Math 115: Elementary Statistics

Rong You

2019-08-12

Contents

Chapter 1

Chapter 1

1.1 Welcome

- Always check Canvas for everything (homework assignments, announcements, grades, and etc..)
- Please send me emails directly using your avc email. Don't send me emails through the Canvas system because Canvas doesn't handle attachments well.
- Syllabus
- Questions?

1.2 Statistical and Critical Thinking

- *Statistics*: The science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data. What does statistics do?
- *Data*: Collections of observations, such as measurements, genders, or survey responses
- *Population*: The complete collection of **all** measurements or data that are being considered. Typically, a population is the complete collection of data that we would like to make inferences about.
- *Census*:The collection of data from every member of a population
- *Sample*: A subcollection of members selected from a population
- *Survey*: The study of samples

1.3 Example 1

The student senate at a university with 15,000 students is interested in the proportion of students who favor a change in the grading system to allow for plus and minus grades (e.g. B+, B, B-, rather than just B). Two hundred students are interviewed to determine their attitude toward this proposed change. What is the population of interest? What group of students constitutes the sample in this problem?

- Population: 15,000 students at that University
- Sample: 200 students are interviewed at that University

1.4 Example 2

In the journal article "Residential Carbon Monoxide Detector Failure Rates in the United States", it was stated that there are 38 million carbon monoxide detectors installed in the United States. When 30 of them were randomly selected and tested, it was found that 12 of them failed to provide an alarm in hazardous carbon monoxide conditions.

- Population: All 38 million carbon monoxide detectors in the United States
- Sample: The 30 carbon monoxide detectors that were selected and tested

1.5 Statistical Thinking Procedure:

- Prepare (the goal of the study, collecting data, what are the data about, cleaning the data)
- Analyze (graph the data, explore the data, apply more statistical methods)
- Conclude (what conclusion can you make? Do they have statistical/pratical significant?).

Section of Plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel System	1.55
Rest of the plane	1.8

-

- Source: *How not to Be Wrong: The Power of Mathematical Thinking*- by Jordan Ellenberg

- Statistical significance is achieved in a study when we get a result that is very unlikely to occur by chance.

1.6 Types of Data

- *Parameter*: a numerical measurement describing some characteristic of a population. (population mean, standard deviation, variance, etc)
- *Statistic*: a numerical measurement describing some characteristic of a sample. (Sample mean, standard deviation, variance, etc)
- *Categorical (or qualitative or attribute) data*: consists of names or labels (representing categories).
- *Quantitative (or numerical) data*: consists of numbers representing counts or measurements.
  - *Discrete data*: result when the number of possible values is either a finite number or a countable number
  - *Continuous (numerical) data*: result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

1.7 Example 3

Classify each of the following attributes as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous

- the appraised value of homes in Frisco
- the color of cars in the teacher's lot
- the number of calculators owned by students at your school
- the zip code of an individual
- the amount of time it takes students to drive to school

- The actual weight of a 1-lb can
- Your height
- Your Weight
- Amount of rain drops in Antelope Valley

1.8 Levels of Measurement

- *Nominal level of measurement*: characterized by data that consist of names, labels, or categories only, and the data cannot be arranged in some order (such as low to high).
  - Survey responses of yes, no, and undecided
- *Ordinal level of measurement*: involves data that can be arranged in some order, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.
  - Course grades A, B, C, D, or F
- *Interval level of measurement*: involves data that can be arranged in order, and the differences between data values can be found and are meaningful. However, there is no natural zero starting point at which none of the quantity is present.
  - Example: Years 1000, 2000, 1776, and 1492
- *Ratio level of measurement*: data can be arranged in order, differences can be found and are meaningful, and there is a natural zero starting point (where zero indicates that none of the quantity is present). Differences and ratios are both meaningful.
  - Example: Class times of 50 minutes and 100 minutes

1.9 Summary - Levels of Measurement

- Nominal: categories only
- Ordinal: categories with some order
- Interval: differences but no natural zero point
- Ratio: differences and a natural zero point

1.10 Big Data and Data Science

Learn more about computer science and statistical programming (Python, SAS, R, STATA). Using a calculator is not enough.

1.11 How do we gather data?

- Census
- Surveys

- Opinion polls
- Interviews
- Studies
  - Observational
    - Retrospective (past)
    - Prospective (future)
- Experiments

1.12 How good a poll would be?

[Daily Show-Poll Bearers]

1.13 Basics of Collecting Data

- *Observational study*: observing and measuring specific characteristics without attempting to modify the subjects being studied.
- *Experiment*: apply some treatment and then observe its effects on the subjects (subjects in experiments are called experimental units)

1.14 Observational Study: Survey

- Sample: A part of the population that we actually examine in order to gather information. We Use sample to generalize to population.

1.15 Sampling Methods:

Let's take a look at a simple dataset:

```
##    Students Sex class GPA Honors
## 1     Alice   F    Fr 3.8    Yes
## 2      Brad   M    Fr 2.6    Yes
## 3     Caleb   M    Fr 2.2     No
## 4     Daisy   F    Fr 2.1     No
## 5      Faye   F    Fr 2.0     No
## 6       Eva   F    Fr 1.8     No
## 7     Georg   M    Fr 1.4     No
## 8    Andrea   F    So 4.0    Yes
```

```
## 9     Betsy   F    So 4.0    Yes
## 10    Chris   M    So 4.0    Yes
## 11    Dylan   M    So 3.5    Yes
## 12   Felipe   M    So 3.0     No
## 13     Eric   M    So 2.1     No
## 14  Gabriel   M    So 2.0     No
## 15     Adam   M    Jr 4.0    Yes
## 16 Brittany   F    Jr 3.9     No
## 17   Cassie   F    Jr 3.8    Yes
## 18    Derek   M    Jr 3.1    Yes
## 19    Faith   F    Jr 2.5    Yes
## 20   Eliott   M    Jr 1.9     No
## 21    Garth   M    Jr 1.1     No
## 22   Angela   F    Sr 4.0    Yes
## 23      Bob   M    Sr 3.8    Yes
## 24     Carl   M    Sr 3.1     No
## 25    Diana   F    Sr 2.9     No
## 26    Frank   M    Sr 2.0     No
## 27       Ed   M    Sr 1.5     No
## 28    Grace   F    Sr 1.4     No
```

The population mean GPA is $\mu = 2.766$

1.16 Sampling Methods: Simple Random Sample

> Simple Random Sample (SRS): consist of n individuals from the population chosen in such a way that every individual has an equal chance of being selected and every set of n individuals has an equal chance of being selected

Let's use SRS to take 7 samples:

```
##    Students Sex class GPA Honors
## 10    Chris   M    So 4.0    Yes
## 2      Brad   M    Fr 2.6    Yes
## 16 Brittany   F    Jr 3.9     No
## 20   Eliott   M    Jr 1.9     No
## 21    Garth   M    Jr 1.1     No
## 23      Bob   M    Sr 3.8    Yes
## 13     Eric   M    So 2.1     No
```

The sample mean using SRS, which is $\bar{x}_{srs} = 2.77$

- Strengths
  - The selection of one element does not affect the selection of others.
  - Each possible sample, of a given size, has an equal chance of being selected.
  - Simple random samples tend to be good representations of the population.
  - Requires little knowledge of the population.
- Weaknesses
  - If there are small subgroups within the population, a SRS may not give an accurate representation of that subgroup. In fact, it may not include it at all! This is especially true if the sample size is small.
  - If the population is large and widely dispersed, it can be costly (both in time and money) to collect the data.

1.17 Samping Methods: Systematic Sampling

In a systematic sample, the members of the population are put in a row. Then 1 out of every k members are selected. The starting point is randomly chosen from the first k elements and then elements are sampled at the same location in each of the subsequent segments of size k. The key word is every…

```
## [1] 4
```

So, we will start with element 4, which is Daisy and choose every 4th element after that for our sample.

The mean GPA of the systematic sample, the sample mean, $\bar{x}_{sys}$, is 2.7714286.

- Strengths
  - Assures an even, random sampling of the population.
  - When the population is an ordered list, a systematic sample gives a better representation of the population than a SRS.
  - Can be used in situations where a SRS is difficult or impossible. It is especially useful when the population that you are studying is arranged in time.
- Weaknesses

  - Not every combination has an equal chance of being selected. Many combinations will never be selected using a systematic sample!
  - Large Variance.
  - Formulas are really complicated

1.18 Sampling Methods: Stratified random sample

Stratified Sampling

In a stratified sample, the population must first be separated into homogeneous groups, or strata. Each element only belongs to one stratum and the stratum consist of elements that are alike in some way. A simple random sample is then drawn from each stratum, which is combined to make the stratified sample.

```
##    Students Sex class GPA Honors
## 1     Alice   F    Fr 3.8    Yes
## 2      Brad   M    Fr 2.6    Yes
## 8    Andrea   F    So 4.0    Yes
## 9     Betsy   F    So 4.0    Yes
## 10    Chris   M    So 4.0    Yes
## 11    Dylan   M    So 3.5    Yes
## 15     Adam   M    Jr 4.0    Yes
## 17   Cassie   F    Jr 3.8    Yes
## 18    Derek   M    Jr 3.1    Yes
## 19    Faith   F    Jr 2.5    Yes
## 22   Angela   F    Sr 4.0    Yes
## 23      Bob   M    Sr 3.8    Yes
```

we take a SRS of size 3 from the Honors students:

```
##    Students Sex class GPA Honors
## 18    Derek   M    Jr 3.1    Yes
## 19    Faith   F    Jr 2.5    Yes
## 23      Bob   M    Sr 3.8    Yes
```

The sample mean for the stratified sample, $\bar{x}_{strat} = 3.13$

- Strengths:
  - Representative of the population, because elements from all strata are included in the sample.
  - Ensures that specific groups are represented, sometimes even proportionally, in the sample.

