MSSC 6010/Comp. Probability

Instructor: Mehdi Maadooliat

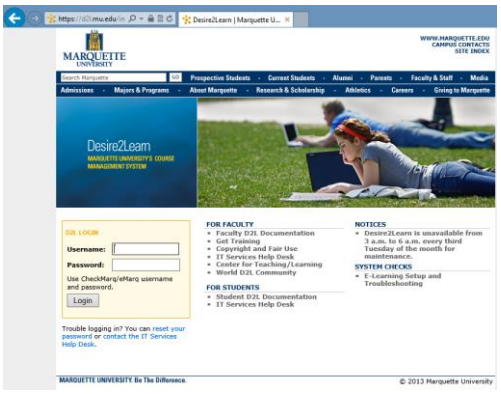
MSSC 6010

Department of Mathematics, Statistics and Computer Science

MARQUETTE UNIVERSITY
Be The Difference.

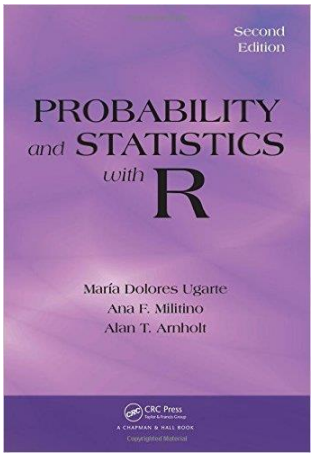
SYLLABUS - D2L - ONLINE MATERIALS

- **Syllabus**
- **Go to https://d2l.mu.edu/index.asp**
- **Enter your account info and click on Log In!**



MSSC 6010

- **Text:** Probability and Statistics with R, 2nd ed.
  (ISBN: 9781466504394) Maria Dolores Ugarte, Ana F. Militino, Alan T. Arnholt,



- **Instructor:** Mehdi Maadooliat, Ph.D.
  - **Office Hours:**
    - M 12:00 - 2:00pm via Skype: https://join.skype.com/ysT7bE6GnHaW
    - W 4:00-5:00pm in CU 311
  - **Email:** mehdi.maadooliat@marquette.edu

- **Homework:** Weekly or biweekly homework assignments will be given.

3

MSSC 6010 CONT...

- **Description:** Foundations of probability for modeling random processes and Bayesian approaches, including: counting techniques, probability of events, random variables, distribution functions, probability functions, probability density functions, expectation, moments, moment generating functions, special discrete and continuous distributions, sampling distributions, prior and posterior distributions, Law of Large Numbers, Central Limit Theorem, Bayesian paradigm.

- **Prereq:** Three semesters of mathematics beyond calculus and MATH 4720 or equiv.

4

WHAT DO WE COVER IN THIS COURS?

MARQUETTE UNIVERSITY
Be The Difference.

- **"Probability and Statistics with R, 1st ed." Covers:**
  - 1. A Brief Introduction to S
  - 2. Exploring Data
  - 3. General Probability and Random Variables
  - 4. Univariate Probability Distributions
  - 5. Multivariate Probability Distributions
  - 6. Sampling and Sampling Distributions
  - 7. Point Estimation
  - 8. Confidence Intervals
  - 9. Hypothesis Testing
  - 10. Nonparametric Methods
  - 11. Experimental Design
  - 12. Regression
- **I may cover some of the following topics**
  - Bayesian Statistics
  - Transformation of variable
  - Hypothesis Testing (LRT)

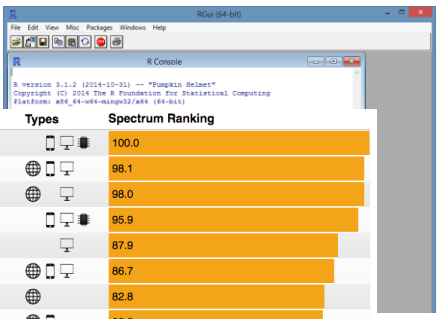
5

R IS THE STATISTICAL SOFTWARE THAT WE MOSTLY USE IN THIS COURSE

MARQUETTE UNIVERSITY
Be The Difference.

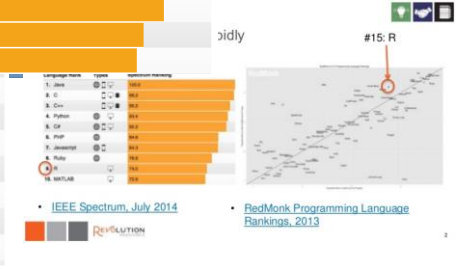
- **It is free!!**

IEEE Spectrum, July 2016



Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

Language Rank	Types		
1. Java		96.5	93.5
2. C		91.3	92.4
3. C++		84.8	84.8
4. Python		84.5	84.5
5. C#		83.0	78.9
6. R		76.2	74.3
7. PHP		72.4	72.8
8. JavaScript			
9. Ruby			
10. Matlab			



- IEEE Spectrum, July 2014
- RedMonk Programming Language Rankings, 2013

IEEE Spectrum, July 2015

6

HISTORY OF R (FROM WIKIPEDIA)

MARQUETTE UNIVERSITY
Be The Difference.

- R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.
- R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered.
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the *R Development Core Team*, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S.
- R is a GNU project. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; there are also several graphical front-ends for it.

7

MSSC 6010 CONT...

MARQUETTE UNIVERSITY
Be The Difference.

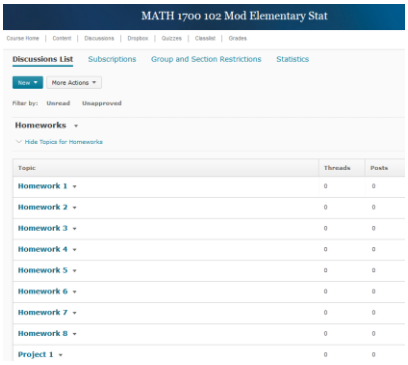
- **Grading:**
  - Weekly homework & class participation
  - Project
  - A midterm (in class), tentatively on Oct 16,
  - plus the final exam (in class or take home) on Dec 11, 5:45PM - 7:45PM.
  - No make-ups, If "unavoidable absence" as defined in Arts and Sci. Bulletin, then % added to final %.
  - Midterm Exams 30%
  - Project 10%
  - Homework & Class Participation 30%
  - Final Exam 30%

8

MSSC 6010 CONT...

MARQUETTE UNIVERSITY
Be The Difference.

- **Questions about Homework, Project and Exams:**
  - **SHOULD** be posted in d2l Discussion Board.
  - I will **NOT** answer general emails about Homework and Exams.



- **Homework:**
  - **Should be submitted as a PDF file:**
    - **How to Combine Images into a PDF file [FREE & EASY + No Software] (Youtube)**
    - **Microsoft Word to PDF in 10 Seconds (Youtube)**
    - **How to: convert Images to PDF in Macbook/IMac (Youtube)**
      - http://apple.stackexchange.com/questions/11163/how-do-i-combine-two-or-more-images-to-get-a-single-pdf-file

9

MSSC 6010/Comp. Probability

Instructor: Mehdi Maadooliat

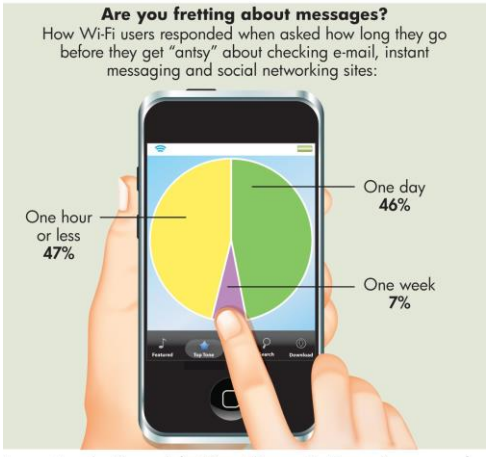
Chapters 2 Review

Department of Mathematics, Statistics and Computer Science

1.1 What Is Statistics?

WHAT IS STATISTICS?

- **Statistics is all around us!**
- **How much time between Internet usage?**



WHAT IS STATISTICS?

- **Statistics is all around us!**

World's Highest Paid Athletes

Forbes' list of the highest-paid athletes looks at earnings derived from salaries, bonuses, prize money, endorsements, and licensing income between June 2008 and June 2009 and does not deduct for taxes or agents' fees.

Here are the Top 5:

Rank	Athlete	Sport	Earnings
1	Tiger Woods	Golf	$110 million
2	Kobe Bryant	Basketball	$45 million
2	Michael Jordan	Basketball	$45 million
2	Kimi Raikkonen	Auto Racing	$45 million
5	David Beckham	Soccer	$42 million

13

WHAT IS STATISTICS?

Many Teens Use Phones in Class

84% teens have cell phones

16% teens do not

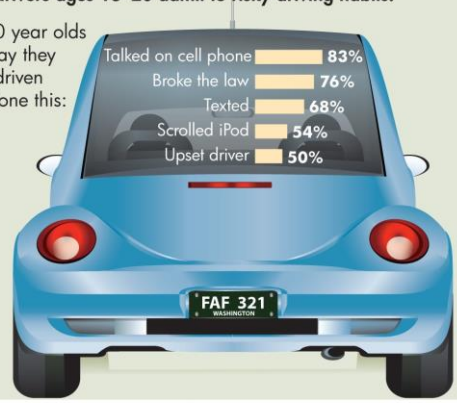
An average of 440 text messages sent per week –110 of them during class. Works out to be 3 text messages per class period.

Source:
Common Sense Media survey of 1013 teens May/June 2009

Busy Behind the Wheel

Most drivers ages 16–20 admit to risky driving habits.

