MATH 1700

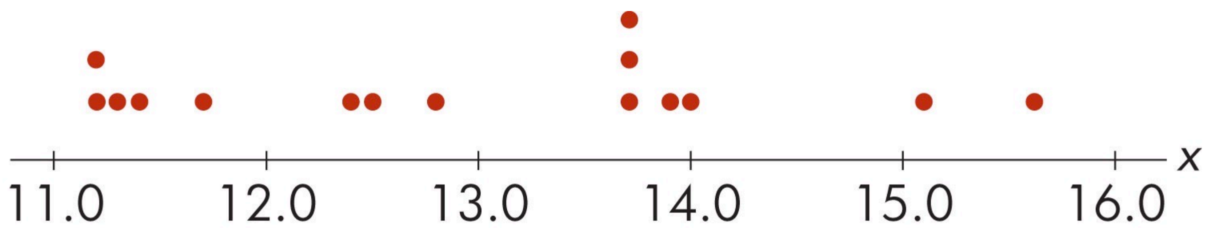
Instructor: Mehdi Maadooliat

Chapter 2A

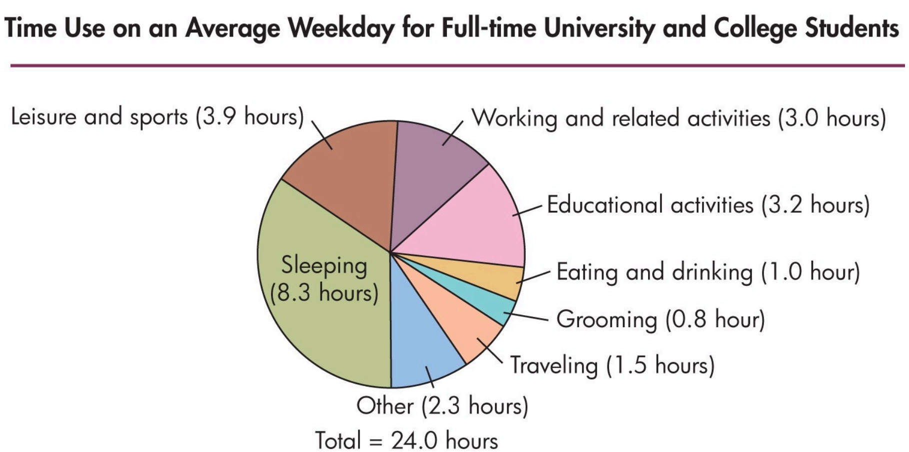
Department of Mathematical and Statistical Sciences

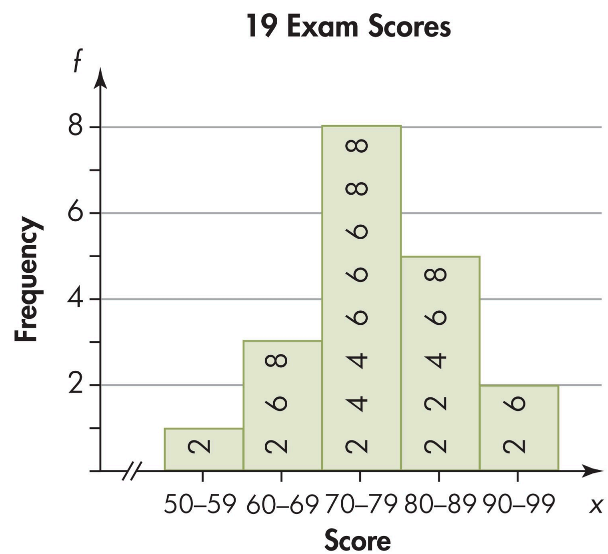
CHAPTER 2A

- Descriptive Analysis
- Presentation of Single-Variable Data
- Graphs for Qualitative Data
  - Pie Chart
  - Bar Graph
- Graphs for Quantitative Data
  - Dot plot
  - Stem and Leaf Plot
  - Histogram
- Measures of Central Tendency
  - Mean, Median, and Mode, Midrange
- Measures of Dispersion
  - Range
  - Sample Variance, and Sample Standard deviation