  - Since each stratified sample will be distributed similarly, the amount of variability between samples is decreased.
  - Allows comparisons to be made between strata, if necessary. For example, a stratified sample allows you to easily compare the mean GPA of Honors students to the mean GPA of non-Honors students.

- Weaknesses

  - Requires prior knowledge of the population. You have to know something about the population to be able to split into strata!

1.19 Sampling Methods: Cluster Sampling

Cluster Sampling

Cluster sampling is a sampling method used when natural groups are evident in the population. The clusters should all be similar each other: each cluster should be a small scale representation of the population. To take a cluster sample, a random sample of the clusters is chosen. All the elements of the randomly chosen clusters make up the sample.

Note: There are a couple of differences between stratified and cluster sampling.

In a stratified sample, the differences between stratum are high while the differences within strata are low. In a cluster sample, the differences between clusters are low while the differences within clusters are high.

In a stratified sample, a simple random sample is chosen from each stratum. So, all of the stratum are represented, but not all of the elements in each stratum are in the sample . In a cluster sample, a simple random sample of clusters is chosen. So, not all of the clusters are represented, but all elements from the chosen clusters are in the sample.

Let's take a cluster sample using the grade level (freshmen, sophomore, junior, senior) of FakeSchool as the clusters. Let's take a random sample of 2 of them.

The sample mean for the clustered sample, $\bar{x}_{clust}$, is 2.7807143.

- Strengths:

  - Makes it possible to sample if there is no list of the entire population, but there is a list of subpopulations. For example, there is not a list of all church members in the United States. However, there is a list of churches that you could sample and then acquire the members list from each of the selected churches.

- Weaknesses:

  - Not always representative of the population. Elements within clusters tend to be similar to one another based on some characteristic(s). This can lead to over-representation or under-representation of those characteristics in the sample.

1.20 Sampling Methods: Convenience sampling

Convenience sampling: Ask people who are easy to ask, friendly, or interviewer is comfortable asking. Produces bias results (Don't use).

1.21 Summary of Sampling Methods

- Simple Random Sample: consists of n individuals from the population chosen in such a way that every individual has an equal chance of being selected and every set of n individuals has an equal chance of being selected.
- Stratified random sample: population is divided into homogeneous groups called strata, then SRS's are pulled from each strata
- Systematic random sample: select sample by following a systematic approach (e.g. every 50th), after randomly selecting where to begin
- Cluster sample: based upon location; randomly pick a location, then sample ALL in that location
- Multistage Sampling: Collect data by using some combination of the basic sampling methods. In a multistage sample design, pollsters select a sample in different stages, and each stage might use different methods of sampling. (will not be on test)
- Convience Sampling: Bad sampling method.

1.22 Example 4

- Identify the sampling design:
  - Modern Managed Hospitals (MMH) is a national for-profit chain of hospitals. Management wants to survey patients discharged this past year to obtain patient satisfaction profiles. They wish to use a sample of such patients. Several sampling techniques are described below. Categorize each as simple random, stratified, systematic, cluster, or convenience sample.
  - Obtain a list of patients discharged from all MMH facilities. Divide the patients according to length of hospital stay (3 days or less, 3 - 7 days, 8 -14 days, more than 14 days). Draw simple random samples from each group.

  - Obtain lists of patients discharged from all MMH facilities. Number these patients, and then use a random number table to obtain the sample.
  - Randomly select some MMH facilities from each of five geographic regions, and then survey all of these hospitals' discharge lists.
  - At the beginning of the year, instruct each MMH facility to survey every 500th patient discharged.
  - Instruct each MMH facility to survey 10 discharged patients this week and send in the results.

- Identify the sampling design:

  - The Educational Testing Service (ETS) needed a sample of colleges. ETS first divided all colleges into groups of similar types (small public, small private, etc.) Then they randomly selected 3 colleges from each group.
  - A county commissioner wants to survey people in her district to determine their opinions on a particular law up for adoption. She decides to randomly select blocks in her district and then survey all who live on those blocks.
  - A local restaurant manager wants to survey customers about the service they receive. Each night the manager randomly chooses a number between 1 & 10. He then gives a survey to that customer and to every 10th customer after that, to fill out before they leave.

1.23 Bias

bias: A systematic error in measuring the estimate ; favors certain outcomes; has to do with center of sampling distributions; if centered over true parameter then considered unbiased.

- sources of bias in surveys:

  - things that can cause bias in your sample
    * voluntary response: people choose to respond; usually only people with very strong opinions respond
    * convenience sampling: ask people who are easy to ask, friendly, or interviewer is comfortable asking; produces bias results
    * undercoverage: some groups of population are left out of the sampling process
    * nonresponse: occurs when an individual chosen for the sample can't be contacted or refuses to cooperate; telephone surveys 70% nonresponse
    * response bias: occurs when the behavior of respondent or interviewer causes bias in the sample; wrong/false answers

    - wording of the questions: wording can influence the answers that are given; leading questions; connotation of words; use of "big" words or technical words
  - cannot do anything with bad data

1.24 Potential Pitfalls

- Misleading Conclusions
- Sample Data Reported Instead of Measured
- Loaded Questions
- Order of Questions
- Nonresponse
- Percentages

1.25 Experimental Design

experiment: actively impose some treatment in order to observe the response

Chapter 2

Graphical Displas

2.1 Frequency Distribution

- *Frequency Distribution (or Frequency Table)* Shows how data are partitioned among several categories (or classes) by listing the categories along with the number (frequency) of data values in each of them.
- Definitions
  - *Lower class limits:* The smallest numbers that can belong to each of the different classes
  - *Upper class limit:* The largest numbers that can belong to each of the different classes
  - *Class boundaries:* The numbers used to separate the classes, but without the gaps created by class limits
  - *Class midpoints:* The values in the middle of the classes Each class midpoint can be found by adding the lower class limit to the upper class limit and dividing the sum by 2.
  - *Class width:* The difference between two consecutive lower class limits in a frequency distribution.

2.2 Procedure for Constructing a Frequency Distribution

1. Select the number of classes, usually between 5 and 20. Calculate the class width. (Will be given on test)

2. Calculate the class width. (**round up the number**)

$$\text{Class Width} \approx \frac{\text{max data value-min data value}}{\text{number of class}}$$

3. Choose the value for the first lower class limit by using either the minimum value or a convenient value below the minimum.
4. Using the first lower class limit and class width, list the other lower class limits.
5. List the lower class limits in a vertical column and then determine and enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to get the frequency.

2.3 Example

Time(Seconds)	Frquency
75-124	11
125-174	24
175-224	10
225-274	3
275-324	2

How do we get this table? Below is the dataset of McDonald's Lunch Drive-Through Service Time from a sample:

107	139	197	209	281	254	163	150	127	308	206	187	169	83	127	133	140
143	130	144	91	113	153	255	252	200	117	167	148	184	123	153	155	154
100	117	101	138	186	196	146	90	144	119	135	151	197	171	190	169	

2.4 Relative Frequency Distribution

- Relative Frequency Distribution or Percentage Frequency Distribution: Each class frequency is replaced by a relative frequency (or proportion) or a percentage

$$\text{Relative Frequency for a class} = \frac{\text{frequency for a class}}{\text{sum of all frequencies}}$$

- Percentage Frequency: just convert the above to percent.

- The sum of the percentages in a relative frequency distribution must be very close to 100% (with a little wiggle room for rounding errors).

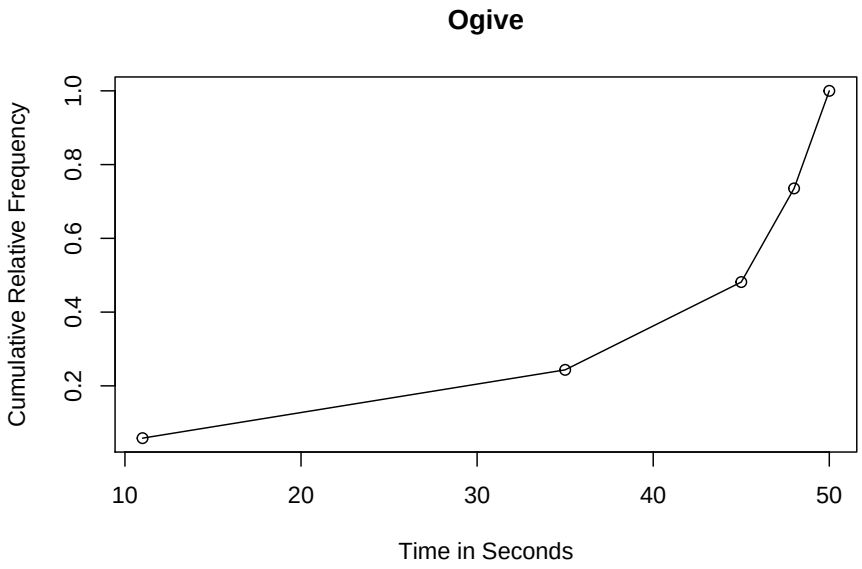
2.5 Cumulative Frequency Distribution

- *Cumulative Frequency Distribution:* The frequency for each class is the sum of the frequencies for that class and all previous classes.

