2 Descriptive Statistics

2.1 Averages

An **average** or **measure of central tendency** is a single value that represents an entire data set. We will be concerned with three versions.

1. **mean**: add'em then divide
2. **median**: middle of ordered list
3. **mode**: most frequently occurring

Sample Data:	**4.9**	**0.7**	**1.8**	**1.1**	**0.6**	**1.8**

1. The **mean** (more accurately, the arithmetic mean) of a set of values is found by adding the values and dividing by the total number of values.

$$\text{Mean} = \frac{4.9 + 0.7 + 1.8 + 1.1 + 0.6 + 1.8}{6} = \frac{10.9}{6} = 1.82$$

Notation:

$\sum$ (sigma) 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 values* in a **sample**.
N represents the *number of values* in a **population**.

$\bar{x} = \dfrac{\sum x}{n}$ is the *mean* of a set of **sample** values. ($\bar{x}$ is spoken '*x-bar*').

$\mu = \dfrac{\sum x}{N}$ is the *mean* of a set of **population** values (μ is a Greek letter pronounced '*myoo*').

2. The **median** of a data set is the measure of center that is the *middle value* when the original data values are arranged in order. It is often denoted $\tilde{x}$ (pronounced *'x-tilde'*).

   (a) If there is an odd number of values the median is the value in the middle of the ordered list.

   (b) If there is an even number of values the median is the mean of the two middle numbers.

   For our example above, the ordered data is 0.6, 0.7, $\boxed{1.1,\ 1.8,}$ 1.8, 4.9 and

$$\text{Median} = \frac{1.1 + 1.8}{2} = \frac{2.9}{2} = 1.45$$

3. The **mode** of a data set is the value that occurs most frequently. For our example the mode is 1.8. When two values tie for the most frequent, each one is a mode and the data set is **bimodal**. When more than two values occur with the greatest frequency, the data set is **multimodal**. When no value is repeated, we say there is **no mode**.

- **Rounding:** When calculating an average that uses a mean (mean and sometimes the median), round your answer to one more decimal place than is present in the original data set.

- The term **average** can be used for any measure of center though it is most often associated with the mean.

- It is seldom that all three measures of center produce the same result as our example demonstrates. Here are brief descriptions of how they differ.

    - The **mean** is sensitive to extreme values.
    - The **median** is not sensitive to extreme values.
    - The **mode** is a good choice for nominal (nonnumeric) data.

- **Your Turn:** Calculate the mean, median, and mode of the given sample data below:

5 8 11 12 13

Mean =

Median =

Mode =

    - Without recalculating these averages, describe what would happen to these if the following changes to the data set were made.

        (a) Suppose the 13 was changed to 23.

        (b) Suppose the 5 was changed to 3 and the 13 to 15.

        (c) Suppose the 11 was changed to 8.

- Loosely speaking, **normally distributed data** has most of the entries bunched around the mean with fewer entries further from the mean. Additionally, mean $\approx$ median $\approx$ mode.

- **Technology:** In practice, one would use software to calculate the values described in this section. See www.StevensStats.com for technology demonstrations.

2.2 Range, Standard Deviation and Variance

A **measure of variation** describes how the data varies.

- **Variation**
  Often, you need more than averages to describe data. The table to the right gives the number of sales made by three different salespeople at the same car company on four randomly selected weeks.
  · They have the same mean number of sales. $\bar{x} = 10$.
  · How would you describe the differences between them?

	Number of Sales		
	Bob	Valerie	Carl
Week 1	8	18	10
Week 2	10	2	10
Week 3	14	16	10
Week 4	8	4	10

- The **range** is the simplest of all measures of variation but it is very sensitive to outliers.

$$\text{range} = \text{max value} - \text{min value}. \tag{2.1}$$

The range for Bob's number of sales is 14-8 = 6. What about Valerie and Carl?

- The **sample standard deviation (denoted by s)** is a measure of how the data varies about the mean. It is a type of average deviation from the mean. The following formula should guide you through the **process** of calculating it.

$$\text{sample standard deviation:} \qquad s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \tag{2.2}$$

Here, n is the sample size, $\bar{x}$ is the sample mean, and the x's are the individual data values.

