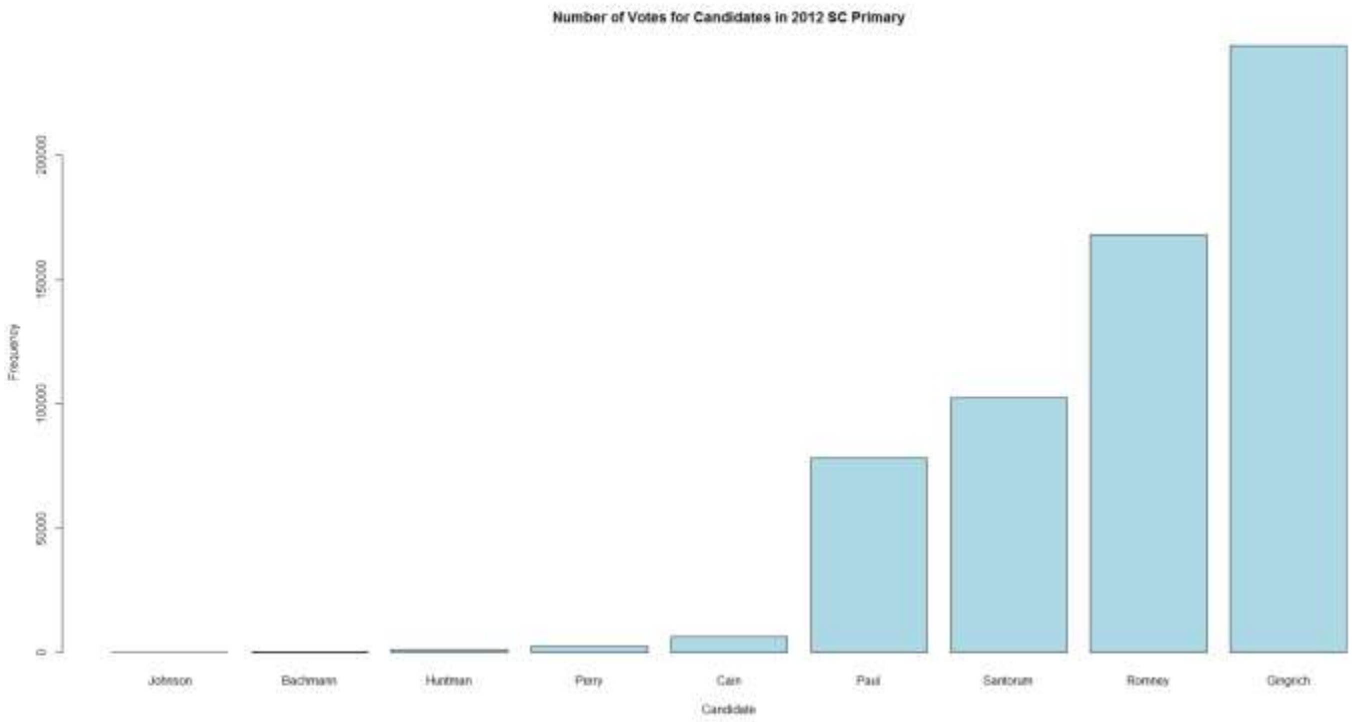
```
######################################################################
###################Creating a Bar Chart in R#######################
######################################################################
#BASIC Class Frequency
barplot(PrimaryVotes)
#Add Title
barplot(PrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary")
#Add X-values
barplot(PrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames)
#Add X-Label
barplot(PrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames,xlab="Candidate")
#Add Y-label
barplot(PrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames,xlab="Candidate", ylab="Frequency")
#Add Colors
barplot(PrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames,xlab="Candidate", ylab="Frequency", col="light blue")

#BASIC Class Relative Frequency
RelPrimaryVotes<-PrimaryVotes/sum(PrimaryVotes)
barplot(RelPrimaryVotes)
#Add Title
barplot(RelPrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary")
#Add X-vaules
barplot(RelPrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames)
#Add X-Label
barplot(RelPrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames,xlab="Candidate")
#Add Y-label
barplot(RelPrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames,xlab="Candidate", ylab="Relative Frequency")
#Add Colors
barplot(RelPrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=PrimaryNames,xlab="Candidate", ylab="Relative Frequency", col="light blue")
```

More Examples: http://www.harding.edu/fmccown/r/

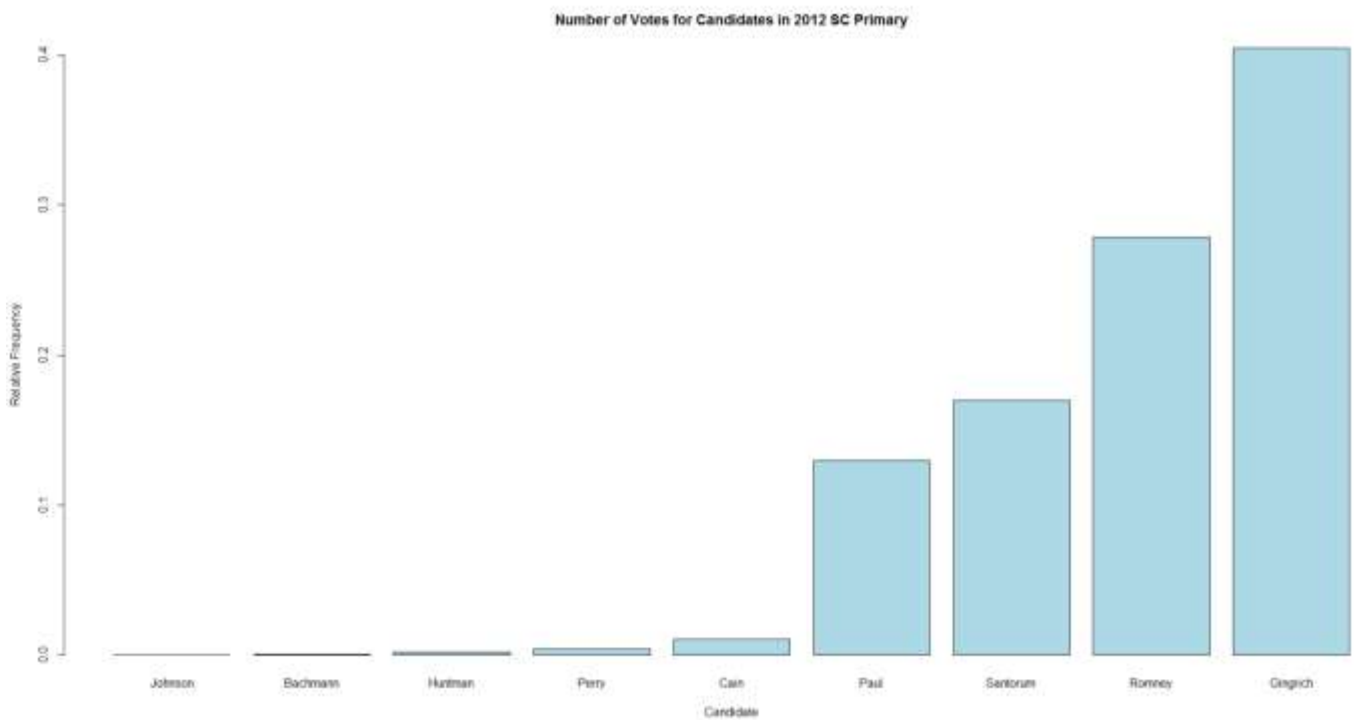
Summarizing Qualitative Data: Pareto Graph

- Same as the bar graph except the bars are ordered by height, making it easier to see what happens ‘most’ or ‘least.’



Summarizing Qualitative Data: Pareto Graph

- **Note:** the relative frequency chart has the same shape but a different y-axis



Summarizing Qualitative Data: Pareto Graph

- **R Commands:**

```
######################################################################
###################Creating a Pareto Chart in R#######################
######################################################################
#Pareto Frequency
#Find out the order of the table
order(PrimaryVotes)
#Reorder Frequency Table
OrdPrimaryVotes<-PrimaryVotes[order(PrimaryVotes)]
OrdPrimaryNames<-PrimaryNames[order(PrimaryVotes)]
#Complete a bar chart using this table
barplot(OrdPrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=OrdPrimaryNames,xlab="Candidate", ylab="Frequency", col="light blue")

#Pareto Relative Frequency
#Find out the order of the table
order(RelPrimaryVotes)
#Reorder Relative Frequency Table
OrdRelPrimaryVotes<-RelPrimaryVotes[order(RelPrimaryVotes)]
OrdPrimaryNames<-PrimaryNames[order(RelPrimaryVotes)]
#Complete a bar chart using this table
barplot(OrdRelPrimaryVotes,main="Number of Votes for Candidates in 2012 SC Primary",names.arg=OrdPrimaryNames,xlab="Candidate", ylab="Relative Frequency", col="light blue")
```

More Examples: http://www.harding.edu/fmccown/r/

End: Summarizing Qualitative Data

Summarizing Quantitative Data

- **Quantitative**: Observations that take on numerical values
  - Quantitative variables cannot be broken up into **classes,** like qualitative variables; there are many more possible values this variable can take on.
  - It should be noted that **discrete** variables can be treated like qualitative variables when there is a small number of observable values

Example

- In the English Premier League(EPL) season spanning 2013 and 2014 matches had between 0 and 9 goals.
  - Even though our variable is measured in numbers we can treat it as a qualitative variable because we can think of each number as a category or class
  - The **classes** of the number of goals for matches in the ’13-’14 EPL season are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 because the minimum number of goals was zero, the maximum number of goals was 9 and they increment by one.

Example

Total Goals in EPL matches '13/'14	Class Frequency - the number of times 'x' goals were scored in a match	Relative Frequency- the proportion of times 'x' goals were scored in a match
Class = X = 0	27	
Class = X = 1	75	
Class = X = 2	82	
Class = X = 3	70	
Class = X = 4	63	
Class = X = 5	39	
Class = X = 6	17	
Class = X = 7	4	
Class = X = 8	1	
Class = X = 9	2	
TOTAL	160	

Example

Total Goals in EPL matches ‘13/’14	Class Frequency - the number of times ‘x’ goals were scored in a match	Relative Frequency- the proportion of times ‘x’ goals were scored in a match
Class = X = 0	27	27/380 = .0711
Class = X = 1	75	75/380 = .1974
Class = X = 2	82	82/380 = .2158
Class = X = 3	70	70/380 = .1842
Class = X = 4	63	63/380 = .1658
Class = X = 5	39	39/380 = .1026
Class = X = 6	17	17/380 = .0447
Class = X = 7	4	4/380 = .0105
Class = X = 8	1	1/380 = .0026
Class = X = 9	2	2/380 = .0053
TOTAL	380	1

Example

Total Goals in EPL matches '13/'14	Class Frequency - the number of times 'x' goals were scored in a match	Relative Frequency- the proportion of times 'x' goals were scored in a match
Class = X = 0	27	27/380 = .0711 = 7.11%
Class = X = 1	75	75/380 = .1974 = 19.74%
Class = X = 2	82	82/380 = .2158 = 21.58%
Class = X = 3	70	70/380 = .1842 = 18.42%
Class = X = 4	63	63/380 = .1658 = 16.58%
Class = X = 5	39	39/380 = .1026 = 10.26%
Class = X = 6	17	17/380 = .0447 = 4.47%
Class = X = 7	4	4/380 = .0105 = 1.05%
Class = X = 8	1	1/380 = .0026 = .26%
Class = X = 9	2	2/380 = .0053 = .53%
TOTAL	380	100%

Example

- **Q:** People complain that soccer is boring - what number of EPL games in the '13-'14 season had **at least one goal**?
- **A:** To get this answer we sum the class frequencies for all games with one goal or more:

75+82+70+63+39+17+4+1+2=353

Note: 380-27 would have given us the same answer

X	Class Frequency	Relative Frequency
0	27	7.11%
1	**75**	19.74%
2	**82**	21.58%
3	**70**	18.42%
4	**63**	16.58%
5	**39**	10.26%
6	**17**	4.47%
7	**4**	1.05%
8	**1**	.26%
9	**2**	.53%
TOTAL	380	100%

Example

- **Q:** 353 doesn't sound like a lot of games, but I'm not familiar with the soccer season – is that a large proportion of the games?
- **A:** To get this answer we sum the class relative frequencies for all games with one goal or more:

19.74+21.58+18.42+16.58+10.26+4.47 +1.05+.26+.53 = 92.89%

Note: 100-7.11 would have given us the same answer

X	Class Frequency	Relative Frequency
0	27	7.11%
1	75	**19.74%**
2	82	**21.58%**
3	70	**18.42%**
4	63	**16.58%**
5	39	**10.26%**
6	17	**4.47%**
7	4	**1.05%**
8	1	**.26%**
9	2	**.53%**
TOTAL	380	100%

English – This is the Hardest Part

- **At least x** – x or any number greater
  - At least 5 = 5, 6, 7, …
- **At most x** – x or any number lesser
  - At most 5 = …, 1, 2, 3 ,4, 5
- **Less than x** – any number smaller than x
  - Less than 5 = … 1, 2 , 3, 4
- **More than x** – any number larger than x
  - More than 5 = 6, 7, 8, 9, …
- **Between x and y** – we will say any number larger than x and less than y excluding x and y
  - Between 5 and 10 = 6, 7, 8, 9

Summarizing Qualitative Data: Frequency Table

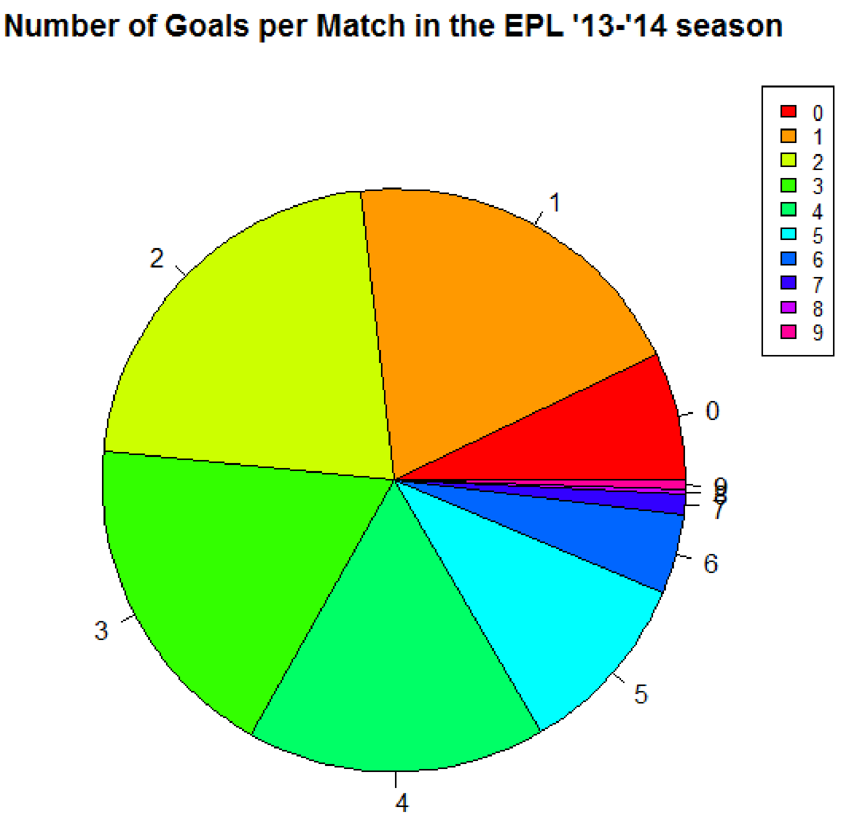
- **R Commands:**

```
#hash marks denote comments like this one
#"<-" can be thought of as an = sign (you can actually use = instead"
#The read.delim function reads data from a text file - you can have tab ("\t"), comma(",") or semicolon(";")
separated
###################################################################
##################Loading and Looking at Data#######################
###################################################################
#file: file location
file<-"E:/Documents/Teaching/USC/515 Course Documents/EPLCSV.csv";
#header: does your data have a header? FALSE
#sep: what are you separating by? ","
EPLdata<-read.delim(file, header = FALSE, sep = ",")
#Calling data is done by typing whatever you called the data
#If you have a lot of data like we do it should be VERY UGLY
EPLdata

###################################################################
##################Creating a Frequency Table in R########################
###################################################################
FreqTable<-table(EPLdata)
RelFreqTable<-FreqTable/sum(FreqTable)
RelFreqTablePER<-FreqTable/sum(FreqTable)*100
```

Summarizing Qualitative Data: Pie Chart

- Useful when there are a small number of categories



Summarizing Qualitative Data: Pie Chart

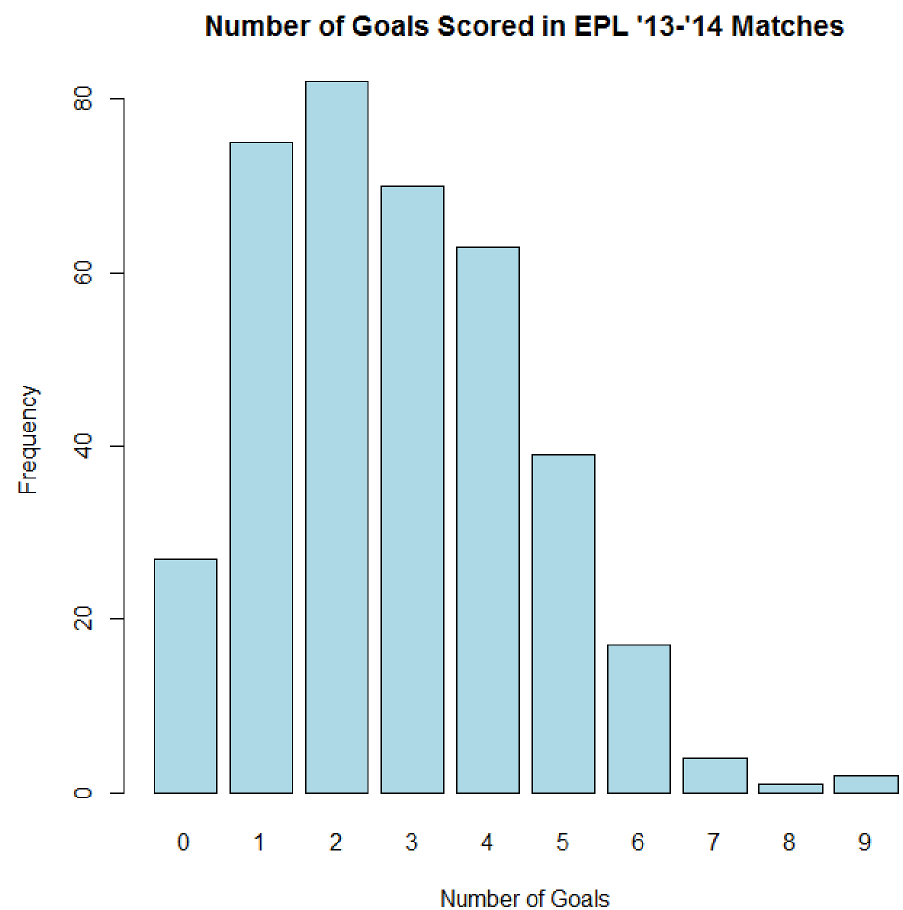
- **R Commands:**

```
##################################################################
########################Creating a Pie Chart in R##########################
##################################################################
#BASIC
pie(FreqTable)
#Add Title
pie(FreqTable,main="Number of Goals per Match in the EPL '13-'14 season")
#Add colors
colors=rainbow(10)#because we have ten classes
pie(FreqTable, col=colors,main="Number of Goals Scored in EPL '13-'14 Matches")
#add legend
legend("topright", names(FreqTable), cex=0.8, fill=colors)
```

More Examples: http://www.harding.edu/fmccown/r/

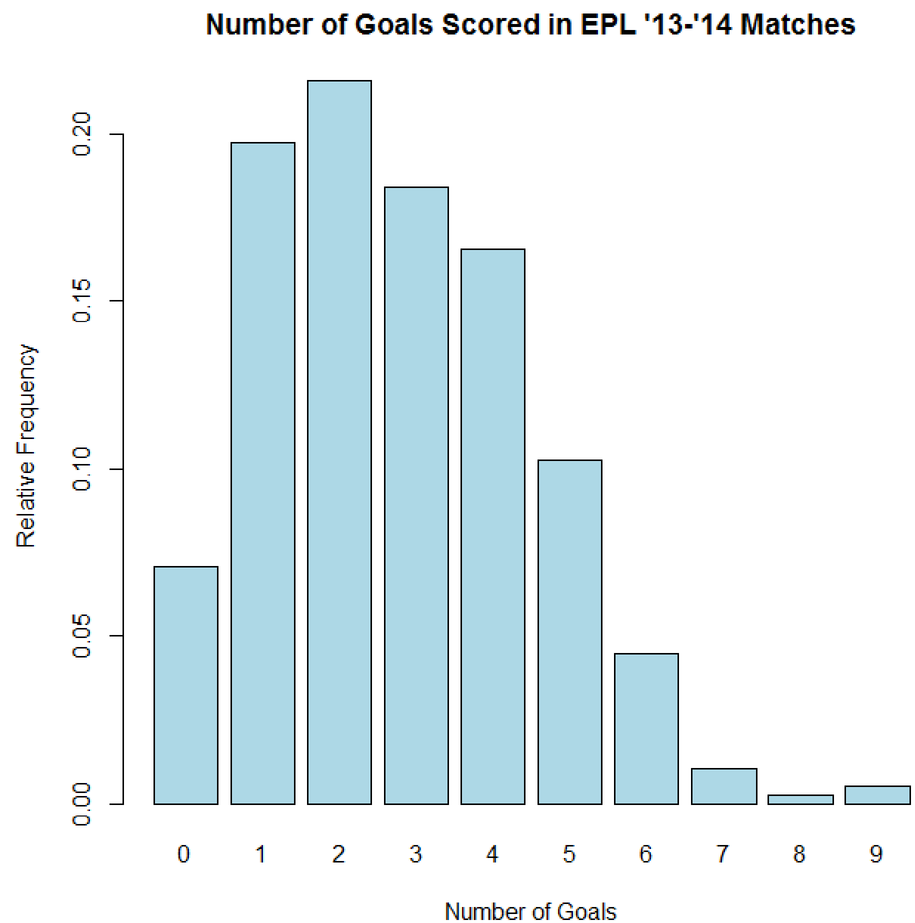
Summarizing Qualitative Data: Bar Graph

- Useful when there are many categories of the variable
- Useful to compare groups