Time (Seconds)	Cumulative Frequency
Less than 125	11
Less than 175	35
Less than 225	45
Less than 275	48
Less than 325	50

2.6 Cumulative Relative Frequence

- Simply divide the cumulative frequency by the total frequency.
- If we plot the cumulative relative frequency, the plot is called the Ogive



2.7 Using Frequency Distributions to Understand Data

We can roughly identify the distribution of a dataset using the frequency distribution. However, the more convinent way to do so is to use a historgram, which we will discuss later.

2.8 Gaps

- The presence of gaps can show that the data are from two or more different populations.
- However, the converse is not true, because data from different populations do not necessarily result in gaps.

2.9 What Do Gaps Tell Us?

The table shown is a frequency distribution of the weights (grams) of randomly selected pennies.

Weight (grams) of Penny	Frequency
2.40-2.49	18
2.50-2.59	19
2.60-2.69	0
2.70-2.79	0
2.80-2.89	0
2.90-2.99	2
3.00-3.09	25
3.10-3.19	8

- Pennies made before 1983 are 95% copper and 5% zinc.
- Pennies made after 1983 are 2.5% copper and 97.5% zinc.

2.10 Histogram

- **Histogram:** A graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data). The horizontal scale represents classes of quantitative data values, and the vertical scale represents frequencies. The heights of the bars correspond to frequency values.

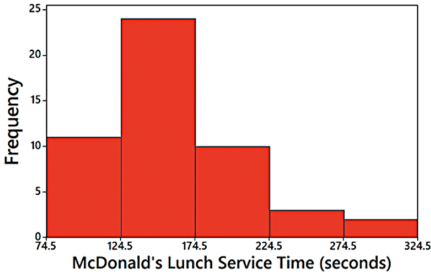
2.11 Important Uses of a Histogram

- Visually displays the shape of the distribution of the data
- Shows the location of the center of the data
- Shows the spread of the data
- Identifies outliers

2.12 Example of A Histogram

- *Relative Frequency Histogram:* It has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies.

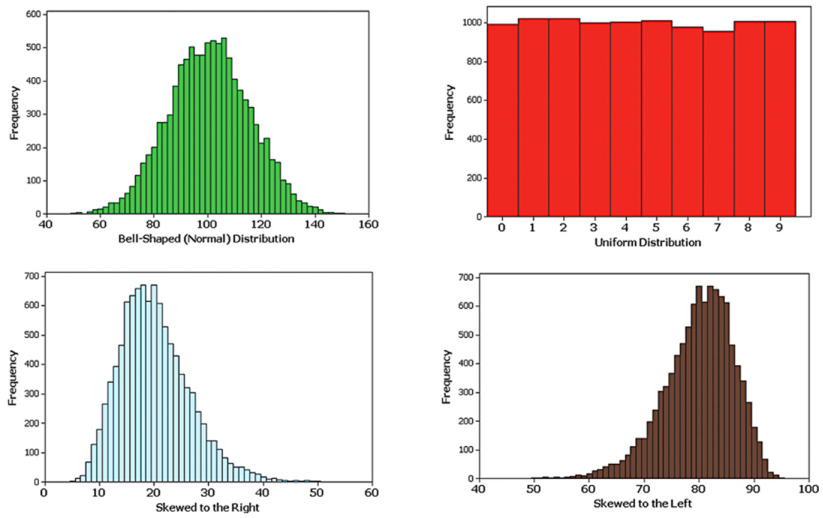
Time (seconds)	Frequency
75-124	11
125-174	24
175-224	10
225-274	3
275-324	2



2.13 Interpreting Histograms

- Explore the data by analyzing the histogram to see what can be learned about "CVDOT":
  - the Center of the data,
  - the Variation,
  - the shape of the Distribution,
  - whether there are any Outliers,
  - and Time. (if possible)

2.14 Common Distribution Shapes



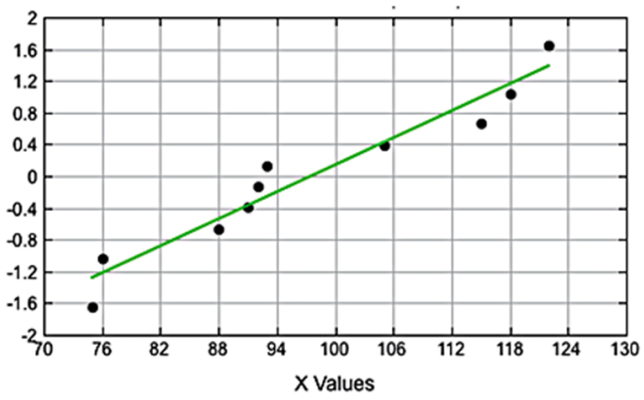
2.15 Assessing Normality with Normal Quantile Plots (Q-Q plots)

- Criteria for Assessing Normality with a Normal Quantile Plot

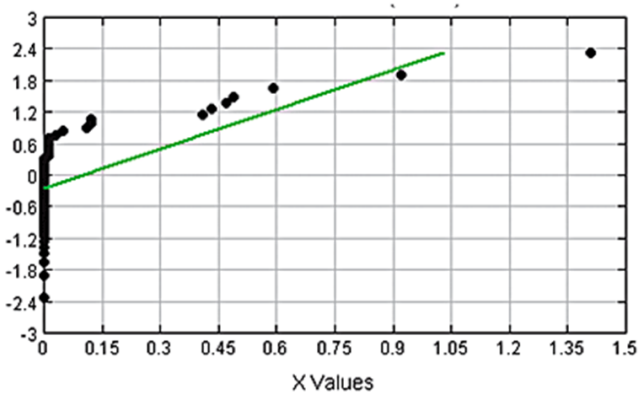
  - Normal Distribution: The pattern of the points in the normal quantile plot is reasonably close to a straight line, and the points do not show some systematic pattern that is not a straight-line pattern.
  - Not a Normal Distribution: The population distribution is not normal if the normal quantile plot has either or both of these two conditions:

    * The points do not lie reasonably close to a straight-line pattern.
    * The points show some systematic pattern that is not a straight-line pattern. (going up and down around the line)

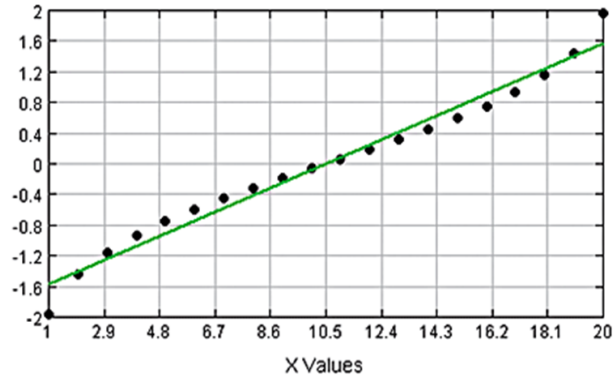
2.16 Examples of Q-Q plots



Normal Distribution: The points are reasonably close to a straight-line pattern, and there is no other systematic pattern that is not a straight-line pattern.



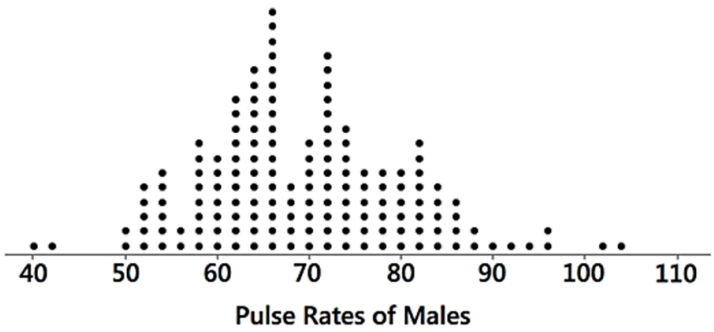
Not a Normal Distribution: The points do not lie reasonably close to a straight line.



Not a Normal Distribution: The points show a systematic pattern that is not a straight-line pattern.

2.17 Dotplots

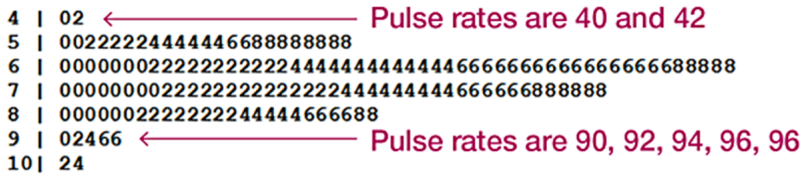
- A graph of quantitative data in which each data value is plotted as a point (or dot) above a horizontal scale of values. Dots representing equal values are stacked.



- Features of a Dotplot
  - Displays the shape of distribution of data.
  - It is usually possible to recreate the original list of data values.

2.18 Stemplots (You need to assign a key to read a stem plot)

Represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit).



- Features of a Stemplot
  - Shows the shape of the distribution of the data.
  - Retains the original data values.
  - The sample data are sorted (arranged in order).

