Chapter 1: Reviews and Introduction

Halima Bensmail
CS502
Monday 4-7pm
LAS Hall C

Statistics: The Science of Data

Collecting Data
- Surveys
- Experiments

Presenting Data
- Chart and Tables

Characterizing data
- Average
- Variances

Data Analysis

Decision Making

Types/categories of data:

–**Cross-sectional** (are not time ordered, collected at the same or approximately the same point in time)

vs

Longitudinal(ordered by time, collected over several time periods)

–**Observational** (are not under the control of the analyst) vs

Experimental (are designed study: agriculture, medicine).

•Data summarization vs data modeling

-**Summary**: describes the data and suggests a link to a model (modeling).

•Two purposes of **modeling**: Understanding vs forecasting.

•Importance of Graphing data

–First and last thing an analyst should do

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Help With R

Getting Help

Documentation

Manuals
FAQs
The R Journal
Books
Certification
Other

Links

Bioconductor

The R Project for Statistical Computing

Getting Started

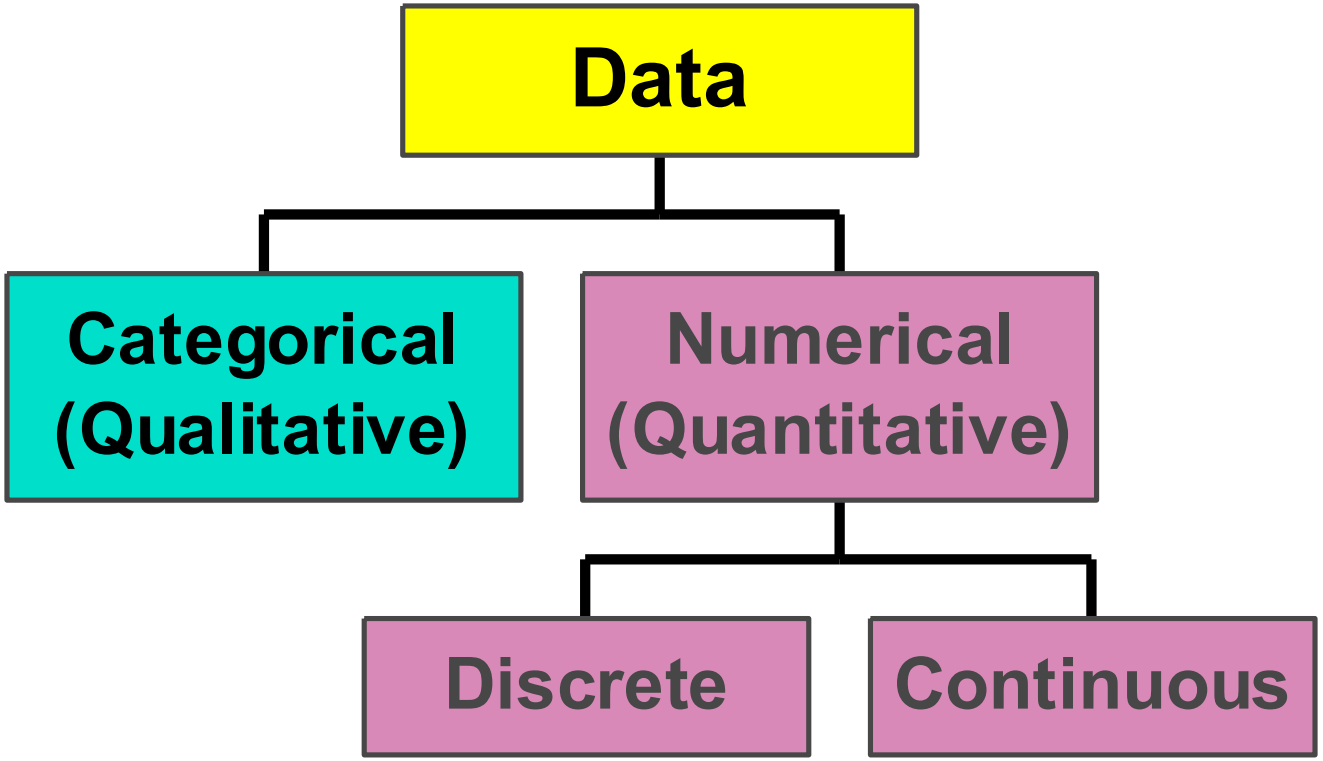
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- The R Journal Volume 8/1 is available.
- The useR! 2017 conference will take place in Brussels, July 4 - 7, 2017, and details will be appear here in due course.
- R version 3.3.1 (Bug in Your Hair) has been released on Tuesday 2016-06-21.
- R version 3.2.5 (Very, Very Secure Dishes) has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- Notice XQuartz users (Mac OS X) A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.
- The R Logo is available for download in high-resolution PNG or SVG formats.
- useR! 2016, have taken place at Stanford University, CA, USA, June 27 - June 30, 2016.
- The R Journal Volume 7/2 is available.
- R version 3.2.3 (Wooden Christmas-Tree) has been released on 2015-12-10.
- R version 3.1.3 (Smooth Sidewalk) has been released on 2015-03-09.

Types of Data



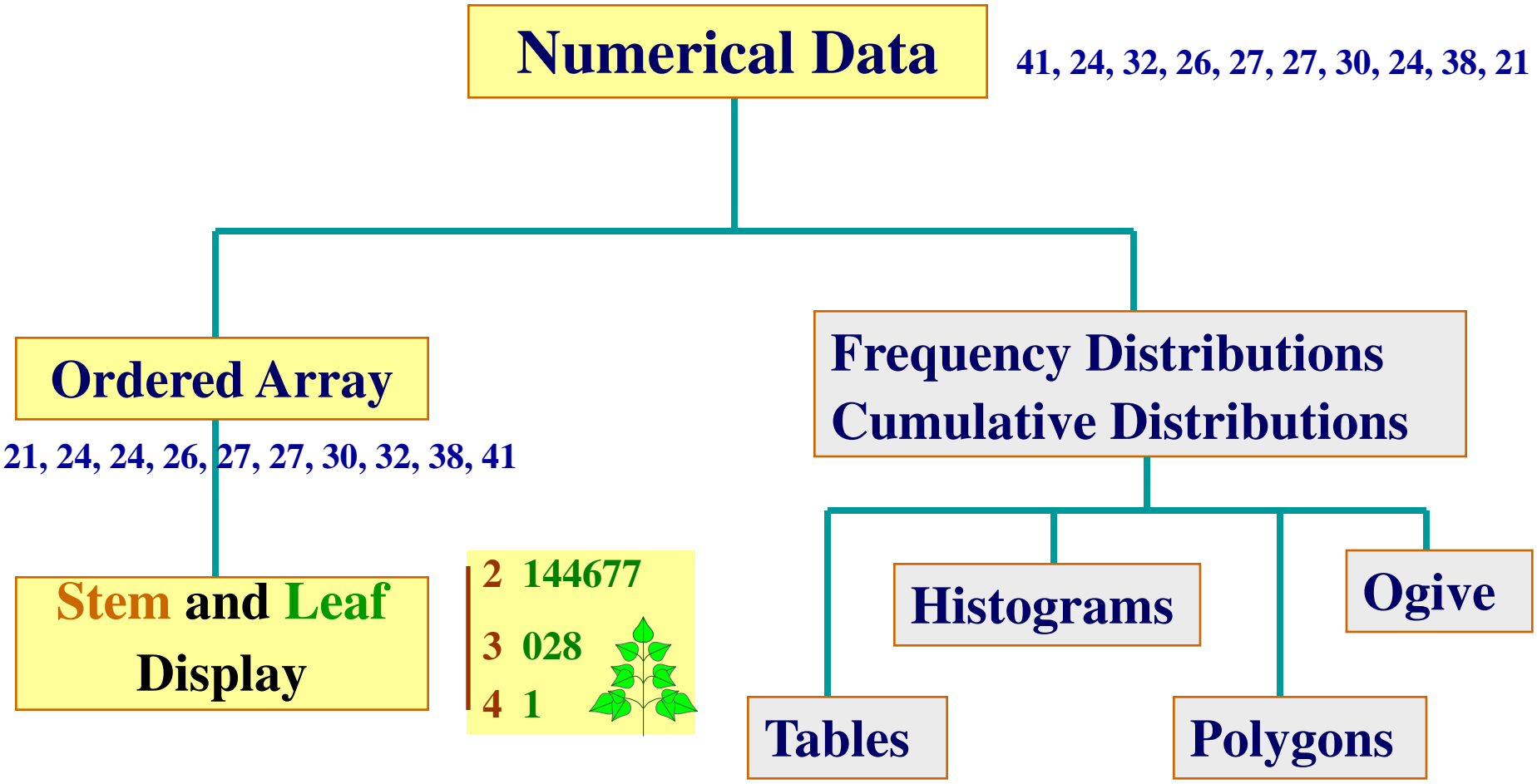
CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

Organizing Numerical Data



Organizing Numerical Data

(continued)

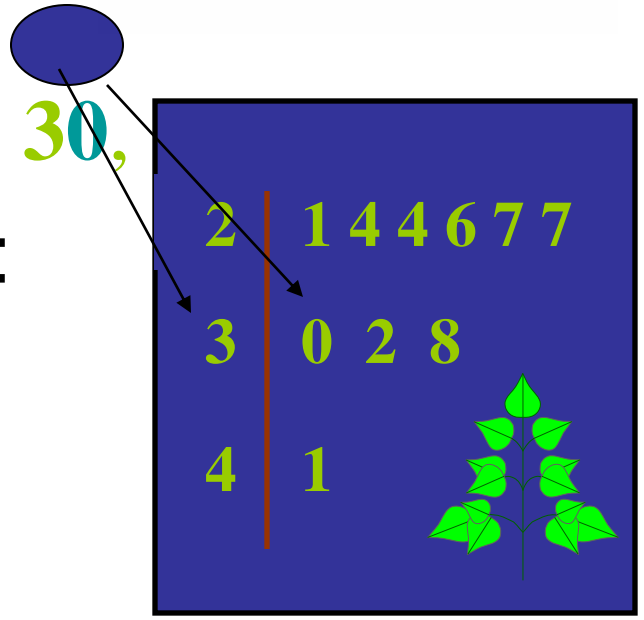
- Data in *Raw* Form (as collected):

  24, 26, 24, 21, 27, 27, 30, 41, 32, 38

- Data in ***Ordered Array*** **from** ***Smallest to Largest*****:**

  21, 24, 24, 26, 27, 27, **30**,

- Stem-and-Leaf Display:



Stem and Leaf Display

- A graphical device that allows one to view the distribution of the data without losing information. The technique of a stem and leaf plot involves separating the data into two parts