Graphical Summaries





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

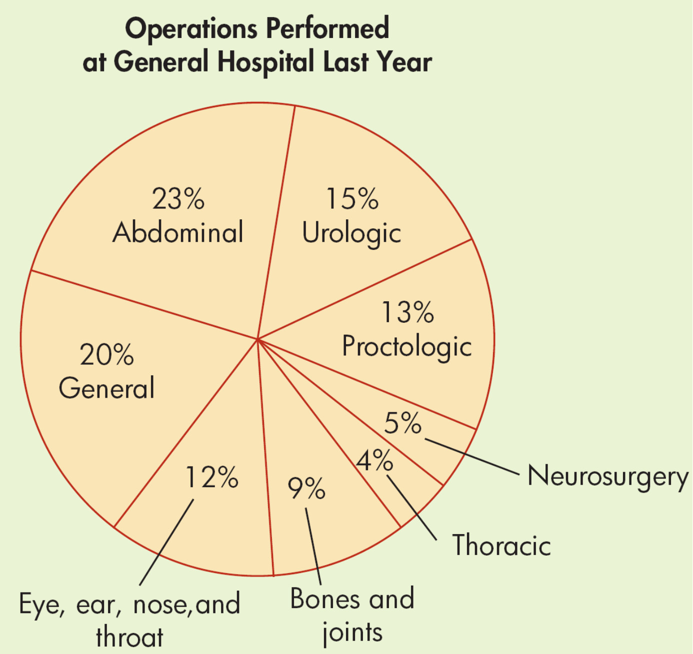
Operations Performed at General Hospital Last Year **[TA02-01]**

Table 2.1

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.**

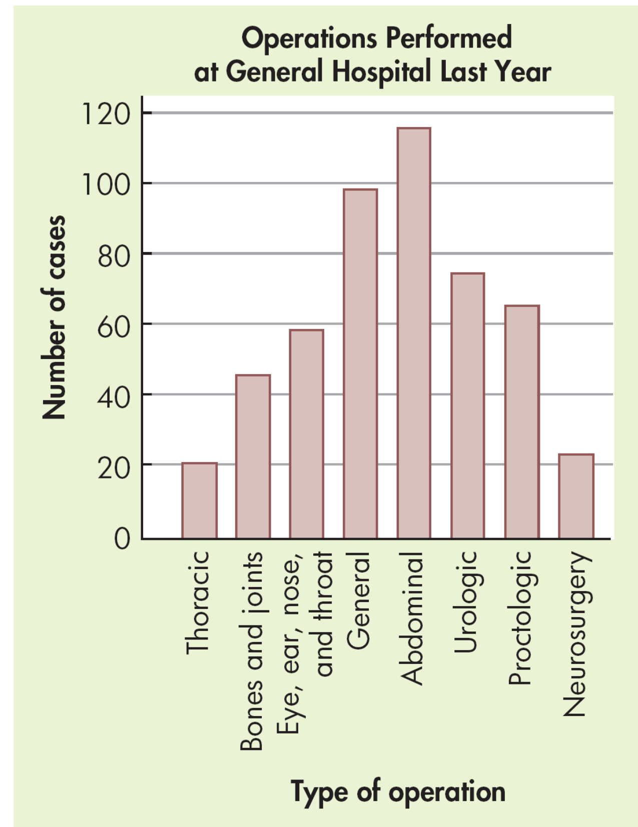
Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23
Total	498



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

- **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 Data

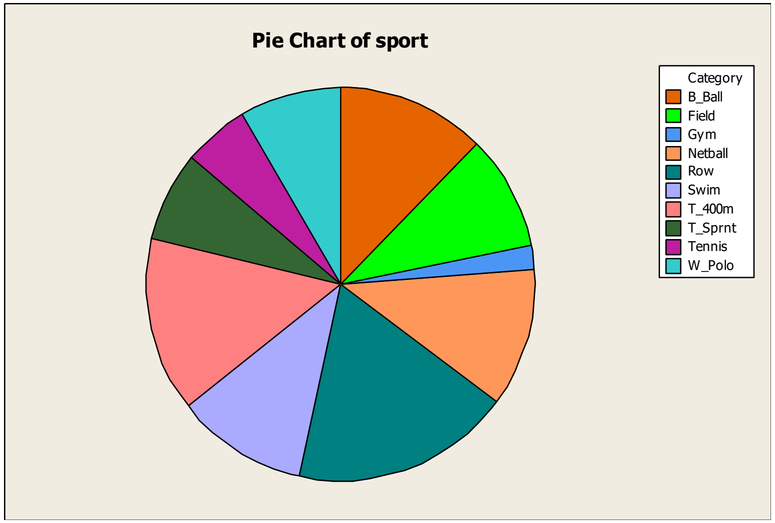
Variable	Description
sex	sex
sport	sport
rcc	red cell count
wcc	white cell count
Hc	Hematocrit
Hg	Hemoglobin
Fe	plasma ferritin concentration
bmi	body mass index, weight/(height)
ssf	sum of skin folds
Bfat	body fat percentage
lbm	lean body mass
Ht	height (cm)
Wt	weight (Kg)