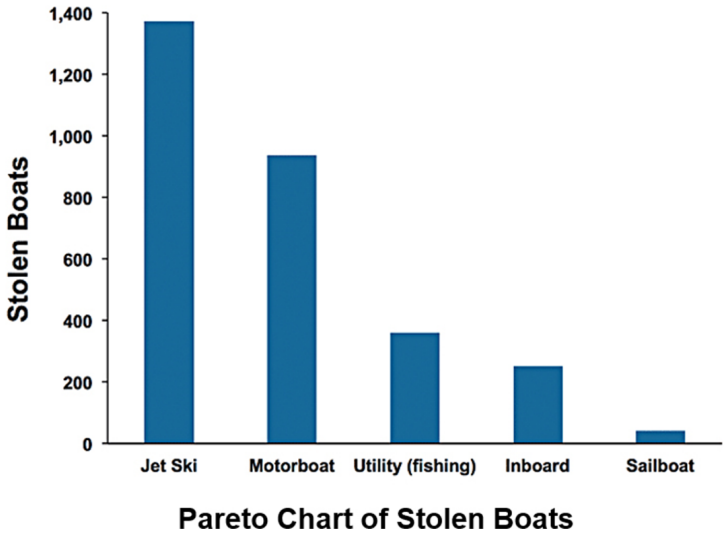
2.19 Bar Graph

Shows the relative distribution of categorical data so that it is easier to compare the different categories.

2.20 Pareto Chart

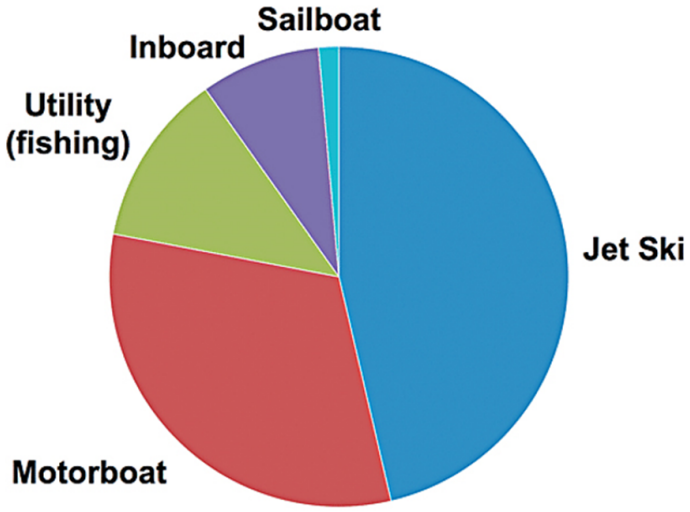
A Pareto chart is a bar graph for categorical data, with the added stipulation that the bars are arranged in descending order according to frequencies, so the bars decrease in height from left to right.

- Features of a Pareto Chart
  - Shows the relative distribution of categorical data so that it is easier to compare the different categories.
  - Draws attention to the more important categories.



2.21 Pie Charts

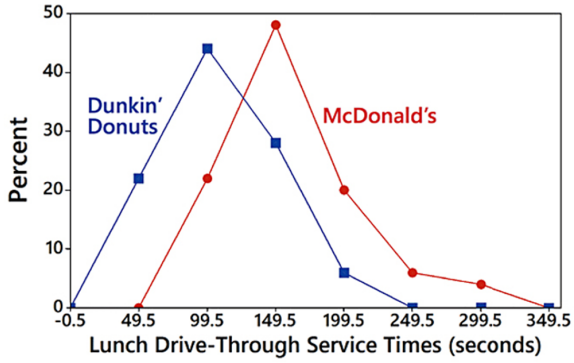
A very common graph that depicts categorical data as slices of a circle, in which the size of each slice is proportional to the frequency count for the category; Shows the distribution of categorical data in a commonly used format.



Pie Chart of Stolen Boats

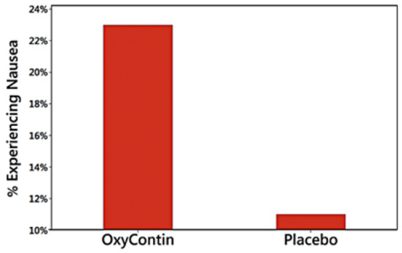
2.22 Frequency Polygon

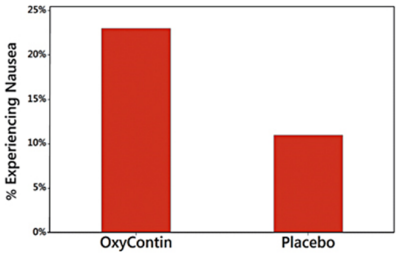
- A graph using line segments connected to points located directly above class midpoint values
- A frequency polygon is very similar to a histogram, but a frequency polygon uses line segments instead of bars.



Relative Frequency Polygons for McDonald's and Dunkin

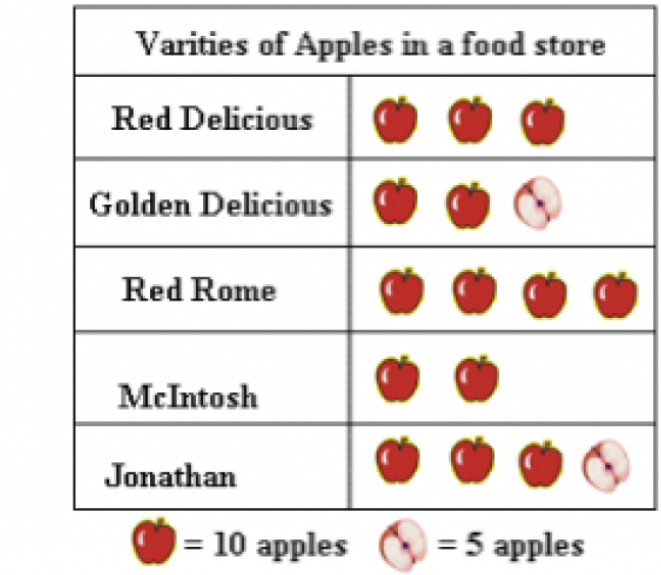
2.23 Graphs That Deceive

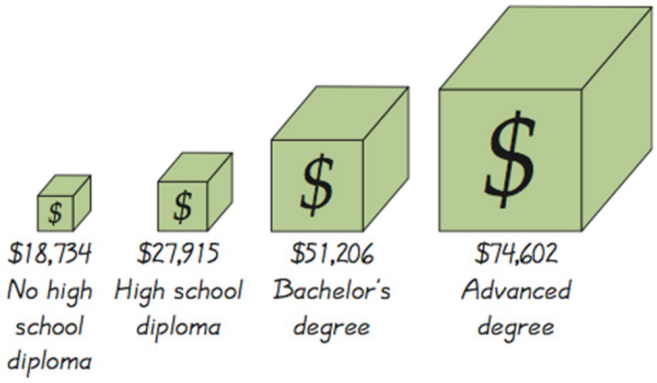




2.24 Pictographs

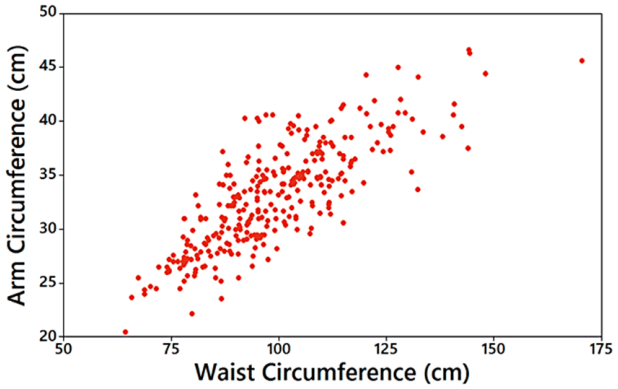
- By using pictographs, artists can create false impressions that grossly distort differences by using these simple principles of basic geometry:
  - When you double each side of a square, its area doesn't merely double; it increases by a factor of four.
  - When you double each side of a cube, its volume doesn't merely double; it increases by a factor of eight.