Stem and Leaf Example

TABLE 1.1 Unemployment rates by state, December 2000

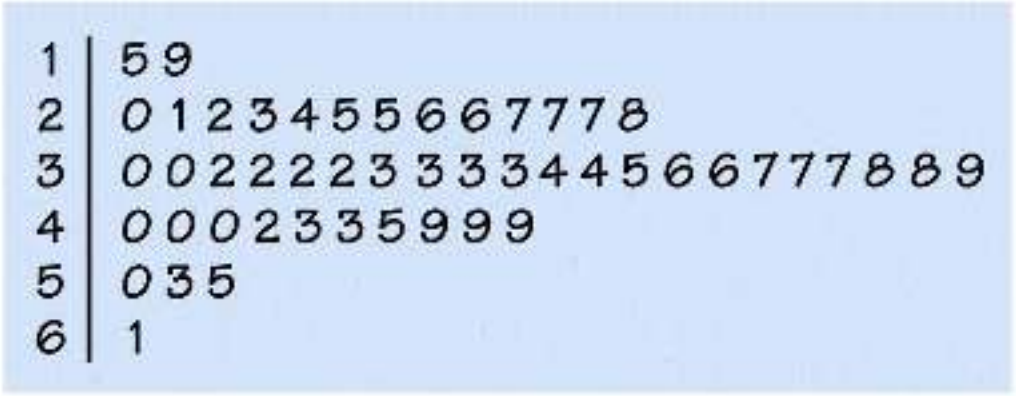
State	Percent	State	Percent	State	Percent
Alabama	4.0	Louisiana	5.3	Ohio	3.7
Alaska	6.1	Maine	2.6	Oklahoma	2.6
Arizona	3.3	Maryland	3.3	Oregon	4.0
Arkansas	3.9	Massachusetts	2.0	Pennsylvania	3.8
California	4.3	Michigan	3.4	Puerto Rico	8.9
Colorado	2.1	Minnesota	2.8	Rhode Island	3.2
Connecticut	1.5	Mississippi	4.3	South Carolina	3.3
Delaware	3.3	Missouri	3.2	South Dakota	2.3
Florida	3.2	Montana	4.9	Tennessee	3.8
Georgia	3.0	Nebraska	2.5	Texas	3.4
Hawaii	3.6	Nevada	4.0	Utah	2.7
Idaho	5.0	New Hampshire	2.2	Vermont	2.4
Illinois	4.5	New Jersey	3.5	Virginia	1.9
Indiana	2.7	New Mexico	4.9	Washington	4.9
Iowa	2.5	New York	4.2	West Virginia	5.5
Kansas	3.2	North Carolina	3.6	Wisconsin	3.0
Kentucky	3.7	North Dakota	2.7	Wyoming	3.7

STEMPLOT

To make a **stemplot:**

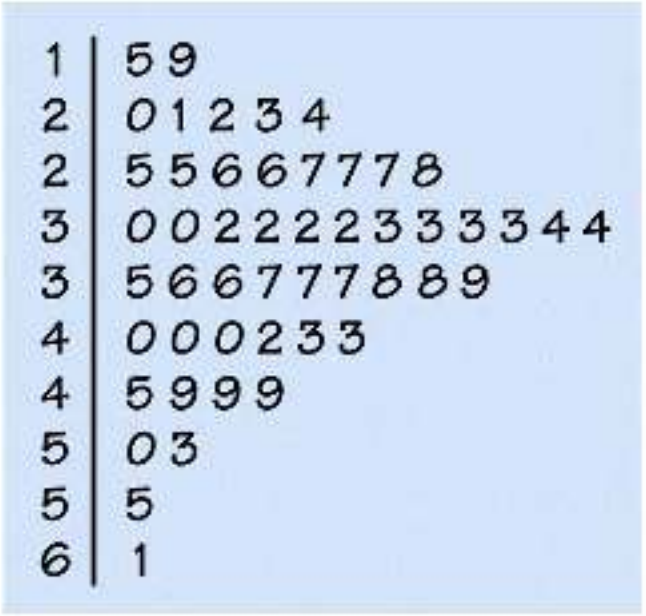
1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf,** the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Stem and Leaf Plot



1	5 9
2	0 1 2 3 4 5 5 6 6 7 7 7 8
3	0 0 2 2 2 2 3 3 3 3 4 4 5 6 6 7 7 7 8 8 9
4	0 0 0 2 3 3 5 9 9 9
5	0 3 5
6	1

(a)



1	5 9
2	0 1 2 3 4
2	5 5 6 6 7 7 7 8
3	0 0 2 2 2 2 3 3 3 3 4 4
3	5 6 6 7 7 7 8 8 9
4	0 0 0 2 3 3
4	5 9 9 9
5	0 3
5	5
6	1

(b)

Tabulating and Graphing Numerical Data

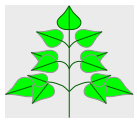
Numerical Data 41, 24, 32, 26, 27, 27, 30, 24, 38, 21

Ordered Array

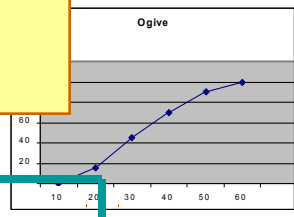
21, 24, 24, 26, 27, 27, 30, 32, 38, 41

Stem and Leaf Display

2	144677
3	028
4	1

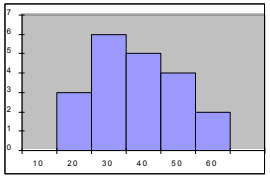


Frequency Distributions
Cumulative Distributions



Ogive

Histograms



Tables

Polygons

Tabulating Numerical Data: Frequency Distributions

- Sort Raw Data in Ascending Order
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find Range: **58 - 12 = 46**
- Select Number of Classes: **5 (usually between 5 and 15)**
- Compute Class Interval (width): **10 (46/5 then round up)**
- Determine Class **Boundaries** (limits): **10, 20, 30, 40, 50, 60**
- Compute Class Midpoints: **15, 25, 35, 45, 55…. (x1+x2)/2**
- Count Observations & Assign to Classes
- The above method has a flaw in it!! For example: Add “8” to the above data set. What do we do?

Frequency Distributions, Relative Frequency Distributions and Percentage Distributions

Data in ordered array:

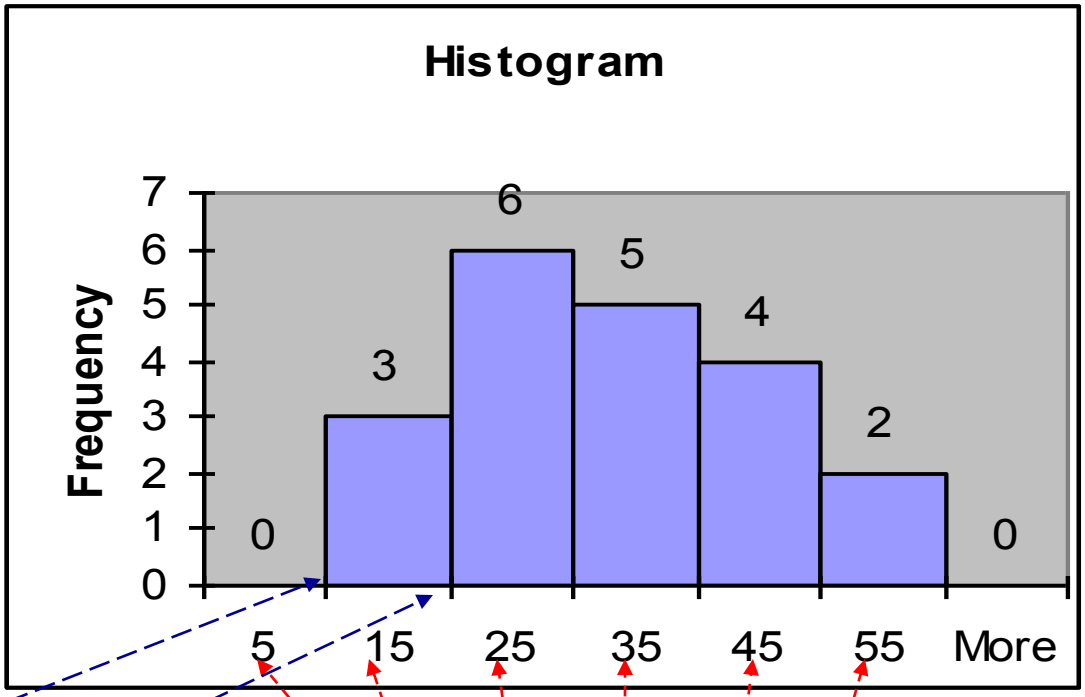
12, 13, 17| 21, 24, 24, 26, 27, 27| 30, 32, 35, 37, 38| 41, 43, 44, 46| 53, 58

Class	Frequency	Relative Frequency	Percentage
10 but under 20	3	.15	15
20 but under 30	6	.30	30
30 but under 40	5	.25	25
40 but under 50	4	.20	20
50 but under 60	2	.10	10
Total	20	1	100

Graphing Numerical Data: The Histogram

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



No Gaps Between Bars

Class Boundaries

Class Midpoints

Tabulating Numerical Data: Cumulative Frequency

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

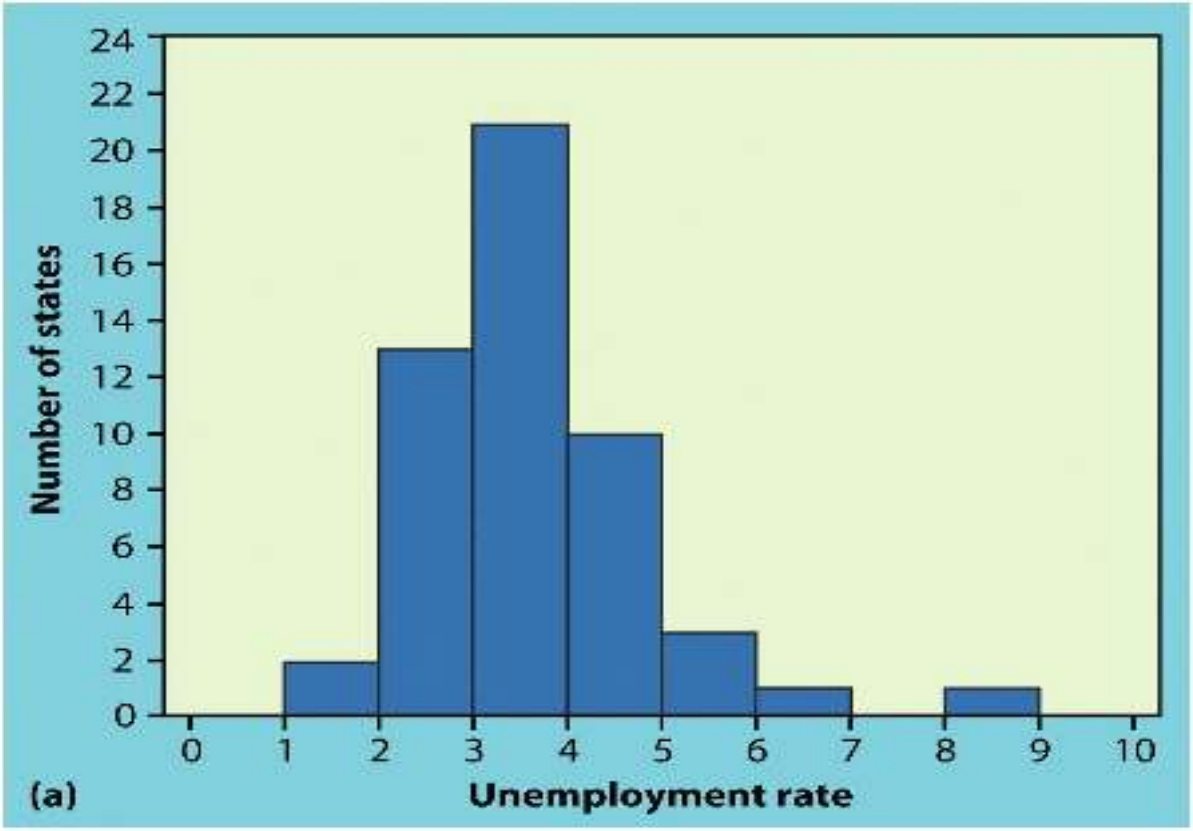
Lower Limit	Cumulative Frequency	Cumulative % Frequency
10	0	0
20	3	15
30	9	45
40	14	70
50	18	90
60	20	100