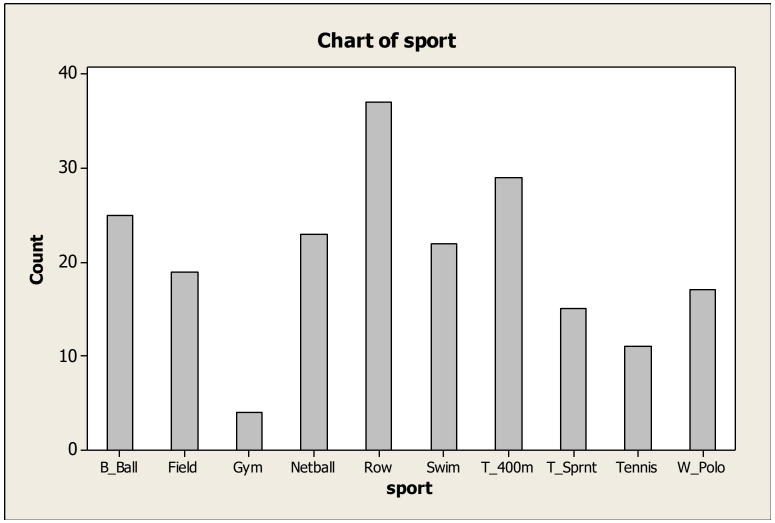
SUMMARIZING A SINGLE CATEGORICAL VARIABLE

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**
- **ais.xls (D2L/Content/Datasets)**

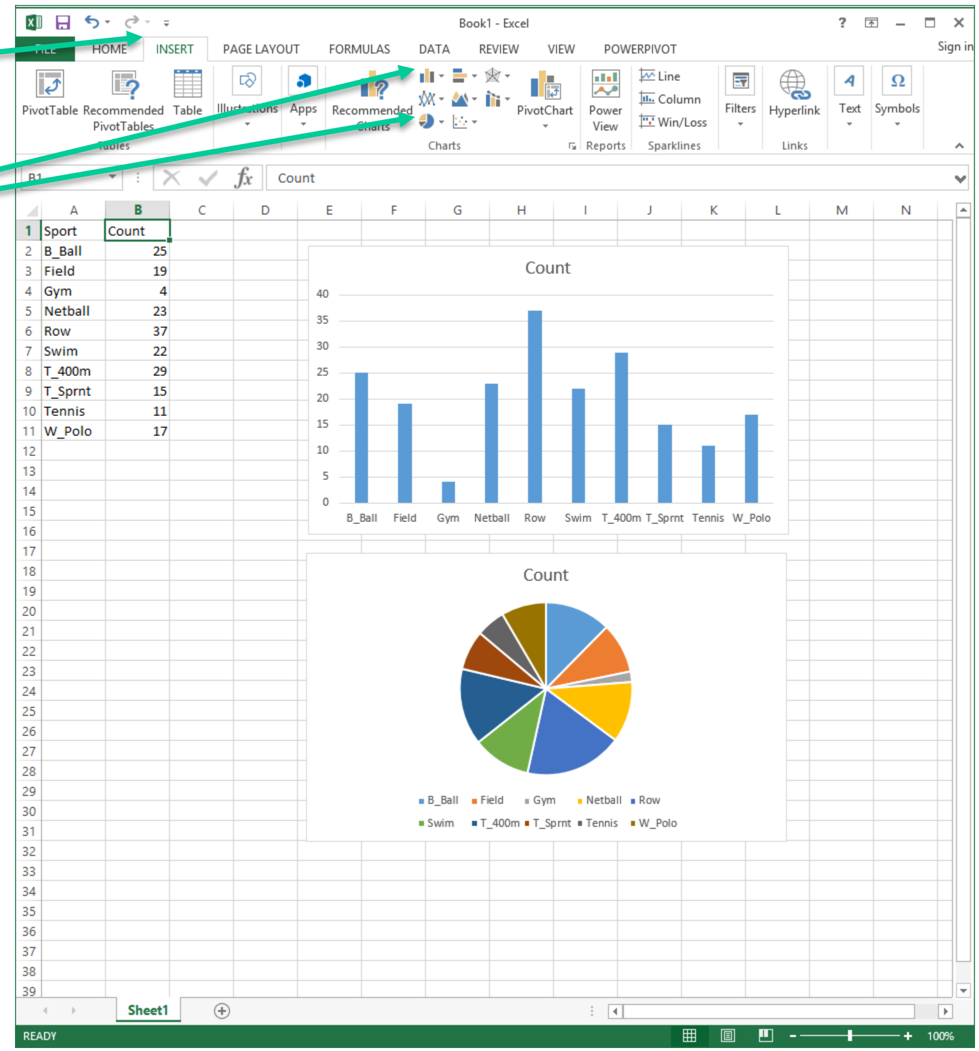
sport	Count	Percent
B_Ball	25	12.38
Field	19	9.41
Gym	4	1.98
Netball	23	11.39
Row	37	18.32
Swim	22	10.89
T_400m	29	14.36
T_Sprnt	15	7.43
Tennis	11	5.45
W_Polo	17	8.42
N=	202	