16–20 year olds who say they have driven and done this:



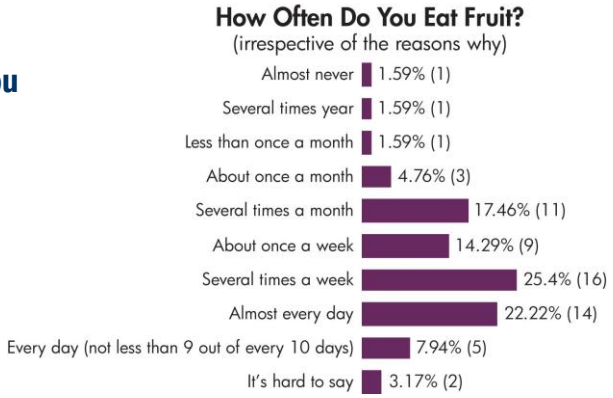
Source: National Organization for Youth Safety, Allstate Foundation online survey of 605 drivers ages 16–20. (6/16/09)

14

WHAT IS STATISTICS?

- **Statistics is all around us!**
- **How often do you eat fruit?**

Postyour.info is a worldwide service where Internet users from arround the world can take part in questionnaires. [http://postyour.info/] Below is a graph depicting the combined summary of how users answered one of the posted questions. Results given in Percent (count).



15

WHAT IS STATISTICS?

- **Statistics has become the universal language of the sciences. As potential users of statistics, we need to master both the "science" and the "art" of using statistical methodology correctly.**
- **Careful use of statistical methods will enable us to obtain accurate information from data. These methods include**
  - **(1) carefully defining the situation,**
  - **(2) gathering data,**
  - **(3) accurately summarizing the data, and**
  - **(4) deriving and communicating meaningful conclusions.**

16

WHAT IS STATISTICS?

MARQUETTE UNIVERSITY
Be The Difference.

- The field of statistics can be divided into two main branches.
- **Descriptive statistics** is what most people think of when they hear the word *statistics*. It includes the collection, presentation, and description of sample data.
- The term **inferential statistics** refers to the technique of interpreting the values resulting from the descriptive techniques and making decisions and drawing conclusions about the population.

17

WHAT IS STATISTICS?

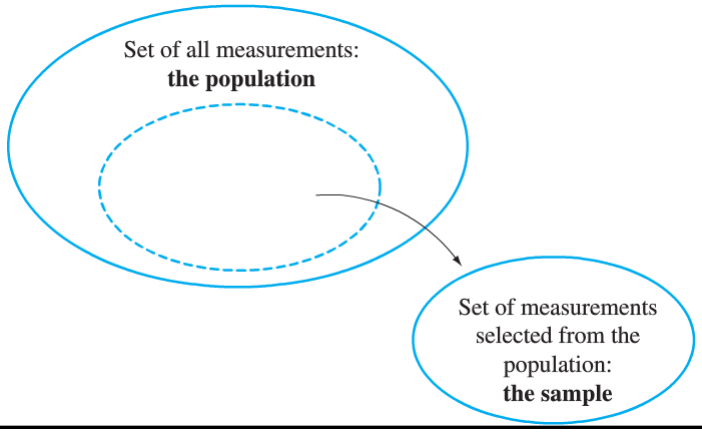
MARQUETTE UNIVERSITY
Be The Difference.

- **Statistics** The science of collecting, describing, and interpreting data.
- **Population** A collection, or set, of individuals, objects, or events whose properties are to be analyzed.
  - The set of "all students who have ever attended a U.S. college" is an example of a well-defined population.
- **Sample** A subset of a population.
- **Variable** A characteristic of interest about each individual element of a population or sample.
  - A student's age at entrance into college, the color of the student's hair, the student's height, and the student's weight are four variables.
- **Data value** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

18

POPULATION VS. SAMPLE

- **Population**: The entire group of interest
- **Sample**: A part of the population selected to draw conclusions about the entire population



19

WHAT IS STATISTICS?

- **Data** The set of values collected from the variable from each of the elements that belong to the sample.
  - The set of 25 heights collected from 25 students is an example of a set of data.
- **Experiment** A planned activity whose results yield a set of data.
  - An experiment includes the activities for both selecting the elements and obtaining the data values.
- **Parameter** A numerical value summarizing all the data of an entire population.
  - A parameter is a value that describes the entire population. Often a Greek letter is used to symbolize the name of a parameter.
- **Statistic** A numerical value summarizing the sample data.
  - Often letters of the English alphabet is used to symbolize the name of a statistic.

20

WHAT IS STATISTICS?

- **Describing a Population**
  - It is common practice to use Greek letters when talking about a population.
  - We call the mean of a population μ.
  - We call the standard deviation of a population σ and the variance σ^2.
  - It is important to know that for a given population there is only **one** true mean and **one** true standard deviation and variance or **one** true proportion.
  - There is a special name for these values: **parameters**.

21

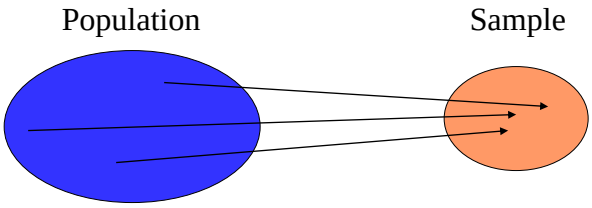
WHAT IS STATISTICS?

- **Describing a Sample**
  - We call the mean of a sample $\bar{x}$.
  - We call the standard deviation of a sample s and the variance s^2.
  - There are many different possible samples that could be taken from a given population. For each sample there may be a **different** mean, standard deviation, variance, or proportion.
  - There is a special name for these values: **statistics**.

22

POPULATION VS SAMPLE

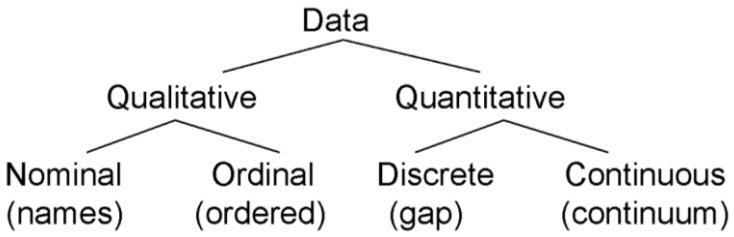
- **We use sample statistics to make inference about population parameters**



Mean:	μ	$\bar{X}$
Standard Deviation:	σ	s

WHAT IS STATISTICS?

- **Data The set of values collected from the variable from each of the elements that belong to the sample.**



WHAT IS STATISTICS?

- **Qualitative (Categorical) variable:** A variable that describes or categorizes an element of a population.
  - **Nominal variable:** A qualitative variable that characterizes an element of a population. No ordering. No arithmetic.
  - **Ordinal variable:** A qualitative variable that incorporates an ordered position, or ranking.

- **Quantitative (Numerical) variable:** A variable that quantifies an element of a population.
  - **Discrete variable:** A quantitative variable that can assume a countable number of values. Gap between successive values.
  - **Continuous variable:** A quantitative variable that can assume an uncountable number of values. Continuum of values.

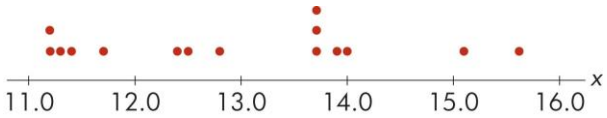
25

TYPES OF VARIABLES

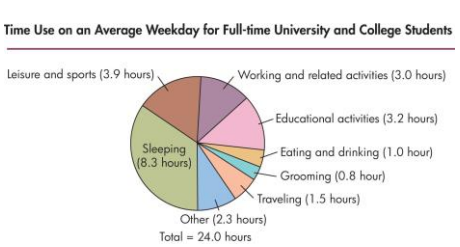
Examples:

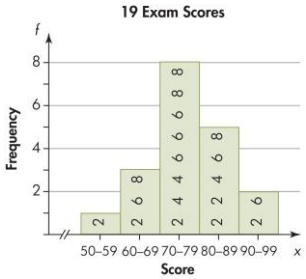
	Numeric		Categorical	
Variable	**Discrete**	**Continuous**	**Nominal**	**Ordinal**
Weight		X		
Hours Enrolled	X			
Major			X	
Zip Code			X	

26



Graphical Summaries





27

EXAMPLE 1 – GRAPHING QUALITATIVE DATA

- **Table 2.1 lists the number of cases of each type of operation performed at General Hospital last year.**

Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23
Total	498

Operations Performed at General Hospital Last Year **[TA02-01]**

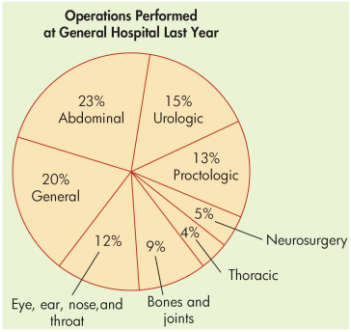
Table 2.1

28

QUALITATIVE DATA (PIE CHART)

- **Pie charts (circle graphs) and bar graphs Graphs that are used to summarize qualitative, or attribute, or categorical data.**
- **Pie charts (circle graphs) show the amount of data that belong to each category as a proportional part of a circle.**

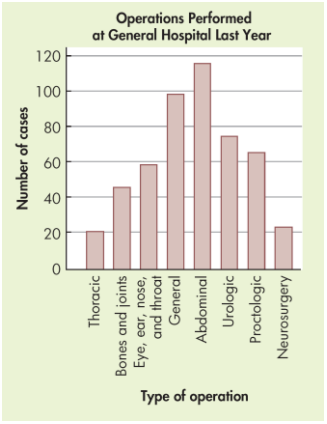
Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23
Total	498



QUALITATIVE DATA (BAR GRAPH)

- **Bar graphs show the amount of data that belong to each category as a proportionally sized rectangular area.**

Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23
Total	498



- **Bar graphs of attribute data should be drawn with a space between bars of equal width.**

EXAMPLE 2- AUSTRALIAN INSTITUTE OF SPORT DATA

MARQUETTE UNIVERSITY
Be The Difference.