- **Example:** Calculate the standard deviation using formula (2.2) for the number of sales by Bob.

x	$x - \bar{x}$	$(x-\bar{x})^2$
8	8 - 10 = -2	4
10	10 - 10 = 0	0
14	14 - 10 = 4	16
8	8 - 10 = -2	4
		24

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{24}{4-1}} = \sqrt{8} = 2.8$$

- **Rounding Rule:** When calculating standard deviation, use one more decimal than the raw data.

- **Your Turn:** Calculate the standard deviation for the number of sales by Valerie and Carl. Before you do, speculate on how these should compare to Bob's standard deviation.

- The **population standard deviation** is given by a similar formula for the sample standard deviation only you do not subtract one from the number of data values and you use the population mean instead of the sample mean. It is denoted with the Greek letter σ (sigma):

$$\boxed{\text{population standard deviation:} \qquad \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \qquad (2.3)}$$

- **Variance:** The variance is the square of the standard deviation. More appropriately, the standard deviation is the square root of the variance.

$$\sigma^2 = \textbf{population variance} = \tfrac{\sum(x-\mu)^2}{N}$$

$$\sigma = \textbf{population standard deviation} = \sqrt{\tfrac{\sum(x-\mu)^2}{N}} = \sqrt{\text{population variance}}$$

$$s^2 = \textbf{sample variance} = \tfrac{\sum(x-\bar{x})^2}{n-1}$$

$$s = \textbf{sample standard deviation} = \sqrt{\tfrac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

Why variance? The sample variance s^2 is an unbiased estimator of the population variance σ^2. This will be useful later when we are doing *inferential* statistics.

- **Technology:** In practice, one would use software to calculate the variance and standard deviation. See www.StevensStats.com for technology demonstrations.

- **Estimating the Standard Deviation:** If you can't get your hands on all the data but have a good idea of the range, a decent estimate for the standard deviation is given by

$$\text{standard deviation} \approx \frac{\text{range}}{4}$$

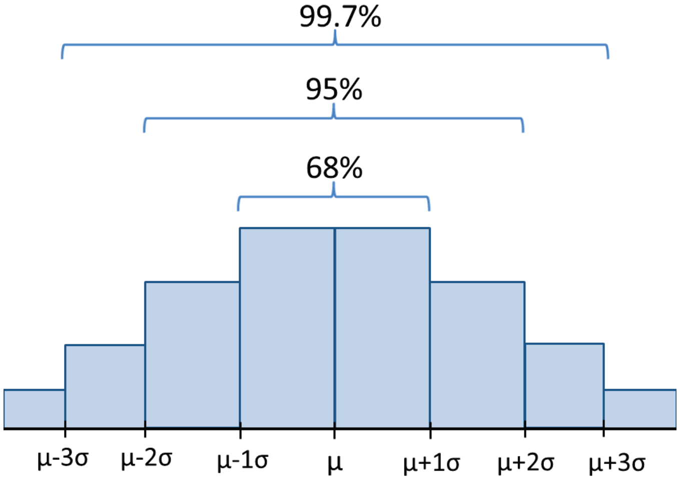
This is a very rough estimate.

- **Chebyshev's Theorem:**
Regardless of the distribution of the data, the proportion of values lying within k standard deviations of the mean is **at least** 1 - $1/k^2$. For example, letting $k = 2$ and $k = 3$ you get

  - At least 3/4 (75%) of all data values fall within 2 standard deviations of the mean.
  - At least 8/9 (89%) of all data values fall within 3 standard deviations of the mean.

- **Empirical Rule:** If the data is approximately **normally distributed**, the following are true.

  - About **68%** of all values fall within **1** standard deviation of the mean.
  - About **95%** of all values fall within **2** standard deviations of the mean.
  - About **99.7%** of all values fall within **3** standard deviations of the mean.



- **Example:** Assume *IQ* scores are normally distributed with a mean of 100 and standard deviation of 15. Use the empirical rule to find the range of *IQ* scores that correspond to the

  (a) middle 68% of scores.

  Answer: We use the empirical rule by subtracting and adding **one** standard deviation from/to the mean. About 68% of *IQ* scores are between **85 and 115**.

  (b) middle 95% of scores.

  (c) middle 99.7% of scores.