HOW TO?

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



- [JAMM: STAT-Calculator]

GRAPHING QUANTITATIVE DATA

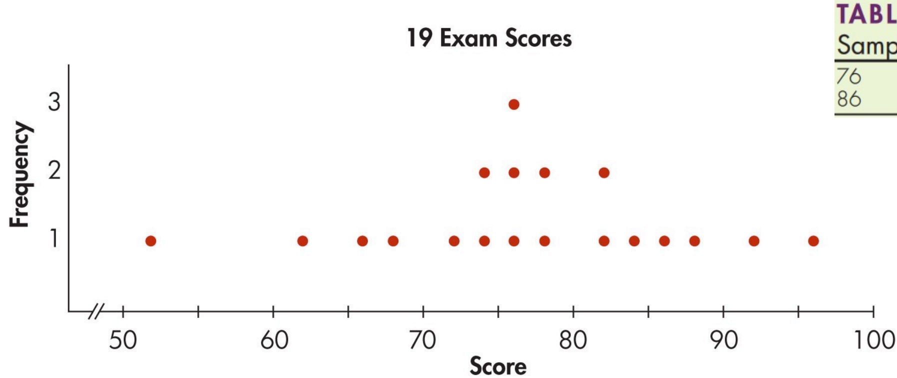
- **Distribution** The pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable.