Example

TABLE 1.1 Unemployment rates by state, December 2000

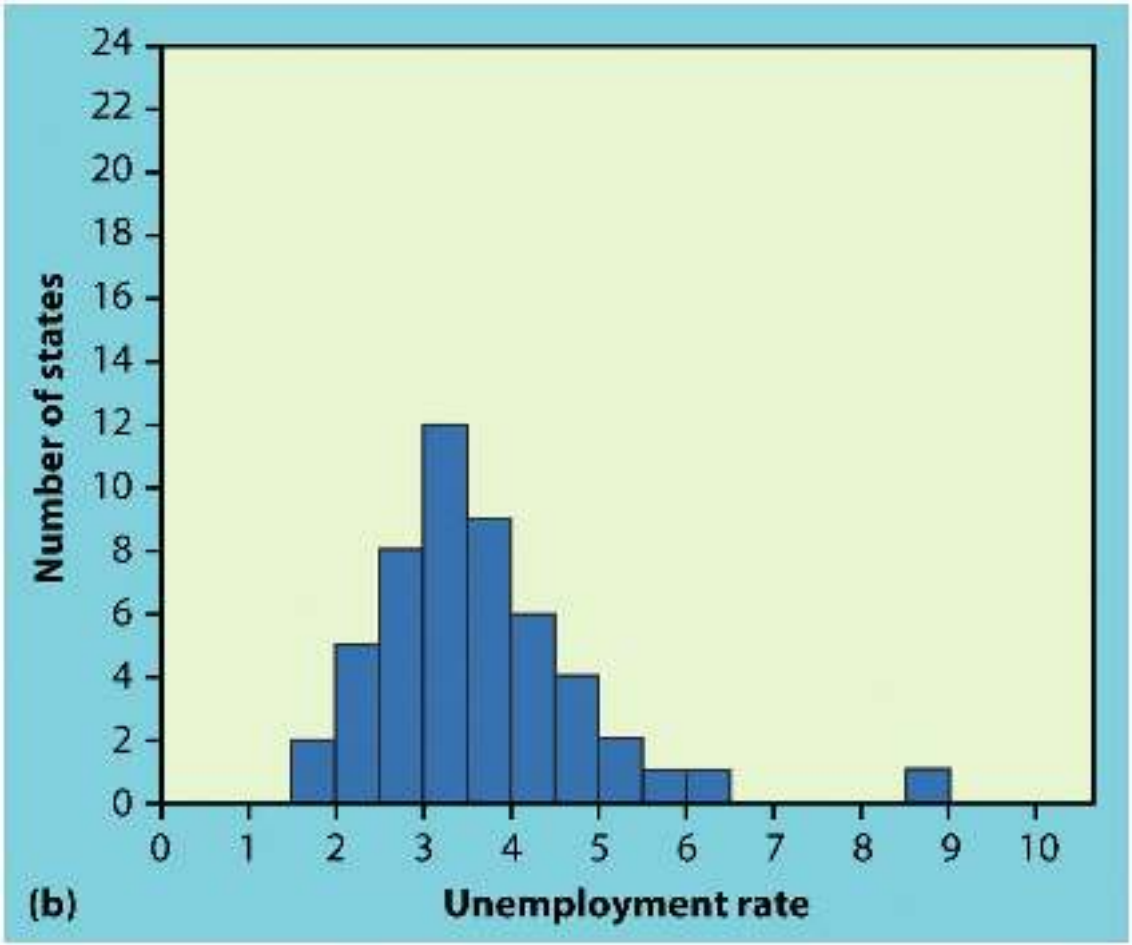
State	Percent	State	Percent	State	Percent
Alabama	4.0	Louisiana	5.3	Ohio	3.7
Alaska	6.1	Maine	2.6	Oklahoma	2.6
Arizona	3.3	Maryland	3.3	Oregon	4.0
Arkansas	3.9	Massachusetts	2.0	Pennsylvania	3.8
California	4.3	Michigan	3.4	Puerto Rico	8.9
Colorado	2.1	Minnesota	2.8	Rhode Island	3.2
Connecticut	1.5	Mississippi	4.3	South Carolina	3.3
Delaware	3.3	Missouri	3.2	South Dakota	2.3
Florida	3.2	Montana	4.9	Tennessee	3.8
Georgia	3.0	Nebraska	2.5	Texas	3.4
Hawaii	3.6	Nevada	4.0	Utah	2.7
Idaho	5.0	New Hampshire	2.2	Vermont	2.4
Illinois	4.5	New Jersey	3.5	Virginia	1.9
Indiana	2.7	New Mexico	4.9	Washington	4.9
Iowa	2.5	New York	4.2	West Virginia	5.5
Kansas	3.2	North Carolina	3.6	Wisconsin	3.0
Kentucky	3.7	North Dakota	2.7	Wyoming	3.7

Histogram



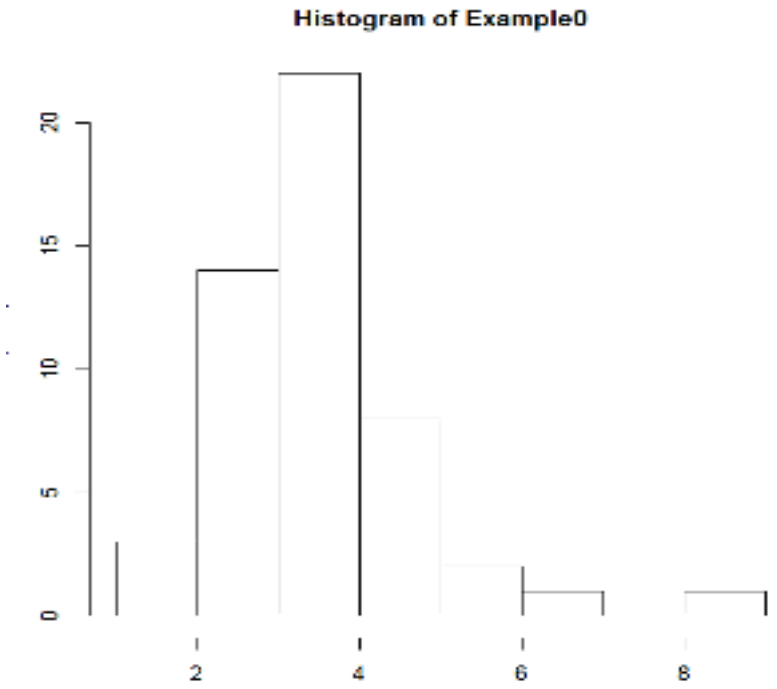
- What do we see?

Same Data, Different Histogram



Histogram in R

- Example 0<- c(4.0, 6.1, 3.3, 3.9, 4.3, 2.1, 1.5, 3.3, 3.2, 3.0, 3.6, 5.0, 4.5, 2.7, 2.5, 3.2,
- 3.7, 5.3, 2.6, 3.3, 2.0, 3.4, 2.8, 4.3, 3.2, 4.9, 2.5, 4.0, 2.2, 3.5, 4.9, 4.2, 3.6, 2.7, 3.7, 2.6, 4.0, 3.8, 8.9, 3.2, 3.3,
- 2.3, 3.8, 3.4, 2.7, 2.4, 1.9, 4.9, 5.5, 3.0, 3.7)
- hist(Example0)



Exporting data outside R

- > dir()
- [1] "~$LungData.xlsx" "Camping-pictures-2013" "CSE502"
- [5] "EuJHumGen2013.pdf" "HalimaCancerGWAS.pptx" "KuwaitiProject" "KuwaitiProject.zip"
- [9] "LungCapData.png" "LungCapData.txt" "LungData.xlsx" "Qatar Bio Bank Project“

- write.csv(Example0, file='dataname.csv')

- >dir()
- "~$LungData.xlsx" "Camping-pictures-2013" "CSE502" **"dataname.csv"**
- [5] "EuJHumGen2013.pdf" "HalimaCancerGWAS.pptx" "KuwaitiProject" "KuwaitiProject.zip"
- [9] "LungCapData.png" "LungCapData.txt" "LungData.xlsx" "Qatar Bio Bank Project"

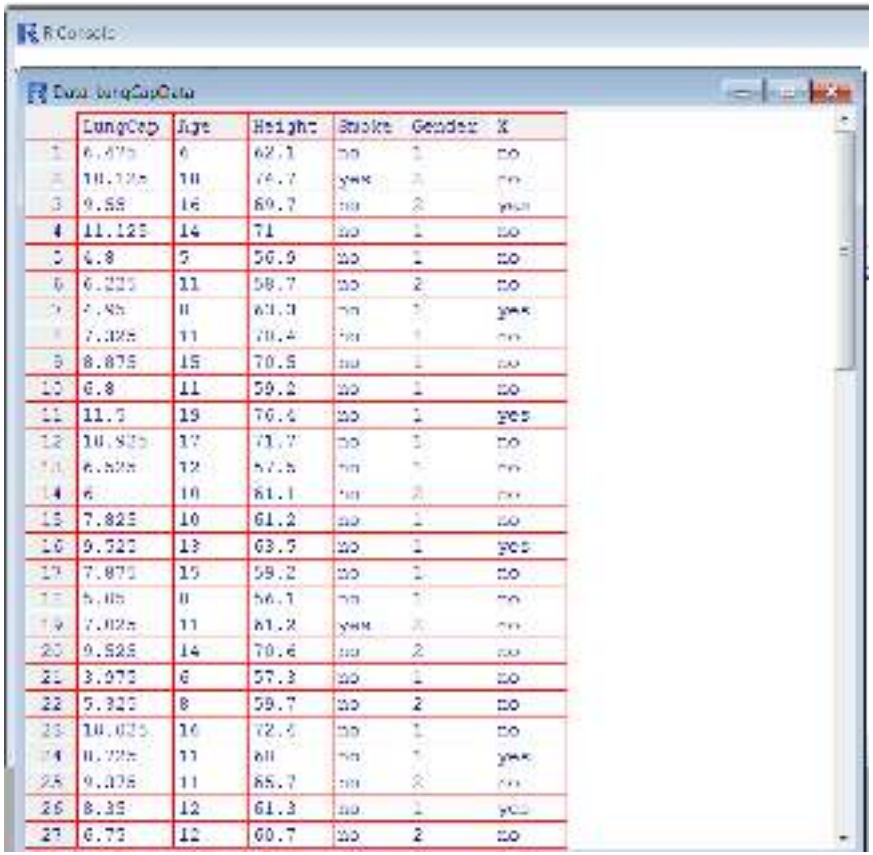
In R

Save LungCapData into your work directory
Change R directory to your work directory