- **Definition of Unusual Values:** If a value lies more than 2 standard deviations away from the mean we call it unusual. Otherwise, it is considered not unusual. **Warning:** The distribution should be approximately normal to use these definitions.

  (a) Is an *IQ* score of 136 unusual?

  Answer: Since 136 is more than two standard deviations above the mean, we categorize this score as **unusual**.

  (b) Is an *IQ* score of 120 unusual?

  (c) Is an *IQ* score of 62 unusual?

2.3 Measures of Relative Standing: z-scores

- **Definition:** A z**-score** is the number of standard deviations that a given value (x) is above or below the mean. It is found using either version of the same formula below.

  sample data $\qquad$ population data

  $$z = \frac{x - \bar{x}}{s} \qquad z = \frac{x - \mu}{\sigma} \tag{2.4}$$

  Here, $\bar{x}$ is the sample mean, s is the sample standard deviation, μ is the population mean, and σ is the population standard deviation.

- z-scores can be used to compare values that come from different populations.

  Example: Based on z-scores, which of following is the highest relative test score?

  (a) A score of 82 on a test with a mean of 75 and standard deviation of 4.
  $z = \frac{82-75}{4} = 1.75.$

  (b) A score of 95 on a test with a mean of 85 and standard deviation of 8.

  (c) A score of 75 on a test with a mean of 80 and a standard deviation of 2.

- Recall our previous definition of unusual values from a normal distribution:
  If a value lies more than 2 standard deviations away from the mean we call it unusual.
  This can now be replaced with the following definition.

  Unusual values have a z-score less than -2 or greater than 2.

- **Example**: IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. Give the z-score of each of the following IQ's and categorize each one as *unusual* or *not unusual.*
  Round z-scores to two decimal places.

  (a) 62
  Answer: $z = \frac{62-100}{15} = -2.53$ which is less than -2. Therefore, 62 would be considered unusual.

  (b) 80

  (c) 101

  (d) 125

  (e) 135

2.4 Measures of Relative Standing: Quartiles, Percentiles, and Box Plots

- **Quartiles** separate the data into 4 parts just like the median separates the data into two parts.
  - Q_1 (First Quartile): Separates the bottom 25% from the top 75%.
  - Q_2 (Second Quartile): Separates the bottom 50% from the top 50% **(same as the median)**.
  - Q_3 (Third Quartile): Separates the bottom 75% from the top 25%.
- **Percentiles** separate the data into 100 different parts.
  - P_k $(0 < k < 100)$ is the k'th percentile.
  - P_{50} is the 50th percentile $= Q_2 =$ median.
  - P_{90} is the 90th percentile. This number separates the bottom 90% of the data from the top 10%.
- **Issues:** There is not complete agreement in how to calculate Q_1 and Q_3. Ideally, you would want $P_{25} = Q_1$ and $P_{75} = Q_3$. This is not always the case and different software packages may result in different values for these terms. There is some agreement on a simple method to calculate percentiles so I will present it here.
- **Procedure for calculating P_k:**
  1. Order the n values from least to greatest.
  2. The index (i) is found by $i = \dfrac{k}{100} \cdot n$. Then,
     - if i is a whole number, you average the i^{th} value and the next to get P_k.
     - if i is not a whole number, you **round up** to get the index of P_k.

- **Example:** Below is a table of 16 quiz scores ordered (and indexed) from least to greatest.

index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
score	15	24	27	31	36	37	37	38	40	41	42	43	44	45	48	50

(a) Calculate P_{20}

Answer: The index is given by $i = \frac{20}{100} \cdot 16 = 3.2$. Since this is not a whole number we **round up** to get $i = 4$ and $P_{20} = 31$.

(b) Calculate P_{25}

Answer: The index is given by $i = \frac{25}{100} \cdot 16 = 4$, and we must average the 4^{th} and 5^{th} values to get $P_{25} = 33.5$.