- **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

- **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

- **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

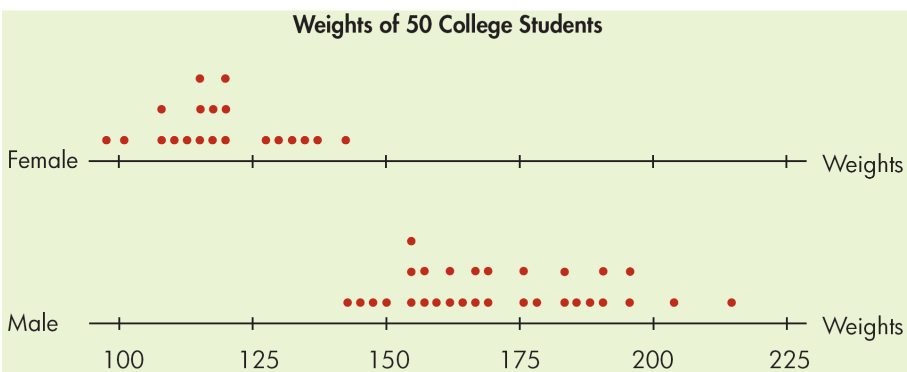
Figure 2.7

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

- **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

FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

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.**

FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

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 ($c = 10$)**

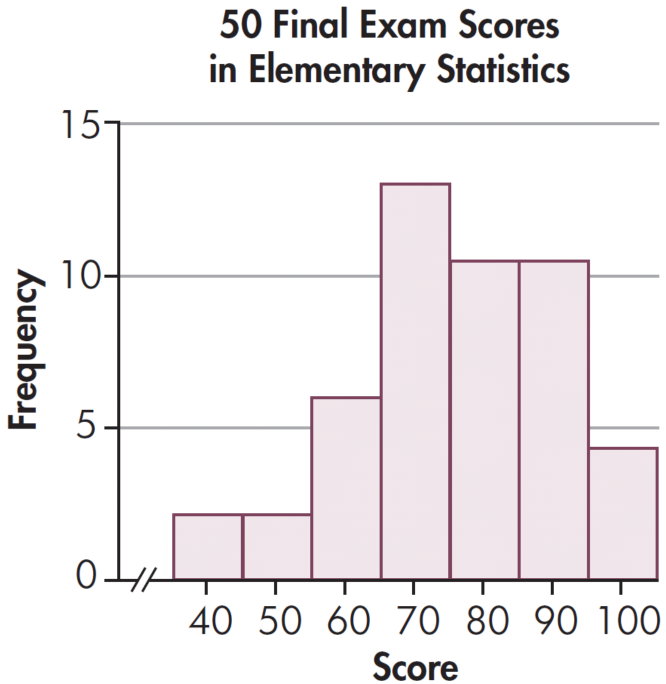
Class Number	Class Tallies	Boundaries	Frequency
1	||	$35 \le x < 45$	2
2	||	$45 \le x < 55$	2
3	||||| ||	$55 \le x < 65$	7
4	||||| ||||| |||	$65 \le x < 75$	13
5	||||| ||||| |	$75 \le x < 85$	11
6	||||| ||||| |	$85 \le x < 95$	11
7	||||	$95 \le x \le 105$	4
			50

Standard Chart for Frequency Distribution

FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

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.



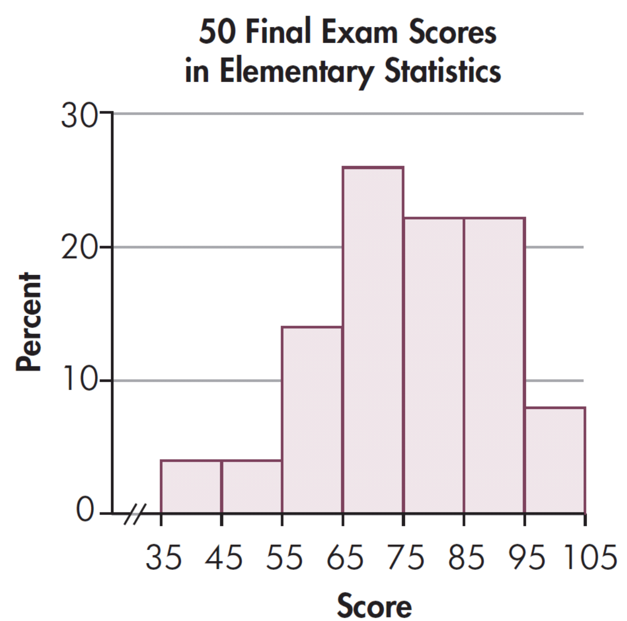
FORM A FREQUENCY DISTRIBUTION AND HISTOGRAM