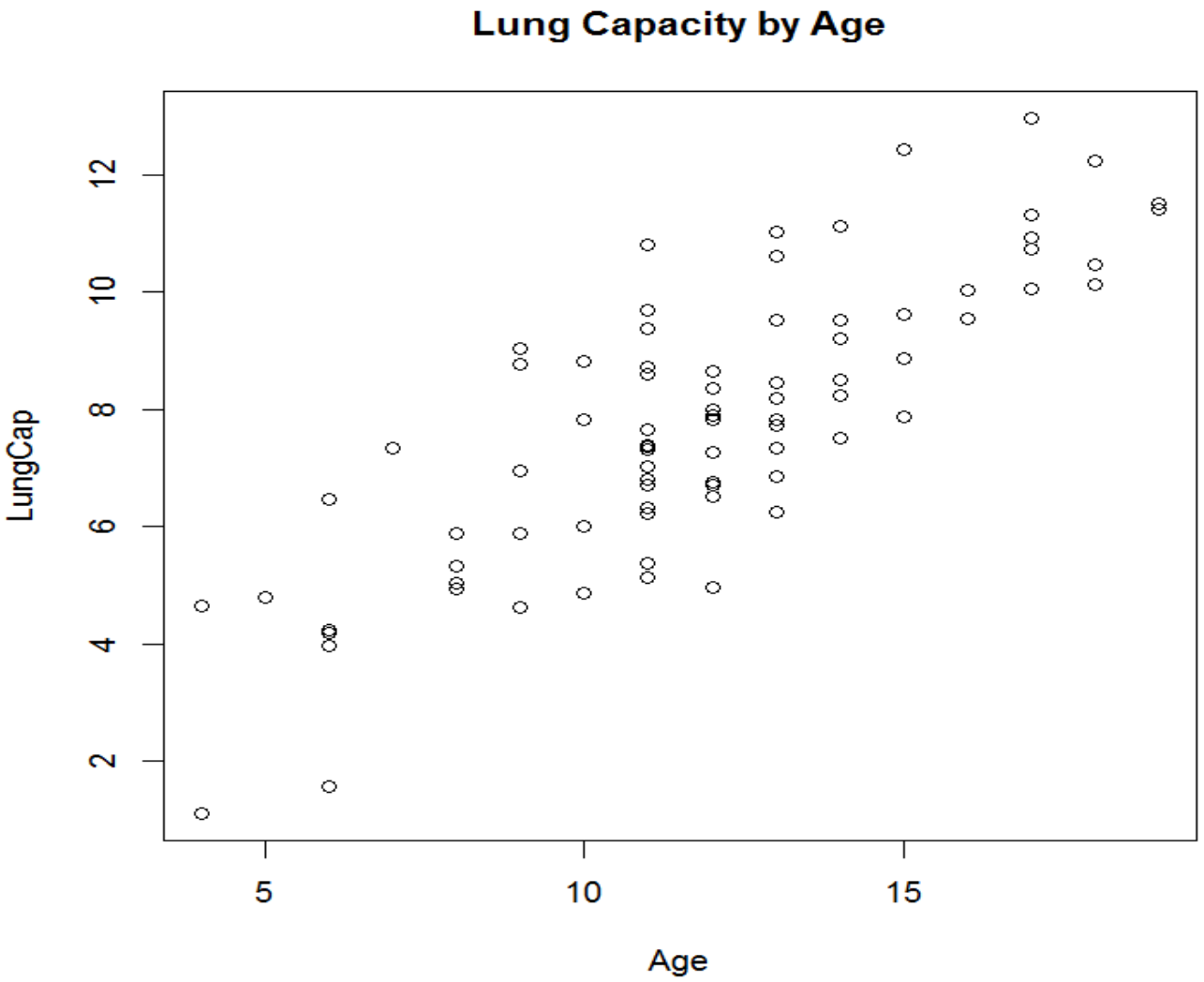
- LungCapData <-read.delim("LungCapData.txt")
- View(LungCapData)
- attach(LungCapData)
- names(LungCapData)
- "LungCap" "Age" "Height" "Smoke" "Gender" "X"
- Plot(Age, LungCap, main="lung Capacity by Age")

Extract LungCap for fenmel only and save in femaleLungCap

- femaleLungCap <-LungCap[Gender==2]
- help(stem)
- ?stem



In R

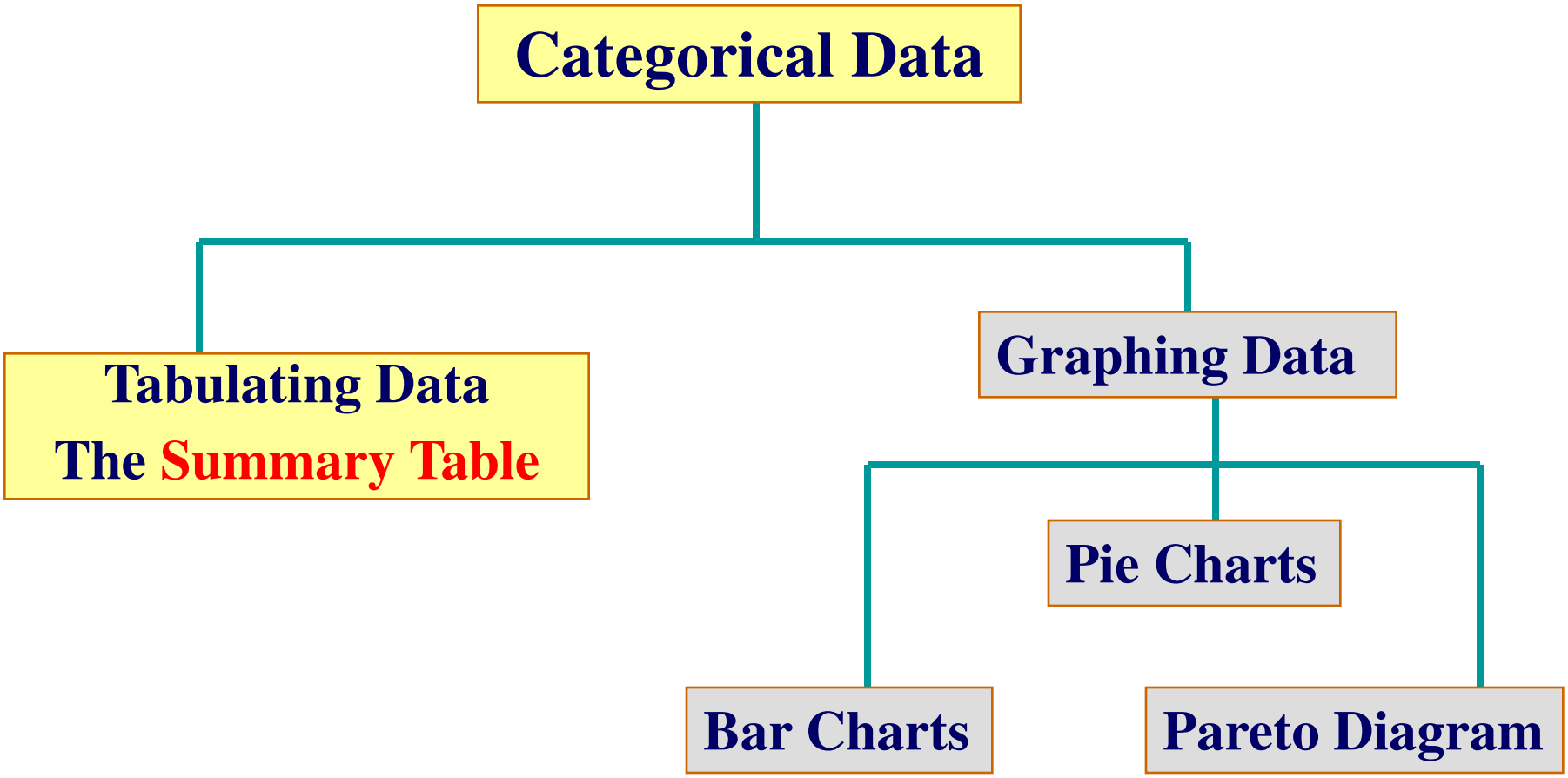


In R

- The decimal point is at the |

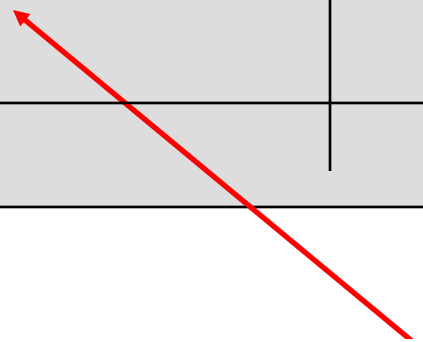
- 1 | 1
- 2 |
- 3 |
- 4 | 67
- 5 | 013499
- 6 | 02337789
- 7 | 00448
- 8 | 0256
- 9 | 456
- 10 | 1157

Tabulating and Graphing Univariate Categorical Data



Summary Table
(for an Investor's Portfolio)

Investment Category	Amount (in thousands $)	Percentage
Stocks	46.5	42.27
Bonds	32	29.09
CD	15.5	14.09
Savings	16	14.55
Total	**110**	**100**



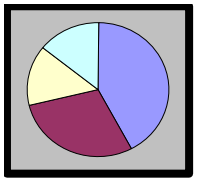
Variables are Categorical.

Graphing Univariate Categorical Data

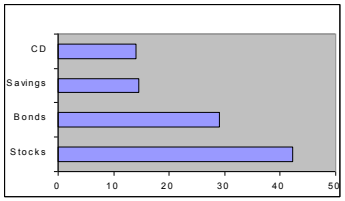
Categorical Data

Tabulating Data
The Summary Table

Graphing Data

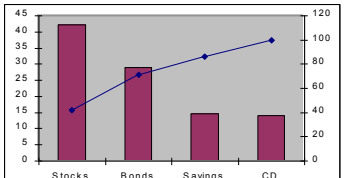


Pie Charts



Bar Charts

Pareto Diagram



Accidents Involving Firestone

Tire Model	Count	Percent
ATX	554	18.7
Firehawk	38	1.3
Firestone	29	1.0
Firestone Wilderness	131	4.4
Radial ATX	48	1.6
Wilderness	1246	42.0
Wilderness AT	709	23.9
Firestone ATX	106	3.6
Wilderness HT	108	3.6

Bar Charts

– Each bar represents a category

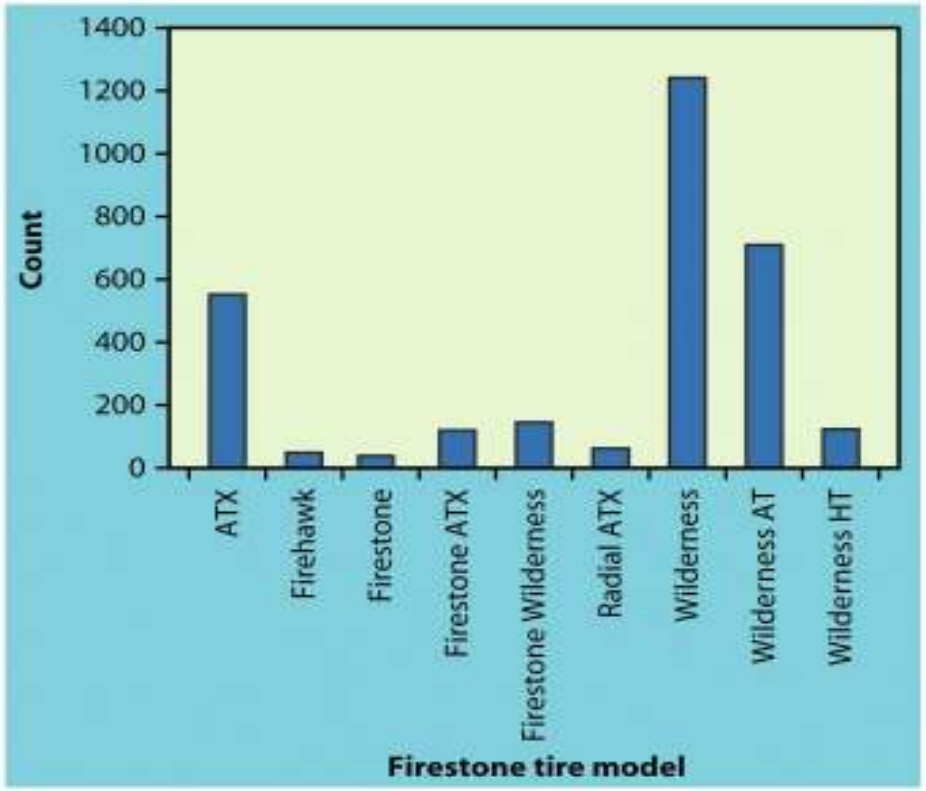
– Height of bar corresponds to percentage or frequency of data in that category.