Summarizing Qualitative Data: Bar Graph

- **Note:** the relative frequency chart has the same shape but a different y-axis



Summarizing Qualitative Data: Bar Graph

- **R Commands:**

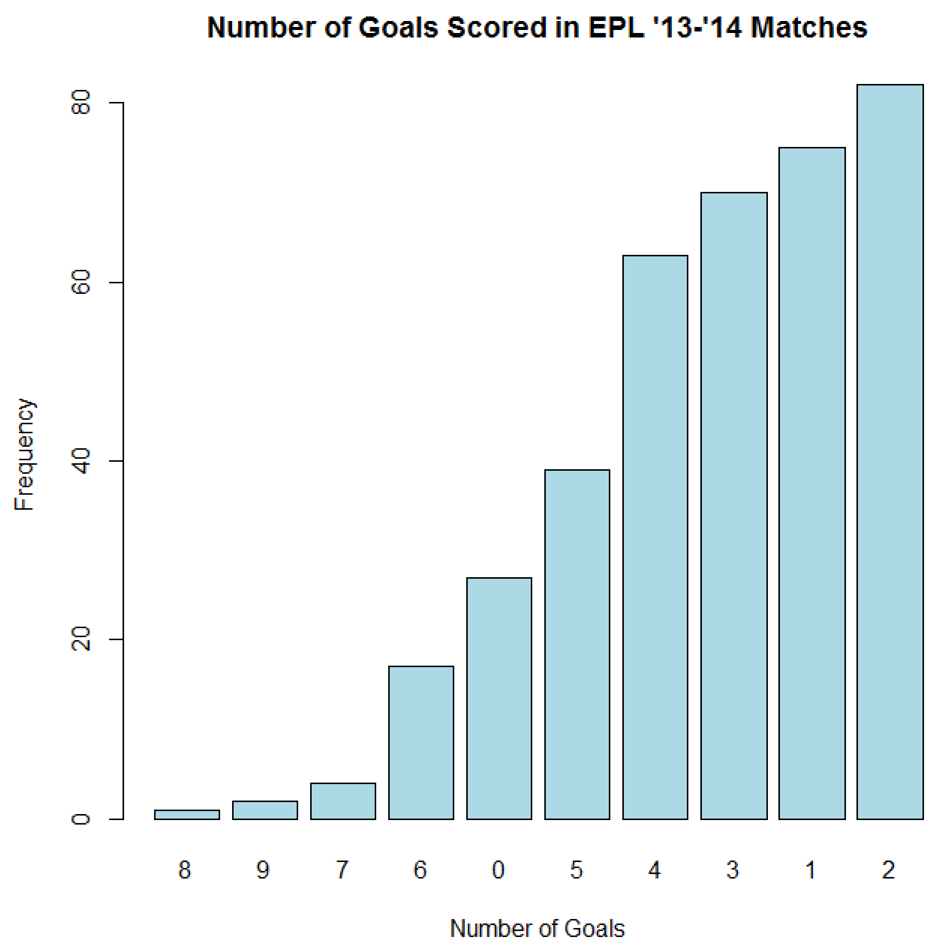
```
#######################################################################
###################Creating a Bar Chart in R##########################
#######################################################################
#BASIC Class Frequency
barplot(FreqTable)
#Add Title
barplot(FreqTable,main="Number of Goals Scored in EPL '13-'14 Matches")
#Add X-Label
barplot(FreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals")
#Add Y-label
barplot(FreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals", ylab="Frequency")
#Add Colors
barplot(FreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals", ylab="Frequency", col="light blue")

#BASIC Class Relative Frequency
barplot(RelFreqTable)
#Add Title
barplot(RelFreqTable,main="Number of Goals Scored in EPL '13-'14 Matches")
#Add X-Label
barplot(RelFreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals")
#Add Y-label
barplot(RelFreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals", ylab="Relative Frequency")
#Add Colors
barplot(RelFreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals", ylab="Relative Frequency", col="light blue")
```

More Examples: http://www.harding.edu/fmccown/r/

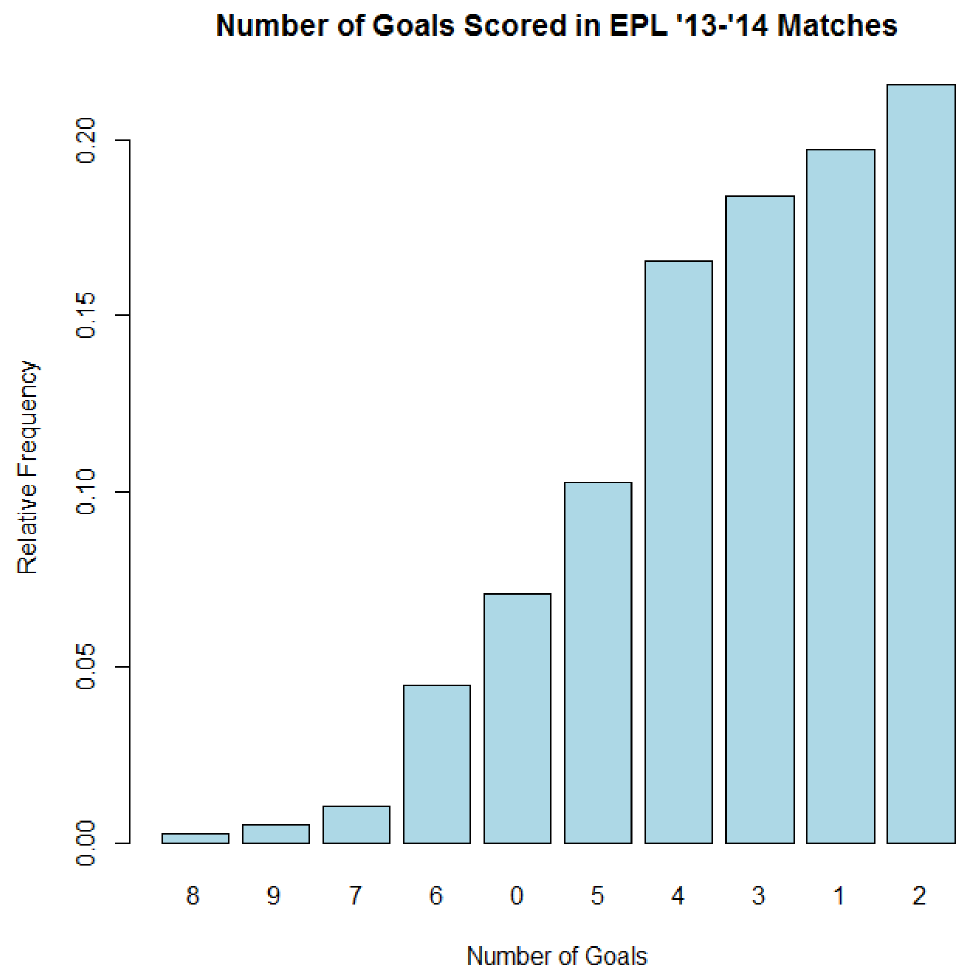
Summarizing Qualitative Data: Pareto Graph

- Useful when there are many categories of the variable
- Same as the bar graph except the bars are ordered by height, making it easier to see what happens 'most' or 'least.'



Summarizing Qualitative Data: Pareto Graph

- **Note:** the relative frequency chart has the same shape but a different y-axis



Summarizing Qualitative Data: Pareto Graph

- **R Commands:**

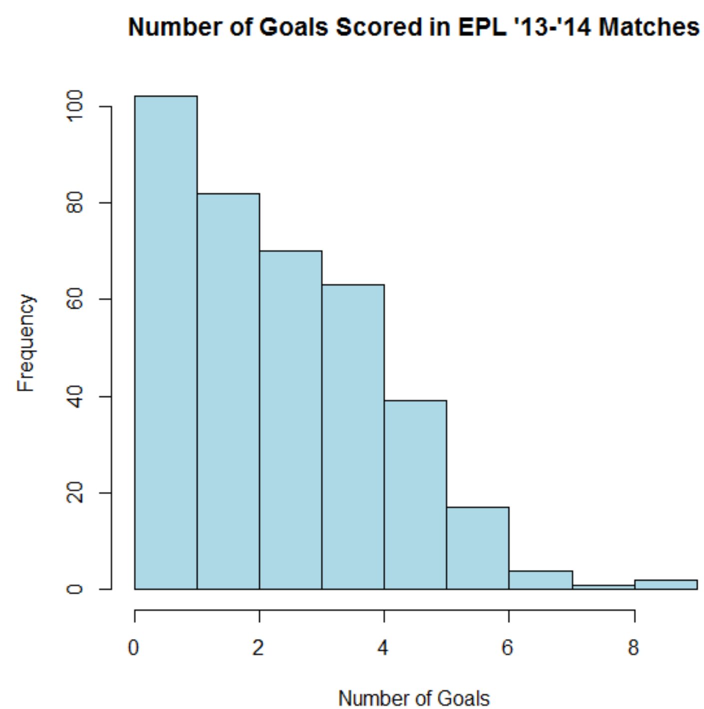
```
######################################################################
###################Creating a Pareto Chart in R#######################
######################################################################
#Pareto Frequency
#Find out the order of the table
order(FreqTable)
#Reorder Frequency Table
OrdFreqTable<-FreqTable[order(FreqTable)]
#Complete a bar chart using this table
barplot(OrdFreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals", ylab="Frequency", col="light blue")

#Pareto Relative Frequency
#Find out the order of the table
order(RelFreqTable)
#Reorder Relative Frequency Table
OrdRelFreqTable<-RelFreqTable[order(RelFreqTable)]
#Complete a bar chart using this table
barplot(OrdRelFreqTable,main="Number of Goals Scored in EPL '13-'14 Matches",xlab="Number of Goals", ylab="Relative Frequency", col="light blue")
```

More Examples: http://www.harding.edu/fmccown/r/

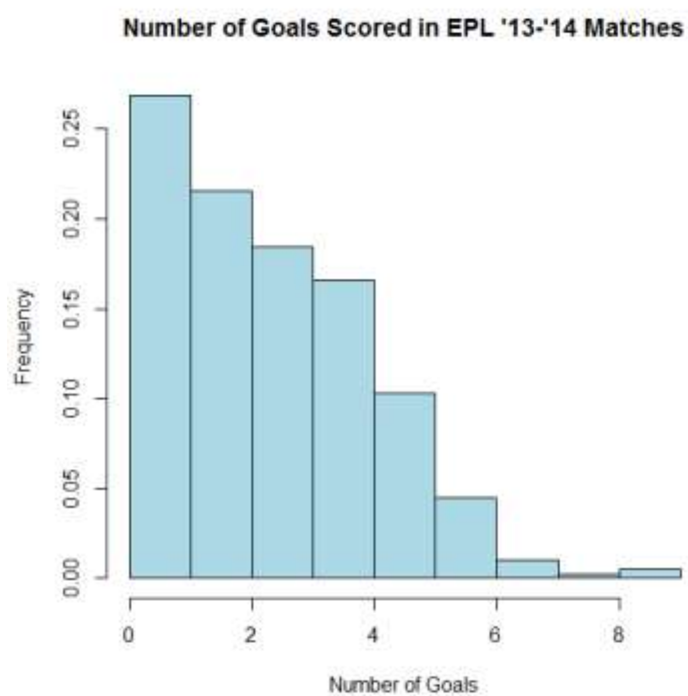
Summarizing Quantitative Data: Histogram

- Histograms are used to summarize quantitative data and will be our main tool for continuous data



Summarizing Quantitative Data: Histogram

- **Note:** the relative frequency chart has the same shape but a different y-axis



Summarizing Quantitative Data: Histogram

- **<u>R Commands:</u>**

```
######################################################################
###################Creating a Histogram in R##########################
######################################################################
#With histograms we no longer use the Frequency Tables as input
#Instead we use the regular data table - but we need to call the column
NumGoals<-EPLdata[,1]
#Basic
hist(NumGoals)
#Add Title
hist(NumGoals,main="Number of Goals Scored in EPL '13-'14 Matches")
#Add X-label
hist(NumGoals,main="Number of Goals Scored in EPL '13-'14 Matches", xlab="Number of Goals")
#Add Y-label
hist(NumGoals,main="Number of Goals Scored in EPL '13-'14 Matches", xlab="Number of Goals", ylab="Frequency")
#Add Color
hist(NumGoals,main="Number of Goals Scored in EPL '13-'14 Matches", xlab="Number of Goals", ylab="Frequency", col="light blue")
#Use Relative Frequency
hist(NumGoals,main="Number of Goals Scored in EPL '13-'14 Matches", xlab="Number of Goals", ylab="Frequency", col="light blue",freq=F)
```

More Examples: http://www.harding.edu/fmccown/r/

Histograms Vs. Bar Charts

- With bar charts, each column represents a group defined by a class of a qualitative (categorical) variable

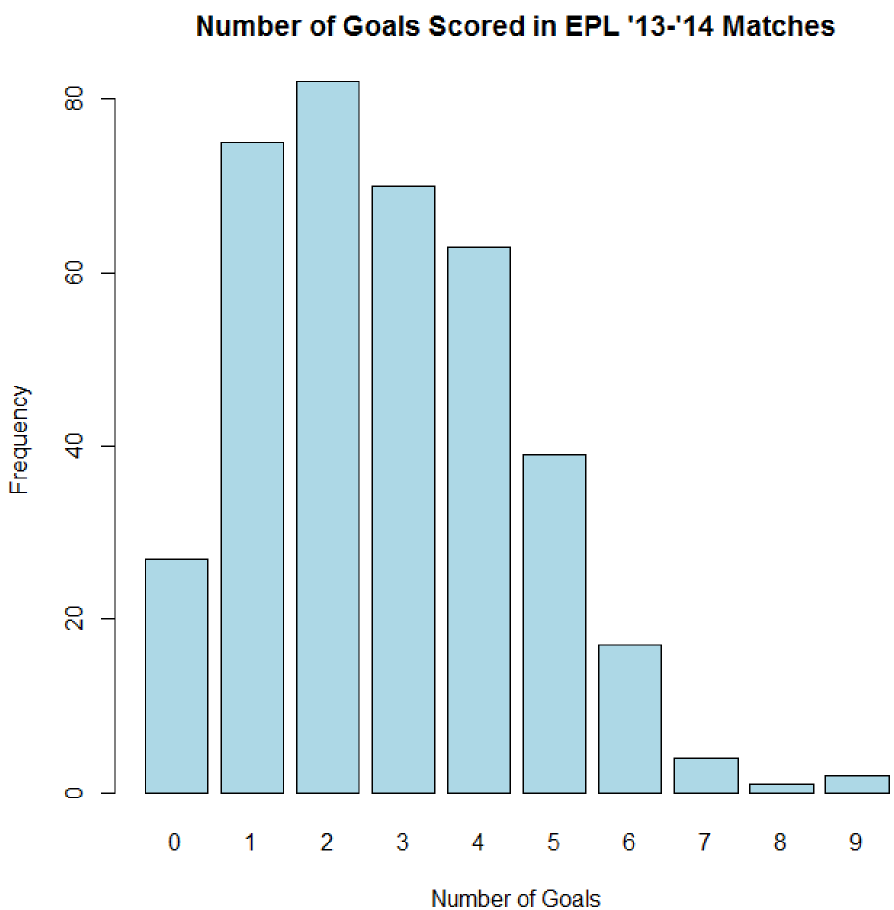
- With histograms, each column represents a group defined by a quantitative variable. R will automatically generate classes for the quantitative data

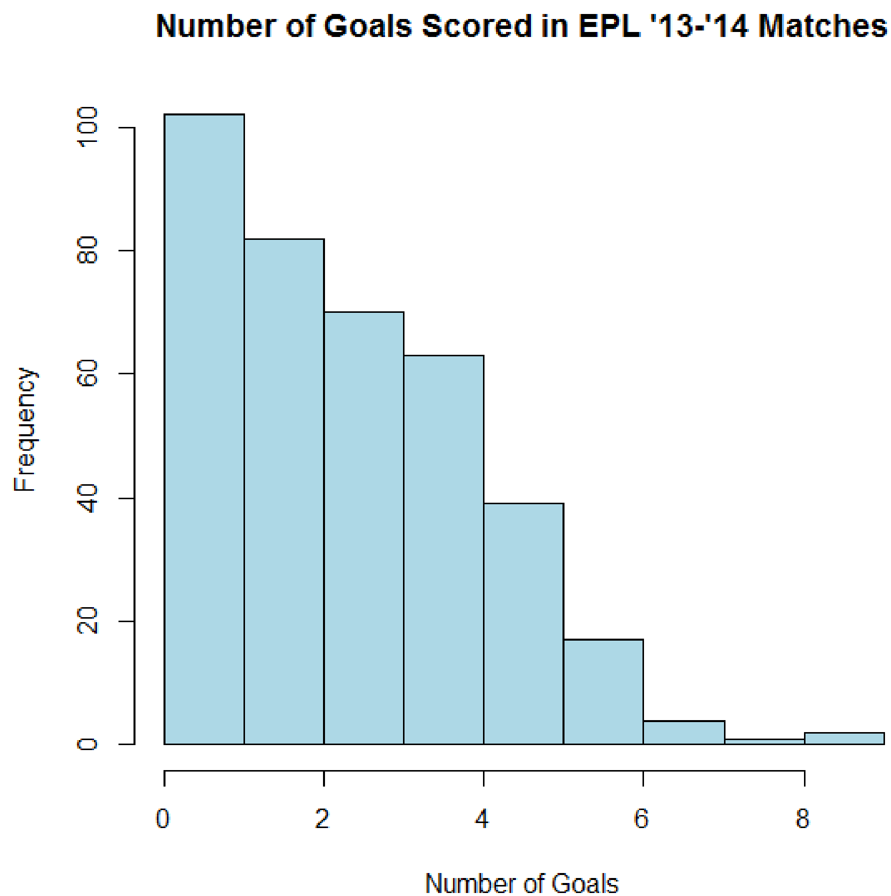
Histograms Vs. Bar Charts

- In our example of EPL goals over the ’13-’14 season the groups that R creates for the histogram are as follow

[0,1]	**102**
(1,2]	**82**
(2,3]	**70**
(3,4]	**63**
(4,5]	**39**
(5,6]	**17**
(6,7]	**4**
(7,8]	**1**
(8,9]	**2**

Histograms Vs. Bar Charts





Histograms Vs. Bar Charts

- In this case, because there are so few observable values the histogram is actually a little misleading – it just combines the bars at 0 and 1 and the rest is the same as the bar plot

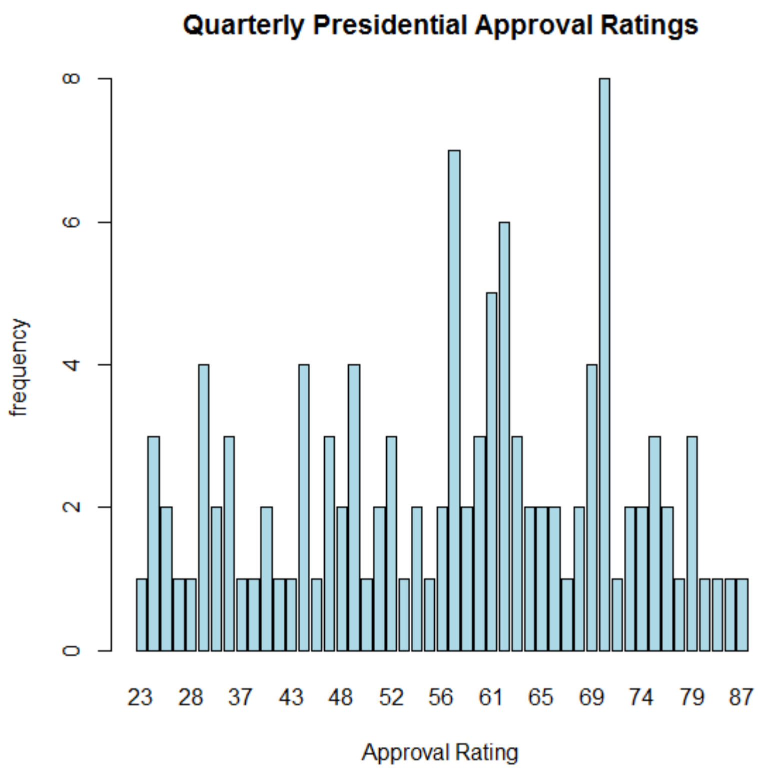
Summarizing Quantitative Data: Histograms

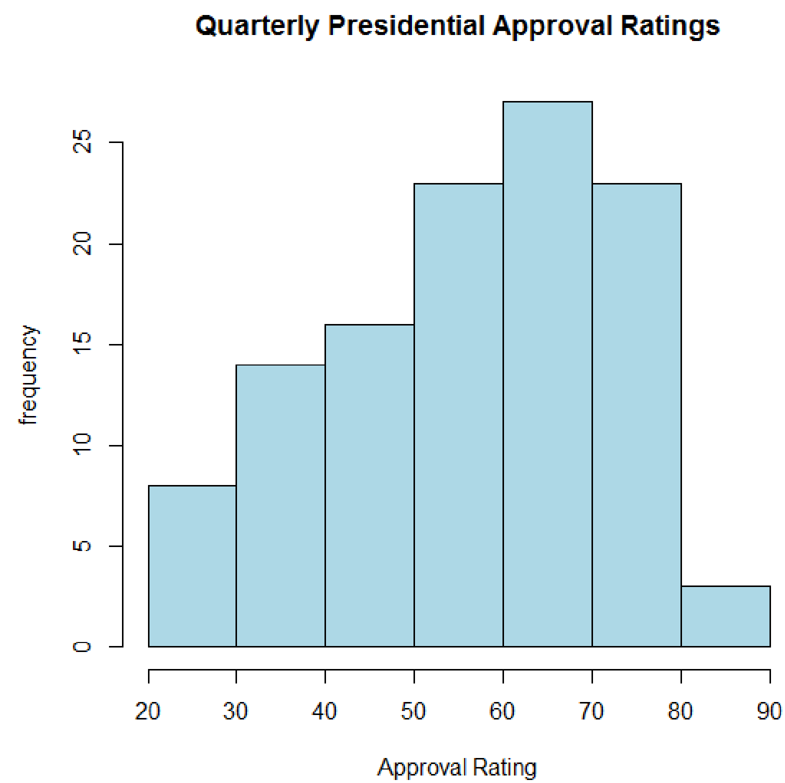
- Let’s consider a different dataset – as we mentioned earlier, the small number of observable values allows us to use the qualitative(categorical) approach with this EPL data
- We will continue looking at histograms by considering the discrete quantitative data considering the quarterly presidential approval ratings from ‘54 to ‘74