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%

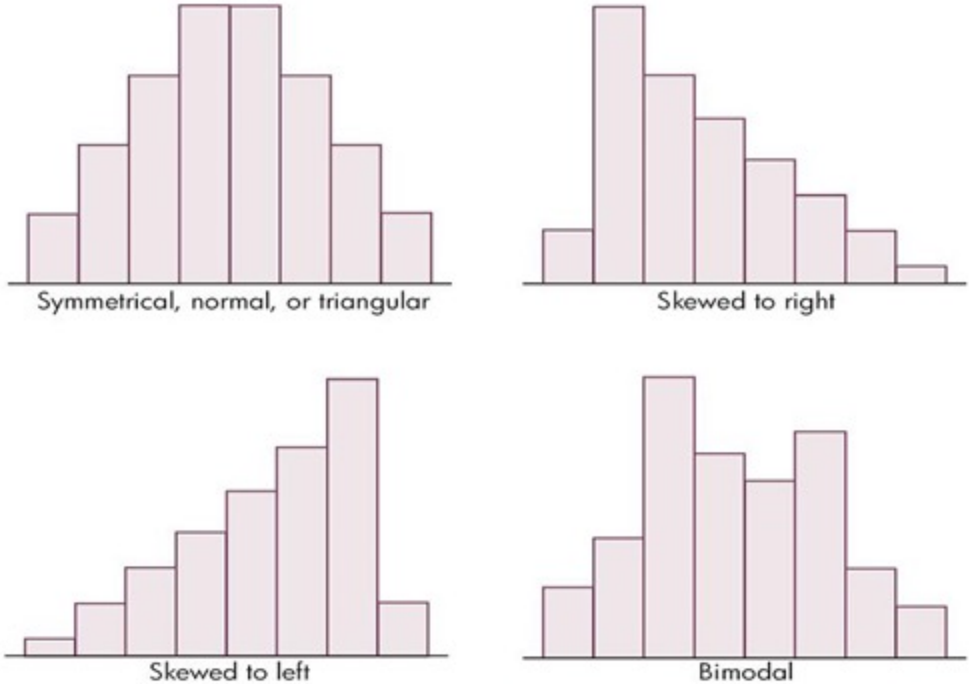
Divide all by 50

- 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.



FREQUENCY DISTRIBUTIONS AND HISTOGRAMS



- **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.

FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

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

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

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

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

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

Numerical Summaries

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

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

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

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

MEASURES OF CENTRAL TENDENCY

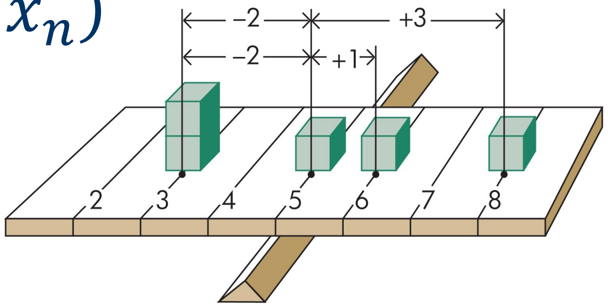
- **The measures of central tendency characterize the center of the distribution of data values. The term *average* is often associated with all measures of central tendency.**

- **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").**

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

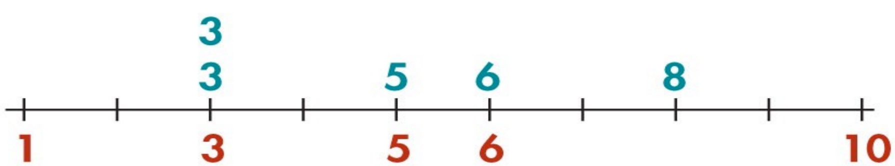
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$

- **The center of gravity or balance point**



MEASURES OF CENTRAL TENDENCY

- **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.



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

$$\text{midrange} = \frac{L + H}{2}$$

- **Sample Midrange**: The number exactly midway between a lowest-valued data, L, and a highest-valued data, H.
- There are other measures called **measures of dispersion** that characterize the spread or variability in the data.

MEASURES OF DISPERSION

- **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

- **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:**

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

- **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.**

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

- **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** $\tilde{x} = \text{middle value} = 5$
  - **Mode** $\hat{x} = \text{the value with the highest count} \Rightarrow \text{There is no mode}$
- **Measures of Dispersion**
  - **Range** $\text{range} = H - L = 8 - 2 = 6$
  - **Sample Variance and Sample Standard Deviation**

EXAMPLE (SAMPLE VARIANCE)