– Percents don’t have to sum to 100.

– Best for comparing categories.

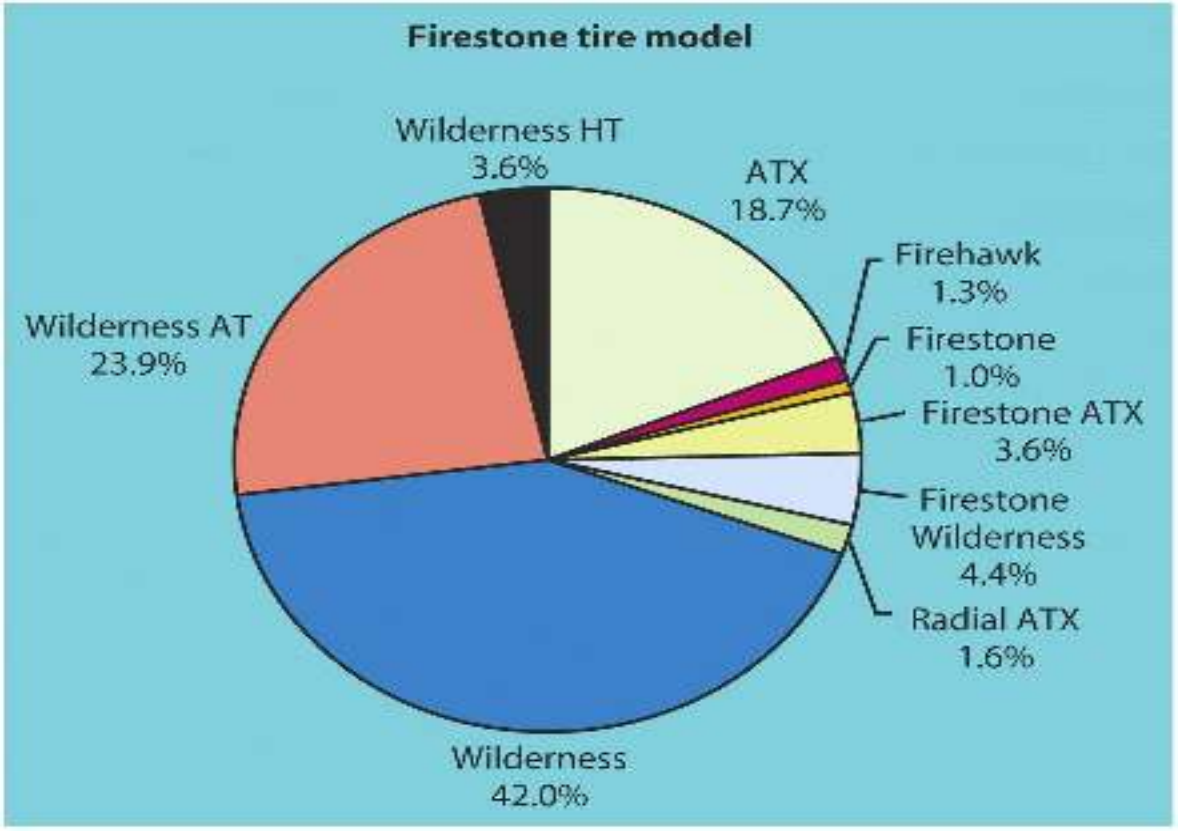
– Pareto charts are bar charts where the bars are arranged in descending order of frequency from left to right. They can identify the “vital few” categories that contains most of the observations

An example for Bar Chart



- A bar graph is a graphical tool that compares the sizes of the response groups. Heights of bars show counts for categories.

An example of Pie Chart



- The Pie Chart is a graphical tool that helps us see what part of the whole each group forms

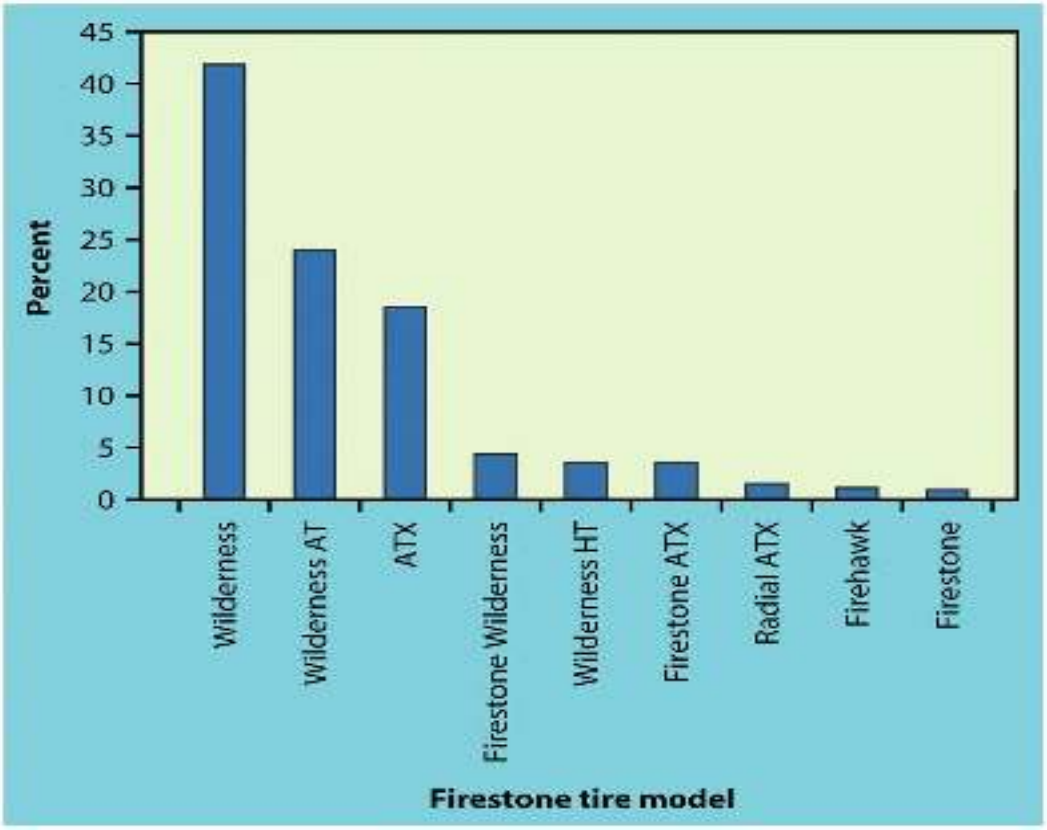
Pie Chart

- Each slice of the pie represents a category

– Size of slices correspond to percentage

– Must sum to 100% – Best for showing percentage of the whole

Pareto Chart Example Distribution of Accidents Involving Firestone Tires



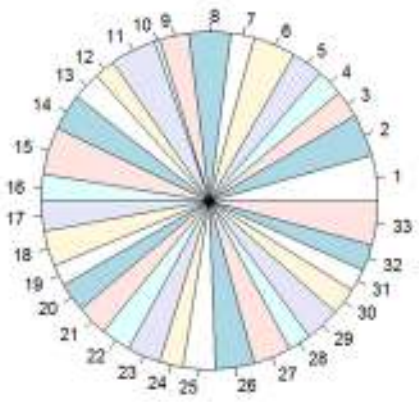
- A Pareto Chart is a Bar Chart whose categories are ordered from most frequent to least frequent (either with counts or percentages)

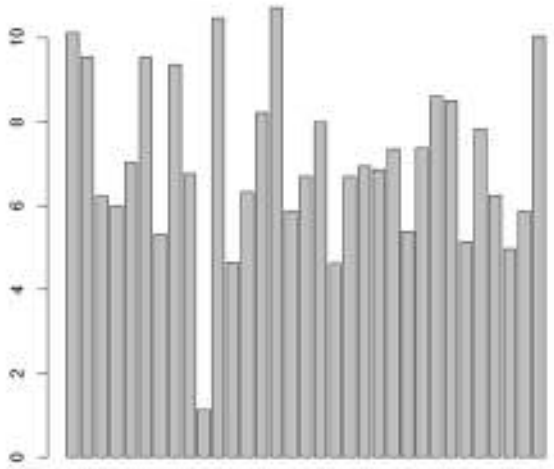
Pie, Bar and Pareto in R

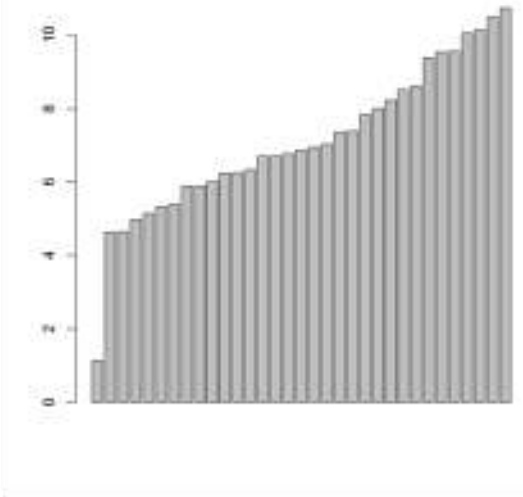
Help(pie)
?(pie)
pie(femaleLungCap)
barplot(femaleLungCap)

For Pareto, try to create a small program to do it in R.