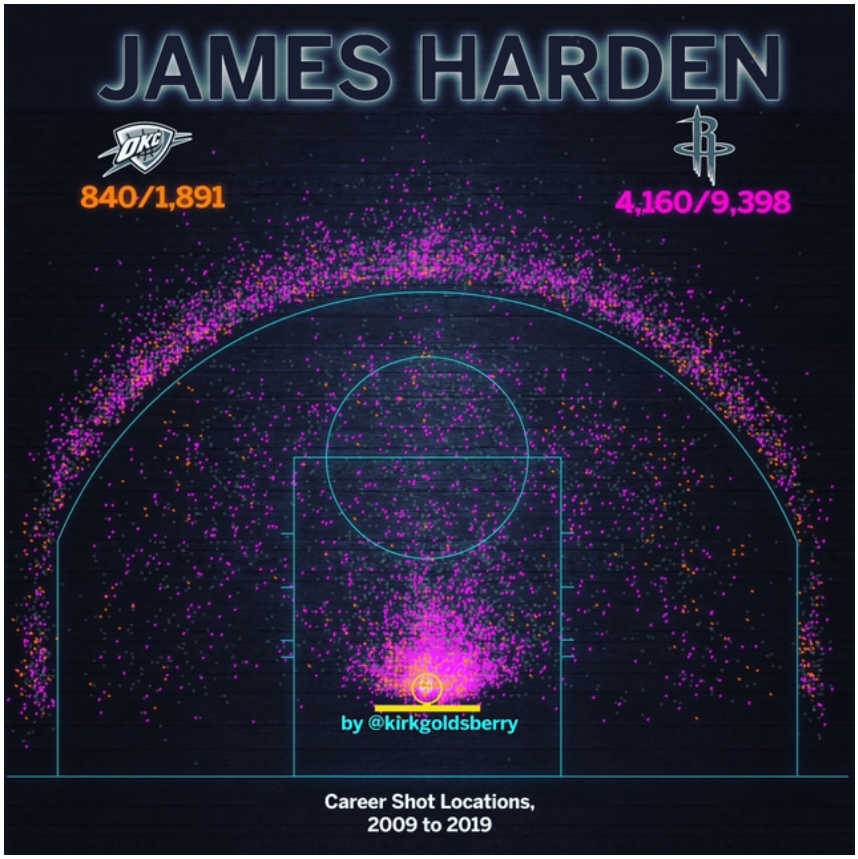


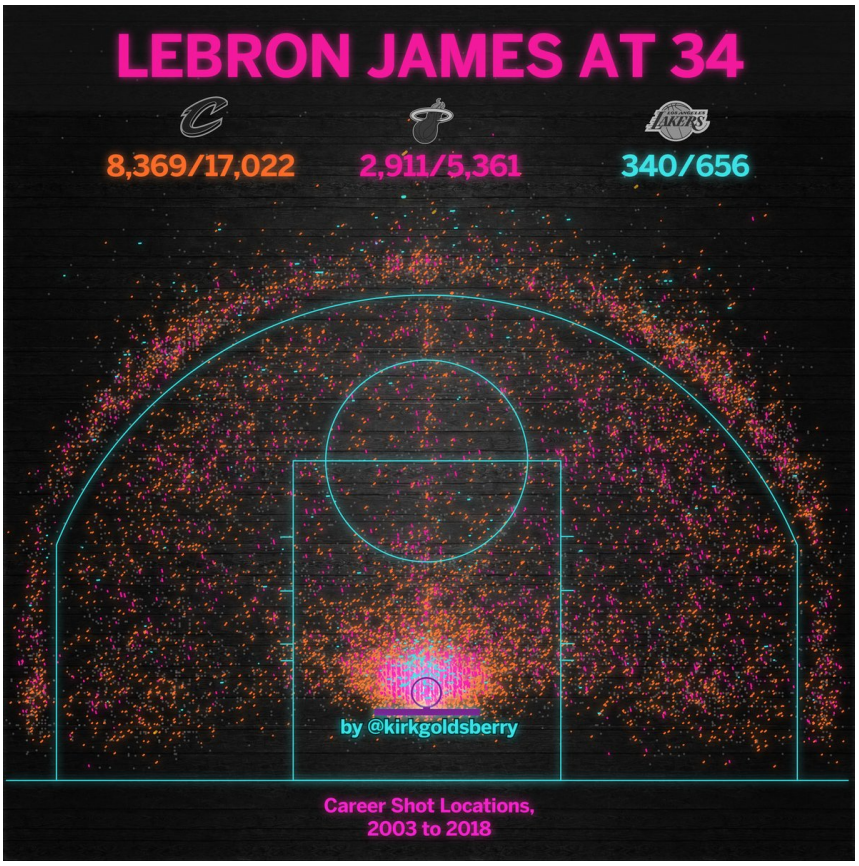
Misleading. Depicts one-dimensional data with three-dimensional boxes. Last box is 64 times as large as first box, but income is only 4 times as large.

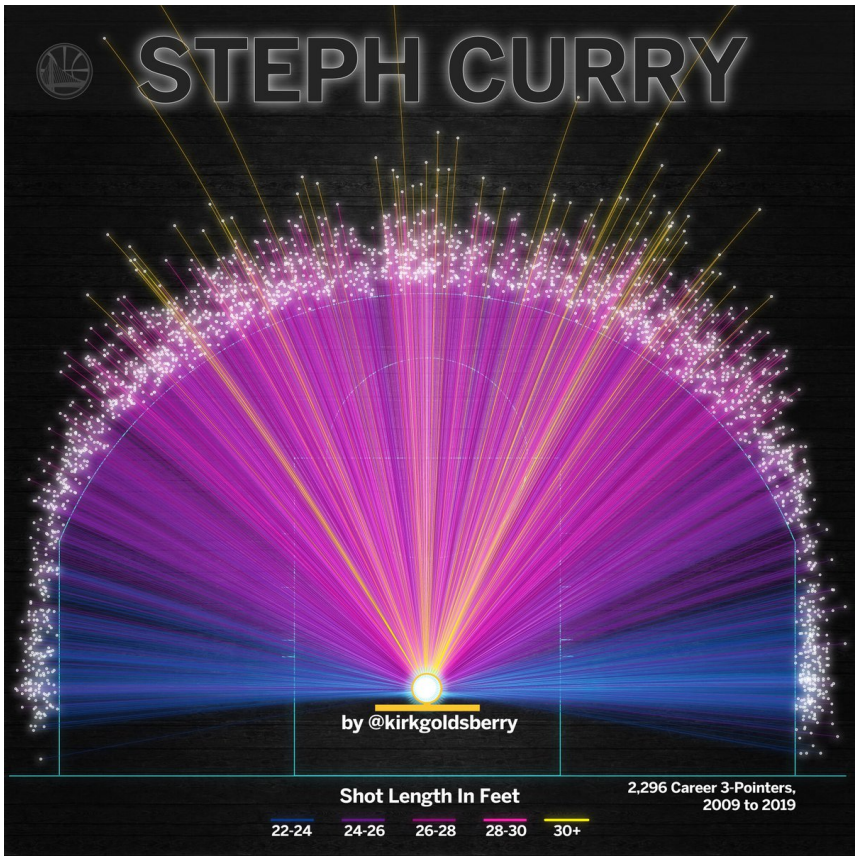
2.25 Scatterplot (Two Numerical Variables)



2.26 NBA

- 
JAMES HARDEN
840/1,891
4,160/9,398
by @kirkgoldsberry
Career Shot Locations,
2009 to 2019

- 
LEBRON JAMES AT 34
8,369/17,022
2,911/5,361
340/656
by @kirkgoldsberry
Career Shot Locations,
2003 to 2018

- 
STEPH CURRY
by @kirkgoldsberry
Shot Length In Feet
22-24
24-26
26-28
28-30
30+
2,296 Career 3-Pointers,
2009 to 2019

Chapter 3

Descriptive Statistics

- In this chapter we'll learn to summarize or describe the important characteristics of a data set (mean, standard deviation, etc.)
- The focus of this section is to obtain a value that measures the center of a data set. In particular, we present measures of center, including mean and median. Our objective here is not only to find the value of each measure of center, but also to interpret those values.
- A measure of center is a value at the center or middle of a data set.

3.1 Mean (or Arithmetic Mean)

- The **mean (or arithmetic mean)** of a set of data is the measure of center found by adding all of the data values and dividing the total by the number of data values.

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

- Sample means drawn from the same population tend to vary less than other measures of center.
- The mean of a data set uses every data value.
- A disadvantage of the mean is that just one extreme value (outlier) can change the value of the mean substantially. (Using the following definition, we say that the mean is not **resistant**.)
- A statistic is resistant if the presence of extreme values (outliers) does not cause it to change very much.
- Caution: We don't use the term average in Statistics. Instead, we use the term mean, or expected value, which will be introduced later.

3.2 Notations

- $\sum$ denotes the sum of a set of values.
- x is the variable usually used to represent the individual data values.
- n represents the number of data values in a sample.
- N represents the number of data values in a population.
- $\bar{x}$: the mean of a set of sample values (sample mean)
- $\bar{x} = \frac{\sum x}{n}$
- μ: the mean of all values in a population (population mean)
- $\mu = \frac{\sum x}{N}$

3.3 Mean-Example

- Find the mean of the first five data speeds for Verizon: 38.5, 55.6, 22.4, 14.1, and 23.1 (all in megabits per second, or Mbps).
- Find the mean of the first five counts for Chips Ahoy regular cookies: 22 chips, 22 chips, 26 chips, 24 chips, and 23 chips.

3.4 Median

- Median is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude
- often denoted by $\tilde{x}$

- is not affected by an extreme value - is a resistant measure of the center
- To find the median, first sort the values (arrange them in order) and then follow one of these two procedures:
  - If the number of data values is odd, the median is the number located in the exact middle of the sorted list.
  - If the number of data values is even, the median is found by computing the mean of the two middle numbers in the sorted list.

3.5 Median-Example

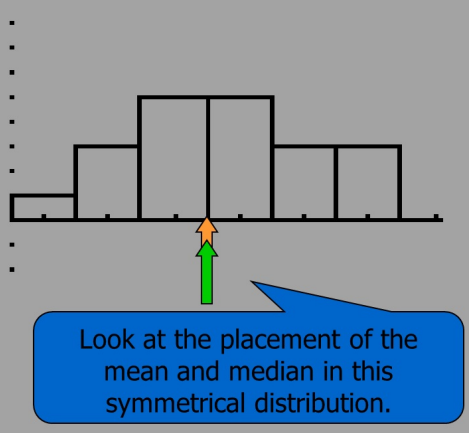
- Find the median of the first five data speeds for Verizon: 38.5, 55.6, 22.4, 14.1, and 23.1 (all in megabits per second, or Mbps).
- Repeat of the previous example after including the sixth data speed of 24.5 Mbps. That is, find the median of these data speeds: 38.5, 55.6, 22.4, 14.1, 23.1, 24.5 (all in Mbps).

3.6 Which one do you use? Mean or Median?

Ex6) Look at the following data set. Find the mean & median.