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

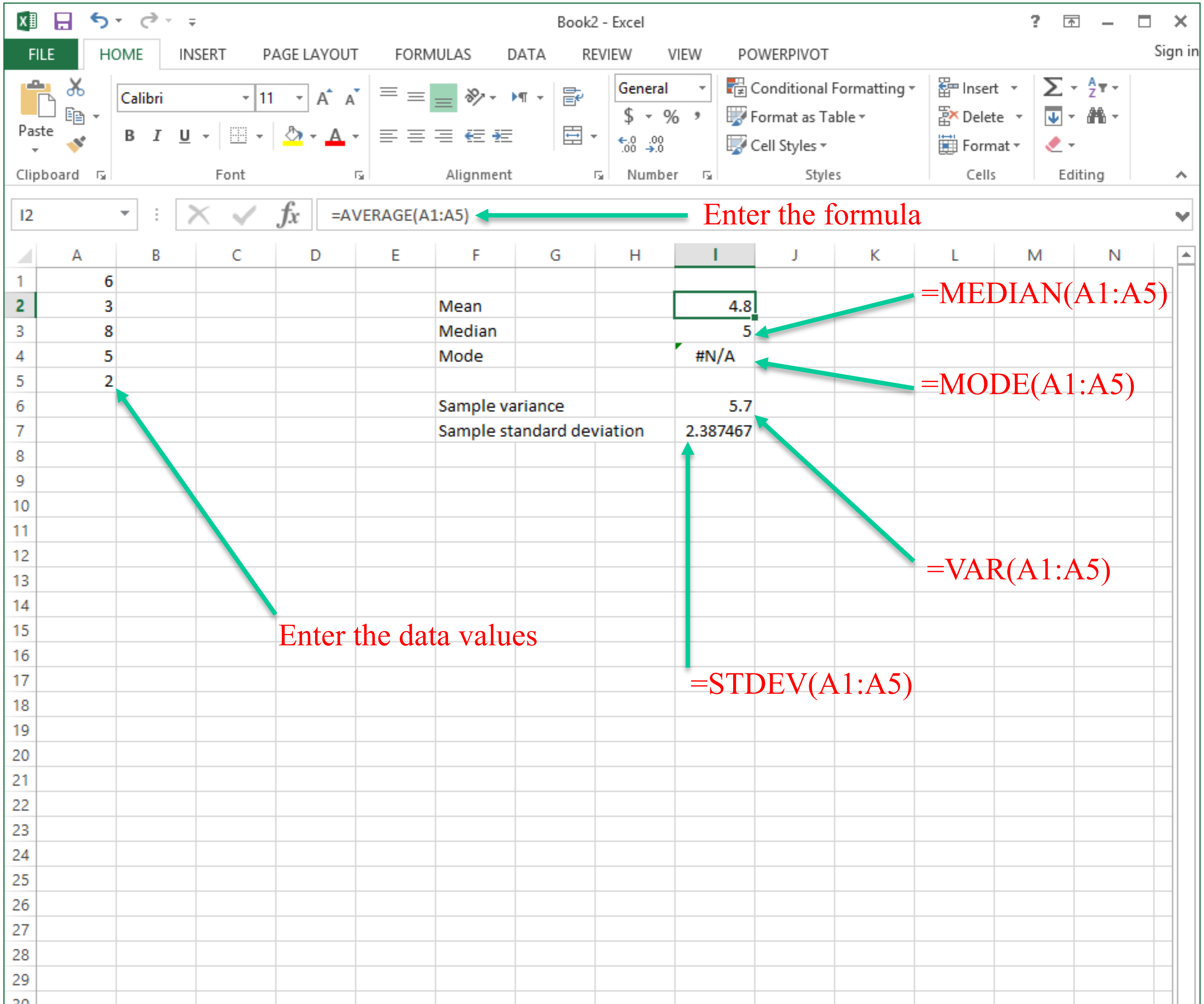
Step 1. Find Σx	Step 2. Find $\bar{x}$	Step 3. Find each $x - \bar{x}$	Step 4. Find $\Sigma(x - \bar{x})^2$	Step 5. Find s^2
6	$\bar{x} = \frac{\Sigma x}{n}$	$6 - 4.8 = 1.2$	$(1.2)^2 = 1.44$	$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$
3		$3 - 4.8 = -1.8$	$(-1.8)^2 = 3.24$	
8		$8 - 4.8 = 3.2$	$(3.2)^2 = 10.24$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

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

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

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	$2^2 = 4$	$SS(x) = 138 - 115.2$		
$\Sigma x = 24$	$\Sigma x^2 = 138$	$SS(x) = 22.8$	$s^2 = 5.7$	

EXAMPLE IN MICROSOFT EXCEL



QUESTIONS?

- ANY QUESTION?