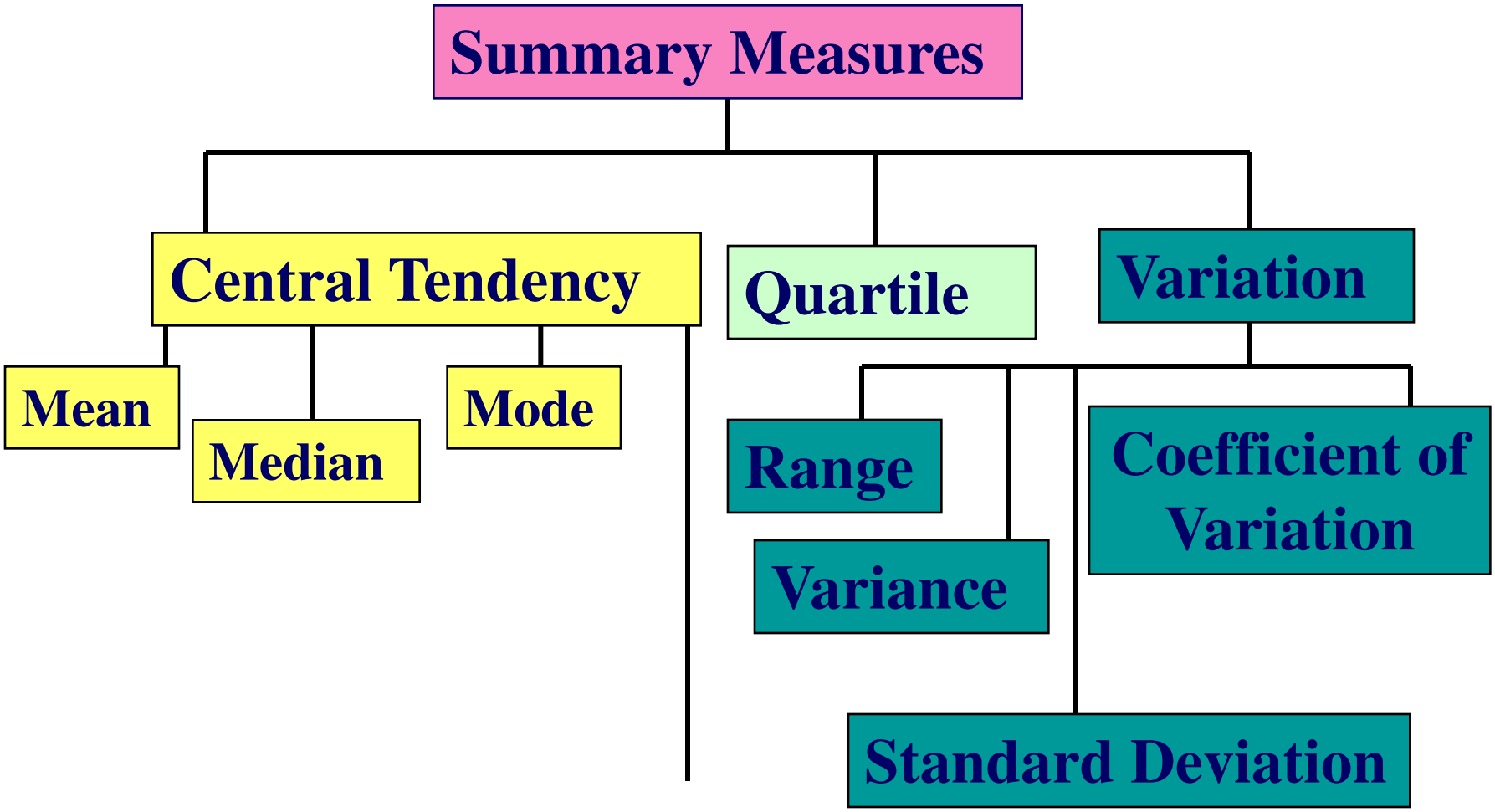
valsort <- femaleLungCap[order(femaleLungCap)]
barplot(valsort)
We can add to this cumulative probabilties



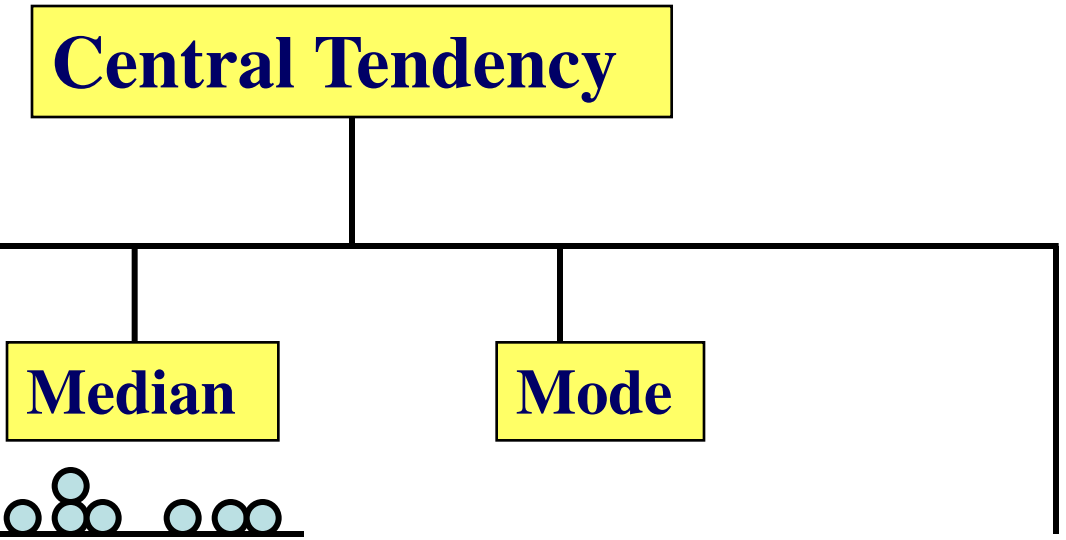




Summary Measures



Measures of Central Tendency



$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Mean (Arithmetic Mean)

- Mean (Arithmetic Mean) of Data Values
  - Sample mean

Sample Size

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

  - Population mean

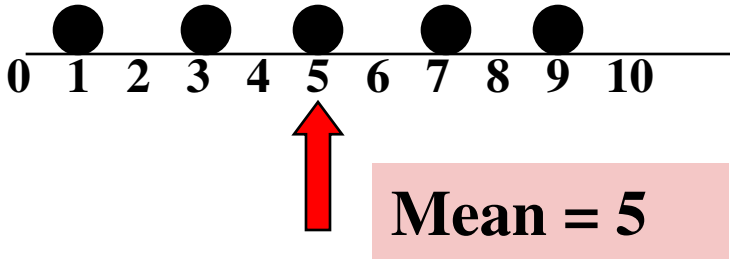
Population Size

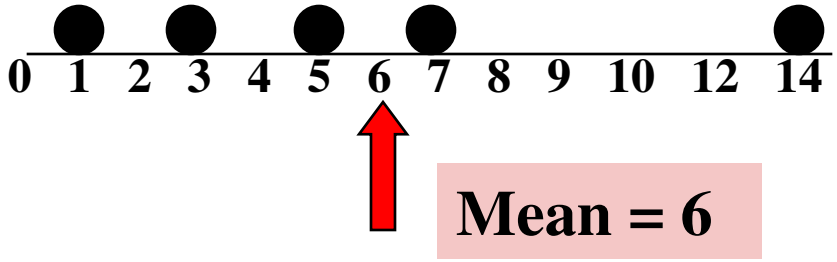
$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

Mean (Arithmetic Mean)

(continued)

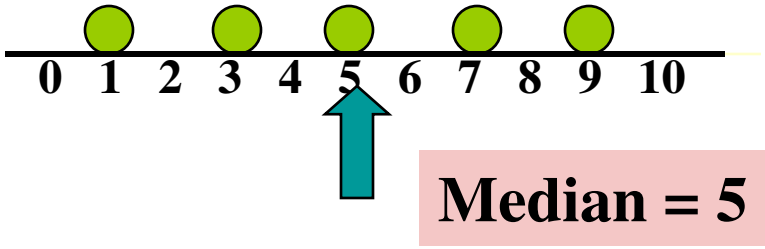
- The Most Common Measure of Central Tendency
- Can Be Affected by Extreme Values (Outliers) (Try -10, 3, 5, 7, 20)

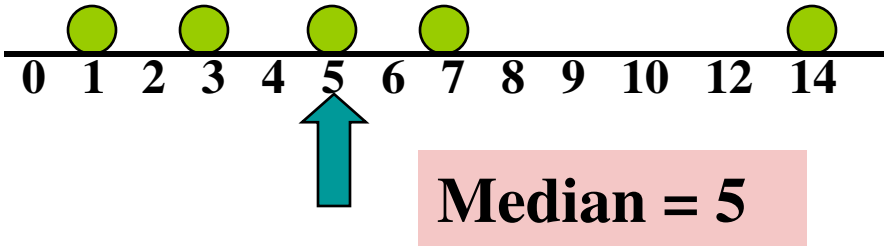




Median

- Robust Measure of Central Tendency
- Not Affected by Extreme Values

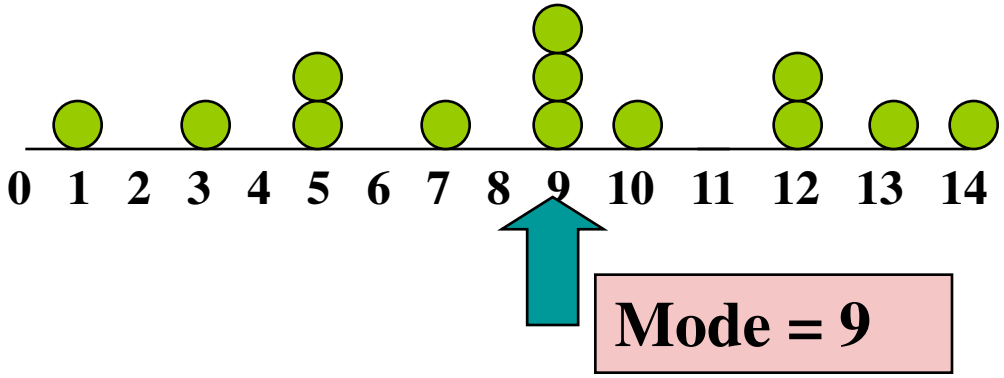


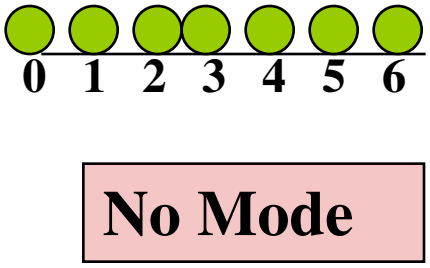


- In an Ordered Array, the Median is the 'Middle' Number
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the 2 middle numbers

Mode

- A Measure of Central Tendency
- Value that Occurs Most Often
- Not Affected by Extreme Values
- There May Not be a Mode
- There May be Several Modes
- Used for Either Numerical or Categorical Data



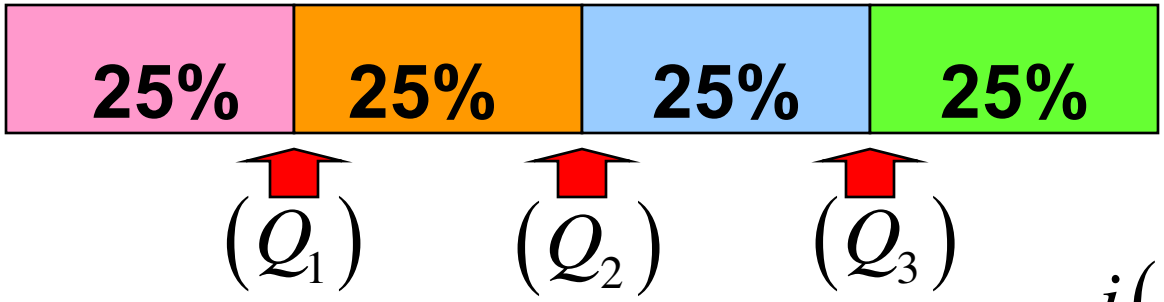


Example

- The following data is the revenue in millions of dollars from 6 different companies:
- 1600, 1157, 937, 800, 707, 700
- Let’s calculate the mean, median, and mode

Quartiles

- Split Ordered Data into 4 Quarters



- Position of i-th Quartile $$(Q_i) = \frac{i(n+1)}{4}$$

Data in Ordered Array: 11 12 13 16 16 17 18 21 22

$$\text{Position of } Q_1 = \frac{1(9+1)}{4} = 2.5 \qquad Q_1 = \frac{(12+13)}{2} = 12.5$$

- Q_1 and Q_3 Are Measures of Noncentral Location
- Q_2 = Median, A Measure of Central Tendency

THE QUARTILES Q_1 and Q_3

To calculate the **quartiles:**

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile** Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The **third quartile** Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

THE FIVE-NUMBER SUMMARY AND BOXPLOTS

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles.
- A line in the box marks the median.
- Lines extend from the box out to the smallest and largest observations.

Boxplots are most useful for side-by-side comparison of several distributions.