Summarizing Quantitative Data: Histograms

- Among the quarterly presidential approval ratings there are 49 observable values ranging from 23 (Truman in ‘51) to 87(Truman in ‘45)
- Here, if we followed what we did for qualitative (categorical data) we would find a frequency table with 49 rows and a bar graph with 49 bars
- Here a histogram is easily a better visual

Summarizing Quantitative Data: Histograms





Histograms Vs. Bar Charts

- In our example of Presidential approval ratings the groups that R creates for the histogram are as follow:

[20,30]	**8**
(30,40]	**14**
(40,50]	**16**
(50,60]	**23**
(60,70]	**27**
(70,80]	**23**
(80,90]	**43**

Summarizing Quantitative Data: Histograms

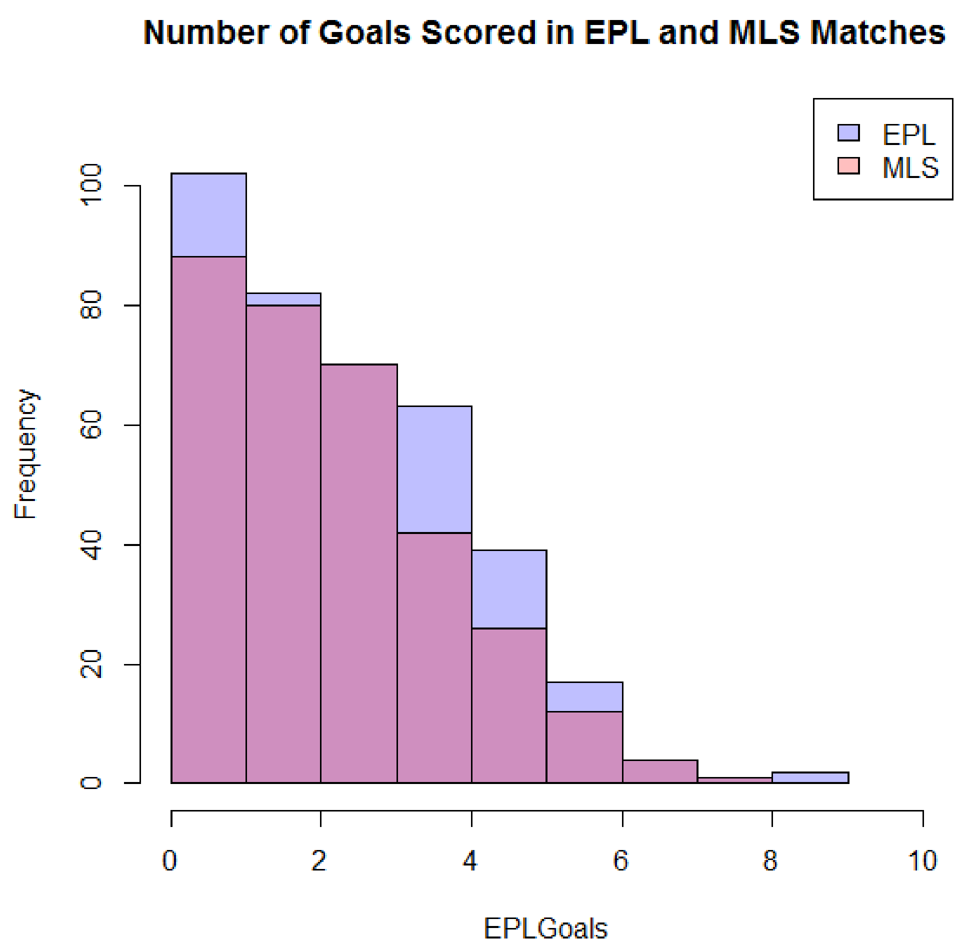
- **R commands:**

```
############################################################
###Load nhtemp and presidential rating data###
install.packages("datasets")
library("datasets")
############################################################
#Load presidential approval ratings data
presidents
table(presidents)
barplot(table(presidents),main="Quarterly Presidential Approval Ratings", xlab="Approval Rating", ylab="frequency", col="light blue") #YUCK
hist(presidents,main="Quarterly Presidential Approval Ratings", xlab="Approval Rating", ylab="frequency", col="light blue") #WAY BETTER
```

Talking about Two Things at Once

- In many cases we're looking at two groups and comparing them.
- Here we consider the EPL goals data and compare it to another league to see if teams score more or less over their season
- The following graphs compare goals in the EPL '13-'14 season and goals in the MLS '13 season

Talking about Two Things at Once



Talking about Two Things at Once

- **R commands:**

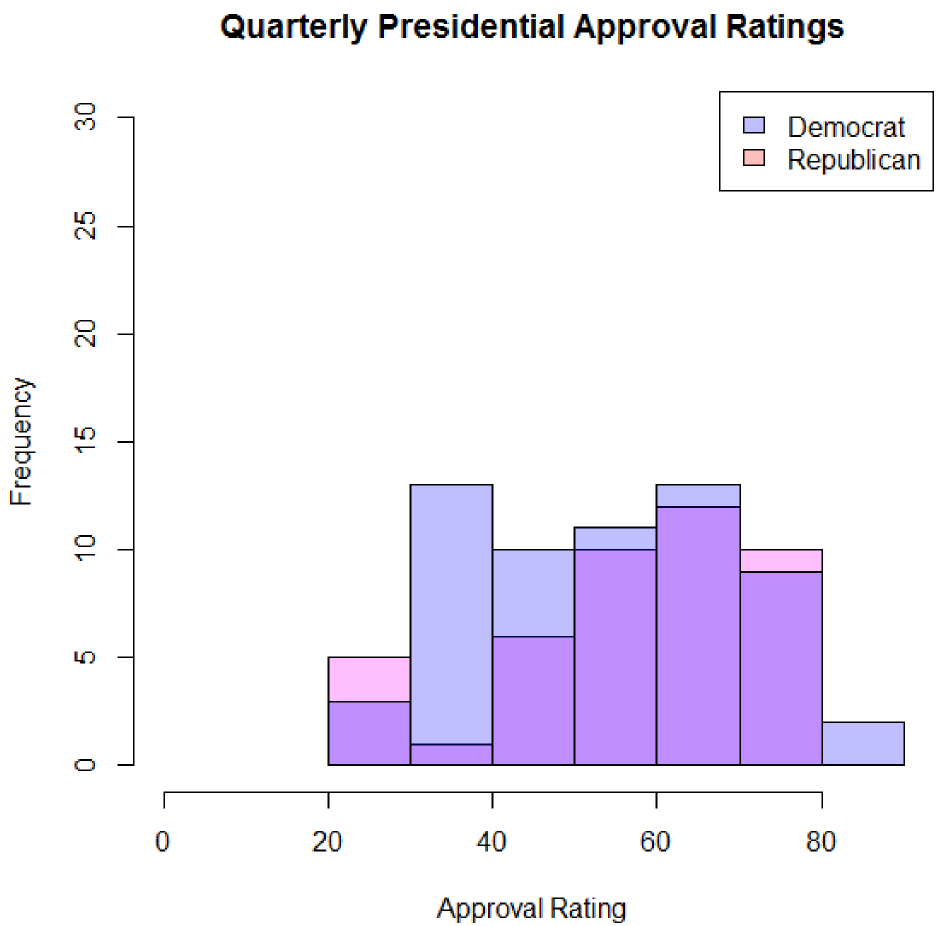
```
######################################################################
###################Loading and Looking at Data########################
######################################################################
#file: file location
file<-"E:/Documents/Teaching/USC/515 Course Documents/MLSCSV.csv";
#header: does your data have a header? FALSE
#sep: what are you separating by? ","
MLSdata<-read.delim(file, header = FALSE, sep = ",")
#Calling data is done by typing whatever you called the data
#If you have a lot of data like we do it should be VERY UGLY
MLSdata
######################################################################
######################################################################

######################################################################
###################Two Histograms in one##############################
######################################################################
EPLGoals<-EPLdata[,1]
MLSGoals<-MLSdata[,1]
#Create separate histograms
EPLhist<-hist(EPLGoals)
MLShist<-hist(rnorm(500,6))
#Plot the first
plot(EPLhist,col=rgb(0,0,1,1/4),xlim=c(0,10),ylim=c(0,110),main="Number of Goals Scored in EPL and MLS Matches") #col=translucent blue
#Add the second to the plot
plot(MLShist,col=rgb(1,0,0,1/4),xlim=c(0,10),ylim=c(0,110),add=T)#col=transulcent red
#Create Legend
legend("topright", c("EPL", "MLS"), fill=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)))
```

Talking about Two Things at Once

- Here. we consider the presidential approval data and split it into democratic and republican presidents to compare the two parties ratings
- The following graphs compare quarterly ratings of republican and democrat presidents

Talking about Two Things at Once



Talking about Two Things at Once

- **R commands:**

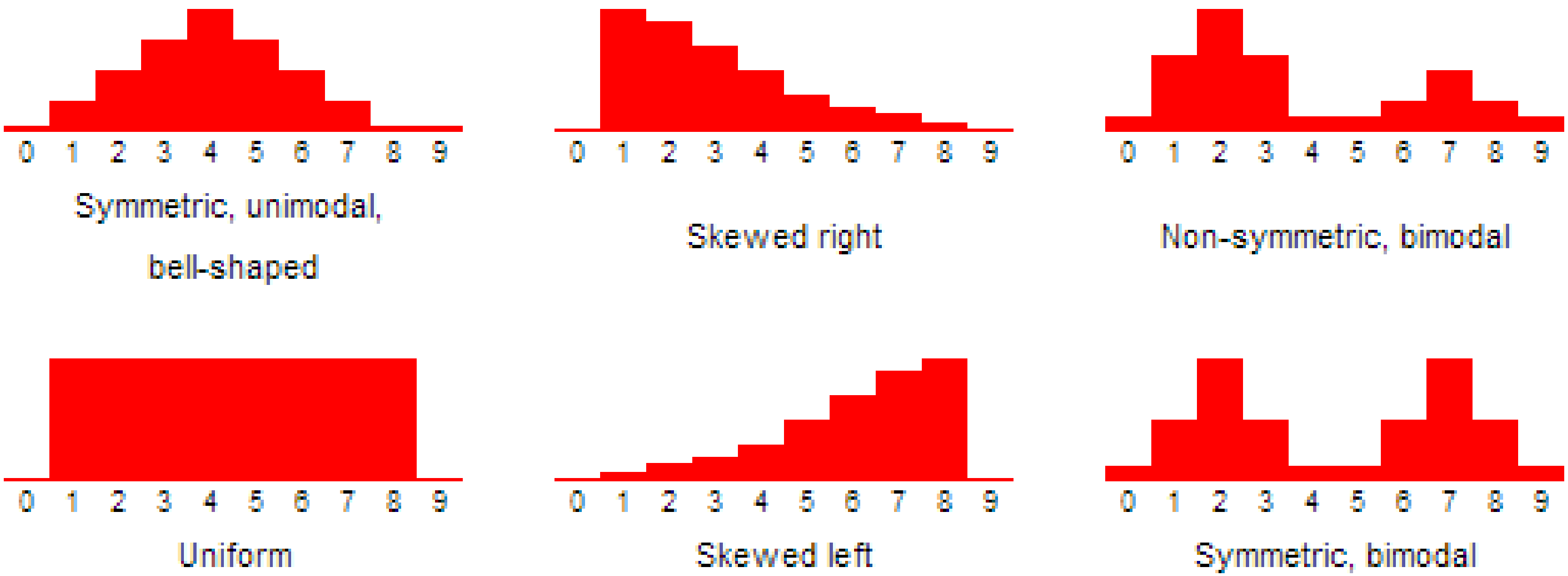
```
######################################################################
###################Two Histograms in one##############################
######################################################################
ryr<-c(36:64,104:120)
dyr<-c(1:32,68:100)
Repub<-presidents[ryr]
Dem<-presidents[dyr]
#Create separate histograms
Rhist<-hist(Repub)
Dhist<-hist(Dem)
#Plot the first
plot(Rhist,col=rgb(1,0,1,1/4),xlim=c(0,90),ylim=c(0,30),main="Quarterly Presidential Approval Ratings",xlab="Approval Rating") #col=translucent blue
#Add the second to the plot
plot(Dhist,col=rgb(0,0,1,1/4),xlim=c(0,90),ylim=c(0,30),add=T)#col=transulcent red
#Create Legend
legend("topright", c("Democrat", "Republican"), fill=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)))
```

Quantitative Summary: Example

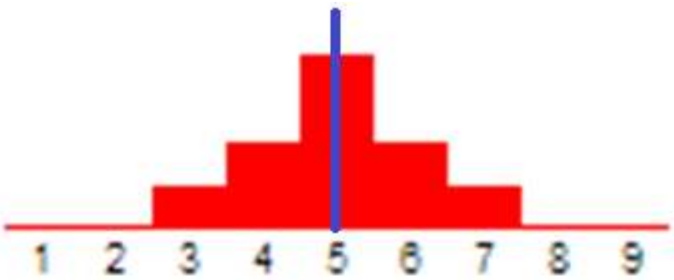
- With histograms we often try to answer the following questions:
  - **What is its shape?**
    - **Is it skewed?**
  - **Where is the center?**
  - **How spread out is it?**
  - **Are there outliers?**

Quantitative Summary: Histogram Shape

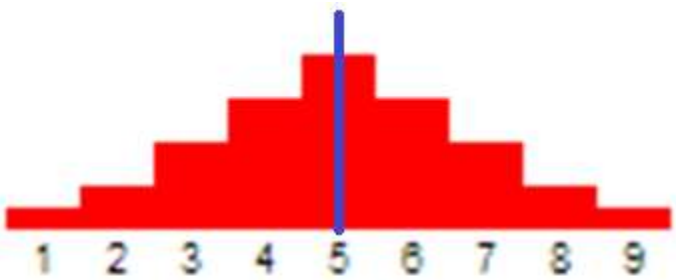
- Shape:



Quantitative Summary: Histogram Center & Spread



Less spread



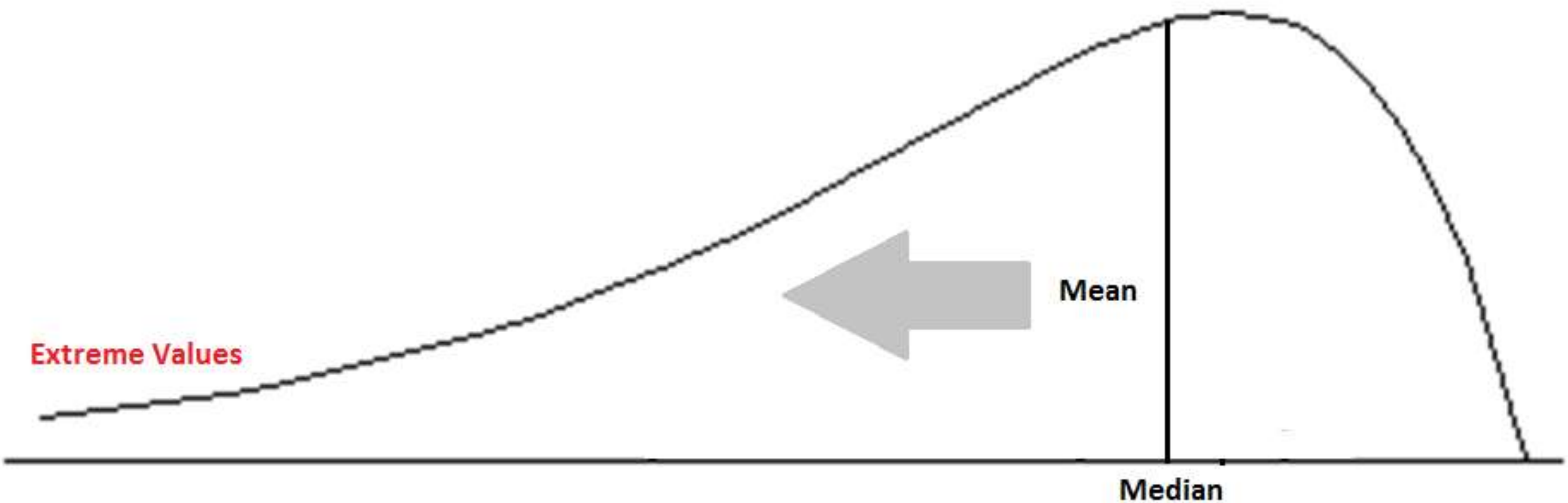
More spread

Quantitative Summary: Histogram Gap vs. Outlier

- Gap vs. Outlier:

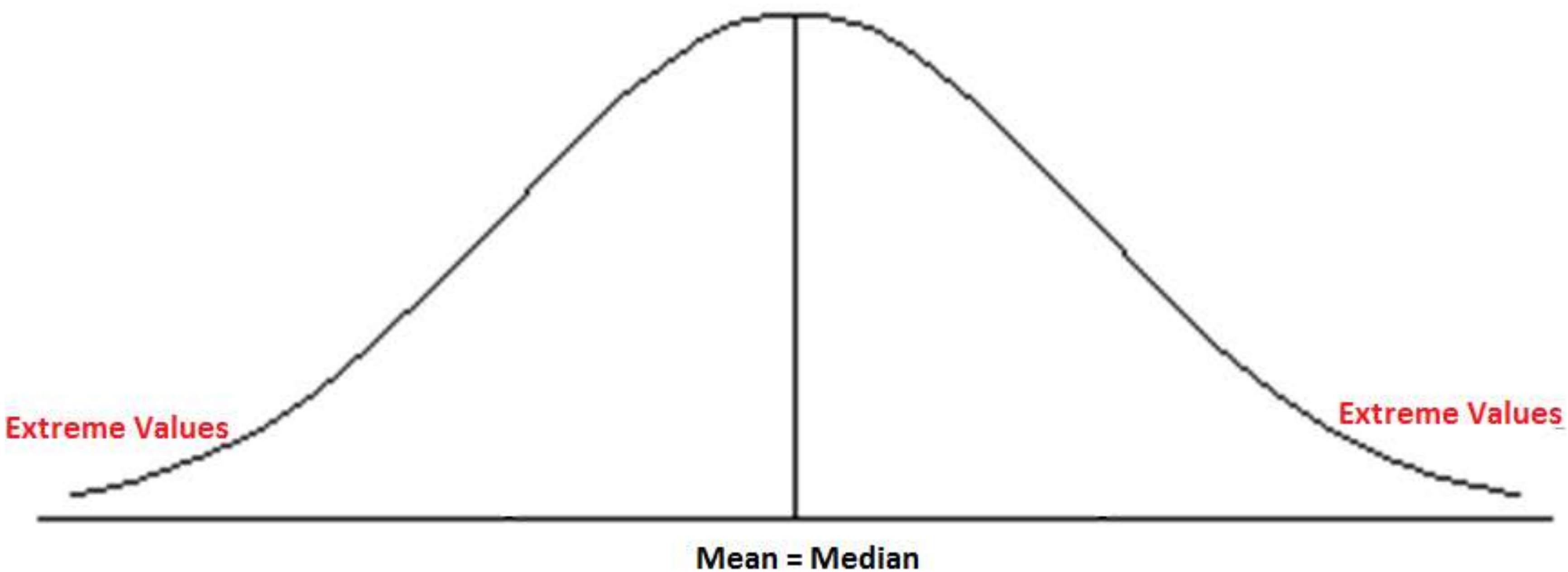
Quantitative Summary: Histograms – Left Skewed

- Here we see a left skewed graph – the extreme values on the left drag the mean to the left tail causing Mean<Median



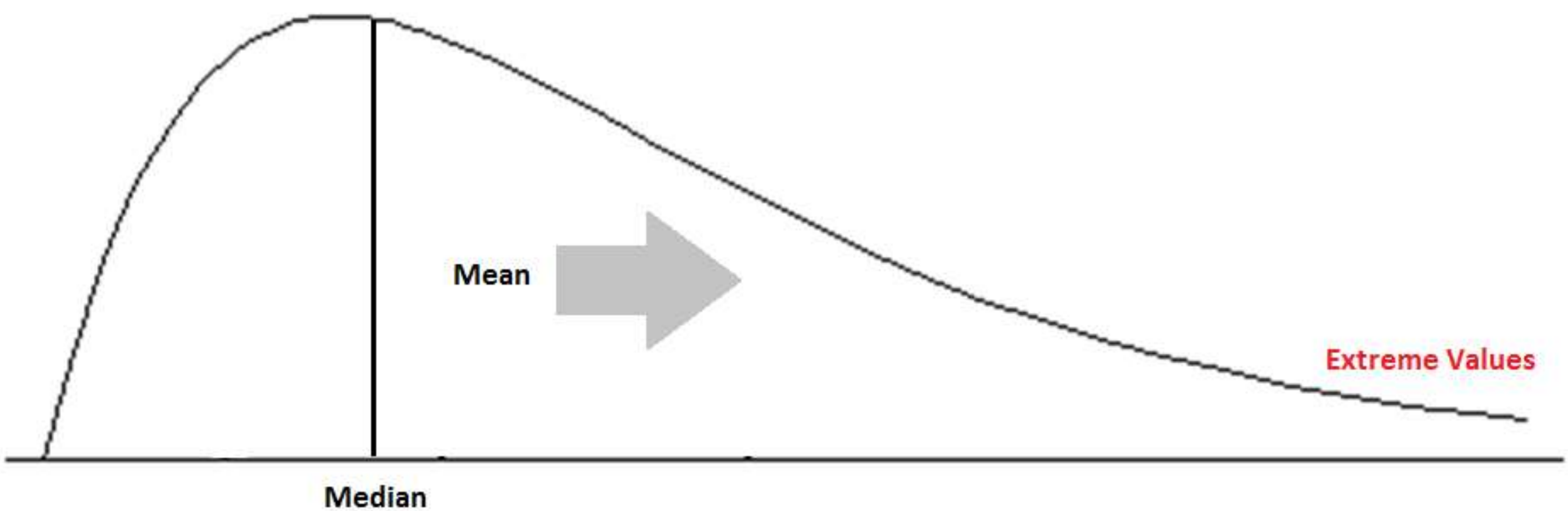
Quantitative Summary: Histograms – Bell Shaped

- Here there is no skew – the extreme values on both side cancel any outlying effect on the mean



Quantitative Summary: Histograms – Left Skewed

- Here we see a right skewed graph – the extreme values on the right drag the mean to the right tail causing Mean>Median



Numerical Measures of Central Tendency

Measure	Computation	R Command	Interpretation	When to Use
Mean Statistic: $\bar{x}$ Parameter: μ	$\bar{x} = \frac{\sum x}{n}$	mean(data)	Center of Gravity	Use for quantitative data when the distribution is roughly symmetric
Median	The point halfway through the data when it is arranged in ascending order.	median(data)	The point which splits the data in half.	Use for quantitative data when the distribution is skewed
Mode	We report the observation with the highest frequency	mode(data)	Most frequent observation	When the most frequent observation is the desired measure or when data is qualitative.