- **Description**
  - Data on 102 male and 100 female athletes collected at the Australian Institute of Sport, courtesy of Richard Telford and Ross Cunningham.
- **Source**
  - Cook and Weisberg (1994), *An Introduction to Regression Graphics*. John Wiley & Sons, New York.

AIS.mjp

Variable	Description
sex	sex
sport	sport
rcc	red cell count
wcc	white cell count
Hc	Hematocrit
Hg	Hemoglobin
Fe	plasma ferritin concentration
bmi	body mass index, weight/(height)
ssf	sum of skin folds
Bfat	body fat percentage
lbm	lean body mass
Ht	height (cm)
Wt	weight (Kg)

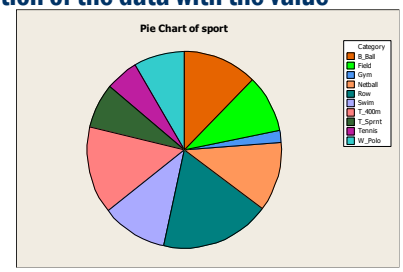
31

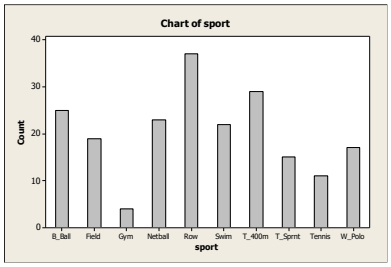
SUMMARIZING A SINGLE CATEGORICAL VARIABLE

MARQUETTE UNIVERSITY
Be The Difference.

- **Frequency** (Count) - number of times the value occurs in the data
- **Relative frequency** (Percent) - proportion of the data with the value
- **als.xls (D2L/Content/Datasets)**

sport	Count	Percent
B_Ball	25	12.38
Field	19	9.41
Gym	4	1.98
Netball	23	11.39
Row	37	18.32
Swim	22	10.89
T_400m	29	14.36
T_Sprnt	15	7.43
Tennis	11	5.45
W_Polo	17	8.42
N=	202	

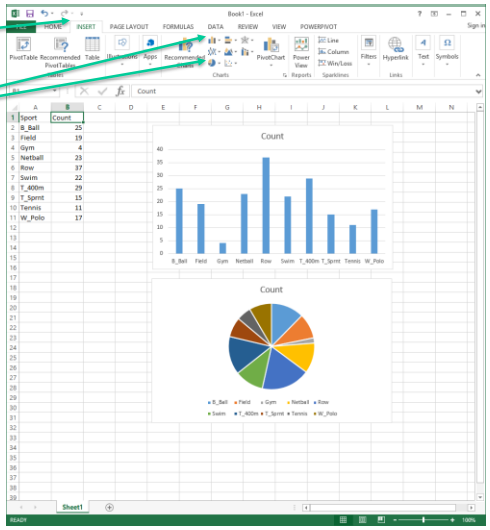




32

HOW TO?

- **Enter the Data in Excel:**
- **Select Insert**
- **Select Pie or Bar Charts**



GRAPHING QUANTITATIVE DATA

- **Distribution** The pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable.

- **Dotplot display** Displays the data of a sample by representing each data value with a dot positioned along a scale. The frequency of the values is represented along the other scale.

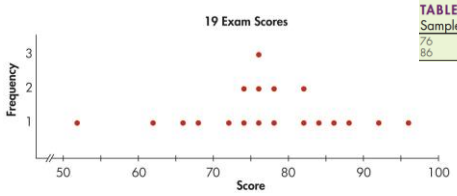


Table 2.2 provides a sample of 19 exam grades randomly selected from a large class.

TABLE 2.2
Sample of 19 Exam Grades [TA02-02]

76	74	82	96	66	76	78	72	52	68
86	84	62	76	78	92	82	74	88	

GRAPHING QUANTITATIVE DATA

MARQUETTE UNIVERSITY
Be The Difference.

- **Stem-and-leaf display** Displays the data of a sample using the actual digits that make up the data values. Each numerical value is divided into two parts: The **leading** digit(s) becomes the **stem**, and the **trailing** digit(s) becomes the **leaf**.

Table 2.2 provides a sample of 19 exam grades randomly selected from a large class.

TABLE 2.2
Sample of 19 Exam Grades [TA02-02]

76	74	82	96	66	76	78	72	52	68
86	84	62	76	78	92	82	74	88	

19 Exam Scores

5	2
6	6 8 2
7	6 4 6 8 2 6 8 4
8	2 6 4 2 8
9	6 2

Or

19 Exam Scores

5	2
6	2 6 8
7	2 4 4 6 6 6 8 8
8	2 2 4 6 8
9	2 6

EXAMPLE 3
OVERLAPPING DISTRIBUTIONS

MARQUETTE UNIVERSITY
Be The Difference.

- **A random sample of 50 college students was selected. Their weights were obtained from their medical records. The resulting data are listed in Table 2.3.**

Student	1	2	3	4	5	6	7	8	9	10
Male/Female	F	M	F	M	M	F	F	M	M	F
Weight	98	150	108	158	162	112	118	167	170	120
Student	11	12	13	14	15	16	17	18	19	20
Male/Female	M	M	M	F	F	M	F	M	M	F
Weight	177	186	191	128	135	195	137	205	190	120
Student	21	22	23	24	25	26	27	28	29	30
Male/Female	M	M	F	M	F	F	M	M	M	M
Weight	188	176	118	168	115	115	162	157	154	148
Student	31	32	33	34	35	36	37	38	39	40
Male/Female	F	M	M	F	M	F	M	F	M	M
Weight	101	143	145	108	155	110	154	116	161	165
Student	41	42	43	44	45	46	47	48	49	50
Male/Female	F	M	F	M	M	F	F	M	M	M
Weight	142	184	120	170	195	132	129	215	176	183

Weights of 50 College Students **[TA02-03]**

Table 2.3

EXAMPLE 3
OVERLAPPING DISTRIBUTIONS

- Notice that the weights range from 98 to 215 pounds. Let's group the weights on stems of 10 units using the hundreds and the tens digits as stems and the units digit as the leaf (see Figure 2.7).

- The leaves have been arranged in numerical order. Close inspection of Figure 2.7 suggests that two overlapping distributions may be involved.

Weights of 50 College Students (lb)

N = 50	Leaf Unit = 1.0
9	8
10	1 8 8
11	0 2 5 5 6 8 8
12	0 0 0 8 9
13	2 5 7
14	2 3 5 8
15	0 4 4 5 7 8
16	1 2 2 5 7 8
17	0 0 6 6 7
18	3 4 6 8
19	0 1 5 5
20	5
21	5

Stem-and-Leaf Display

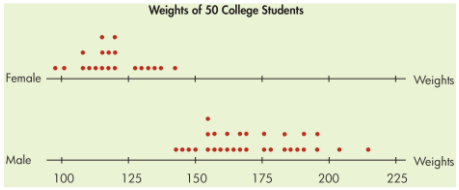
Figure 2.7

EXAMPLE 3
OVERLAPPING DISTRIBUTIONS

cont'd

- That is exactly what we have: a distribution of female weights and a distribution of male weights.

- Figure 2.8 shows a "back-to-back" stem-and-leaf display of this set of data and makes it obvious that two distinct distributions are involved.



Weights of 50 College Students (lb)

Female		Male
8	09	
1 8 8	10	
0 2 5 5 6 8 8	11	
0 0 0 8 9	12	
2 5 7	13	
2	14	3 5 8
	15	0 4 4 5 7 8
	16	1 2 2 5 7 8
	17	0 0 6 6 7
	18	3 4 6 8
	19	0 1 5 5
	20	5
	21	5

"Back-to-Back" Stem-and-Leaf Display

Figure 2.8

FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

MARQUETTE UNIVERSITY
Be The Difference.

- **Frequency distribution** A listing, often expressed in chart form, that pairs values of a variable with their frequency.

- Let's use a sample of 50 final exam scores taken from last semester's elementary statistics class.

60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	95	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

Statistics Exam Scores **[TA02-06]**

Table 2.6

39

FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

MARQUETTE UNIVERSITY
Be The Difference.

60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	95	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

1. Identify the high score ($H = 98$) and the low score ($L = 39$), and find the range:
   - range = $H - L = 98 - 39 = 59$
2. Select a number of classes ($m = 7$) and a class width ($c = 10$) so that the product ($mc = 70$) is a bit larger than the range (range $= 59$).
3. Pick a starting point. This starting point should be a little smaller than the lowest score, L.

40

FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

MARQUETTE UNIVERSITY
Be The Difference.

60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	95	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

- **Let the starting point to be 35. Given class width ($c = 10$)**

Class Number	Class Tallies	Boundaries	Frequency
1	||	$35 \le x < 45$	2
2	||	$45 \le x < 55$	2
3	|||| ||	$55 \le x < 65$	7
4	|||| |||| |||	$65 \le x < 75$	13
5	|||| |||| |	$75 \le x < 85$	11
6	|||| |||| |	$85 \le x < 95$	11
7	||||	$95 \le x \le 105$	4
			50

Standard Chart for Frequency Distribution

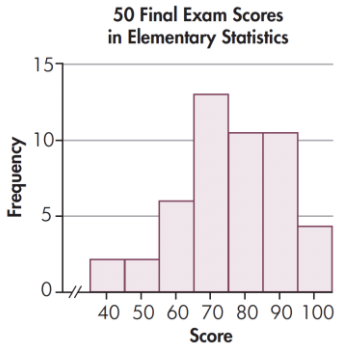
41

FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

MARQUETTE UNIVERSITY
Be The Difference.