Quantile In R using type 8

- Example <- c(11, 12, 13, 16, 16, 17, 18, 21, 22)
- quantile(Example)
- 0% 25% 50% 75% 100%
- 11   13   16   18   22

Quantile in general

- Order the data $x_1 \le x_2 \le \cdots \le x_n$

- For any set of data the median is its middle value when there are an odd number of values; otherwise it is the average of the two middle values when there are an even number of values.
- R's median function calculates this.

- The index of the middle value is $m=(n+1)/2$.

- When m it is not an integer, $(x_l + x_u)/2$ is the median, where l and u are m
- rounded down and up.

- Otherwise when m is an integer, x_m is the median.

- In that case take $l=m-1$ and $u=m+1$
- In either case l is the index of the data value immediately to the left of the median
- and u is the index of the data value immediately to the right of the median.

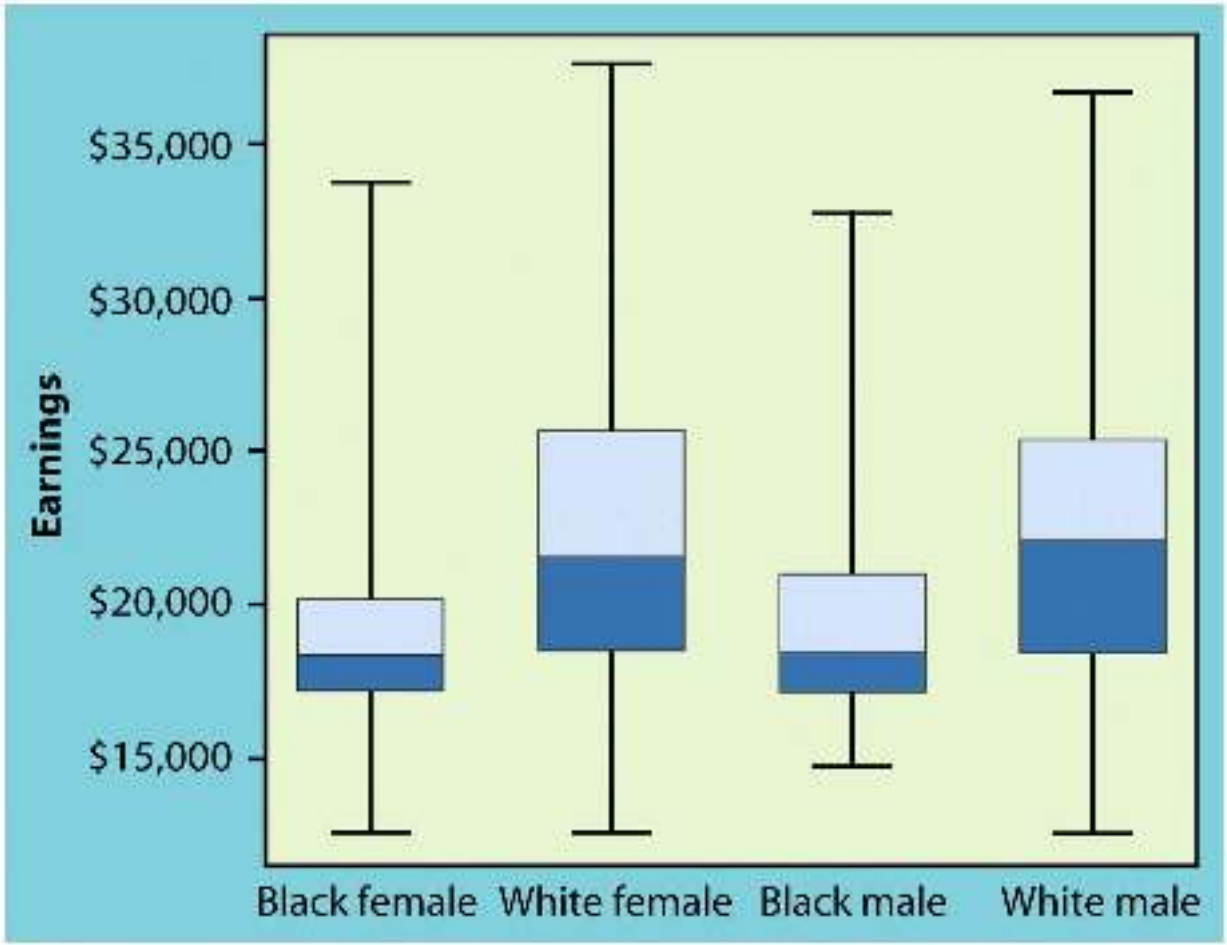
- The "first quartile" is the median of all x_i for which $i \le l$
- The "third quartile" is the median of (x_i) for which $i \ge u$

Implementation in R

```
quart <- function(x) {
  x <- sort(x)
  n <- length(x)
  m <- (n+1)/2
  if (floor(m) != m) {
    l <- m-1/2; u <- m+1/2
                        }
else {
       l <- m-1; u <- m+1
      }
  c(Q1=median(x[1:l]), Q3=median(x[u:n]))
}
```

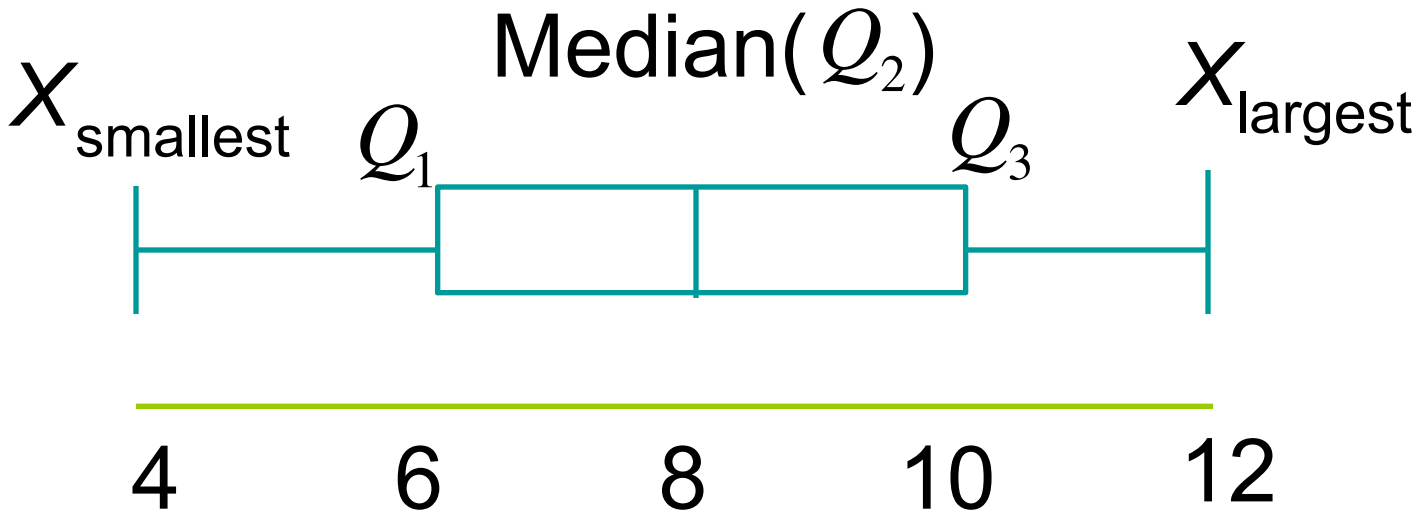
Let's apply this function to the data we have: **11, 12, 13, 16, 16, 17, 18, 21, 22**