Mean = 27

Median = 27

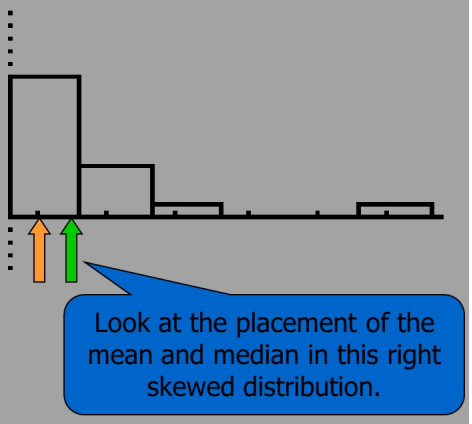


21	23	23	24	25	25	26	26	26	27
27	27	27	28	30	30	30	31	32	32

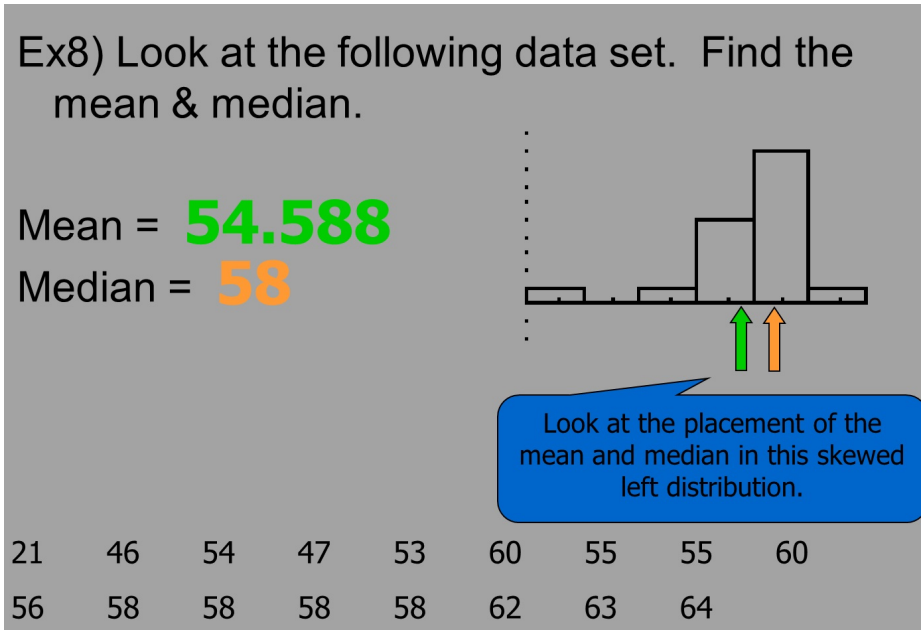
Ex7) Look at the following data set. Find the mean & median.

Mean = 28.176

Median = 25



22	29	28	22	24	25	28	21	25
23	24	23	26	36	38	62	23	



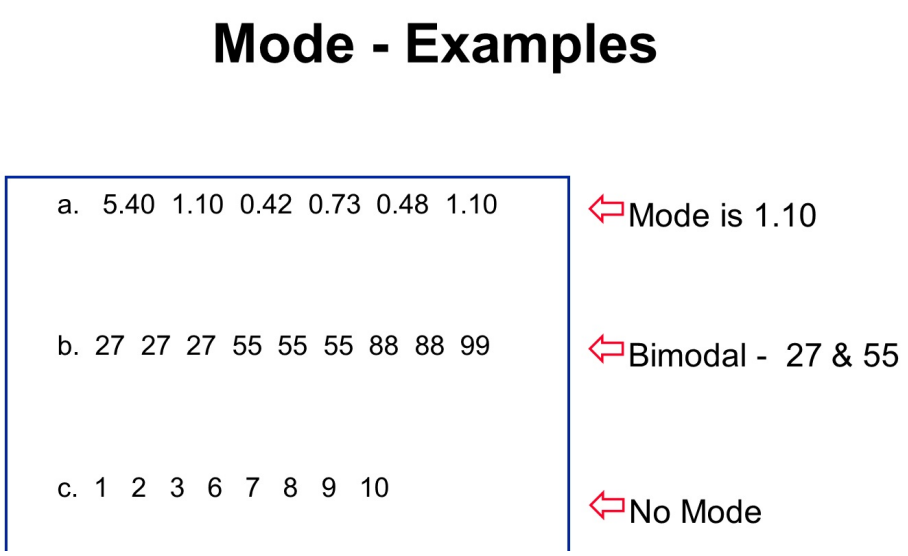
3.7 Which one do you use? Mean or Median?

- In a symmetrical distribution, the mean and median are equal.
- In a skewed distribution, the mean is pulled in the direction of the skewness.
- In a symmetrical distribution, you should report the mean!
- In a skewed distribution, the median should be reported as the measure of center!

3.8 Mode

- Mode is the value that occurs with the greatest frequency
- Data set can have one, more than one, or no mode
- Bimodal two data values occur with the same greatest frequency
- Multimodal more than two data values occur with the same greatest frequency
- No Mode no data value is repeated
- Mode is the only measure of central tendency that can be used with **nominal** data.

3.9 Mode-Example



3.10 Midrange

- Midrange is the value midway between the maximum and minimum values in the original data set
- $Midrange = \frac{Maximum\ \ Value + Minimum\ \ Value}{2}$
- Because the midrange uses only the maximum and minimum values, it is very sensitive to those extremes so the midrange is not resistant.

3.11 Round-Off Rules for Measures of Center

- For the mean, median, and midrange, carry one more decimal place than is present in the original set of values.
- For the mode, leave the value as is without rounding (because values of the mode are the same as some of the original data values).

3.12 Calculating a Weighted Mean

- When data values are assigned different weights, w, we can compute a weighted mean.

$$\bar{x} = \frac{\sum (w \cdot x)}{\sum w}$$

- In her first semester of college, a student of the author took five courses. Her final grades along with the number of credits for each course were A (3 credits), A (4 credits), B (3 credits), C (3 credits), and F (1 credit). The grading system assigns quality points to letter grades as follows: A = 4; B = 3; C = 2; D = 1; F = 0.

3.13 More Examples About Finding The Means

- Angelique made scores of 85, 56, and 91 on her first three statistic tests. What does she need to make on her next test to have an 80 test average?
- Mr. Plum's math class of 25 students had an average of 85 on a test. Miss Scarlet's class of 22 students had an average of 87 on the same test. What is the average of the two classes combined?
- Consider the time that it takes the faculty of AVC to drive to school. The mean and median times are calculated. Of the times, 40 minutes and 25 minutes, which is the mean and which is the median? Why?

3.14 Measures of Variation

Variation is the single most important topic in statistics, so this is the single most important section in this book. This section presents three important measures of variation: range, standard deviation, and variance. These statistics are numbers, but our focus is not just computing those numbers but developing the ability to interpret and understand them.

3.15 Round-off Rule for Measures of Variation

When rounding the value of a measure of variation, carry one more decimal place than is present in the original set of data.

3.16 Range

- The range of a set of data values is the difference between the maximum data value and the minimum data value.
- Range = (maximum value) - (minimum value)
- It is very sensitive to extreme values; therefore, it is not as useful as other measures of variation.
- Because the range uses only the maximum and minimum values, it does not take every value into account and therefore does not truly reflect the variation among all of the data values.

- Find the range of these Verizon data speeds (Mbps): 38.5, 55.6, 22.4, 14.1, 23.1.

3.17 Standard Deviation

- The **standard deviation** of a set of sample values, denoted by s, is a measure of how much data values deviate away from the mean.
- **s**: sample standard deviation
- σ: population standard deviation

3.18 Formula

- Sample standard deviation:

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

or

$$s = \sqrt{\frac{n\sum(x_i^2) - \sum(x_i)^2}{n(n-1)}}$$

- Population standard deviation

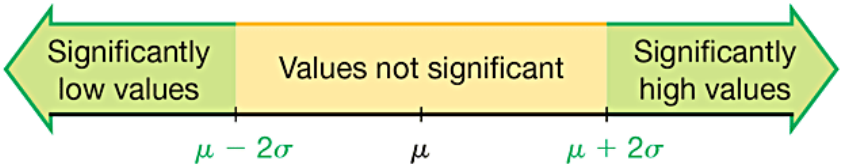
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

3.19 Interpretation of Standard Deviation

- The standard deviation is a measure of variation of all values from the mean; or a measure of how much data values deviate away from the mean.
- The value of the standard deviation s is never negative. It is zero only when all of the data values are exactly the same.
- The value of the standard deviation s can increase dramatically with the inclusion of one or more outliers (data values far away from all others).
- The units of the standard deviation s are the same as the units of the original data values.
- Larger values of s indicate greater amounts of variation.
- The sample standard deviation s is a **biased estimator** of the population standard deviation σ, which means that values of the sample standard deviation s do not center around the value of σ.
- Use sample standard deviation formula to find the standard deviation of these Verizon data speed times (in Mbps): 38.5, 55.6, 22.4, 14.1, 23.1.

3.20 Range Rule of Thumb for Understanding Standard Deviation

- The range rule of thumb is a crude but simple tool for understanding and interpreting standard deviation. The vast majority (such as 95%) of sample values lie within 2 standard deviations of the mean.
- Significantly low values are $\mu - 2\sigma$ or lower. (Minimum "usual" value = (mean) - 2 $\times$ (standard deviation))
- Significantly high values are $\mu + 2\sigma$ or higher. (Maximum "usual" value = (mean) + 2 $\times$ (standard deviation))
- Values not significant are between $\mu - 2\sigma$ and $\mu + 2\sigma$.



- Using the 40 chocolate chip counts for the Chips Ahoy cookies, the mean is 24.0 chips and the standard deviation is 2.6 chips.
  - Use the range rule of thumb to find the minimum and maximum "usual" numbers of chips.\
  - Would a cookie with 30 chocolate chips be "unusual"?