Class Number	Class Tallies	Boundaries	Frequency
1	||	$35 \le x < 45$	2
2	||	$45 \le x < 55$	2
3	|||| ||	$55 \le x < 65$	7
4	|||| |||| |||	$65 \le x < 75$	13
5	|||| |||| |	$75 \le x < 85$	11
6	|||| |||| |	$85 \le x < 95$	11
7	||||	$95 \le x \le 105$	4
			50

- **Histogram A bar graph that represents a frequency distribution of a quantitative variable.**



42

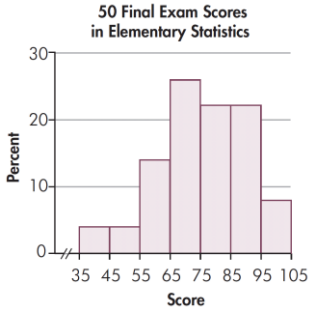
FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

MARQUETTE UNIVERSITY
Be The Difference.

Class Number	Class Tallies	Boundaries	Frequency	Percentage
1	||	$35 \le x < 45$	2	4%
2	||	$45 \le x < 55$	2	4%
3	|||| ||	$55 \le x < 65$	7	14%
4	|||| |||| |||	$65 \le x < 75$	13	26%
5	|||| |||| |	$75 \le x < 85$	11	22%
6	|||| |||| |	$85 \le x < 95$	11	22%
7	||||	$95 \le x \le 105$	4	8%
			50	100%

Divide all by 50

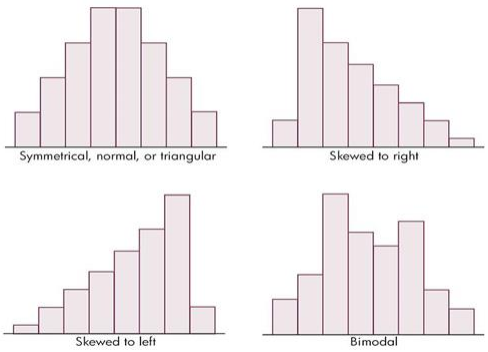
- The **relative frequency (percentage)** is a proportional measure of the frequency for an occurrence.
- It is found by dividing the class frequency by the total number of observations.



43

FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

MARQUETTE UNIVERSITY
Be The Difference.



- **Symmetrical** Both sides of this distribution are identical (halves are mirror images).
- **Skewed** One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail.

44

FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

MARQUETTE UNIVERSITY
Be The Difference.

Bimodal **The two most populous classes are separated by one or more classes. This situation often implies that two populations are being sampled. (See Figure 2.7)**

Weights of 50 College Students (lb)

$N = 50$	Leaf Unit = 1.0
9	8
10	1 8 8
11	0 2 5 5 6 8 8
12	0 0 0 8 9
13	2 5 7
14	2 3 5 8
15	0 4 4 5 7 8
16	1 2 2 5 7 8
17	0 0 6 6 7
18	3 4 6 8
19	0 1 5 5
20	5
21	5

Stem-and-Leaf Display

Figure 2.7

MARQUETTE UNIVERSITY
Be The Difference.

$$\text{midrange} = \frac{\text{low value} + \text{high value}}{2}$$

$$\text{midrange} = \frac{L + H}{2} \qquad (2.3)$$

$$\text{Sample mean: } \text{x-bar} = \frac{\text{sum of all } x}{\text{number of } x}$$

$$\bar{x} = \frac{\Sigma x}{n} \qquad (2.1)$$

Numerical Summaries

$$s \text{ squared} = \frac{(\text{sum of } x^2) - \left[\frac{(\text{sum of } x)^2}{\text{number}}\right]}{\text{number} - 1}$$

$$\text{sample variance: } s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1} \qquad (2.9)$$

$$\text{sample variance: } s \text{ squared} = \frac{\text{sum of (deviations squared)}}{\text{number} - 1}$$

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} \qquad (2.5)$$

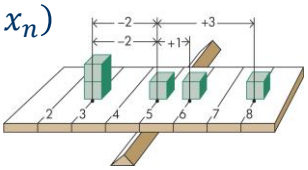
MEASURES OF CENTRAL TENDENCY

MARQUETTE UNIVERSITY
Be The Difference.

- The measures of central tendency characterize the center of the distribution of data values. The term *average* is often associated with all measures of central tendency.
- **Mean (arithmetic mean)** The average with which you are probably most familiar. The sample mean is represented by $\bar{x}$ (read "*x*-bar" or "sample mean").

$$\text{Sample mean: } x\text{-bar} = \frac{\text{sum of all } x}{\text{number of } x}$$

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$
- The center of gravity or balance point

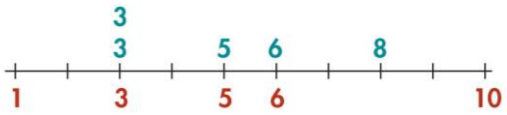


47

MEASURES OF CENTRAL TENDENCY

MARQUETTE UNIVERSITY
Be The Difference.

- **Sample Median:** Middle value when data ordered. 50% above, 50% below. Represented by $\tilde{x}$ called "*x-tilde.*"
  - Order data from smallest to largest.
  - If n odd, $\tilde{x}$ = middle value
  - If n even, $\tilde{x}$ = average of middle two values
- **Sample Mode:** The value that happens most often in sample.
- Represented by $\hat{x}$ called "*x-hat.*"
  - If two or more values in a sample are tied for the highest frequency, we say that there is **no mode**.



- There are other measures called **measures of dispersion** that characterize the spread or variability in the data.

48

MEASURES OF DISPERSION

MARQUETTE UNIVERSITY
Be The Difference.

- **Range** The difference in value between the highest-valued data, ***H***, and the lowest-valued data, ***L***:
  - ***Range = high value – low value =*** $H - L$
- **Deviation from the mean**: The difference between the data value x_i and the sample mean $\bar{x}$
  - i^{th} deviation from the mean = $x_i - \bar{x}$
- The sum of the deviations, $\sum_{i=1}^{n}(x_i - \bar{x})$ is always zero because the deviations of x_i values smaller than the mean (which are negative) cancel out those x_i values larger than the mean (which are positive).

MEASURES OF DISPERSION

MARQUETTE UNIVERSITY
Be The Difference.

- **Sample Variance**: The mean of the squared deviations using $n - 1$ as a divisor.

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

  - where x_i is the i^{th} data value, $\bar{x}$ is the sample mean, and n is the sample size.

$SS(x)$: *sum of squares for x*

- This is equivalent to:

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \left[\frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]\right\}$$

- Therefore

$$s^2 = \frac{1}{n-1}SS(x)$$

MEASURES OF DISPERSION

- **Sample Standard Deviation: Square root of the sample variance.**
  - **Has same units data values and sample mean.**

$$s = \sqrt{s^2}, \qquad \text{where} \qquad s^2 = \frac{1}{n-1} SS(x)$$

- **Example: Consider a second set of data: $\{6, 3, 8, 5, 2\}$. Find the followings:**
- **Measures of Central Tendency**
  - **Mean** $\bar{x} = \frac{1}{5}(6 + 3 + 8 + 5 + 2) = 4.8$
  - **Meadian** $\tilde{x} =$ middle value $= 5$
  - **Mode** $\hat{x} =$ the value with the highest count $\Rightarrow$ There is no mode
- **Measures of Dispersion**
  - **Range** range $= H - L = 8 - 2 = 6$
  - **Sample Variance and Sample Standard Deviation**

51

EXAMPLE (SAMPLE VARIANCE)

- **Consider a second set of data: $\{6, 3, 8, 5, 2\}$**
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Step 1. Find Σx	Step 2. Find $\bar{x}$	Step 3. Find each $x - \bar{x}$	Step 4. Find $\Sigma(x - \bar{x})^2$	Step 5. Find s^2
6	$\bar{x} = \frac{\Sigma x}{n}$	$6 - 4.8 = 1.2$	$(1.2)^2 = 1.44$	$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$
3		$3 - 4.8 = -1.8$	$(-1.8)^2 = 3.24$	
8		$8 - 4.8 = 3.2$	$(3.2)^2 = 10.24$	
5	$\bar{x} = \frac{24}{5}$	$5 - 4.8 = 0.2$	$(0.2)^2 = 0.04$	$s^2 = \frac{22.80}{4}$
2		$2 - 4.8 = -2.8$	$(-2.8)^2 = 7.84$	
$\Sigma x = 24$	$\bar{x} = 4.8$	$\Sigma(x - \bar{x}) = 0$	$\Sigma(x - \bar{x})^2 = 22.80$	$s^2 = 5.7$