The Greek Letter Sigma in Math

- Before the Sigma was famous for representing Greek organizations on campus it was used by those developing mathematics
- This is a mathematical operator just like +, -, etc.
- This weird looking E, capital sigma, is the notation for a summation – essentially it tells you to add everything up

$$\Sigma$$

The Greek Letter Sigma in Math

- X = {1,2,3,4,5,6,7,8,9)
- $\sum x$ = 1+2+3+4+5+6+7+8+9

  = 45

$$\Sigma$$

- This is easy, you could have learned this in first grade – don't make it harder than it actually is
- You can add, I have faith in you

Quantitative Summary: Mean

- **Mean (Average) –** The mean is the sum of observations divided by the number of observations
  - **Properties:** Sensitive to outliers, pulled in direction of the longer tail of a skewed distribution

$$\bar{x} = \frac{\sum x}{n}$$

- X are the **variable** values for our sample
- n is the size of the sample

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9}
- $\bar{x} = \frac{\sum x}{n} = \frac{1+2+3+4+5+6+7+8+9}{9} = \frac{45}{9} = 5$

Quantitative Summary: Median

- **Median** – the median is the midpoint of the observations when they are ordered from the smallest to largest
  - Properties: Resistant to outliers
  - In position .5(n+1) when the data is in ascending order

Is the position value a whole number	The Median
Yes	The number in that position
No	The average of the numbers in the above and below positions

Quantitative Summary: Example

- X = {0,1,2,3,4,5,6,7,8) **(n is odd)**
- Position = .5*(n+1) = .5*(9+1) = 5^{th} position
- Median = 4

- X = {0,1,2,3,4,5,6,7,8,9) **(n is even)**
- Position = .5*(n+1) = .5*(10+1) = 5.5^{th} position
- Median = (4+5)/2 = 4.5

Quantitative Summary: Mode

- **Mode**– the mode is the observation that shows up the most in the data set.
  - We allow up to three ties, if there are more we say that there is no mode

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)
  - There is no mode; all observations are tied with one occurrence
- X ={1, 1, 2, 3, 4, 5, 5, 5, 5, 6, 10}
  - Mode = 5 because 5 is the observation that occurred most.
- X ={1, 1, 1, 2, 3, 4, 5, 5, 5, 6, 10, 10}
  - Mode = 5 and 1 because 5 and 1 are the observations that occurred most.
  - **We will allow up to three ties before we revert to the first answer – There is no mode.**

Measures of Dispersion

Measure	Computation	R command	Interpretation
Range	Max – Min	max(data) – min(data)	The difference between the largest and smallest data point
Standard Deviation Statistic: s Parameter: σ	$\sqrt{Variance}$	sd(data)	The square root of the mean of squared deviations from the mean in the original units – this usually makes the standard deviation easier to interpret
Variance Statistic: s^2 Parameter: σ^2	$\frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^2}{\boldsymbol{n}-\mathbf{1}}$	var(data)	The square root of the mean of squared deviations from the mean in units squared

Quantitative Summary: Range

- **Range –** The range is the difference between the maximum and minimum observations
  - **Properties:** easy to calculate but relies on only two values, which may be outliers

Range = Maximum - Minimum

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)

- Range = max – min = 9 – 1 = 8

Quantitative Summary: Variance

- **Variance –** the average, squared deviation of each observation from the mean
  - The idea is that it measures the spread of the data about the mean
  - **Properties:** difficult to interpret because it's in squared units, cannot be negative and is only zero when all data points are equal

Variance = $s^2 = \frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{n}-\mathbf{1}}$

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)
- $\bar{x} = 5$
- **variance**= $s^2 = \frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\boldsymbol{2}}}{\boldsymbol{n}-\boldsymbol{1}}$

$$= \frac{\mathbf{60}}{\mathbf{9-1}} = \frac{\mathbf{60}}{\mathbf{8}} = \mathbf{7.5}$$

X	$(\overline{\boldsymbol{x}} - \boldsymbol{x})$	$(\overline{\boldsymbol{x}} - \boldsymbol{x})^{\boldsymbol{2}}$
1	(1-5)=-4	$(-4)^2 = 16$
2	(2-5)=-3	$(-3)^2 = 9$
3	(3-5)=-2	$(-2)^2 = 4$
4	(4-5)=-1	$(-1)^2 = 1$
5	(5-5)=0	$0^2 = 0$
6	(6-5)=1	$1^2 = 1$
7	(7-5)=2	$2^2 = 4$
8	(8-5)=3	$3^2 = 9$
9	(9-5)=4	$4^2 = 16$
	Total:	60

Quantitative Summary: Standard Deviation

- **Standard Deviation** – the standard deviation is an adjusted average deviation of each observations' distance from the mean
  - The idea is that it measures the spread of the data about the mean
  - **We prefer this to the variance because it isn't in squared units.**
  - **Properties:** The larger the value the more spread or variability in the data, influenced by outliers and it's always positive.

Standard Deviation = $s = \sqrt{Variance} = \sqrt{\frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{2}}{\boldsymbol{n}-\mathbf{1}}}$

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)
- $\bar{x} = 5$
- **variance**= $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$

$$= \frac{60}{9-1} = \frac{60}{8} = 7.5$$

- **Standard Deviation** = $s = \sqrt{Variance}$

$$= \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{7.5} = 2.7386$$

Interpreting the Standard Deviation

- The next two topics we talk about – the Empirical Rule and Chebyshev's Rule – show how valuable the standard deviation is

- These two results are very powerful in the sense that they give us a good idea about how the data is spread out

The Empirical Rule

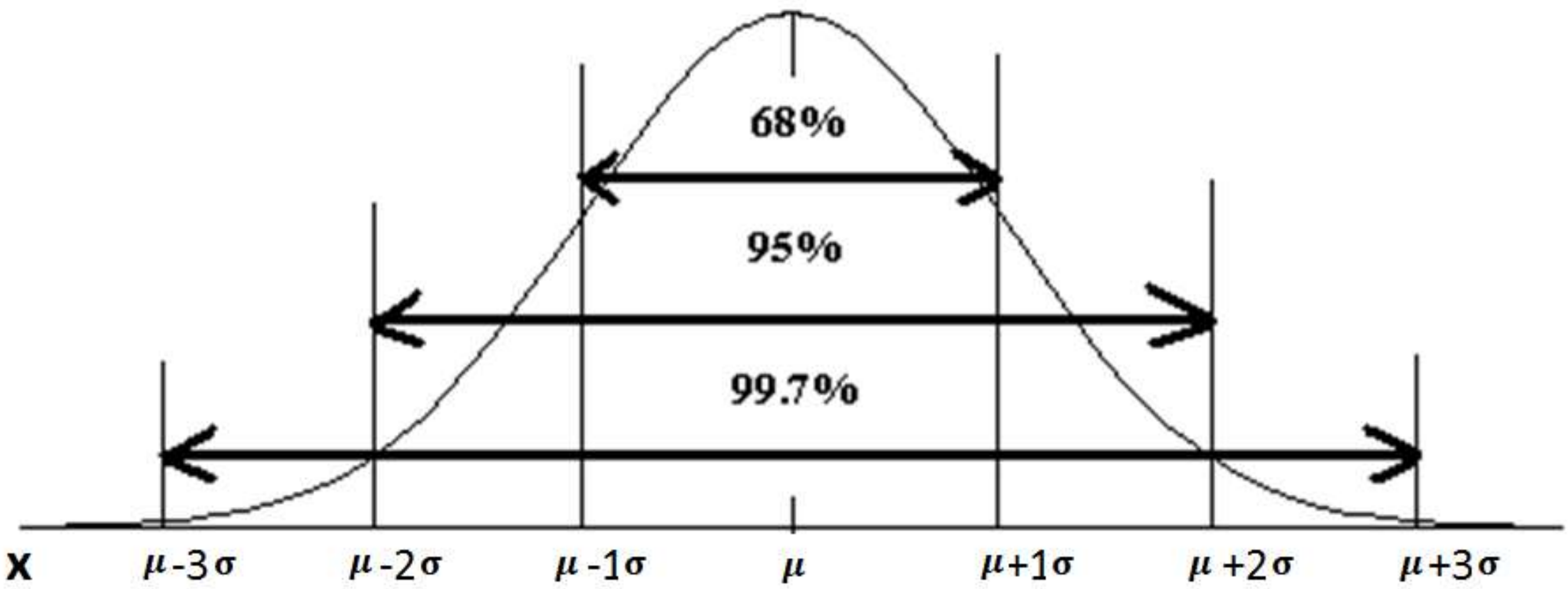
- A VERY basic Introduction to the Empirical Rule:
  – https://www.youtube.com/watch?v=Vt8ZoT3eTmY

- Introductory problems to the Empirical Rule:
  – https://www.youtube.com/watch?v=cgxPcdPbujI
  – https://www.youtube.com/watch?v=2fzYE-Emar0
  – https://www.youtube.com/watch?v=itQEwESWDKg

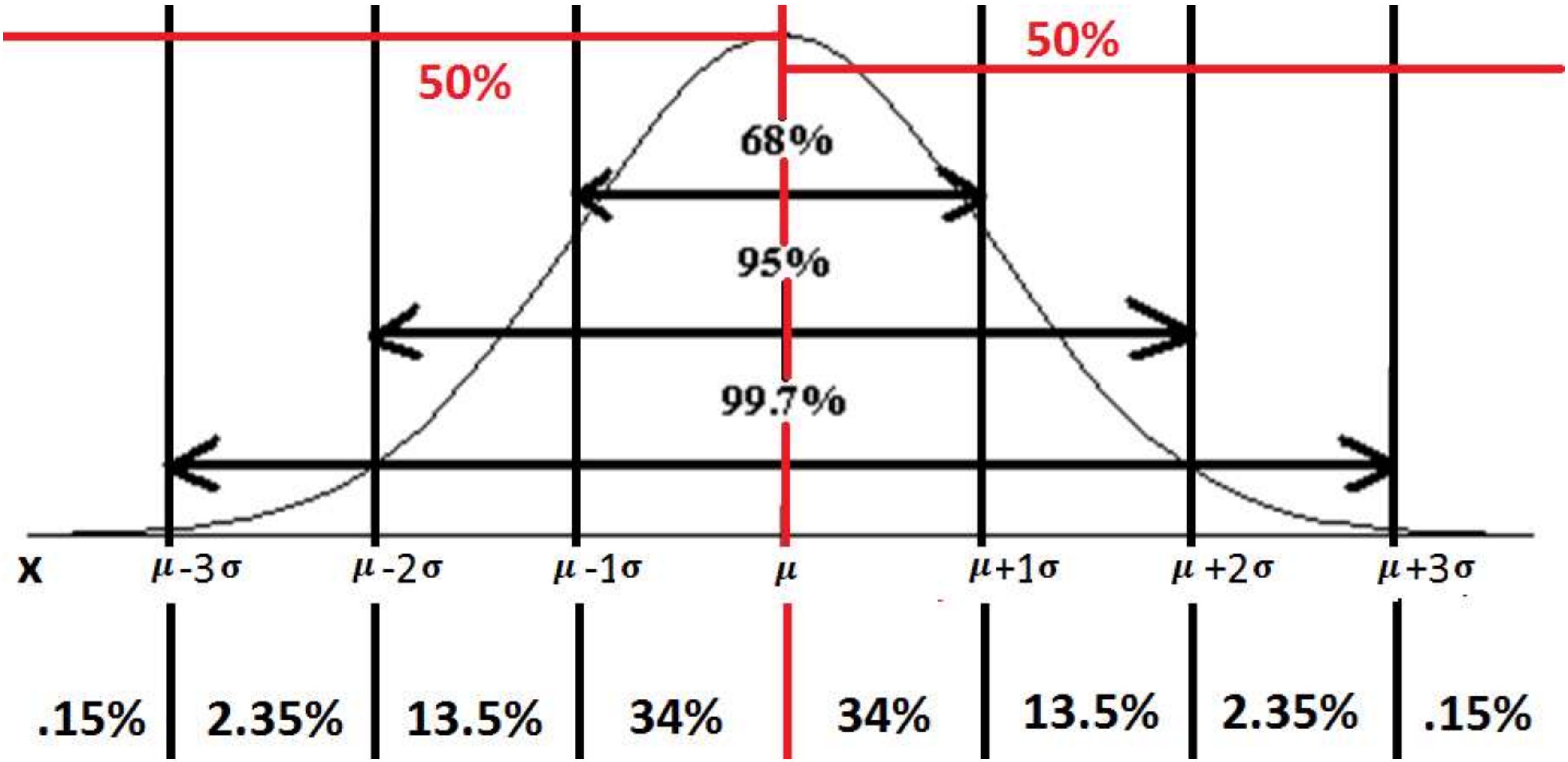
The Empirical Rule

- About 68% of data fall within 1 standard deviation of the mean
- About 95% of data fall within 2 standard deviation of the mean
- About 99.7% of data fall within 3 standard deviation of the mean
- **The distribution must be symmetric and bell shaped to use this Rule**

The Empirical Rule

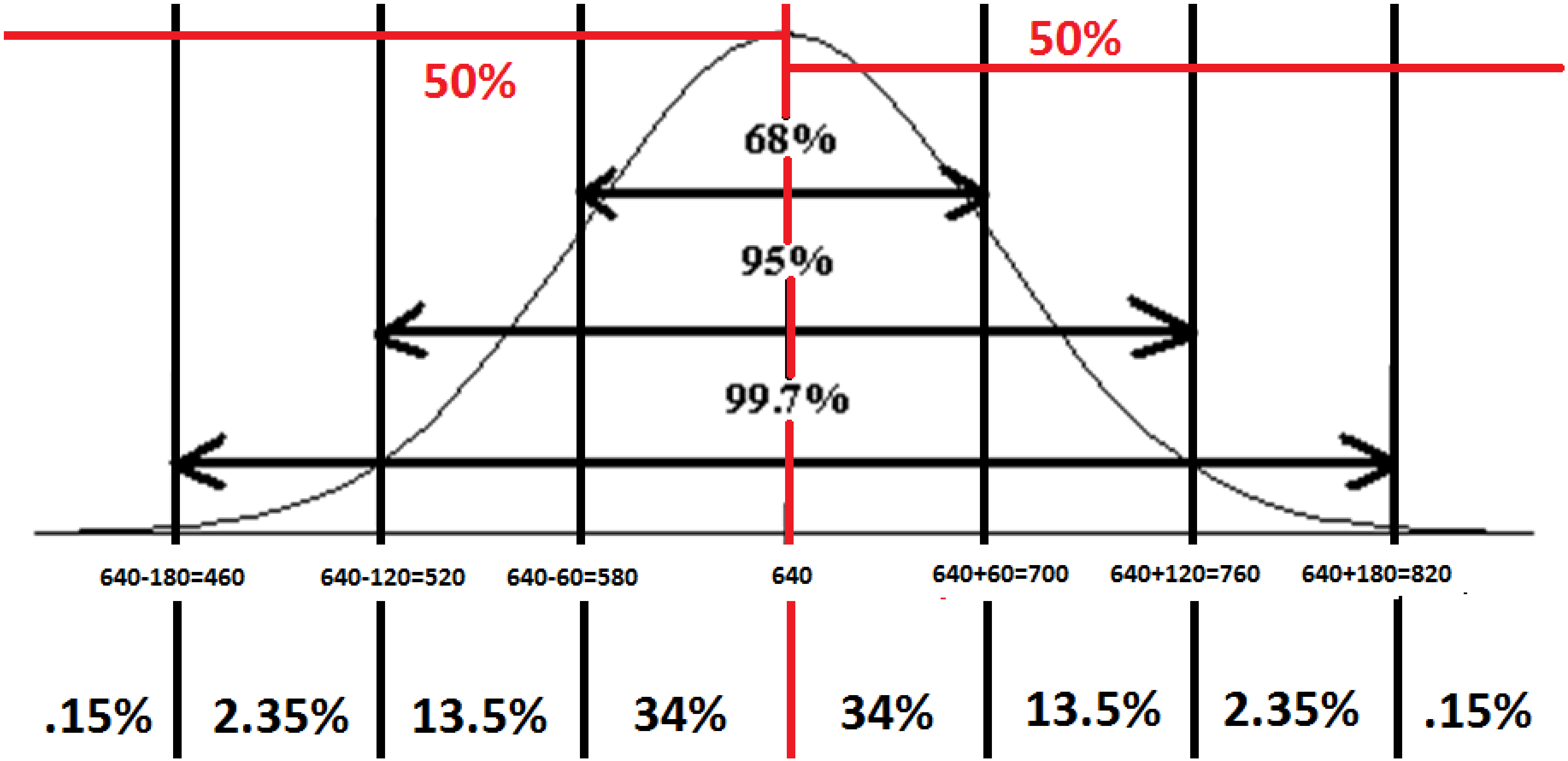


The Empirical Rule



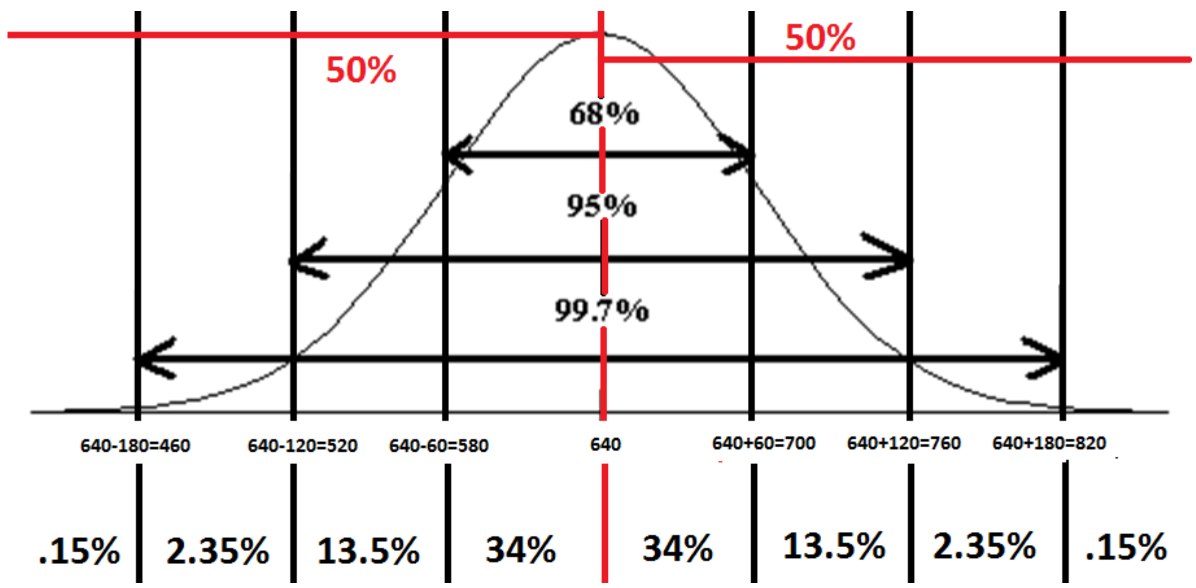
The Empirical Rule: Example

- The average college student consumes 640 cans of beer each year. Assume the distribution of cans of beers consumed per college student is **bell-shaped** with a **mean of 640 cans** and a **standard deviation of 60 cans**.


50%
50%
68%
95%
99.7%
640-180=460
640-120=520
640-60=580
640
640+60=700
640+120=760
640+180=820
.15%
2.35%
13.5%
34%
34%
13.5%
2.35%
.15%

The Empirical Rule: Example

- What percent of students consume less than 700 cans of beer per year?