The five number summary statistics Box-and-Whisker

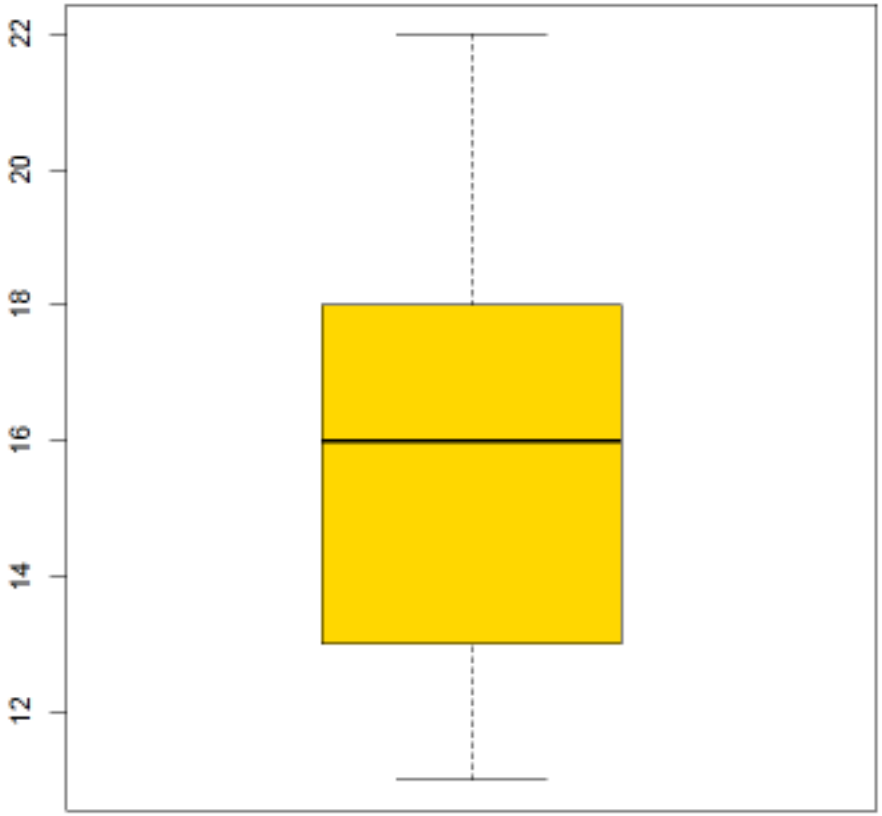


Exploratory Data Analysis

- Box-and-Whisker
  - Graphical display of data using 5-number summary



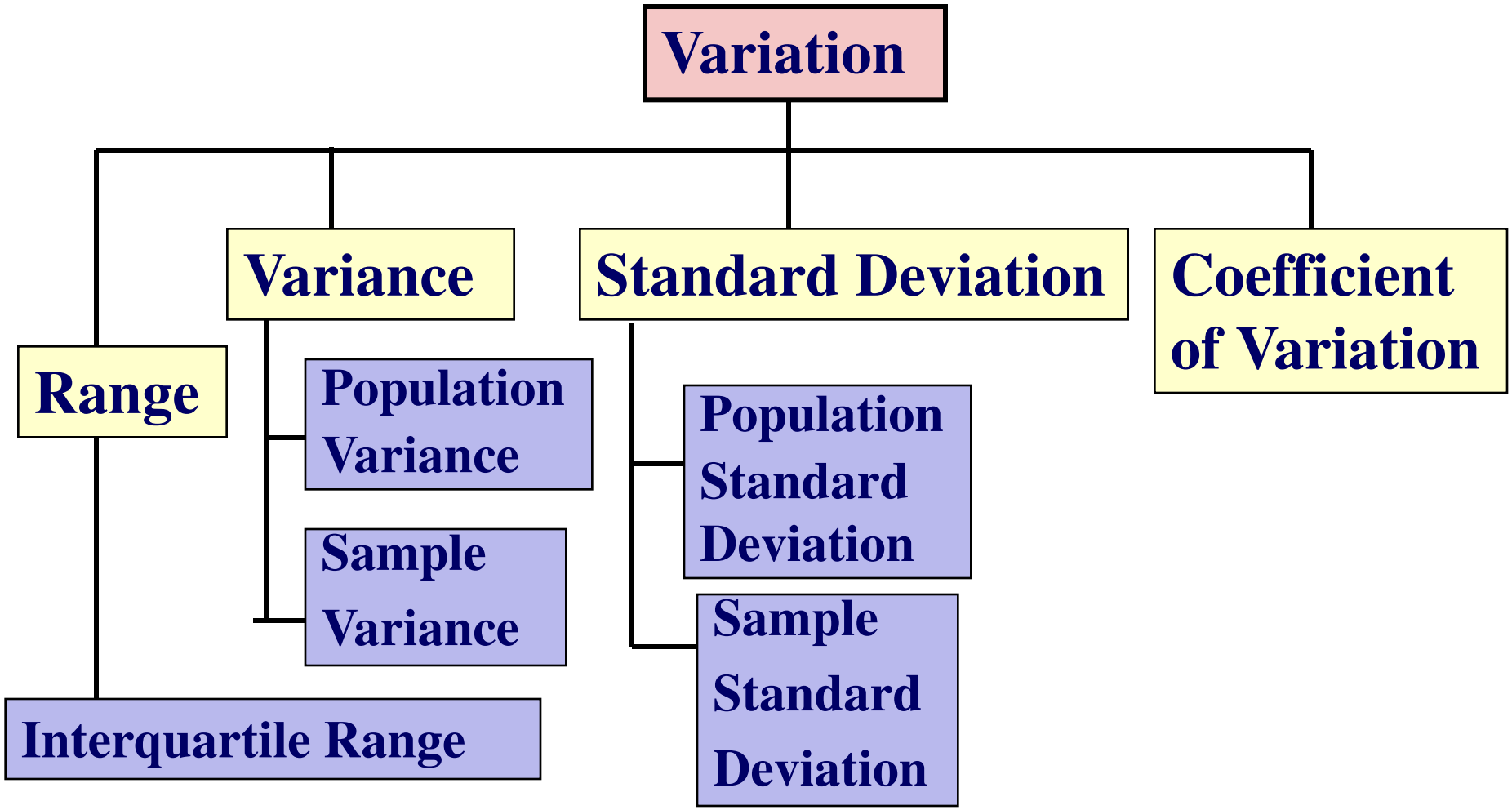
Boxplot in R



Example

- Let’s calculate the 5-number summary for our revenue data
- That is, 1600, 1157, 937, 800, 707, 700

Measures of Variation



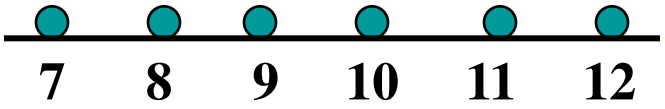
Range

- Measure of Variation
- Difference between the Largest and the Smallest Observations:

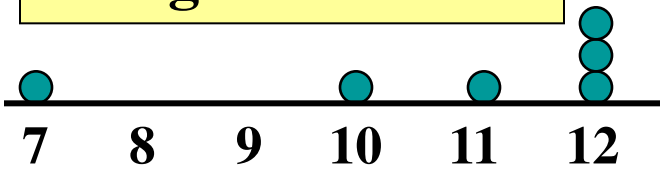
$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$

- Ignores How Data are Distributed

Range = 12 - 7 = 5



Range = 12 - 7 = 5



Interquartile Range

- Measure of Variation
- Also Known as Midspread
  - Spread in the middle 50%
- Difference between the First and Third Quartiles

Data in Ordered Array: 11 12 13 16 16 17 17 18 21

$$\text{Interquartile Range} = Q_3 - Q_1 = 17.5 - 12.5 = 5$$

- Not Affected by Extreme Values

Variance

- Important Measure of Variation
- Shows Variation About the Mean
  - Sample Variance:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{n-1}$$

  - Population Variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(X_i - \mu \right)^2}{N}$$

Standard Deviation

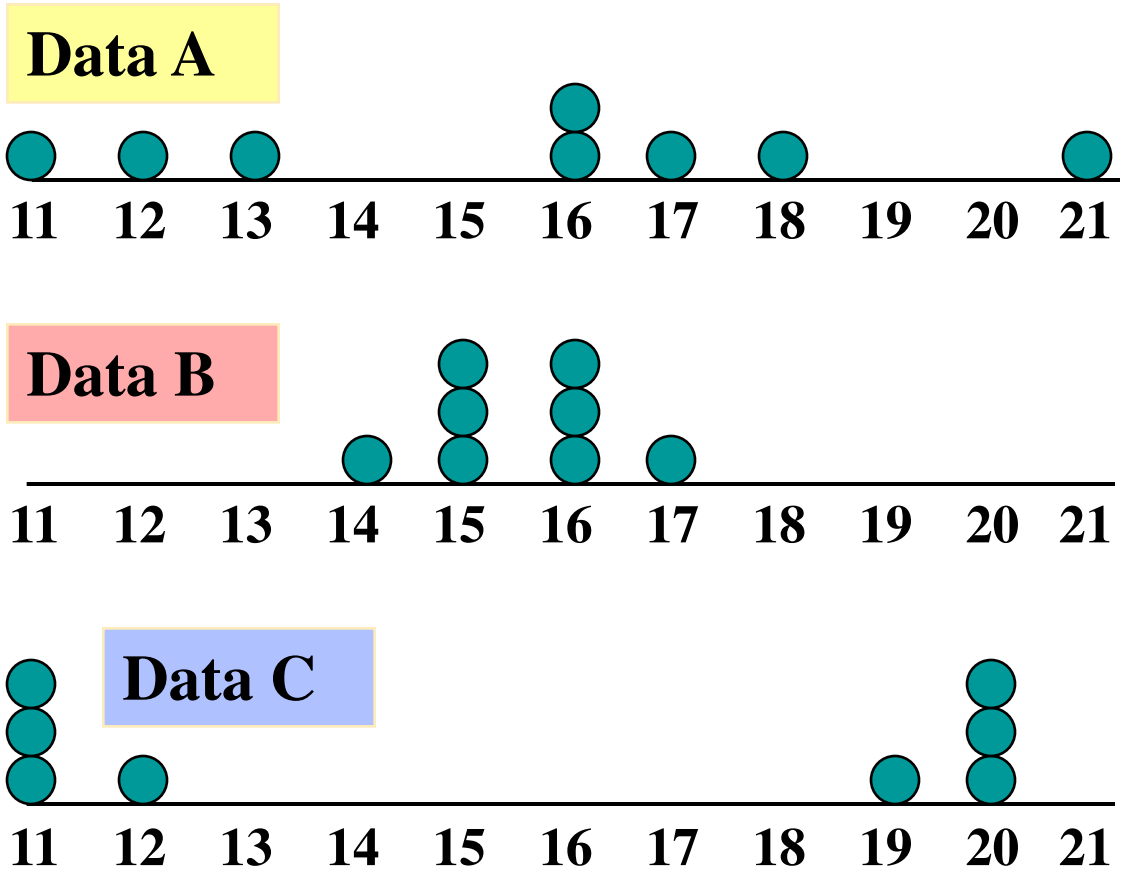
- Most Important Measure of Variation
- Shows Variation about the Mean
- Has the Same Units as the Original Data
  - Sample Standard Deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$$

  - Population Standard Deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \mu\right)^2}{N}}$$

Comparing Standard Deviations



Mean = 15.5
s = 3.338

Mean = 15.5
s = .9258

Mean = 15.5
s = 4.57

Standard deviation in R

- In R, we use “sd” to calculate standard deviation

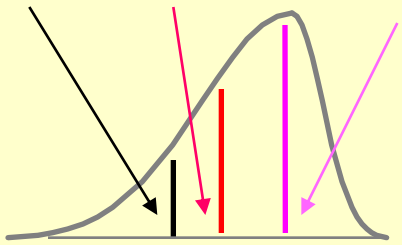
- Example2 <- c(11,12,13,16,16,17,18,21)
- sd(Example2)
- 3.338092

Shape of a Distribution

- Describe How Data are Distributed
- Measures of Shape
  - Symmetric or skewed

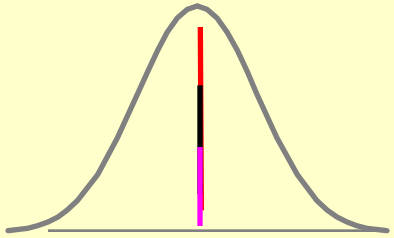
Left-Skewed

Mean < Median < Mode



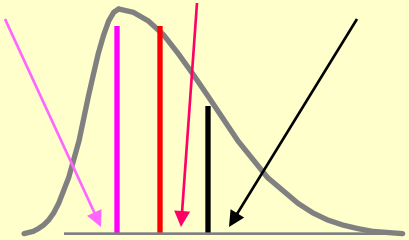
Symmetric

Mean = Median =Mode



Right-Skewed

Mode < Median < Mean



Example

- Let’s calculate the range, variance, and standard deviation for our revenue data
- That is, 1600, 1157, 937, 800, 707, 700

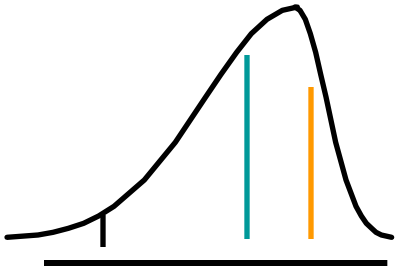
SYMMETRIC AND SKEWED DISTRIBUTIONS

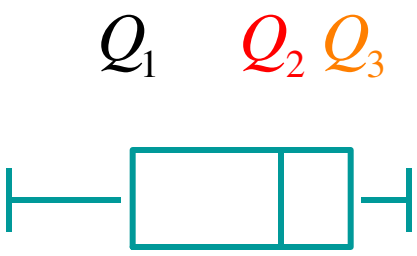
A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

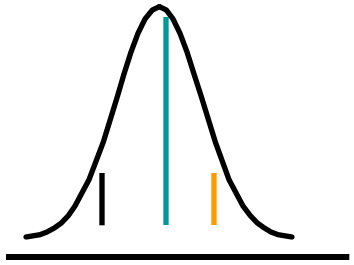
Distribution Shape & Box-and-Whisker

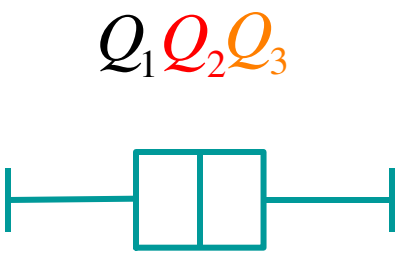
Left-Skewed



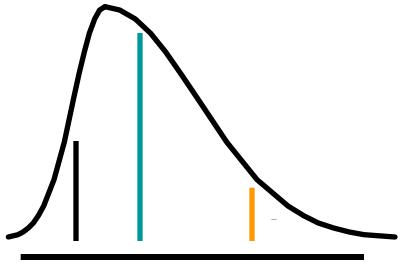


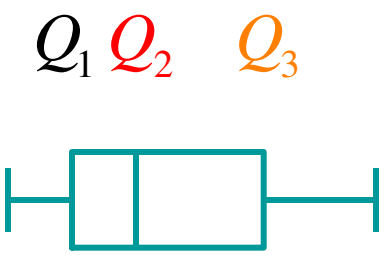
Symmetric





Right-Skewed





Numerical summary in JMP

- JMP Starter →Open Data Table → open data file or:
- JMP Starter → New Data Table → then enter your data
- select Analyze → select Distribution → input Y columns → click O.K.