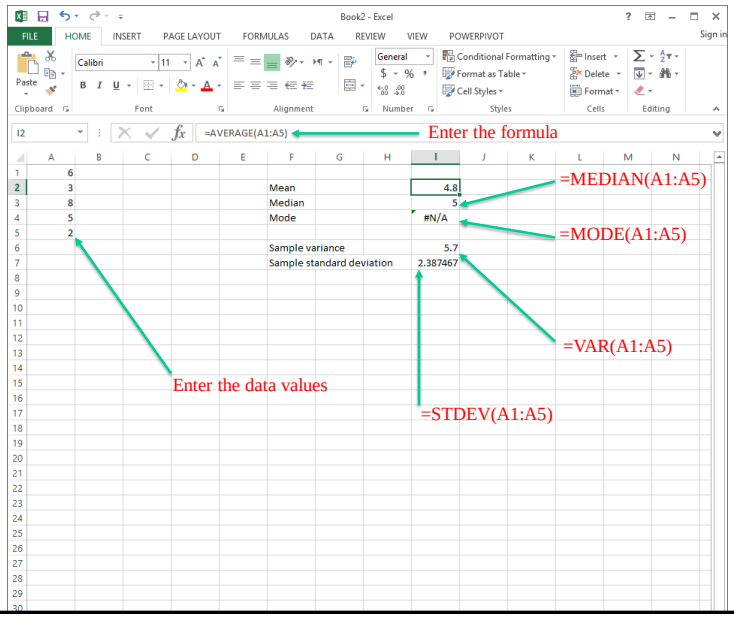
- **Or** $s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right\}$

$$s = \sqrt{s^2} = \sqrt{5.7}$$

Step 1. Find Σx	Step 2. Find Σx^2	Step 3. Find SS(x)	Step 4. Find s^2	Step 5. Find s
6	$6^2 = 36$	$SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$		$s = \sqrt{s^2}$
3	$3^2 = 9$		$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$	$s = \sqrt{5.7}$
8	$8^2 = 64$	$SS(x) = 138 - \frac{(24)^2}{5}$		$s = 2.4$
5	$5^2 = 25$		$s^2 = \frac{22.8}{4}$	
2	$2^2 = 4$	$SS(x) = 138 - 115.2$		
$\Sigma x = 24$	$\Sigma x^2 = 138$	$SS(x) = 22.8$	$s^2 = 5.7$	

52

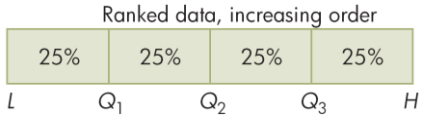
EXAMPLE IN MICROSOFT EXCEL



MEASURES OF POSITION

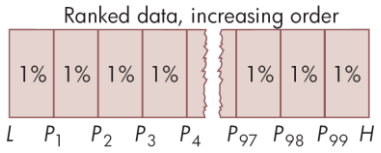
- **Measures of position** are used to describe the position a specific data value possesses in relation to the rest of the data when in ranked order. ***Quartiles*** and ***percentiles*** are two of the most popular measures of position.

- **Quartiles** Values of the variable that divide the ranked data into quarters; each set of data has three quartiles.
  - L = lowest value
  - Q_1 = data value where 25% are smaller
  - Q_2 = $\tilde{X}$ =median
  - Q_3 = data value where 75% are smaller
  - H = highest value



MEASURES OF POSITION

- **Percentiles Values of the variable that divide a set of ranked data into 100 equal subsets.**



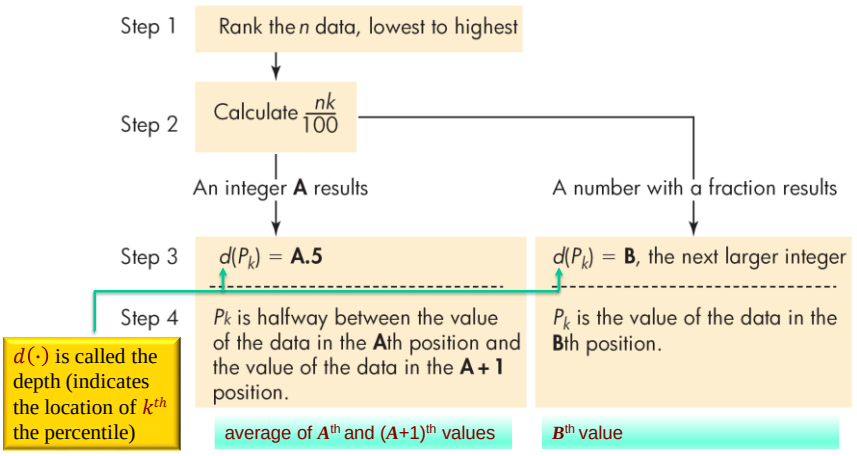
- **The k^{th} percentile,**
  - **P_k = value where $k\%$ are smaller**

- **Quartiles are special percentiles.**

55

MEASURES OF POSITION

- **The Percentile Process: Finding k^{th} percentile.**



56

EXAMPLE 12 - FINDING QUARTILES AND PERCENTILES

- **Using the sample of 50 elementary statistics final exam scores listed in Table 2.15, find the first quartile, Q_1; the 58^{th} percentile, P_{58}.**

60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	95	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

Raw Scores for Elementary Statistics Exam **[TA02-06]**
Table 2.15

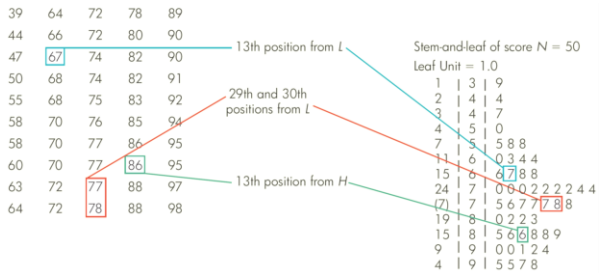
- **Step 1:**
  - **Rank the data from lowest to highest**

39	64	72	78	89
44	66	72	80	90
47	67	74	82	90
50	68	74	82	91
55	68	75	83	92
58	70	76	85	94
58	70	77	86	95
60	70	77	86	95
63	72	77	88	97
64	72	78	88	98

EXAMPLE 12 – *SOLUTION*

- **Find Q_1:**
  - **Step 2: Find** $\frac{nk}{100} : \frac{nk}{100} = \frac{(50)(25)}{100} = 12.5$
  - **Step 3: B is the next larger integer,** 13.
  - **Step 4: Find Q_1: Q_1 is the 13^{th} value, $Q_1 = 67$**

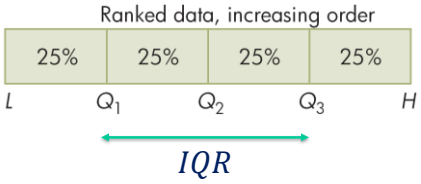
- **Find P_{58}:**
  - **Step 2: Find** $\frac{nk}{100} : \frac{nk}{100} = \frac{(50)(58)}{100} = 29$
  - **Step 3: Since $A = 29$, an integer, add 0.5 and use 29.5.**
  - **Step 4: Find P_{58}: P_{58} is the average of 29^{th} and 30^{th} values, $P_{58} = 77.5$**



MEASURES OF POSITION

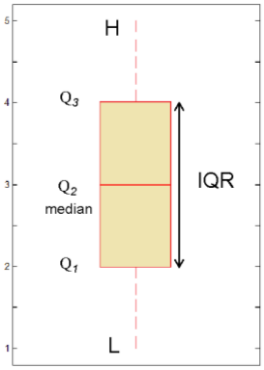
- **Five Number Summary**
  - L = lowest value
  - Q_1= data value where 25% are smaller
  - Q_2 = $\tilde{x}$ =median
  - Q_3 = data value where 75% are smaller
  - H = highest value

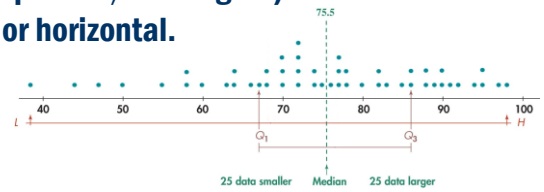
- **Interquartile range: The difference between the first and third quartiles. It is the range of the middle 50% of the data.**
  - $IQR = Q_3 - Q_1$

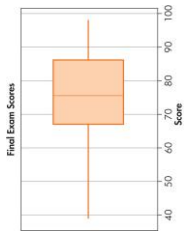


MEASURES OF POSITION

Box-and-whiskers display A graphic representation of the 5-number summary. The five numerical values (smallest, first quartile, median, third quartile, and largest) are located on a scale, either vertical or horizontal.

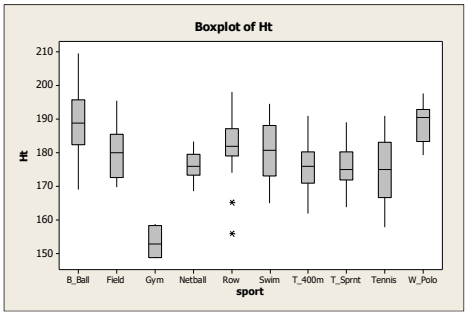


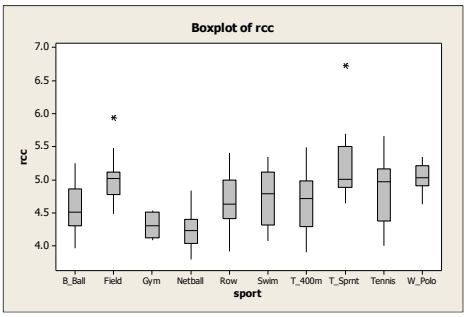




SIDE-BY-SIDE BOX PLOT FOR AIS DATA (AUSTRALIAN INSTITUTE OF SPORT)

MARQUETTE UNIVERSITY
Be The Difference.





61

MEASURES OF POSITION

MARQUETTE UNIVERSITY
Be The Difference.

- The position of a specific value can also be measured in terms of the mean and standard deviation using the ***standard score***, commonly called the ***z-score***.