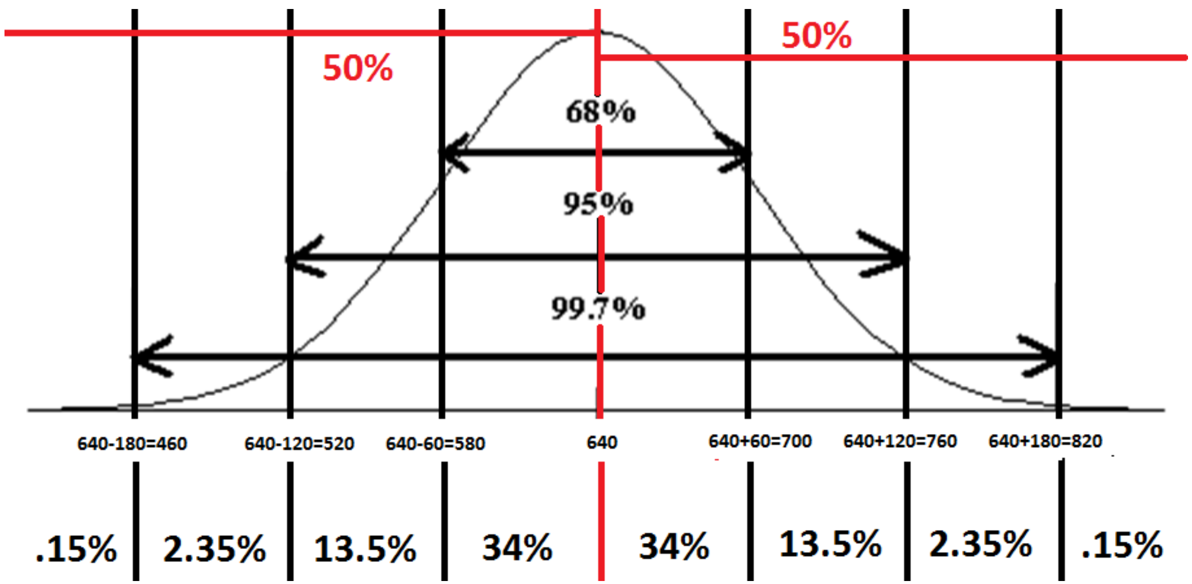


The Empirical Rule: Example

- What percent of students consume less than 700 cans of beer per year?
- We can add up the area under the curve as we go left

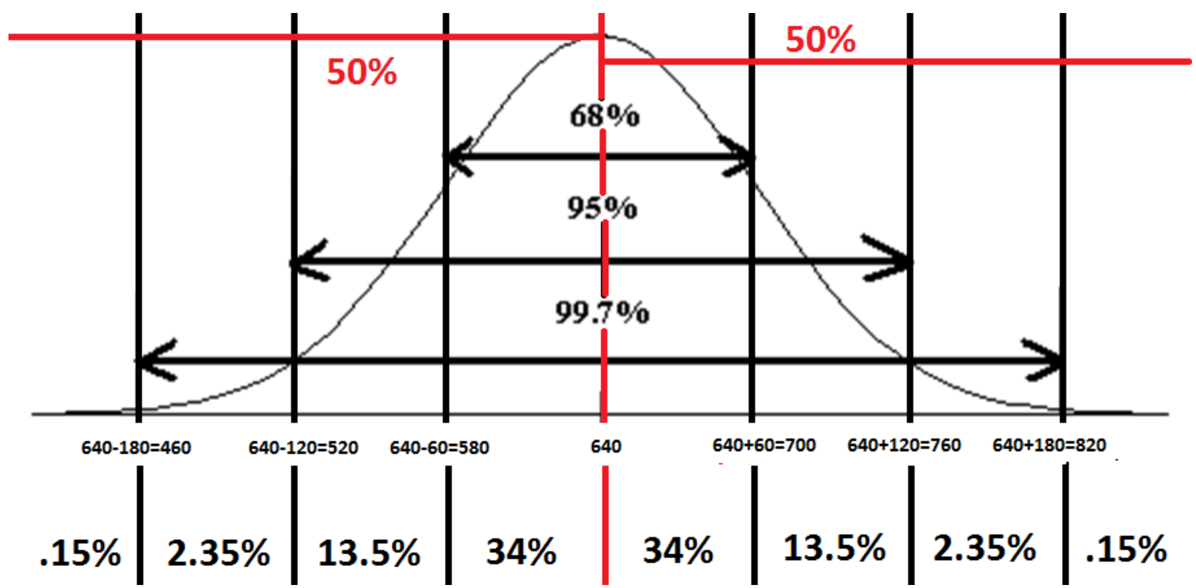
  2.5%+13.5%+34%+34%+.15%

  = 84%



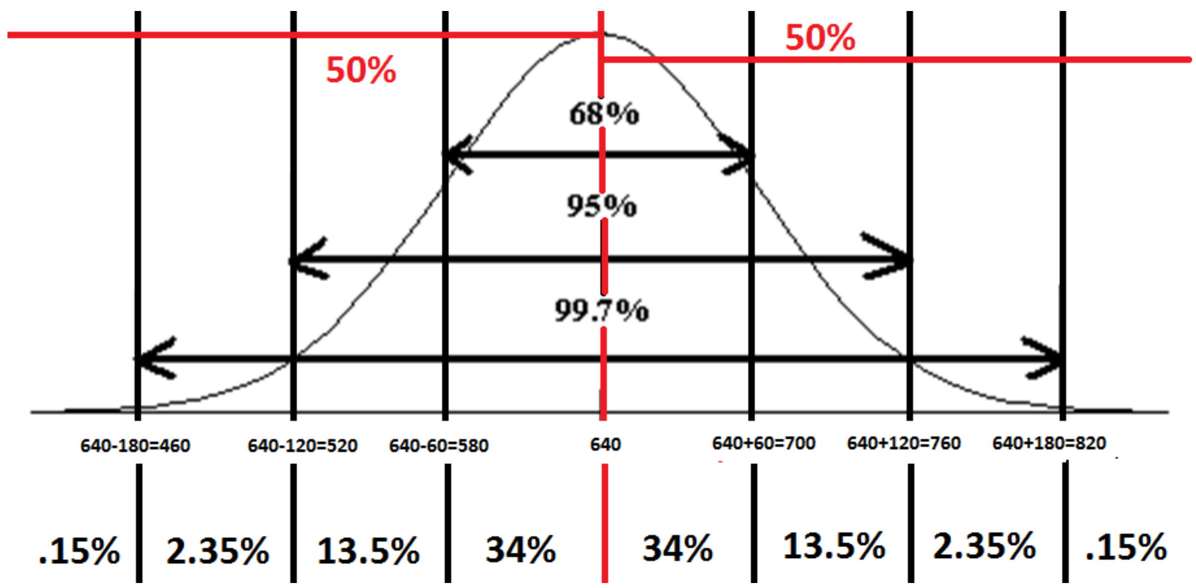
The Empirical Rule: Example

- What percent of students consume less than 700 cans of beer per year?
- We can subtract the area from 100% as we go right
100%-13.5%-2.5%-.15%
= 84%



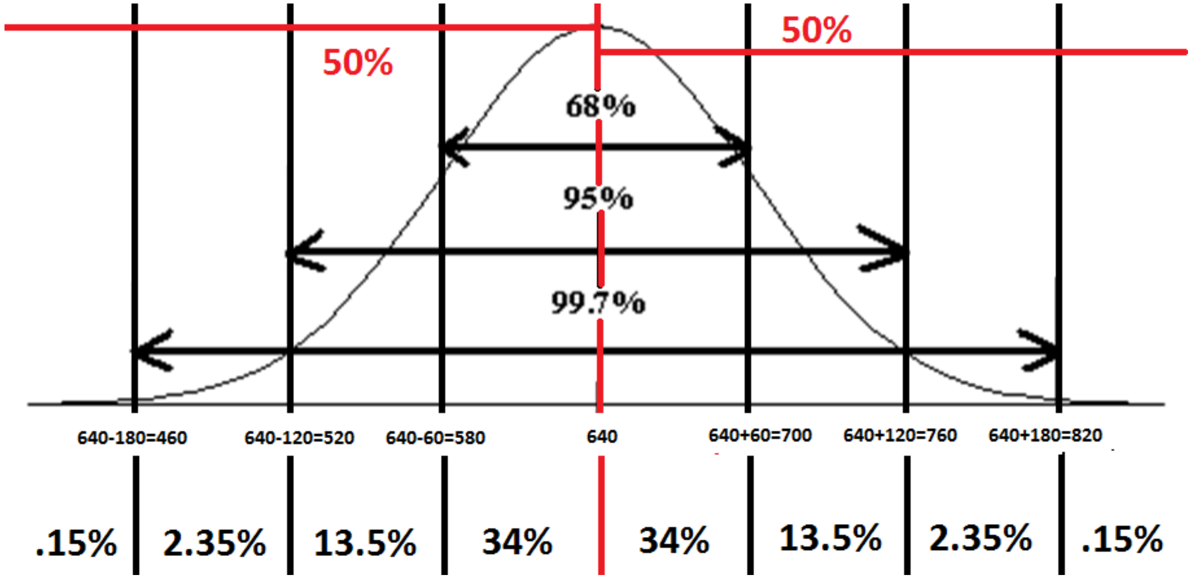
The Empirical Rule: Example

- What percent of students consume more than 700 cans of beer per year?



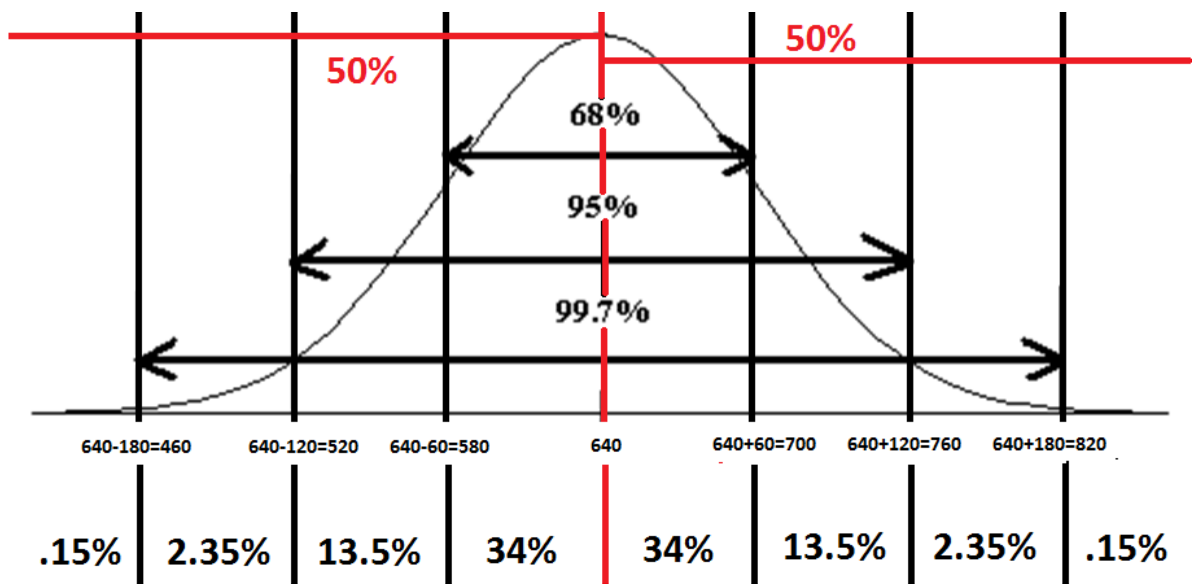
The Empirical Rule: Example

- What percent of students consume more than 700 cans of beer per year?
- We can add up the area under the curve as we go right
  13.5% + 2.35% + .15% = 16%



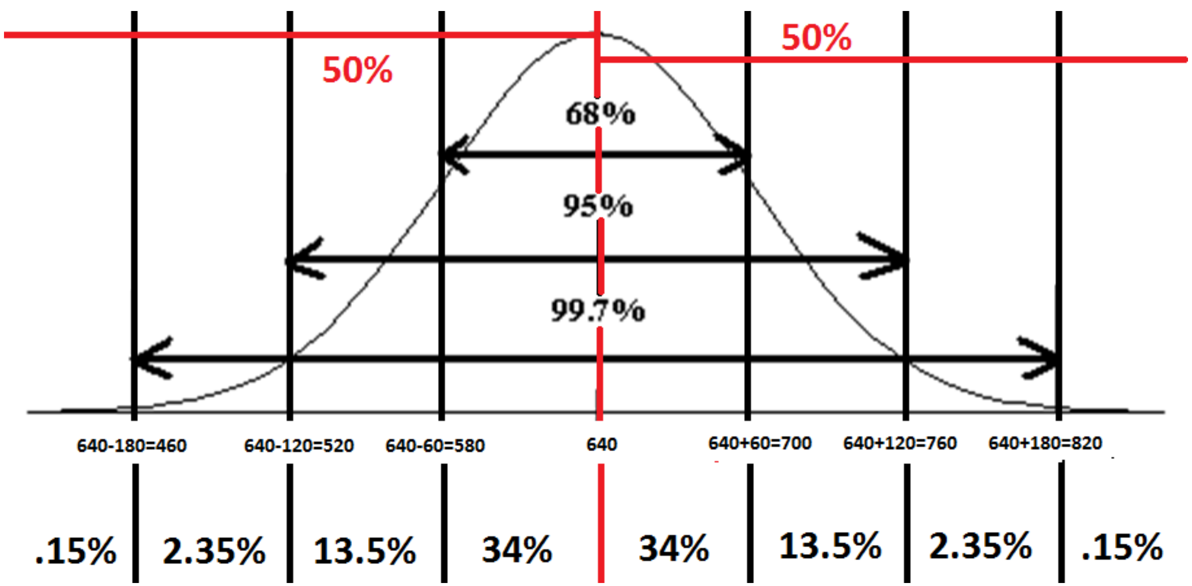
The Empirical Rule: Example

- What percent of students consume more than 700 cans of beer per year?
- We can subtract the area from 100% as we go left
  100%-34%-34%-13.5%-2.5%-.15%
  100%-84% (we know 84% from the last question)
  = 16%



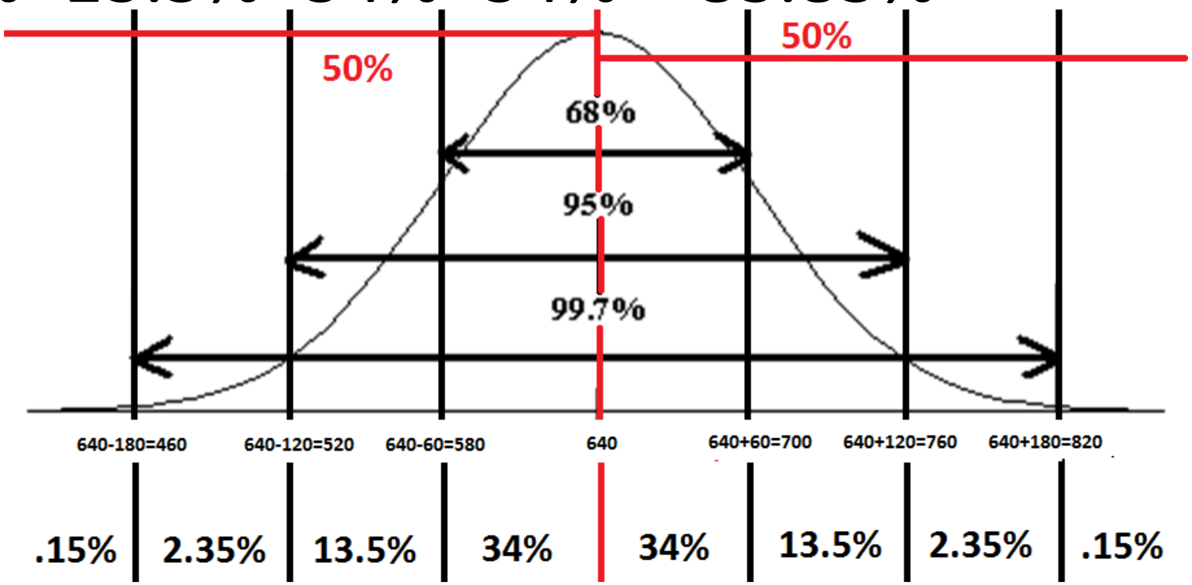
The Empirical Rule: Example

- What percent of students consume between 460 and 700 cans of beer per year?



The Empirical Rule: Example

- What percent of students consume between 460 and 700 cans of beer each year?
- We can add up the area under the curve as we go from 460 to 700
  2.35%+13.5%+34%+34% = 83.85%



Chebyshev’s Rule

- Chebyshev’s Rule is very similar to the Empirical Rule **except** we don’t require the distribution must be symmetric and bell shaped to use this Rule

Chebyshev's Rule

- It is possible that very few observations fall within 1 standard deviation of the mean
- At least 75% of the data fall within 2 standard deviation of the mean
- At least $88.\overline{88}$% of the data fall within 3 standard deviation of the mean
- In general, at least $\left[\left(1-\frac{1}{\mathrm{k}^2}\right)*100\right]$% of the data will fall within k standard deviations of the mean

Chebyshev’s Rule: Example

- Let’s say this time that the average college student consumes 640 cans of beer each year. Assume the distribution of cans of beers consumed per college student is **not bell-shaped** with a **mean of 640 cans** and a **standard deviation of 60 cans**.

Chebyshev’s Rule Example

- **It is possible that very few observations fall within 1 standard deviation of the mean**

- It is possible that few students drink between (640-60)=580 and (640+60)=700 cans of beer each year

Chebyshev's Rule Example

- **At least 75% of the data fall within 2 standard deviation of the mean**

- At least 75% of students drink between (640-2*60)=520 and (640+2*60)=760 cans of beer each year

Chebyshev's Rule Example

- **At least $88.\overline{88}\%$ of the data fall within 3 standard deviation of the mean**

- At least $88.\overline{88}\%$ of students drink between (640-3*60)=460 and (640+3*60)=820 cans of beer each year

Chebyshev's Rule Example

- **At least $88.\overline{88}\%$ of the data fall within 3 standard deviation of the mean**

- At least $88.\overline{88}\%$ of students drink between (640-3*60)=460 and (640+3*60)=820 cans of beer each year

Chebyshev's Rule Example

- **In general, at least $\left[\left(\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{k}^{\mathbf{2}}}\right)*\mathbf{100}\right]$% of the data will fall within k standard deviations of the mean**
- This allows us to choose a "between k standard deviations" and find the percent of the data that should fall on that interval
- This also allows us to choose a percentage and solve for k

Chebyshev's Rule Example

- **In general, at least $\left[\left(\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{k}^{\mathbf{2}}}\right)*\mathbf{100}\right]$% of the data will fall within k standard deviations of the mean**
- **This allows us to choose a "between k standard deviations" and find the percent of the data that should fall on that interval**

    1. $\left(1-\frac{1}{1^2}\right)*100=0\%$
    2. $\left(1-\frac{1}{2^2}\right)*100=75\%$
    3. $\left(1-\frac{1}{3^2}\right)*100=88.\overline{88}\%$

Chebyshev's Rule Example

- **In general, at least $\left[\left(\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{k}^{\mathbf{2}}}\right) * \mathbf{100}\right]$% of the data will fall within k standard deviations of the mean**
- **This also allows us to choose a percentage and solve for k**
- Say we wanted to find an interval where 90% of the data lies we solve:

$$90 = \left(1 - \frac{1}{k^2}\right) * 100$$
$$.9 = \left(1 - \frac{1}{k^2}\right)$$
$$\frac{1}{k^2} = .1$$
$$k^2 = 10$$
$$k = \sqrt{10} \approx 3.162278$$

  - Here we can say that at least 90% of students drink within 3.162278 standard deviations of the mean

Z Score: If you don’t know what it is you can’t afford it.

- What happens when we’re interested in percentiles and x values that aren’t perfectly spaced according to the Empirical Rule or Chebyshev’s Rule?
- We note that in most scenarios the data we’re concerned with will fit this scenario.
- We can also use z scores to provide a numerical method for finding outliers

Z Score: What are we doing here?

- What did we do with the Empirical and Chebyshev's Rules?
  - We looked at how many whole standard deviations away the data values were
- The idea here is to be able to find out how many standard deviations the data values we're looking at are from the mean but we allow fractional answers
  - answers outside of -3, -2,-1, 0 , 1, 2, 3 which the Empirical and Chebyshev's Rules cover

Z Score: How Do We Calculate It?

- $z = \frac{observation - mean}{standard\ deviation} = \frac{x - \mu_x}{\sigma_x}$

- This gives us the number of standard deviations from the mean the observation is

- **Note: we consider any observation with a Z score above 3 or below -3 an outlier**

Z Score: Example

- The average college student consumes 640 cans of beer per year. Assume the distribution of beers consumed per year per college student is **bell-shaped** with a **mean of 640 cans** and a **standard deviation of 60 cans**.

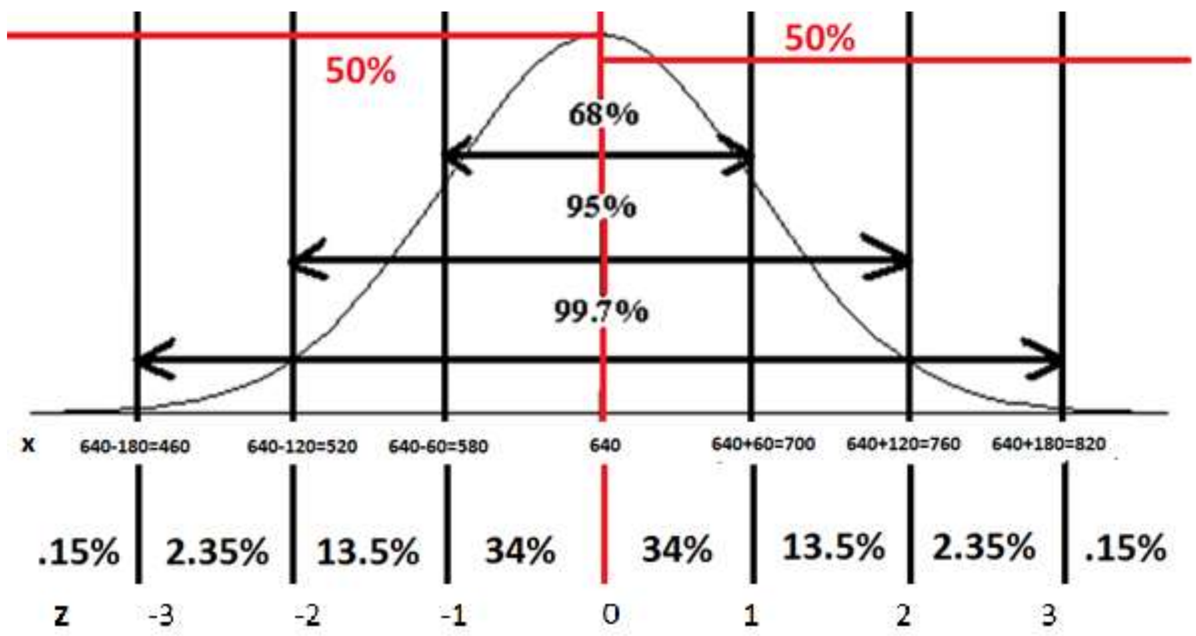
Z Score: Example

- $z_{460} = \dfrac{460 - 640}{60} = \dfrac{-180}{60} = -3$
- $z_{820} = \dfrac{820 - 640}{60} = \dfrac{180}{60} = 3$

- Note the Z score has given us the correct number of standard deviations from the mean for each case!