3.21 Range Rule of Thumb for Estimating a Value of the Standard Deviation s

- To roughly estimate the standard deviation from a collection of known sample data use

$$s \approx \frac{Range}{4}$$

where range = (maximum value) - (minimum value)

3.22 Variance of a Sample and a Population

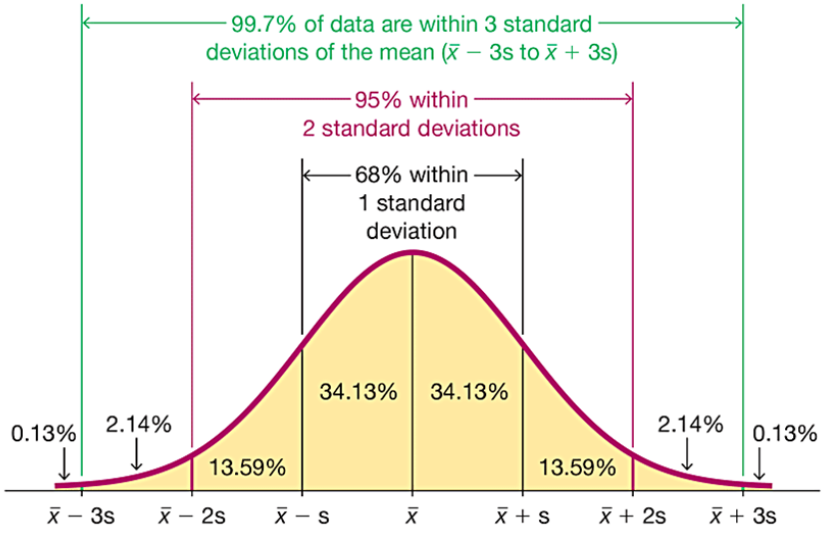
- The **variance** of a set of values is a measure of variation equal to the square of the standard deviation.
  - Sample variance: s^2 - Square of the sample standard deviation s
  - Population variance: σ^2 - Square of the population standard deviation σ

- The units of the variance are the squares of the units of the original data values.
- The value of the variance can increase dramatically with the inclusion of outliers. (The variance is not resistant.)
- The value of the variance is never negative. It is zero only when all of the data values are the same number.
- The sample variance s^2 is an unbiased estimator of the population variance σ^2.

3.23 Notations

- s: samle standard deviation
- n: sample size
- s^2: sample variance
- σ: population standard deviation
- N: population size
- σ^2: population variance

3.24 The Empirical Rule



- IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15. What percentage of IQ scores are between 70 and 130?

3.25 Chebyshev's Theorem

The proportion of any set of data lying within K standard deviations of the mean is always **at least** $1 - \frac{1}{k^2}$, where K is any positive number greater than 1.

For $K = 2$ and $K = 3$, we get the following statements:

- At least $\frac{3}{4}$ (or 75%) of all values lie within 2 standard deviations of the mean.
- At least $\frac{8}{9}$ (or 89%) of all values lie within 3 standard deviations of the mean.

3.26 Extra Example

- A list has 10 numbers. Each number is a 1, 2, or 3. The average is 2 and the SD is 0. What is the list?

3.27 Measures of Relative Standing and Boxplots

- This section introduces measures of relative standing, which are numbers showing the location of data values relative to the other values within a data set.
- They can be used to compare values from different data sets, or to compare values within the same data set.

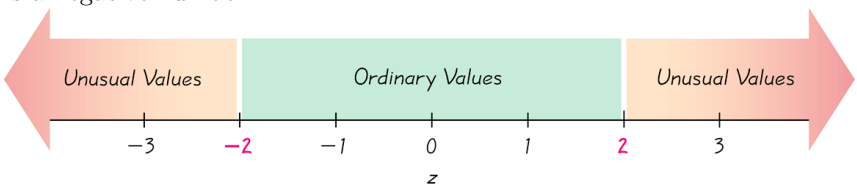
- The most important concept is the z score.
- We will also discuss percentiles and quartiles, as well as a new statistical graph called the boxplot.

3.28 z Scores

- z-score (or standardized value) is the number of standard deviations that a given value x is above or below the mean.
- z-score of a sample data value: $z = \frac{x - \bar{x}}{s}$

- z-score of a population data value: $z = \frac{x-\mu}{\sigma}$
- Round z scores to two decimal places

3.29 Important Properties of z Scores

- A z-score is the number of standard deviations that a given value x is above or below the mean.
- z-scores are expressed as numbers with no units of measurement.
- A data value is significantly low if its z-score is less than or equal to -2 or the value is significantly high if its z-score is greater than or equal to +2.
- If an individual data value is less than the mean, its corresponding z score is a negative number.
- 

3.30 Examples

- Suppose the mean and standard deviation of a distribution are $\mu = 50$ and $s = 5$
  - If the x-value is 55, what is the z-score?
  - If the x-value is 45, what is the z-score?
  - If the x-value is 60, what is the z-score?
  - So what does the z-score tell you?
  - The author of the text measured his pulse rate to be 48 beats per minute. Is that pulse rate unusual if the mean adult male pulse rate is 67.3 beats per minute with a standard deviation of 10.3?
  - Sally is taking two different math achievement tests with different means and standard deviations. The mean score on test A was 56 with a standard deviation of 3.5, while the mean score on test B was 65 with a standard deviation of 2.8. Sally scored a 62 on test A and a 69 on test B. On which test did Sally score the best? Why?

3.31 Percentiles

- **Percentiles** are measures of location, denoted P_1, P_2, . . . , P_99, which divide a set of data into 100 groups with about 1% of the values in each group.

- The process of finding the percentile that corresponds to a particular data value x is given by the following (round the result to the nearest whole number)

$$\text{Percentile of a data value x} = \frac{\text{number of values less than x}}{\text{total number of values}} \times 100$$

- The airport Verizon cell phone data speeds listed below are arranged in increasing order. Find the percentile for the data speed of 11.8 Mbps.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

3.32 Interpretation of Percentiles

- From the previous example: A data speed of 11.8 Mbps is in the 40th percentile. This can be interpreted loosely as this: A data speed of 11.8 Mbps separates the lowest 40% of values from the highest 60% of values. We have $P_{40} = 11.8$ Mbps.
- Better interpretation: 40% of the data values are less than 11.8 Mbps or 60% of the data values are greater than 11.8 Mbps
- What is the percentile for the median? How would you interpret the median better?

3.33 Example from a dataset

```
##   height ideal_ht sleep fastest
## 1     76       78   9.5     119
## 2     74       76   7.0     110
## 3     64       NA   9.0      85
## 4     62       65   7.0     100
## 5     72       72   8.0      95

## 20% 50% 80% 90%
##  90 102 120 130
```

- Interpet the meanings of the above perentiles
  - About 20% of the students drove slower than 90 mph

- About 50% of the students drove slower than 102 mph (median!)
- About 80% of the students drove slower than 120 mph
- About 90% of the students drove slower than 130 mph

3.34 Notations

- n total number of values in the data set
- k percentile being used (Example: For the 25th percentile, k = 25.)
- L locator that gives the position of a value (Example: For the 12th value in the sorted list, L = 12.)
- P_k kth percentile (Example: P_{25} is the 25th percentile.)

3.35 Converting a Percentile to a Data Value

1. Find L, where $L = \frac{k}{100} \times n$
2. If L is not a whole number, round up to the next whole number and find the data value in that position. If L is a whole number, average the data values in position L and L+1. (Notes: You need to put the numbers in order from small to large first.)

3.36 Example

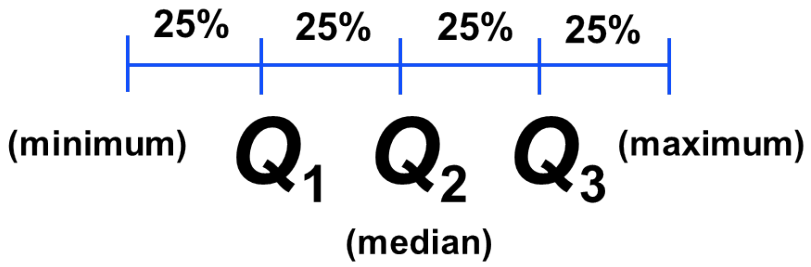
Refer to the sorted data speeds below. Find the 40th and 60th percentile, denoted by P_{40} and P_{60}, respectively.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

3.37 Quartiles

- Quartiles are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with about 25% of the values in each group
- Q_1 (First quartile): Same value as P_{25}. It separates the bottom 25% of the sorted values from the top 75%.

- Q_2 (Second quartile): Same as P_{50} and same as the **median**. It separates the bottom 50% of the sorted values from the top 50%.
- Q_3 (Third quartile): Same as P_{75}. It separates the bottom 75% of the sorted values from the top 25%.
- Different technologies often yield different results.



-

3.38 Statistics defined using quartiles and percentiles

Interquartile range (or IQR)

Semi-interquartile range

Midquartile range

10 – 90 quartile range

Just focus on the Interquartile Range (IQR)

3.39 5-Number Summary

- For a set of data, the 5-number summary consists of these five values:
  - Minimum
  - Q_1
  - Q_2 or median

- Q_3
- Maximum

3.40 Example: Finding a 5-Number Summary

Use the Verizon airport data speeds to find the 5-number summary.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

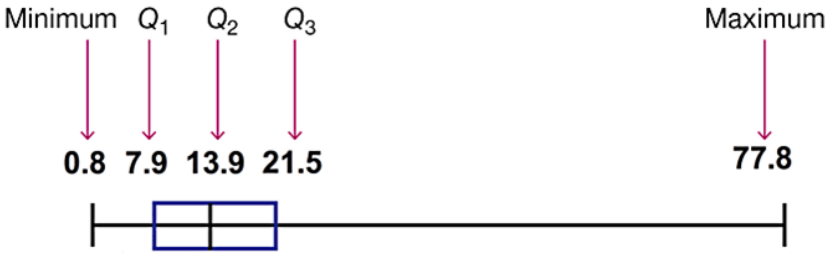
3.41 Boxplot (or Box-and-Whisker Diagram)

- A boxplot (or box-and-whisker diagram) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile Q_1, the median, and the third quartile Q_3.