- **Standard score, or *z*-score** The position a particular value of x has relative to the mean, measured in standard deviations. The z-score is found by the formula

$$z = \frac{\text{value} - \text{mean}}{\text{st.dev.}} = \frac{x - \bar{x}}{s} \qquad (2.11)$$

- **Example: Find the standard scores for (a)** 92 **and (b)** 72 **with respect to a sample of exam grades that have a mean score of** 74.92 **and a standard deviation of** 14.20**.**

62

EXAMPLE 14 – FINDING Z-SCORES

- **a.** $x_1 = 92,\ \bar{x} = 74.92, s = 14.20$. **Thus,**

$$z = \frac{x - \bar{x}}{s} = \frac{92 - 74.92}{14.20} = \frac{17.08}{14.20} = 1.20.$$

- **b.** $x_2 = 72,\ \bar{x} = 74.92, s = 14.20$. **Thus,**

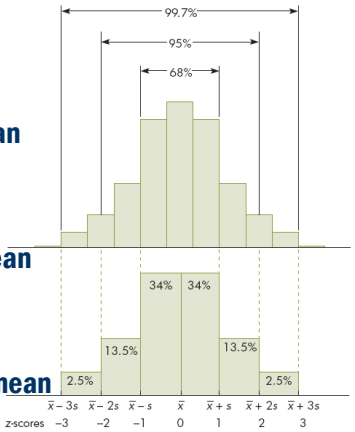
$$z = \frac{x - \bar{x}}{s} = \frac{72 - 74.92}{14.20} = \frac{-2.92}{14.20} = -0.21.$$

- **This means that the score 92 is approximately 1.2 standard deviations above the mean and**
- **that the score 72 is approximately one-fifth of a standard deviation below the mean.**

63

EMPIRICAL RULE (THE 68-95-99.7 RULE)

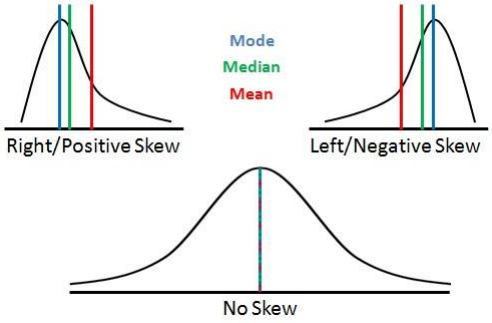
- **If the distribution is mound-shaped, then**
  - **Approximately 68% of the data falls within one standard deviation of the mean**
  - **Approximately 95% of the data falls within two standard deviations of the mean**
  - **Approximately 99.7% of the data falls within three standard deviations of the mean**



64

COMPARING MEASURES OF CENTER AND SPREAD

- The **sample mean** and the **sample standard deviation** are good measures of center and spread, respectively, for **symmetric** data
- If the data set is **skewed** or has **outliers**, the **sample median** and the **interquartile range** are more commonly used
- **Mean versus median**



65

BIVARIATE DATA

- **Bivariate data** The values of two different variables that are obtained from the same population element.

- Each of the two variables may be either *qualitative* or *quantitative*. As a result, three combinations of variable types can form bivariate data:

1. Both variables are qualitative (categorical).
2. One variable is qualitative (categorical), and the other is quantitative (numerical).
3. Both variables are quantitative (numerical).

66

TWO QUALITATIVE VARIABLES

- **Qualitative vs Qualitative:**
  - **A cross-tabulation or contingency table will be used**

- **Example: Thirty students from our college were randomly identified and classified according to two variables: gender (M/F) and major (liberal arts (LA), business administration (BA), technology(T))**

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

67

EXAMPLE 1 – CONSTRUCTING CROSS-TABULATION TABLES

- **We can construct a 2×3 table.**

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

- **Given:**
  M = male
  F = female
  LA = liberal arts
  BA = business admin
  T = technology

- **The entry in each cell is found by determining how many students fit into each category. Adams is male (M) and liberal arts (LA) and is classified in the cell in the first row, first column.**

	Major		
Gender	LA	BA	T
M	|||| (5)	|||| | (6)	|||| || (7)
F	|||| | (6)	||| (4)	|| (2)

68

EXAMPLE 1 CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

Gender	Major LA	BA	T
M	|||| (5)	|||| | (6)	|||| || (7)
F	|||| | (6)	|||| (4)	|| (2)

- **The resulting 2 × 3 contingency (cross-tabulation) table is:**

Gender	Major LA	BA	T	Row Total
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

- **Percentages Based on the Grand Total (Entire Sample)**

Gender	Major LA	BA	T	Row Total
M	17%	20%	23%	60%
F	20%	13%	7%	40%
Col. Total	37%	33%	30%	100%

- **Percentages Based on Row Totals**
- **(Marginal: within Gender)**

Gender	Major LA	BA	T	Row Total
M	28%	33%	39%	100%
F	50%	33%	17%	100%
Col. Total	37%	33%	30%	100%

Percentages Based on Column Totals
(Marginal: within Major)

Gender	Major LA	BA	T	Row Total
M	45%	60%	78%	60%
F	55%	40%	22%	40%
Col. Total	100%	100%	100%	100%

69

QUANTITATIVE VS QUALITATIVE: SIDE-BY-SIDE COMPARISONS

MARQUETTE UNIVERSITY
Be The Difference.

- **Quantitative vs Qualitative:**
- **Example 2: The distance required to stop a 3000-pound automobile on wet pavement was measured to compare the stopping capabilities of three tire tread designs.**

Design A ($n = 6$)	Design B ($n = 6$)	Design C ($n = 6$)
37 36 38	33 35 38	40 39 40
34 40 32	34 42 34	41 41 43

- **5-Number Summary for Each Design**

	Design A	Design B	Design C
High	40	42	43
Q_3	38	38	41
Median	36.5	34.5	40.5
Q_1	34	34	40
Low	32	33	39

Mean and Standard Deviation for Each Design

	Design A	Design B	Design C
Mean	36.2	36.0	40.7
Standard deviation	2.9	3.4	1.4

70

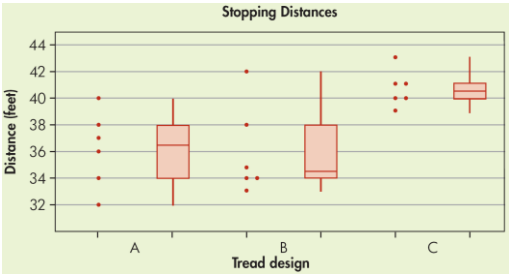
EXAMPLE 2 – CONSTRUCTING SIDE-BY-SIDE COMPARISONS

MARQUETTE UNIVERSITY
Be The Difference.

- **The distance required to stop a 3000-pound automobile on wet pavement was measured to compare the stopping capabilities of three tire tread designs**

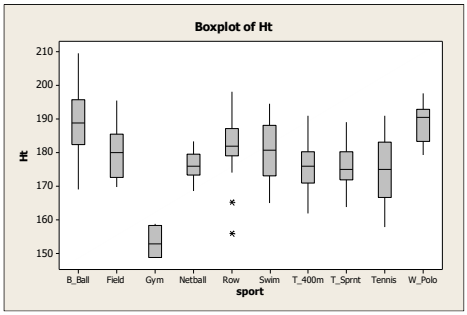
Design A ($n = 6$)	Design B ($n = 6$)	Design C ($n = 6$)
37 36 38	33 35 38	40 39 40
34 40 32	34 42 34	41 41 43

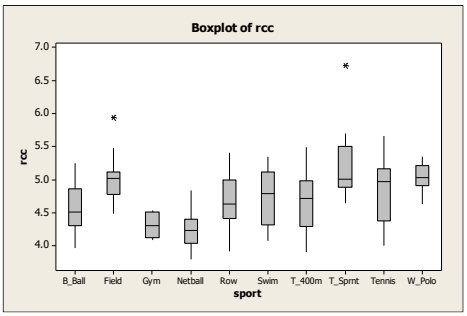
- **Side by side Box -and-Whiskers**



SIDE-BY-SIDE BOX PLOT FOR AIS DATA (AUSTRALIAN INSTITUTE OF SPORT)

MARQUETTE UNIVERSITY
Be The Difference.





TWO QUANTITATIVE VARIABLES

MARQUETTE UNIVERSITY
Be The Difference.

- It is customary to express the data mathematically as **ordered pairs** (x, y), where
  - x is the **input variable** (called the **independent variable**)
  - y is the **output variable** (called the **dependent variable**).
- The data are said to be ***ordered*** because one value, x, is always written first.
- They are called ***paired*** because for each x value, there is a corresponding y value from the same source.

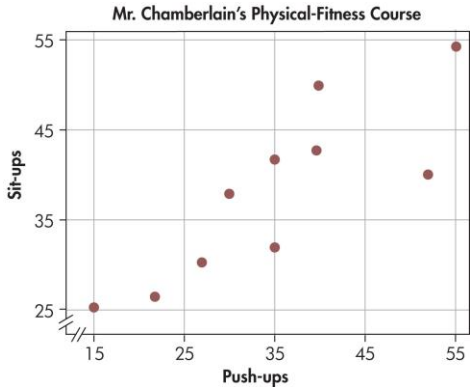
- **Scatter diagram** A plot of all the ordered pairs of bivariate data on a coordinate axis system. The input variable, x, is plotted on the horizontal axis, and the output variable, y, is plotted on the vertical axis.

TWO QUANTITATIVE VARIABLES

MARQUETTE UNIVERSITY
Be The Difference.