Z Score: Example

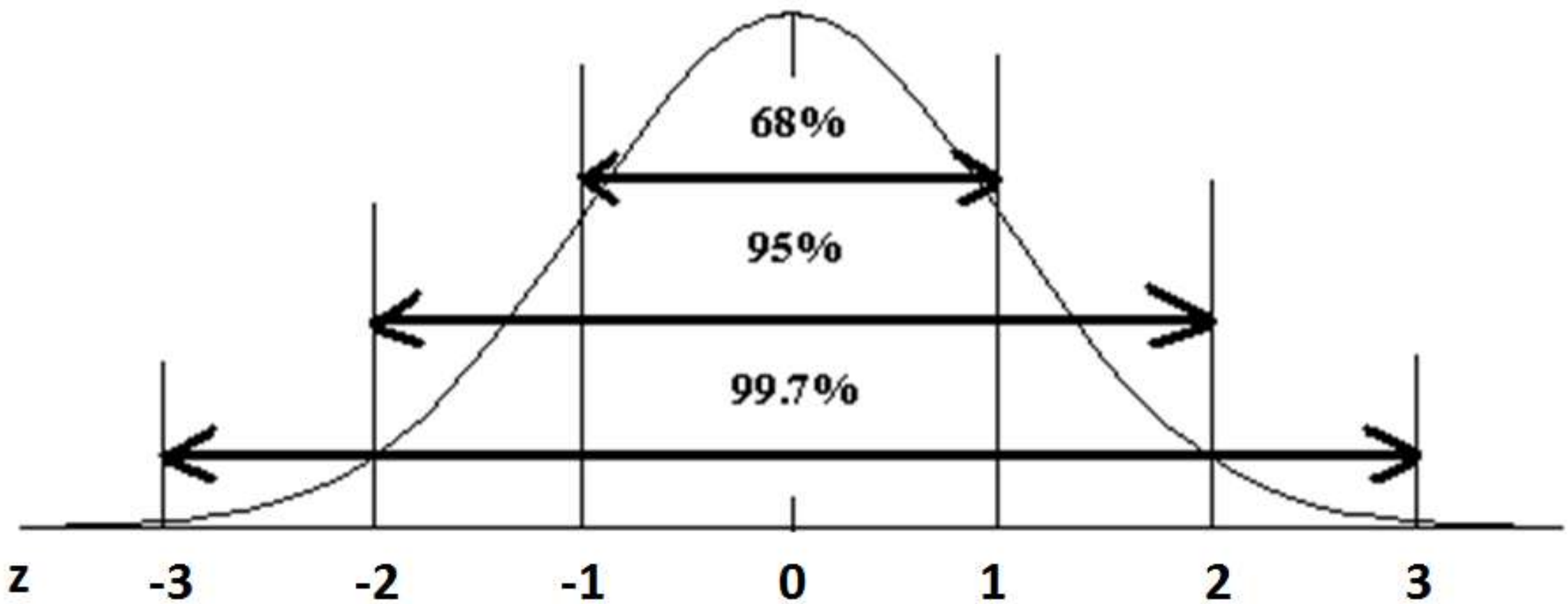
- Recall from the Empirical Rule that about 99.7% of college students consume between 460 and 820 cans of beer per year (+- 3 standard deviations)



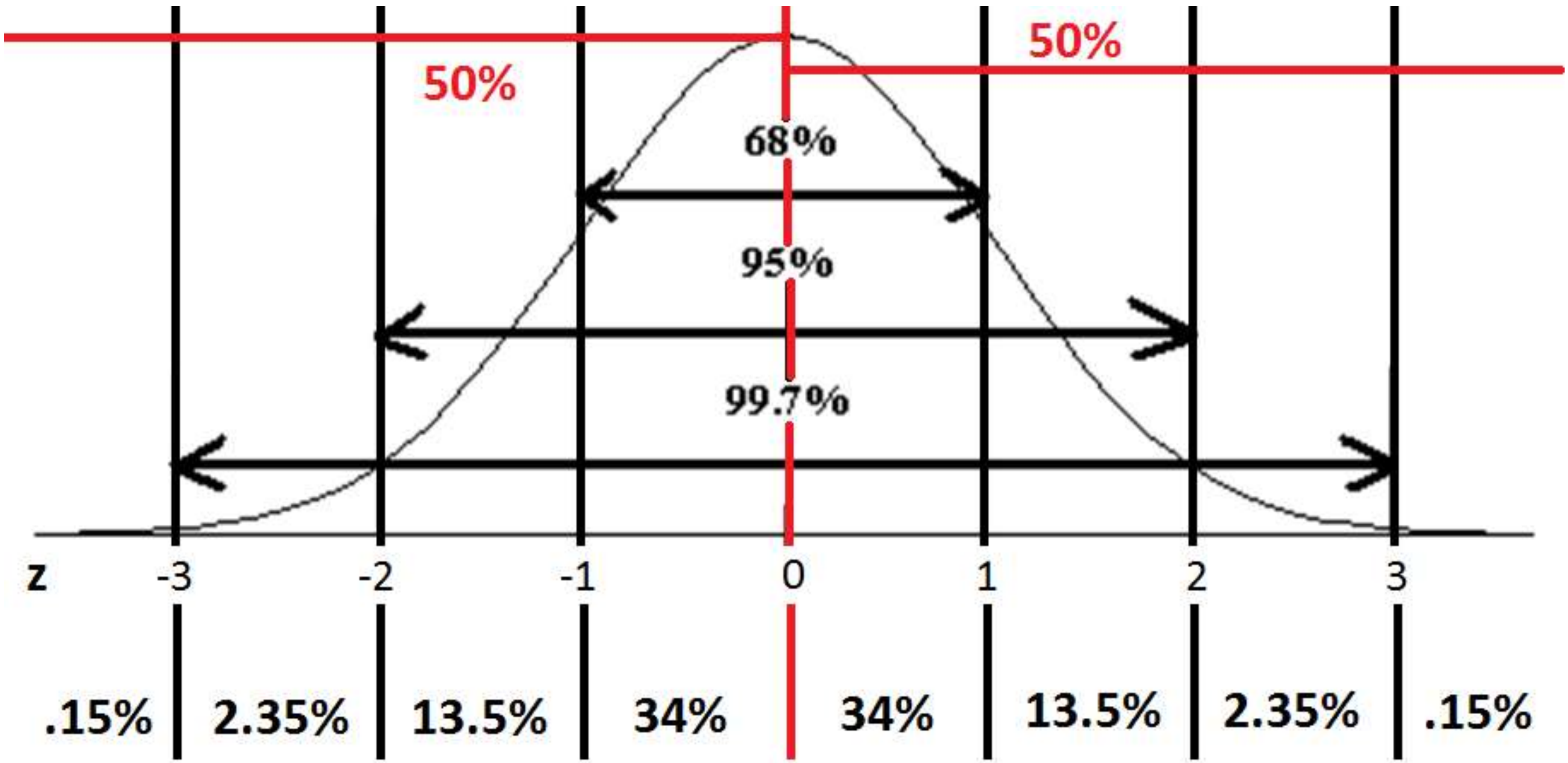
The Empirical Rule with z-scores

- About 68% of data fall between z=-1 and z=1
- About 95% of data fall between z=-2 and z=2
- About 99.7% of data fall between z=-3 and z=3

- **The distribution must be symmetric and bell shaped to use this Rule**


68%
95%
99.7%
z
-3
-2
-1
0
1
2
3

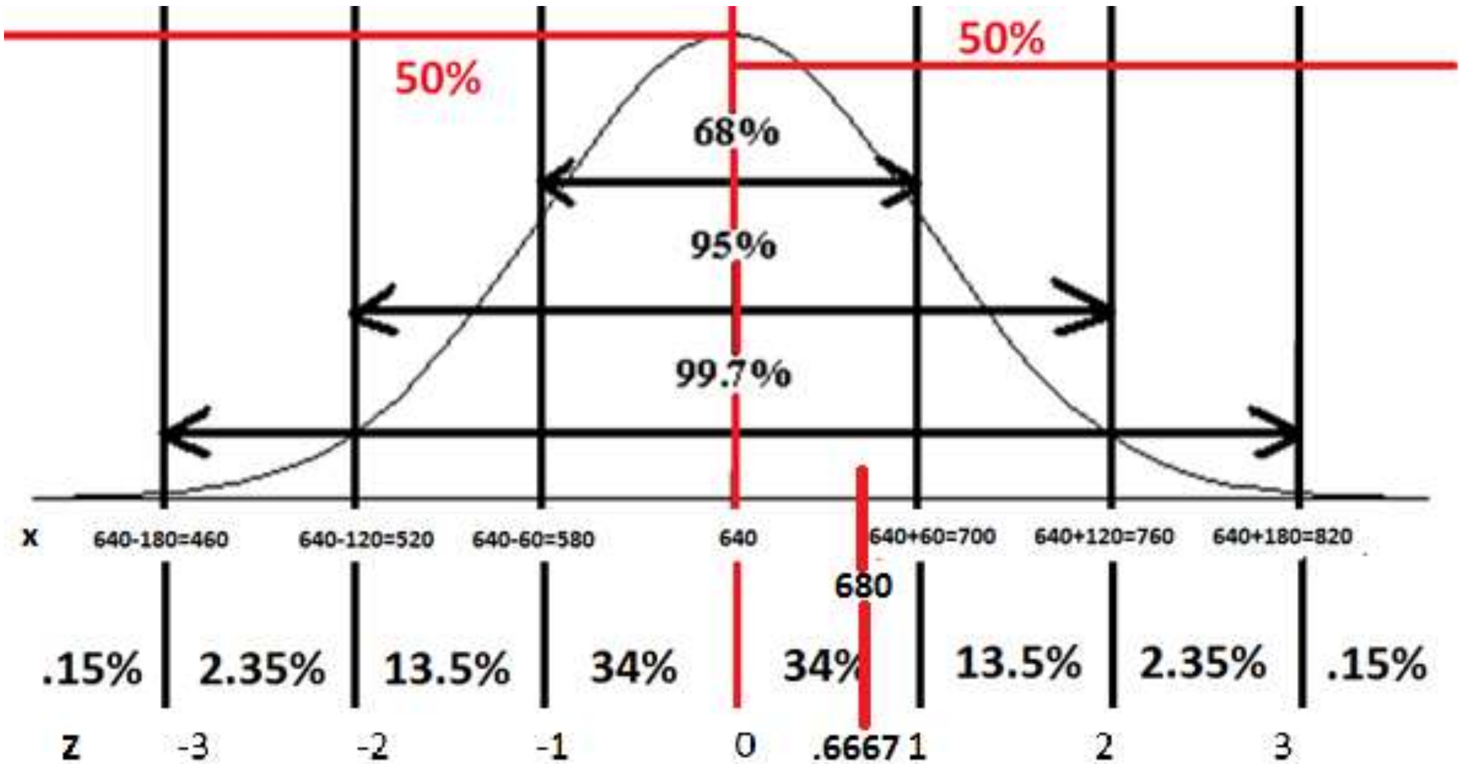
Empirical Rule



Z Score: Example 2

- Let's consider an observation of 680 cans of beer.
  - 680 is not 1, 2, or 3 standard deviations away
  - $z = \frac{680-640}{60} = .6667$
    - X=680 is .6667 standard deviations above the mean
    - .6667 indicates this observation is not an outlier because .6667<3 and .6667>-3
    - We will be able to find these percentages in chapter 6 so don't forget z-scores!

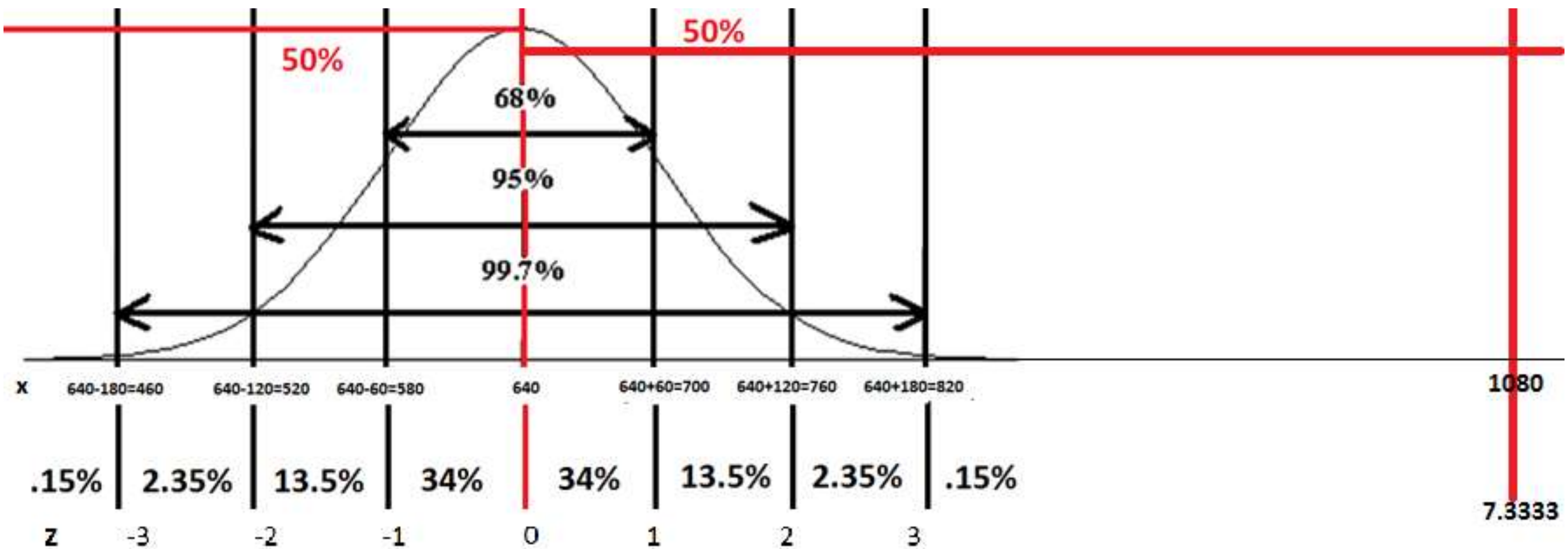
Z Score: Example 2



Z Score: Example 3

- Let's consider an observation of 1080 cans of beer.
  - 1080 is not 1, 2, or 3 standard deviations away
  - $z = \frac{1080-640}{60} = 7.3333$
    - X=1080 is 7.3333 standard deviations above the mean
    - Z=.7333 indicates this observation is an outlier because 7.3333>3

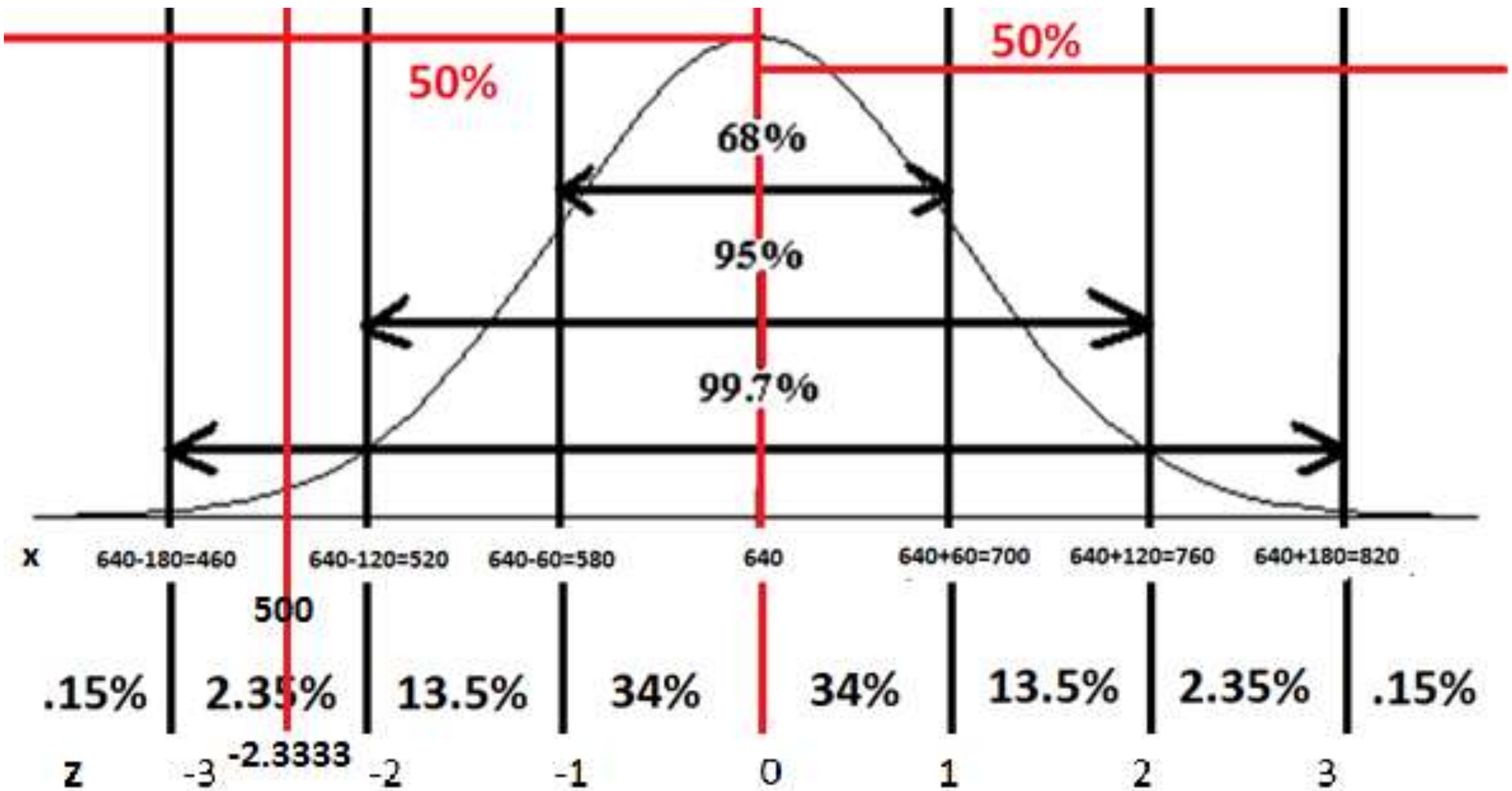
Z Score: Example 3



Z Score: Example 4

- Let's consider an observation of 500 cans of beer.
  - 500 is not 1, 2, or 3 standard deviations away
  - $z = \frac{500-640}{60} = -2.3333$
    - X=500 is 2.3333 standard deviations below the mean
    - -2.3333 indicates this observation isn't an outlier because -2.3333<3 and -2.3333>-3

Z Score: Example 4

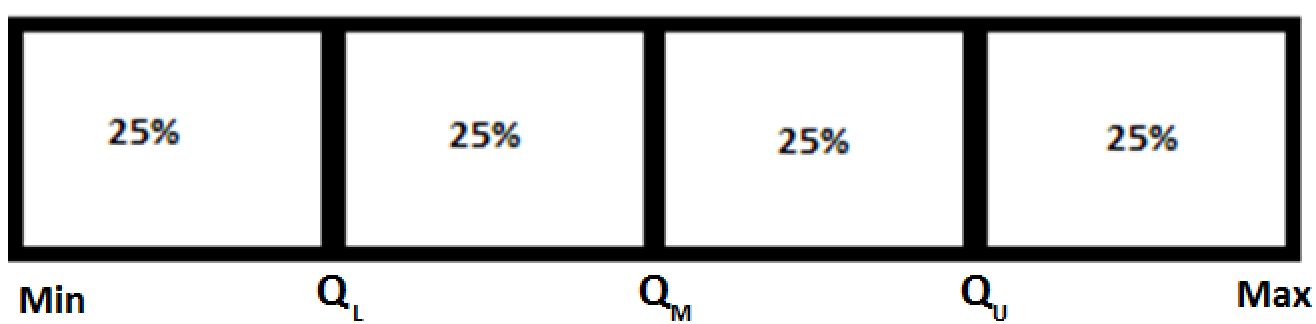


Percentiles

- How many of you have heard this term before?
  - Testing
  - Medical terminology
  - Etc
- **Percentiles -** the pth percentile is a value such that p percent of the observations fall below or at that value.

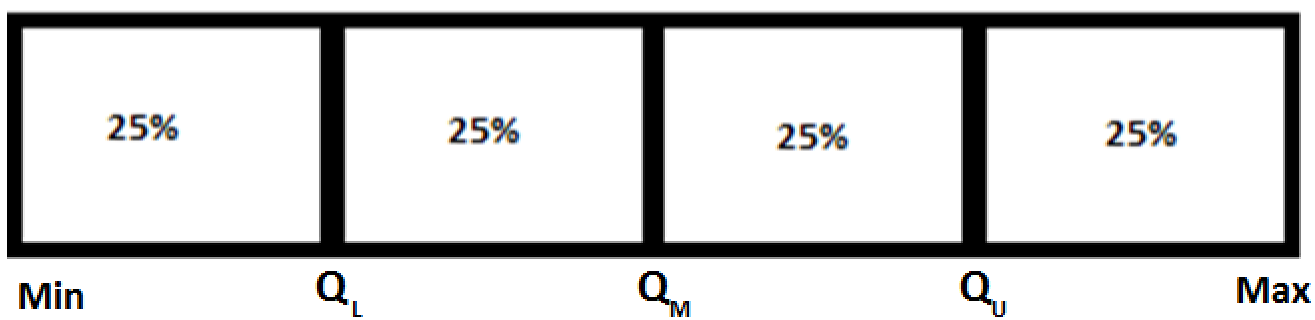
Five Number Summary: Important Percentiles

- We call these quartiles because they split the data into quarters
  - Q_L: the observation at the 25^{th} percentile
  - Q_M: the observation at the 50^{th} percentile
    - This is the same as the median
  - Q_U: the observation at the 75^{th} percentile
- Min: the smallest observation – the 0^{th} percentile
- Max: the largest observation – the 100^{th} percentile



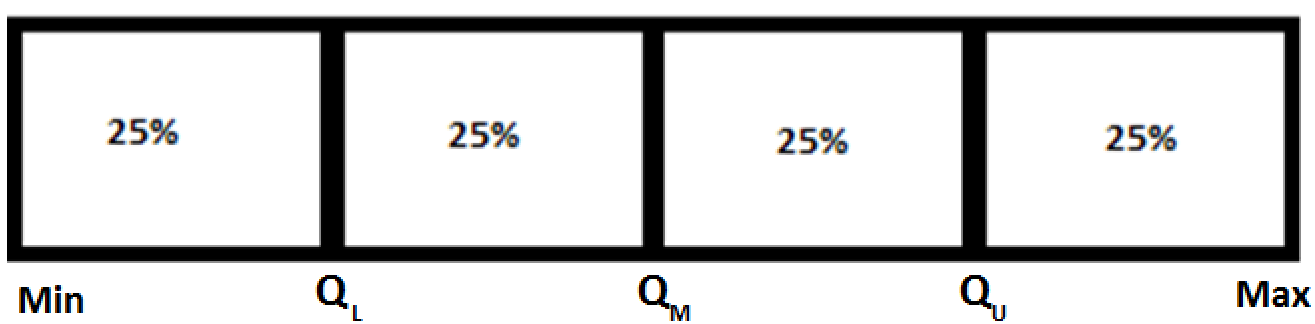
Five Number Summary: Interquartile Range

- **IQR**=Q_U-Q_L: another measure of spread used in place of standard deviation w/ skewed data
  - IQR gives the range of the middle 50% of the data



Five Number Summary: Finding Outliers with Quartiles

- Lower Fence= Q_L – (1.5)*IQR
  = 1.5 – (1.5)*5 = -6
- Upper Fence= Q_U + (1.5)*IQR
  = 6.5 + (1.5)*5 = 14
- **We consider any observation with a value outside of the interval (Lower Fence, Upper Fence) an outlier**



Five Number Summary: Where to Find Them

- The five number summary, of n items, that we use to draw a box plot includes the following:

Name	Position in Ascending Order
Minimum	1st
Q_L	.25*(n+1)th
Q_M (This is the median)	.5*(n+1)th
Q_U	.75*(n+1)th
Maximum	nth

Example: The Lower (1st) Quartile

Is the position value a whole number	The Quartile
Yes	The number in that position
No	The weighted average of the numbers in the above and below positions

- X = {0,1,2,3,4,5,6,7,8)
- Position of Q_L = .25*(n+1) = .25*(9+1)

  = 2.5th position (the remainder is .5)
- Q_L= (.5)*(# In the 3rd pos.) +(1-.5)*(# in the 2nd pos.)