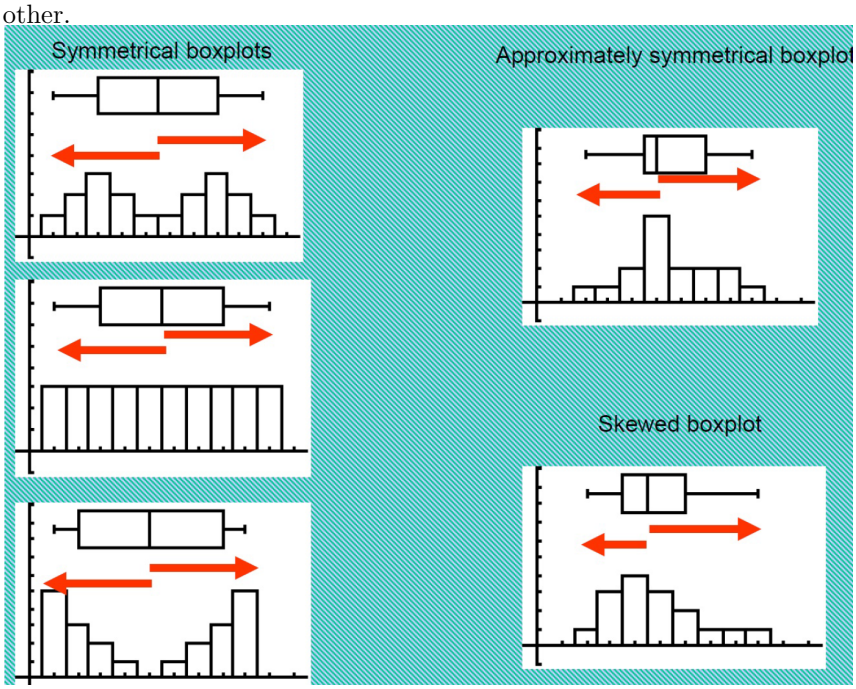
3.42 Procedure for Constructing a Boxplot

1. Find the 5-number summary (minimum value, Q_1, Q_2, Q_3, maximum value).
2. Construct a line segment extending from the minimum data value to the maximum data value.
3. Construct a box (rectangle) extending from Q1 to Q3, and draw a line in the box at the value of Q2 (median).

- Use the Verizon airport data speeds to construct a boxplot.
-



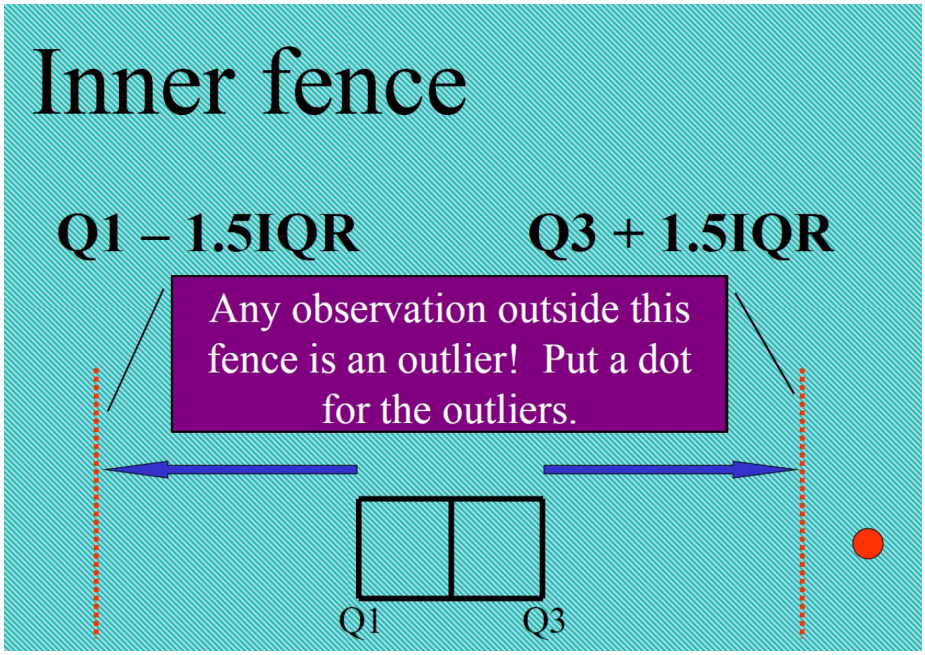
3.43 Skewness

- A boxplot can often be used to identify skewness. A distribution of data is skewed if it is not symmetric and extends more to one side than to the other.
- 

3.44 Identifying Outliers for Modified Boxplots

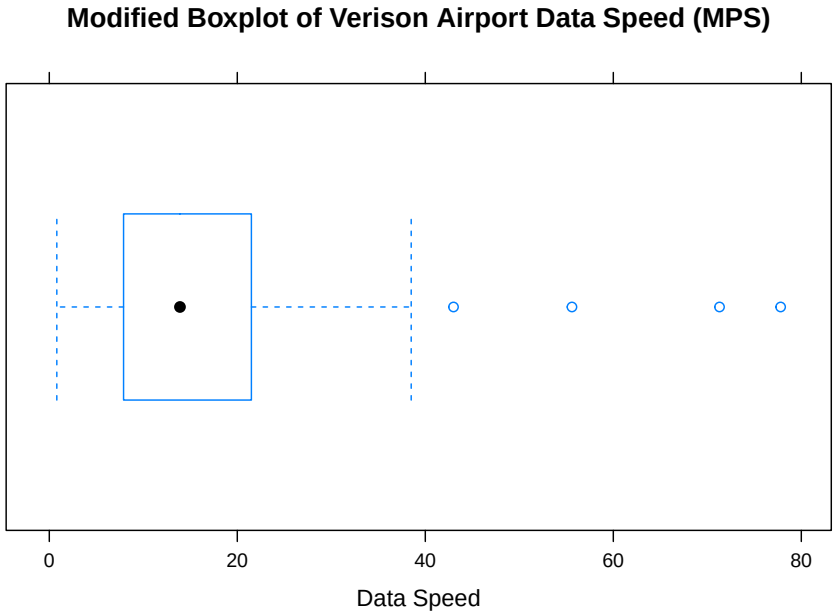
1. Find the quartiles Q_1, Q_2, and Q_3.
2. Find the interquartile range (IQR), where IQR $= Q_3 - Q_1$.
3. Evaluate $1.5 \times$ IQR.
4. In a modified boxplot, a data value is an outlier if it is above Q_3, by an amount greater than $1.5 \times$ IQR or below Q_1, by an amount greater than $1.5 \times$ IQR.

3.45 Modified Boxplot



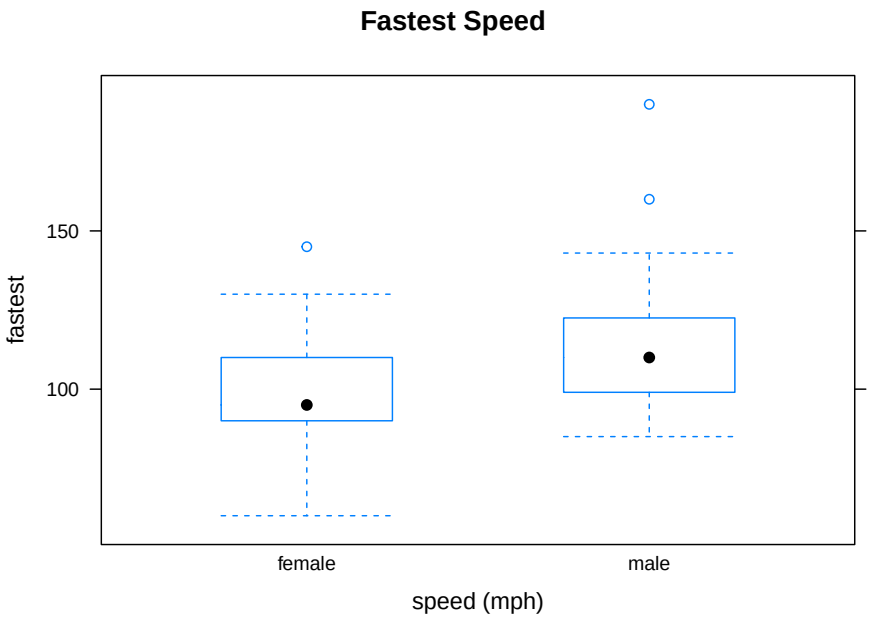
3.46 Example

```
##  min  Q1 median Q3 max mean sd  n missing
##  0.8 8.4     14 21  78   18 16 50       0
```




```
## $stats
## [1]  0.8  7.9 13.9 21.5 38.5
##
## $n
## [1] 50
##
## $conf
## [1] 11 17
##
## $out
## [1] 43 56 71 78
```

3.47 Comparative Boxplots



Chapter 4

Methods

We describe our methods in this chapter.

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.