- Example: Push-ups vs Sit-ups

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43



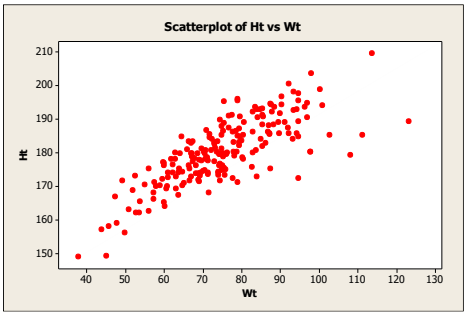
TWO QUANTITATIVE VARIABLES

MARQUETTE UNIVERSITY
Be The Difference.

- There is not always an explanatory-response (dependent-independent) relationship.

- More examples:
  - Height and Weight
  - Income and Age
  - SAT scores on math exam and on verbal exam
  - Amount of time spent studying for an exam and exam score

Australian Institute of Sport (AIS.xlsx)



75

TWO QUANTITATIVE VARIABLES

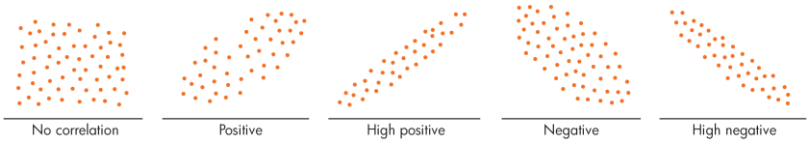
MARQUETTE UNIVERSITY
Be The Difference.

- Why Scatterplots?
- Look for overall pattern and any striking deviations from that pattern.
- Look for outliers, values falling outside the overall pattern of the relationship
- You can describe the overall pattern of a scatterplot by the form, direction, and strength of the relationship.
  - Form: Linear or clusters
  - Direction
    - Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and likewise below-average values also tend to occur together.
    - Two variables are **negatively associated** when above-average values of one variable accompany below-average values of the other variable, and vice-versa.
  - Strength-how close the points lie to a line

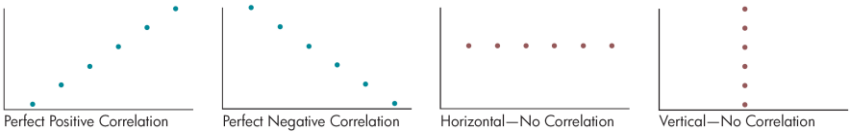
76

LINEAR CORRELATION

- **Linear Correlation, r, is a measure of the strength of a linear relationship between two variables x and y.**
- $-1 \leq r \leq 1$ **Will discuss the definition in a minute**



- $r \approx 0$ $r \approx 0.5$ $r \approx 0.8$ $r \approx -0.5$ $r \approx -0.8$

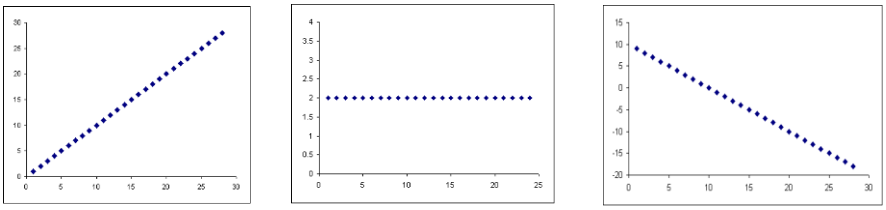


77

LINEAR CORRELATION (FORMULA 1)

- $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y}$, **where**
  - s_x **and** s_y**: are the standard deviations of** x**'s and** y**'s**
  - $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
  - $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$
- **Examples of extreme cases**

$r = 1$ $r = 0$ $r = -1$



78

LINEAR CORRELATION (FORMULA 2)

- $r = \dfrac{SS(xy)}{\sqrt{SS(x)SS(y)}}$ **(*), where**
  - $SS(x) = \sum_{i=1}^{n} x_i{}^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$
  - $SS(y) = \sum_{i=1}^{n} y_i{}^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2$
  - $SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_i x_i)(\sum_i y_i)$

OR:

- $SS(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- $SS(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $SS(xy) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

- **(*) is equivalent to the first formula:** $r = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$
- **Example 5: In Mr. Chamberlain's physical-fitness course, several fitness scores were taken. The following sample is the numbers of push-ups and sit-ups done by 10 randomly selected students:**

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

79

EXAMPLE 5 – *SOLUTION*

- **Find the linear correlation coefficient for the push-up/sit-up data.**

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1,225	42	1,764	1,470
5	30	900	38	1,444	1,140
6	52	2,704	40	1,600	2,080
7	35	1,225	32	1,024	1,120
8	55	3,025	54	2,916	2,970
9	40	1,600	50	2,500	2,000
10	40	1,600	43	1,849	1,720
	$\Sigma x = 351$ sum of x	$\Sigma x^2 = 13{,}717$ sum of x^2	$\Sigma y = 380$ sum of y	$\Sigma y^2 = 15{,}298$ sum of y^2	$\Sigma xy = 14{,}257$ sum of xy

$$SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 13{,}717 - \frac{(351)^2}{10} = 1396.9$$

$$SS(y) = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 15{,}298 - \frac{(380)^2}{10} = 858.0$$

$$SS(xy) = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 14{,}257 - \frac{(351)(380)}{10} = 919.00$$

80

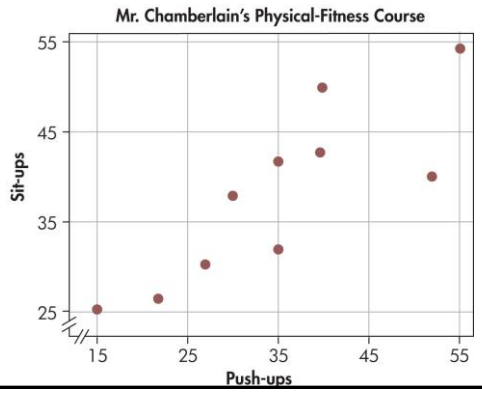
EXAMPLE 5 – *SOLUTION*

MARQUETTE UNIVERSITY
Be The Difference.

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

$$= \frac{919.0}{\sqrt{(1396.9)(858.0)}}$$

$$= 0.8394 = \mathbf{0.84}$$



RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

MARQUETTE UNIVERSITY
Be The Difference.

- **Correlation or *r* : measures the direction and strength of the linear relationship between two numeric variables**

- **General Properties**
  - **It must be between -1 and 1, or ($-1 \le r \le 1$).**
  - **If *r* is negative, the relationship is negative.**
  - **If $r = -1$, there is a perfect negative linear relationship (extreme case).**
  - **If *r* is positive, the relationship is positive.**
  - **If $r = 1$, there is a perfect positive linear relationship (extreme case).**
  - **If *r* is 0, there is no linear relationship.**
  - ***r* measures the strength of the linear relationship.**

- **Correlation Applet**

CAUSATION AND LURKING VARIABLES

MARQUETTE UNIVERSITY
Be The Difference.

- The **cause-and-effect relationship:** Correlation does not necessarily imply causation. Just because two things are highly related does not mean that one causes the other.
- A perceived relationship between a dependent(response) variable and an independent(explanatory) variable that has been misestimated due to the failure to account for a confounding factor (lurking variable) is termed a spurious relationship
- Examples of spurious relationship

83

LURKING VARIABLE AND SIMPSON'S PARADOX

MARQUETTE UNIVERSITY
Be The Difference.

- **Lurking variable:** A variable that is not included in a study but has an effect on the variables of the study and makes it appear that those variables are related.
- **Simpson's Paradox:** An association or comparison that holds for all of several groups can reverse direction when a lurking variable is present.
- Example: **Kidney stone treatment**(Br Med J (Clin Res Ed) 292 (6524): 879–882)

	Treatment A	Treatment B
Small Stones	*Group 1* 93% (81/87)	*Group 2* 87% (234/270)
Large Stones	*Group 3* 73% (192/263)	*Group 4* 69% (55/80)
Both	78% (273/350)	83% (289/350)

- http://en.wikipedia.org/wiki/Simpson's_Paradox

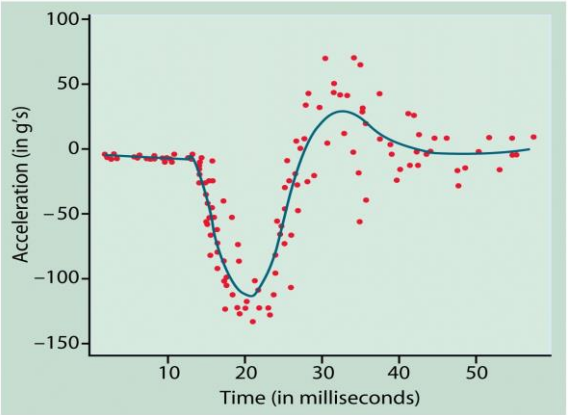
84

RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

MARQUETTE UNIVERSITY
Be The Difference.

It is possible for there to be a strong relationship between two variables and still have $r \approx 0$.

EX.



LINEAR REGRESSION

MARQUETTE UNIVERSITY
Be The Difference.

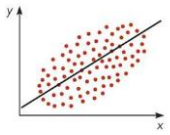
- **Regression analysis finds the equation of the line that best describes the relationship between two variables.**

- **Here are some examples of various possible relationships, called *models* or prediction equations:**
  - **Linear (straight-line):** $\hat{y} = b_0 + b_1 x$ **What we cover in this book**
  - **Quadratic:** $\hat{y} = a + bx + cx^2$
  - **Exponential:** $\hat{y} = a(b^x)$
  - **Logarithmic:** $\hat{y} = a \log_b x$

LINEAR REGRESSION

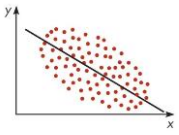
MARQUETTE UNIVERSITY
Be The Difference.