  = .5*2 + .5*1 = 1 + .5 = 1.5

Example: The Middle (2nd) Quartile

Is the position value a whole number	The Quartile
Yes	The number in that position
No	The average of the numbers in the above and below positions

- X = {0,1,2,3,4,5,6,7,8)
- Position of the Median = .5*(n+1) = .5*(9+1)

  = 5th position
- Q_M= 4

Example: The Upper (3^{rd}) Quartile

Is the position value a whole number	The Quartile
Yes	The number in that position
No	The average of the numbers in the above and below positions

- X = {0,1,2,3,4,5,6,7,8)
- Position of Q_U = .75*(n+1) = .75*(9+1)

  = 7.5^{th} position (.5 is the remainder)
- Q_U = = (.5)*(# In the 8^{th} pos.) +(1-.5)*(# in the 7^{th} pos.)

  = .5*7 + .5*6 = 1 + 1.5 = 6.5

Example: Interquartile Range

X = {0,1,2,3,4,5,6,7,8)

- Q_L = (1+2)/2 = 1.5
- Q_M = 4
- Q_U = (6+7)/2 = 6.5

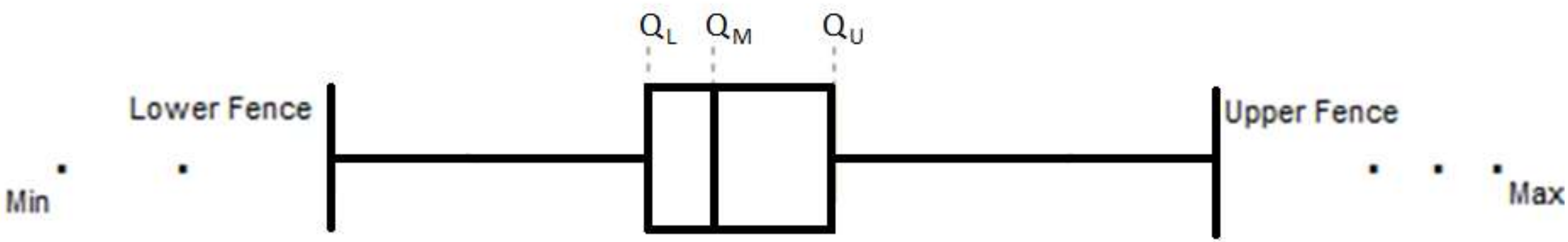
- IQR = $Q_U - Q_L$ = 6.5 - 1.5 = 5
  – 50% of the data lies between 1.5 and 6.5
  – 50% of the data lies on a range of size 5

Example: Using Quartiles to find Outliers

X = {0,1,2,3,4,5,6,7,8)

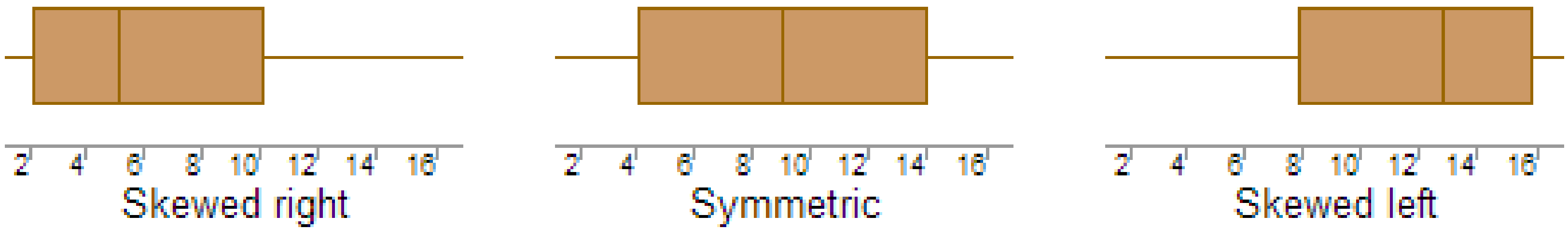
- Q_L = (1+2)/2 = 1.5
- Q_U = (6+7)/2 = 6.5
- IQR = $Q_U - Q_L$ = 6.5 - 1.5 = 5
- Lower Fence= Q_L – (1.5)*IQR
  = 1.5 – (1.5)*5 = -6
- Upper Fence= Q_U + (1.5)*IQR
  = 6.5 + (1.5)*5 = 14
- **In this case anything smaller than -6 or greater than 14 would be an outlier**

Box Plots:
The Graph of a Five Number Summary

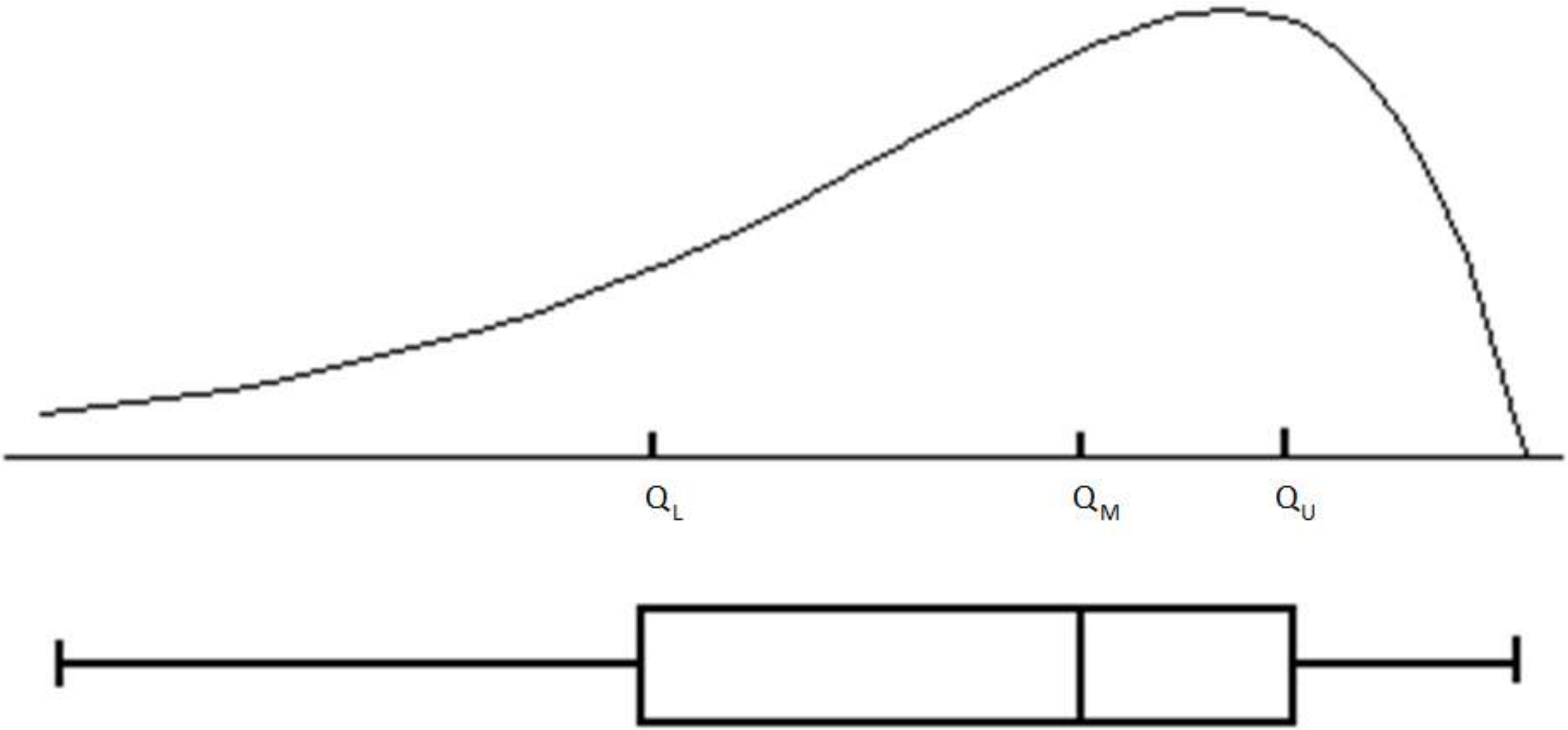


- The box plot utilizes the five number summary
  - The box is created using quartiles
  - The whiskers are created using the fences
  - The points are the outlying points –if there are any

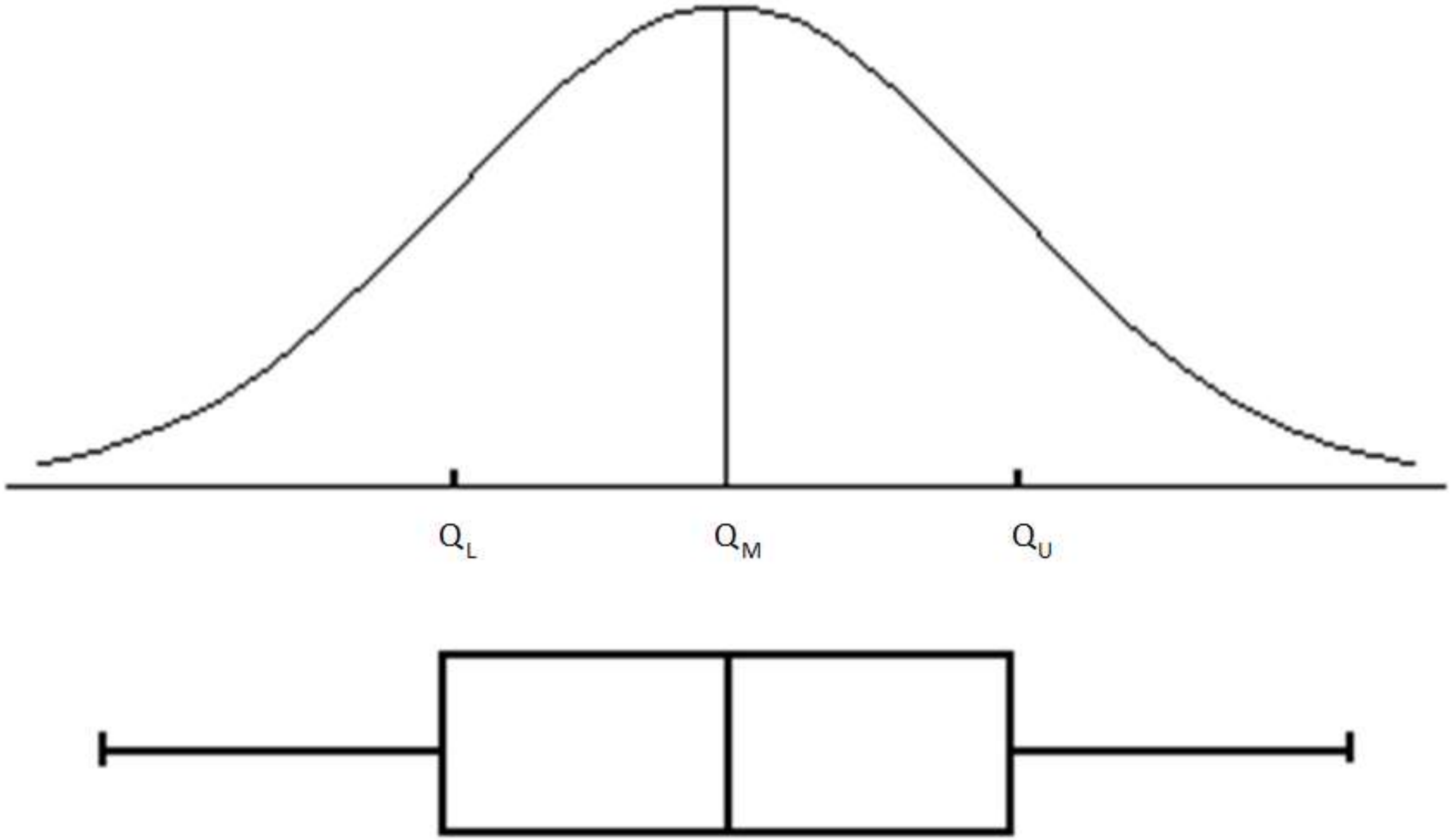
Skewness in Boxplots



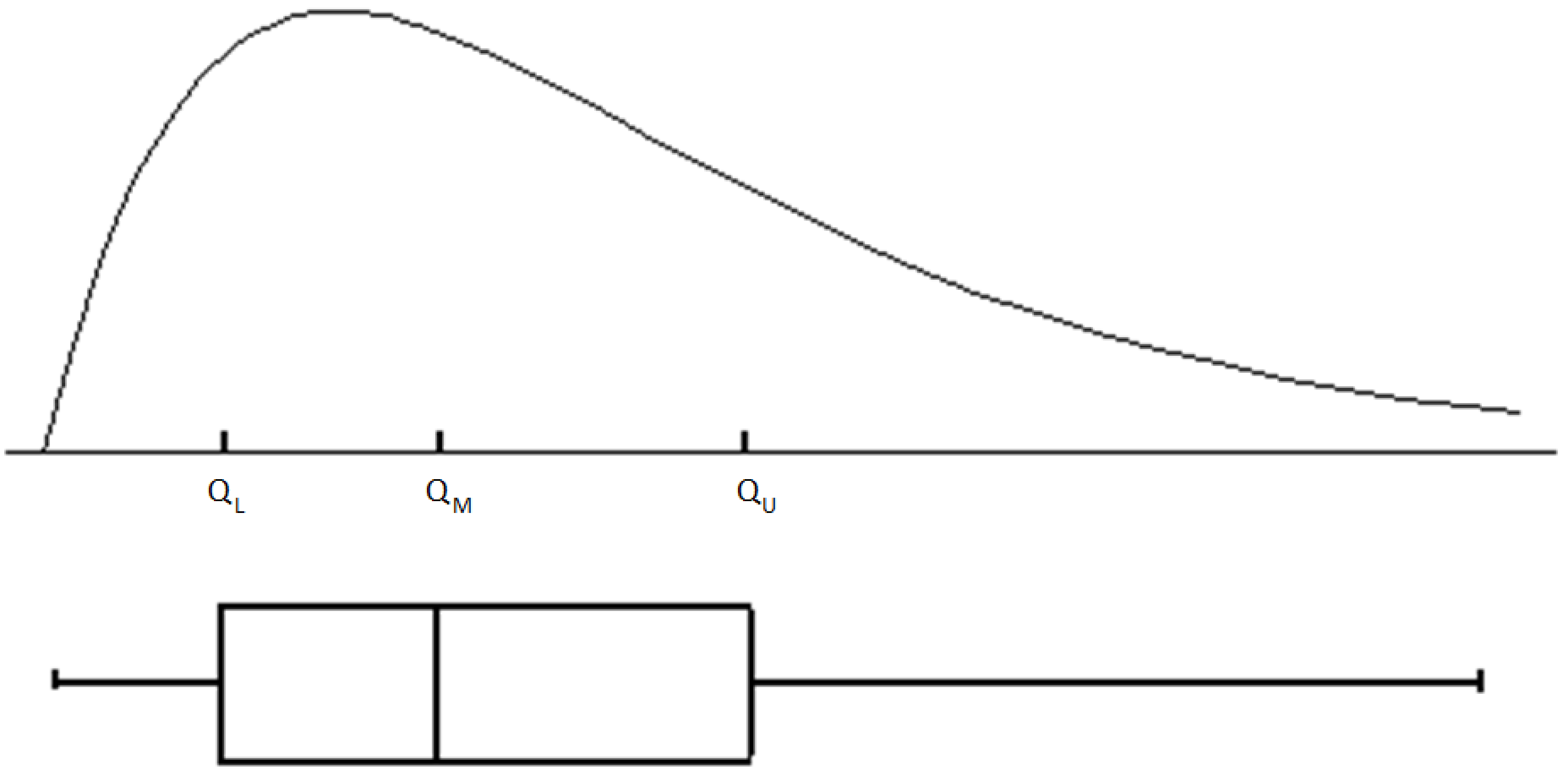
Left Skewed w/ Boxplots



Bell Shaped w/ Boxplots



Right Skewed w/ Boxplots



Watch This!

- Sample Vs. Population*
  - https://www.youtube.com/watch?v=lnDPVBp-1_A
- Mean median and mode
  - https://www.youtube.com/watch?v=5C9LBF3b65s
- Dispersion Walkthrough*
  - https://www.youtube.com/watch?v=9mnjDp6tg-4

Summarizing Quantitative Data: Numerical Summaries

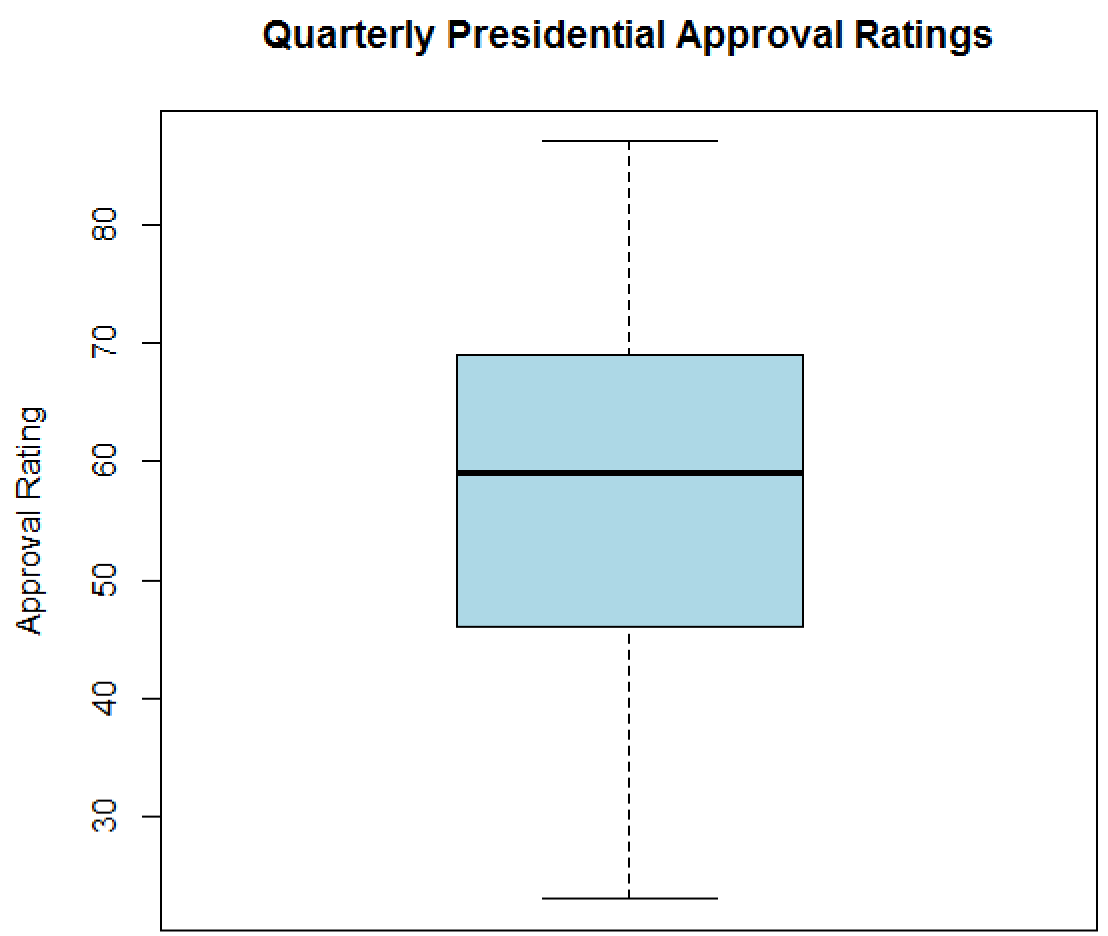
- **R Commands:**

```
##################################################################################
####################Numerical Summaries#####################################
##################################################################################
presidents <- presidents[!is.na(presidents)] #REMOVES NA
#mean
mean(presidents)
#median
median(presidents)
#mode
mode(presidents)
#range
max(presidents)-min(presidents)
#standard deviation
sd(presidents)
#variance
var(presidents)
#Five Number Summary - we can calculate IQR and fences from here
summary(presidents)
###################################################################################
###################################################################################
```

Summarizing Quantitative Data: Box Plot

- **R Commands:**

```
############################################################################
##############################Box Plots######################################
############################################################################
#R automatically takes care of the fencing
boxplot(presidents)
#add title
boxplot(presidents,main="Quarterly Presidential Approval Ratings")
#add y-label
boxplot(presidents,main="Quarterly Presidential Approval Ratings",ylab="Approval Rating")
#add color
boxplot(presidents,main="Quarterly Presidential Approval Ratings",ylab="Approval Rating", col="light blue")
############################################################################
############################################################################
```

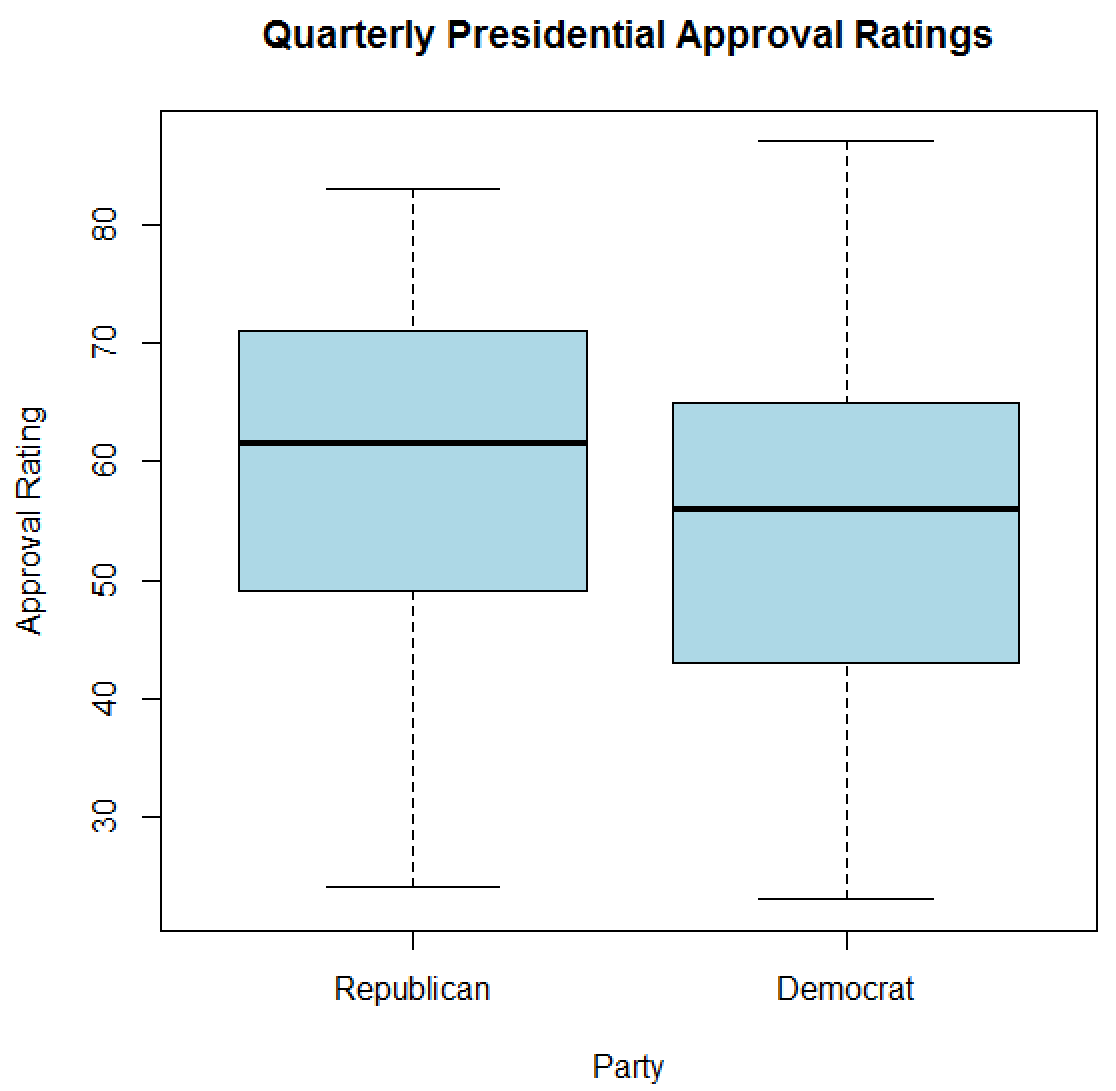


Quarterly Presidential Approval Ratings
Approval Rating
30
40
50
60
70
80

Summarizing Quantitative Data: Side by Side Box Plots

- **R Commands:**

```
##########################################################################
######################Two Box Plots####################################
##########################################################################
ryr<-c(36:64,104:120)
dyr<-c(1:32,68:100)
Repub<-presidents[ryr]
Dem<-presidents[dyr]
#Basic
boxplot(Repub,Dem)
#Add Title
boxplot(Repub,Dem,main="Quarterly Presidential Approval Ratings")
#Add y-labels
boxplot(Repub,Dem,main="Quarterly Presidential Approval Ratings",ylab="Approval Rating")
#Add x-labels
boxplot(Repub,Dem,main="Quarterly Presidential Approval Ratings",ylab="Approval Rating",
xlab="Party",names=c("Republican","Democrat"))
#Add color
boxplot(Repub,Dem,main="Quarterly Presidential Approval Ratings",ylab="Approval Rating",
xlab="Party",names=c("Republican","Democrat"),col="light blue")
```

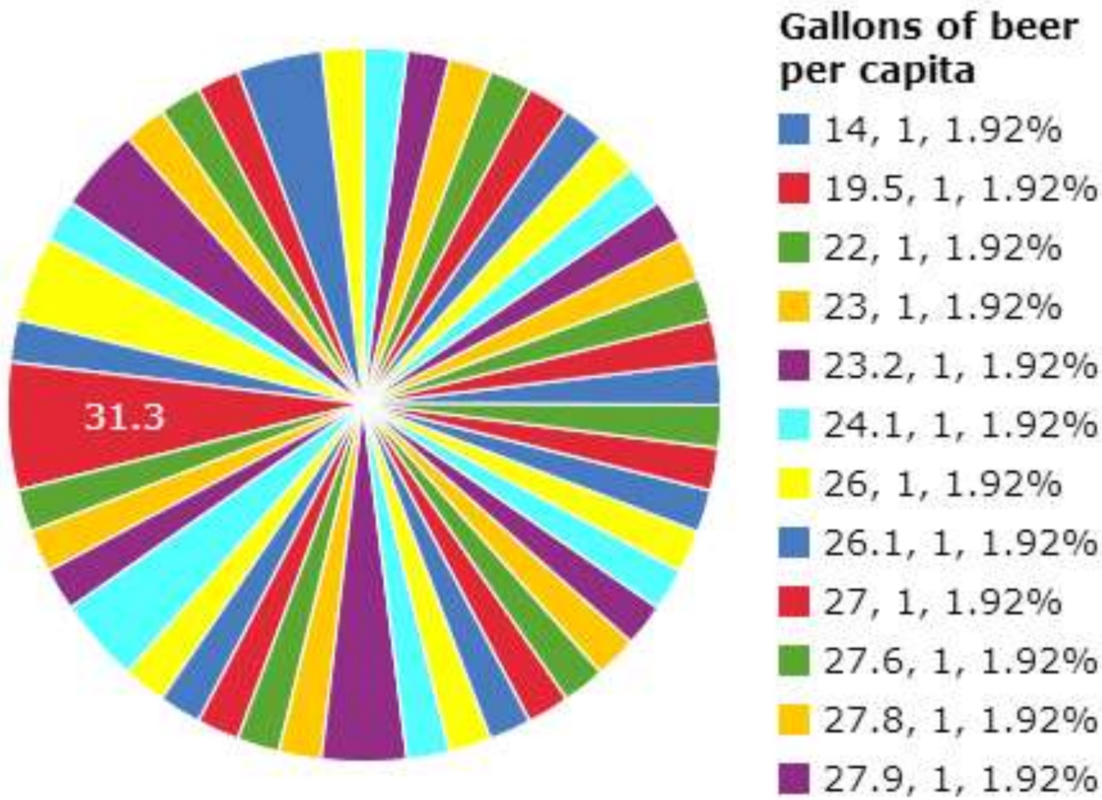

Quarterly Presidential Approval Ratings
Approval Rating
30
40
50
60
70
80
Republican
Democrat
Party

Summary!

Graphical Displays

Variable Type	Graphical Display	Numerical Summary
Categorical	Pie chart or bar graph	Frequency table
Quantitative	Histogram or box plot – can also try dotplot or stem & leaf	Quantitative Summary
1-Categorical and 1-Quantitative	Side by Side boxplots	Quantitative Summary for groups
2-Categorical	Side by side pie charts or bar graphs best: stacked bar chart	Contingency Table or side by side frequency tables
2-Quantitative	Scatter plot	Side by side Quantitative Summaries

Remember: With graphs, if it's ugly it's probably not right.



Misrepresentation of Data

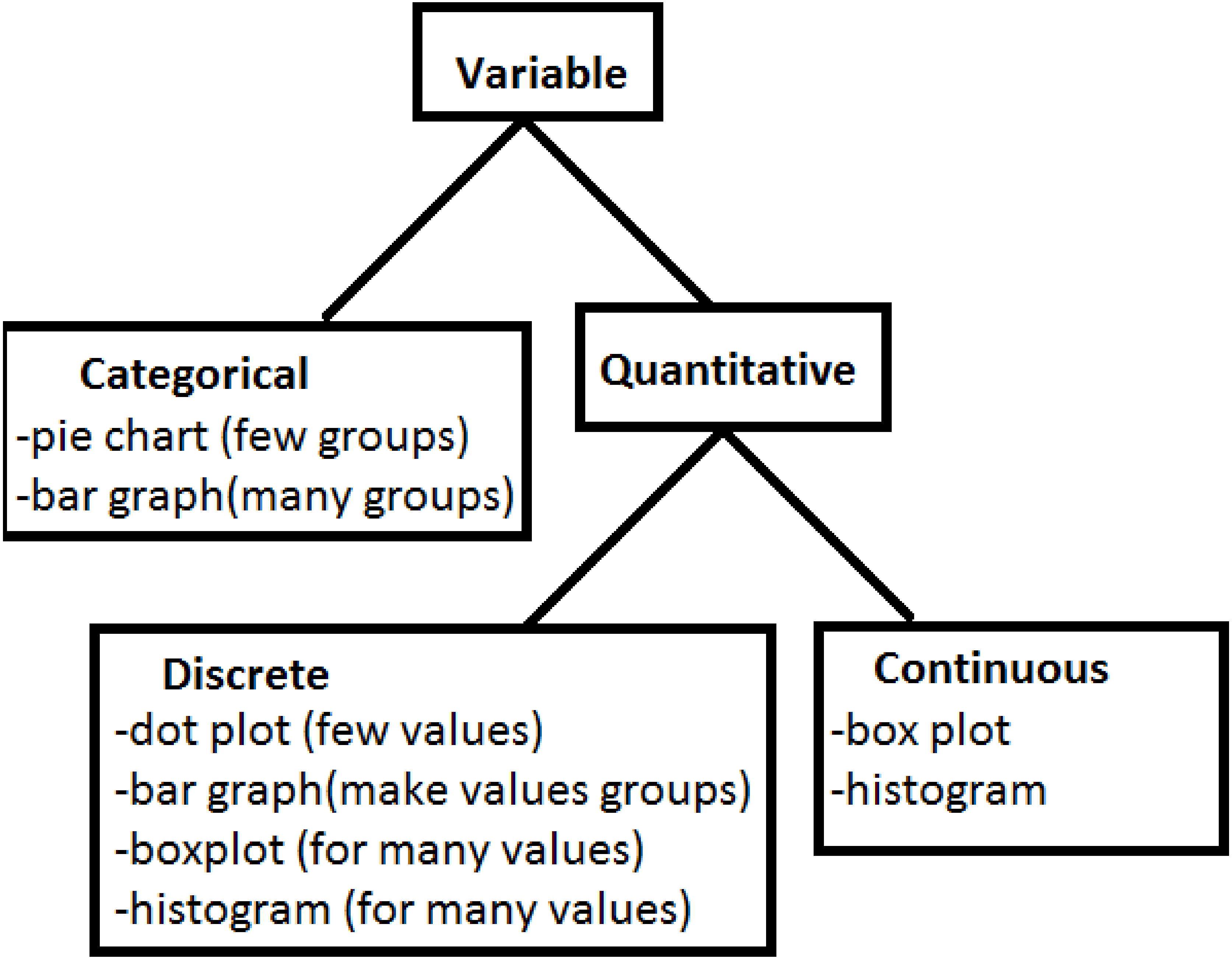
- You should be able to look at your graphs and realize when you’ve made a mistake

    -The percentages of all relative frequency graphs should add to 1 or 100%

    -The scale should be understandable and constant

    -Consider whether or not you need to start your y axis at zero or caution against misreading the graph

    -Graphs should be simple and easy to interpret correctly in just a few moments.


Variable
Categorical
-pie chart (few groups)
-bar graph(many groups)
Quantitative
Discrete
-dot plot (few values)
-bar graph(make values groups)
-boxplot (for many values)
-histogram (for many values)
Continuous
-box plot
-histogram

Measures of Central Tendency

Measure	Computation	R Command	Interpretation	When to Use
Mean Statistic: $\bar{x}$ Parameter: μ	$\bar{x} = \frac{\sum x}{n}$	mean(data)	Center of Gravity	Use for quantitative data when the distribution is roughly symmetric
Median	The point halfway through the data when it is arranged in ascending order.	median(data)	The point which splits the data in half.	Use for quantitative data when the distribution is skewed
Mode	We report the observation with the highest frequency	mode(data)	Most frequent observation	When the most frequent observation is the desired measure or when data is qualitative.

* Denotes robustness to outliers – to be used when data is not bell-shaped

Measures of Dispersion

Measure	Computation	R command	Interpretation
Range	Max – Min	max(data) – min(data)	The difference between the largest and smallest data point
Standard Deviation Statistic: s Parameter: σ	$\sqrt{Variance}$	sd(data)	The square root of the mean of squared deviations from the mean in the original units – this usually makes the standard deviation easier to interpret
Variance Statistic: s^2 Parameter: σ^2	$\frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\boldsymbol{2}}}{\boldsymbol{n}-\boldsymbol{1}}$	var(data)	The square root of the mean of squared deviations from the mean in units squared
IQR*	$Q_U - Q_L$	Calculated from summary(data)	The range of the middle 50%

* Denotes robustness to outliers – to be used when data is not bell-shaped

The Empirical Rule

- About 68% of data fall within 1 standard deviation of the mean
- About 95% of data fall within 2 standard deviation of the mean
- About 99.7% of data fall within 3 standard deviation of the mean
- **The distribution must be symmetric and bell shaped to use this Rule**

The Empirical Rule with z-scores

- About 68% of data fall between z=-1 and z=1
- About 95% of data fall between z=-2 and z=2
- About 99.7% of data fall between z=-3 and z=3

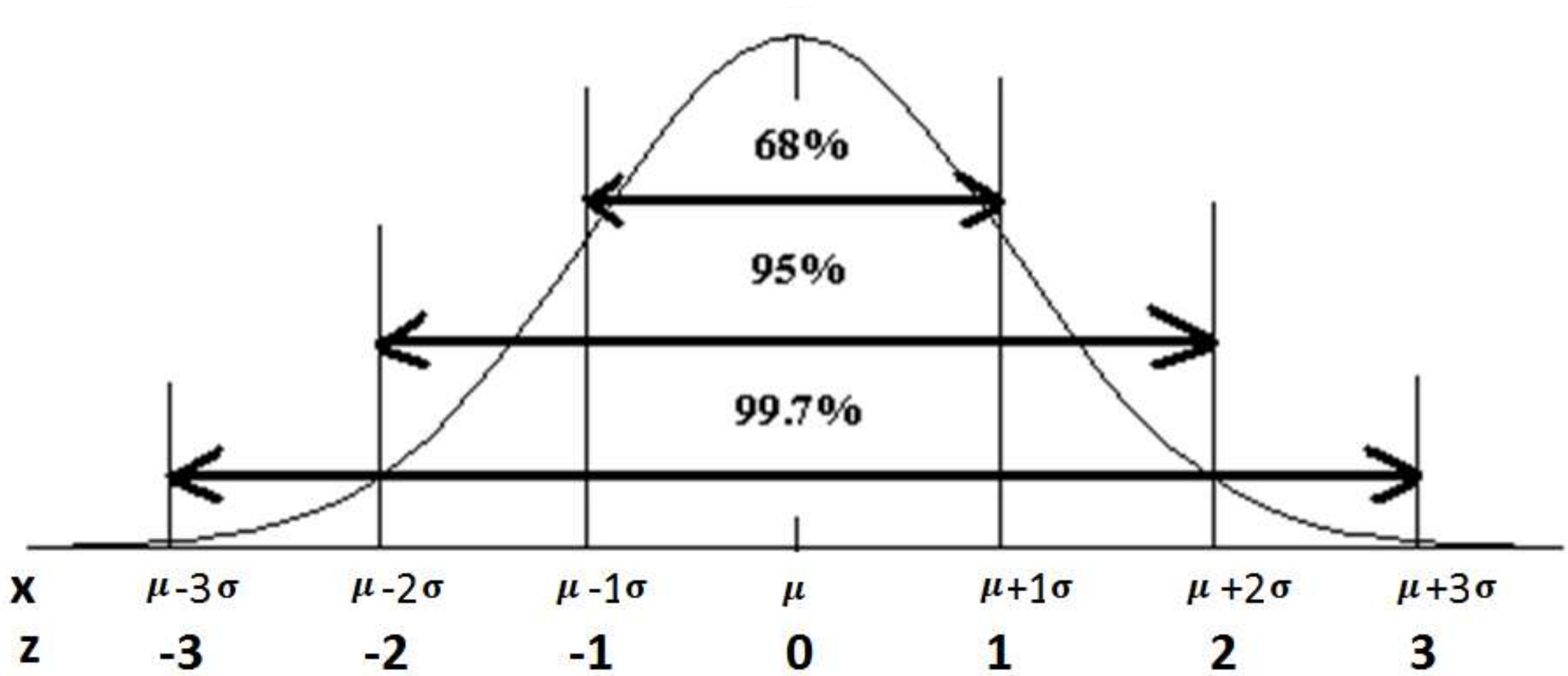
- **The distribution must be symmetric and bell shaped to use this Rule**

Z Score: How Do We Calculate It?

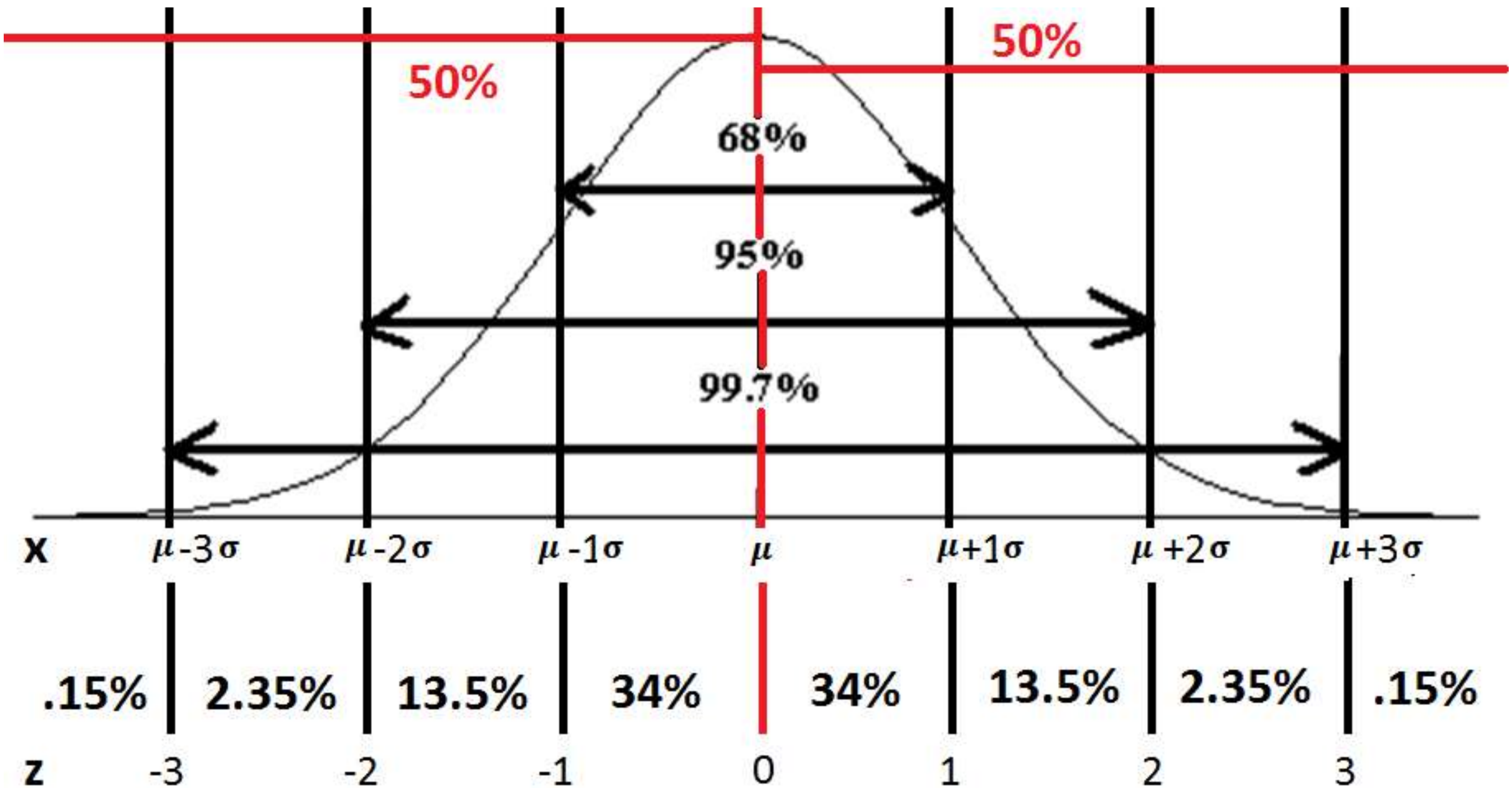
- $z = \frac{observation - mean}{standard\ deviation}$

- This gives us the number of standard deviations from the mean the observation is

- **Note: we consider any observation with a Z score above 3 or below -3 an outlier**


68%
95%
99.7%
x
μ-3σ
μ-2σ
μ-1σ
μ
μ+1σ
μ+2σ
μ+3σ
z
-3
-2
-1
0
1
2
3

Empirical Rule



Chebyshev's Rule

- It is possible that very few observations fall within 1 standard deviation of the mean
- At least 75% of the data fall within 2 standard deviation of the mean
- At least $88.\overline{88}$% of the data fall within 3 standard deviation of the mean
- In general, at least $\left[\left(1-\frac{1}{k^2}\right)*100\right]$% of the data will fall within k standard deviations of the mean