- **Suppose that $\hat{y} = b_0 + b_1 x$ is the equation of a straight line, where $\hat{y}$ (read "y-hat") represents the predicted value of y that corresponds to a particular value of x.**



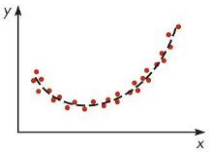
Linear Regression with Positive Slope

Figure 3.17



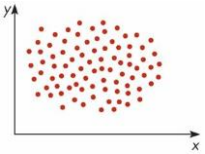
Linear Regression with Negative Slope

Figure 3.18



Curvilinear Regression (Quadratic)

Figure 3.19



No Relationship

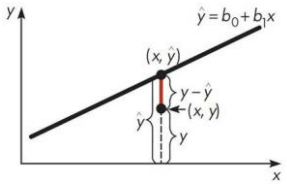
Figure 3.20

87

LINEAR REGRESSION

MARQUETTE UNIVERSITY
Be The Difference.

- **The least squares criterion requires that we find the constants b_0 and b_1 such that $\sum_{i=1}^{n}(y_i - \hat{y})^2$ is as small as possible.**



Observed and Predicted Values of *y*

- **The length of this distance represents the value $(y_i - \hat{y})$ (shown as the red line segment in the Figure. We call it residual).**
- **Note that $(y_i - \hat{y})$ is positive when the point (x, y) is above the line and negative when (x, y) is below the line**

88

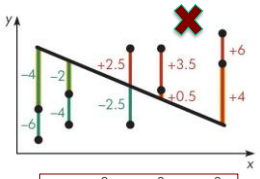
LINEAR REGRESSION

- **Figure 3.22 shows a scatter diagram with what appears to be the line of best fit, along with 10 individual $(y_i - \hat{y})$ values. (Positive values are shown in red; negative, in green.)**

$\Sigma (y - \hat{y})^2 = (-1)^2 + (+1)^2 + \ldots + (+1)^2 = 23.0$

The Line of Best Fit

Figure 3.22



$\Sigma (y - \hat{y})^2 = (-6)^2 + (-4)^2 + \ldots + (+6)^2 = 149.0$

Not the Line of Best Fit

Figure 3.23

- **Figure 3.23 shows the same data points as Figure 3.22. The 10 individual values $(y_i - \hat{y})$ are plotted with a line that is definitely not the line of best fit.**

89

LINEAR REGRESSION

- **Our job is to find the one line that will make $\sum_{i=1}^{n}(y_i - \hat{y})^2$ the smallest possible value.**

- **The equation of the line of best fit is determined by its slope (b_1) and its y-intercept (b_0).**

- **Slope:** $b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SS(xy)}{SS(x)}$, **where**

  - $SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_i x_i)(\sum_i y_i)$
  - $SS(x) = \sum_{i=1}^{n} {x_i}^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$

- **y-intercept:** $b_0 = \bar{y} - b_1 \bar{x}$

90

LINEAR REGRESSION

- **Example-5: push-up/sit-up data.**

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1,225	42	1,764	1,470
5	30	900	38	1,444	1,140
6	52	2,704	40	1,600	2,080
7	35	1,225	32	1,024	1,120
8	55	3,025	54	2,916	2,970
9	40	1,600	50	2,500	2,000
10	40	1,600	43	1,849	1,720
	$\Sigma x = 351$ sum of x	$\Sigma x^2 = 13{,}717$ sum of x^2	$\Sigma y = 380$ sum of y	$\Sigma y^2 = 15{,}298$ sum of y^2	$\Sigma xy = 14{,}257$ sum of xy

$$SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 13{,}717 - \frac{(351)^2}{10} = 1396.9$$

$$SS(xy) = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 14{,}257 - \frac{(351)(380)}{10} = 919.00$$

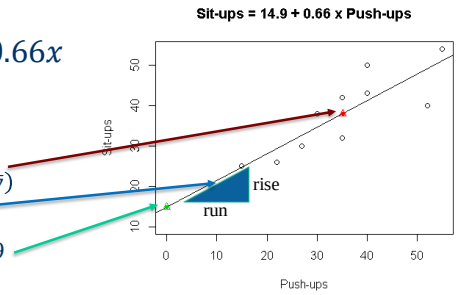
- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{351}{10} = 35.1$
- $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{380}{10} = 38$

LINEAR REGRESSION (EXAMPLE)

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

- Give $\bar{x} = 35.1$, $\bar{y} = 38$, $SS(x) = 1396.9$ and $SS(xy) = 919$

- **We want to find the line of best fit, $\hat{y} = b_0 + b_1 x$, where**
- $b_1 = \frac{SS(xy)}{SS(x)} = \frac{919}{1396.9} = 0.6579 = 0.66$
- $b_0 = \bar{y} - b_1 \bar{x} = 38 - 0.6579 \times 35.1 = 14.9077 = 14.9$
- **Therefore**
- $\hat{y} = 14.9 + 0.66x$



- **Important Notes:**
  - **The line goes through $(\bar{x}, \bar{y})$**
  - **The slope: $b_1 = 0.66$**
  - **The y-intercept: $b_0 = 14.9$**

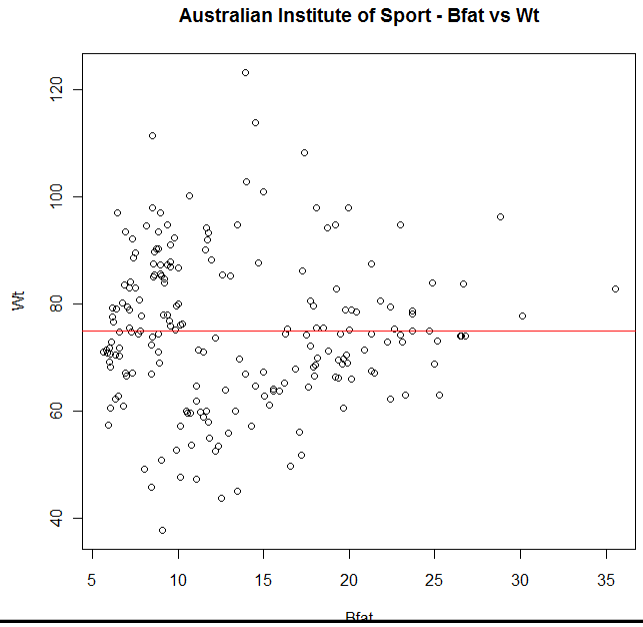
EXAMPLE – AUSTRALIAN INSTITUTE OF SPORT

- **Data on 102 male and 100 female athletes collected at the Australian Institute of Sport, (courtesy of Richard Telford and Ross Cunningham.)**

		Gender	Bfat	Wt
•	1	female	19.75	78.9
•	2	female	21.30	74.4
•	3	female	19.88	69.1
•	4	female	23.66	74.9
•	5	female	17.64	64.6
		:	:	:
•	198	male	11.79	93.2
•	199	male	10.05	80.0
•	200	male	8.51	73.8
•	201	male	11.50	71.1
•	202	male	6.26	76.7

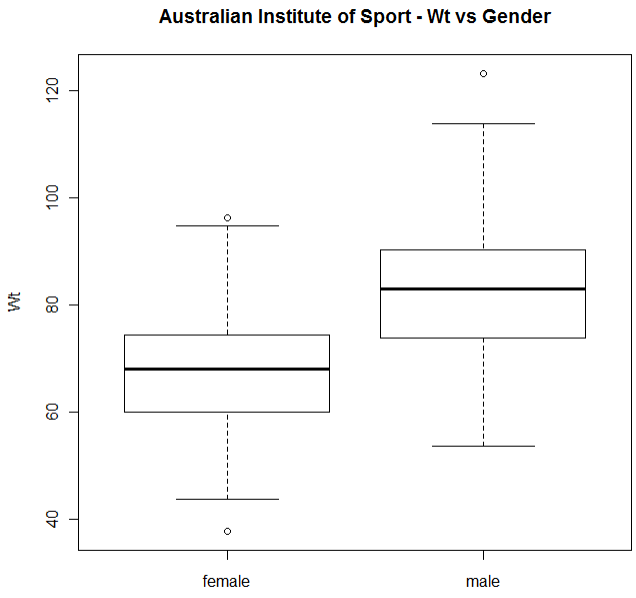
93

EXAMPLE CONT'D



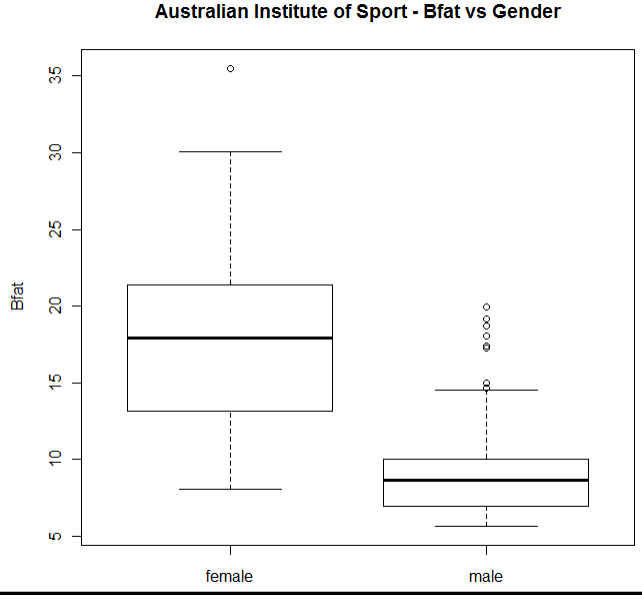
94

EXAMPLE CONT'D



95

EXAMPLE CONT'D



96

EXAMPLE CONT'D