(c) **Your Turn:** Calculate P_{75}

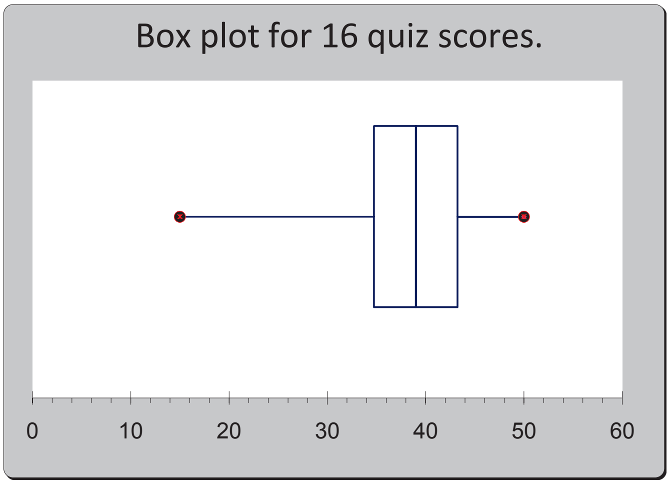
(d) **Your Turn:** Calculate P_{90}

- **Procedure for Calculating Quartiles**. This is where different procedures produce different results. To avoid this whole mess, let's go with the following:

$$\boxed{Q_1 = P_{25} \qquad Q_2 = \text{the median} \qquad Q_3 = P_{75}}$$

- **The 5-number summary and box plots:** In statistics, the 5-number summary includes the minimum value, Q_1, Q_2, Q_3, and the maximum value. These values are used to create a box plot of the data. A box plot is often called a **box and whisker plot**.

  Example: Below is the box plot as well as the 5-number summary for the 16 quiz scores listed on the previous page.



5-number summary		
min	**15**	left *whisker*
Q_1	**33.5**	left boundary of box
Q_2	**39.0**	line in middle of box
Q_3	**43.5**	right boundary of box
max	**50**	right *whisker*

$Q_1 = P_{25}$ from previous page
$Q_2 =$ median
$Q_3 = P_{75}$ from previous page

- **Your Turn:** Use the data below for 18 quiz scores to create the 5-number summary and sketch a box plot of the data. How does it compare to the set of 16 scores from the previous page?

index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
score	20	21	21	27	29	30	30	32	33	38	40	41	44	46	46	48	50	50

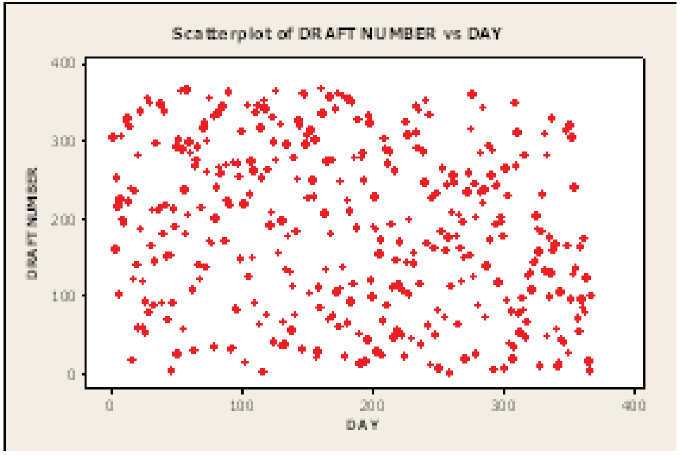
- **Notes on percentiles, quartiles, and box plots**
  - Box plots are often presented vertically.
  - The **interquartile range (IQR)** is given by IQR $= Q_3 - Q_1$.
  - **Outliers:** Some texts are bold enough to claim that any data value more than 1.5 IQR's below Q_1 or above Q_3 are outliers. There is no consensus as to whether this is a good definition or not.
  - A **modified box-and-whisker plot** displays outliers as well. In this case, the *whiskers* end at the most extreme values not considered outliers.
  - The **discrepancy** with regards to Q_1 and Q_3 goes like this: First you get the median to divide your data into two segments. Some methods keep the median in the two remaining halves, some don't. If you have a large set of distinct values there should not be a big difference. Either way, you might get different values for the quartiles and their associated percentiles. I don't know of a single method that resolves all of the possible conflicts so I presented the fastest way out.
  - The **percentile of a score**: $\text{percentile of x} = \dfrac{\text{number of values less than x}}{\text{total number of values}} \cdot 100$

    Problem: Sometimes if you calculate the percentile of a given score, it will not match the same percentile of the data set. Again, there is no easy way around this.

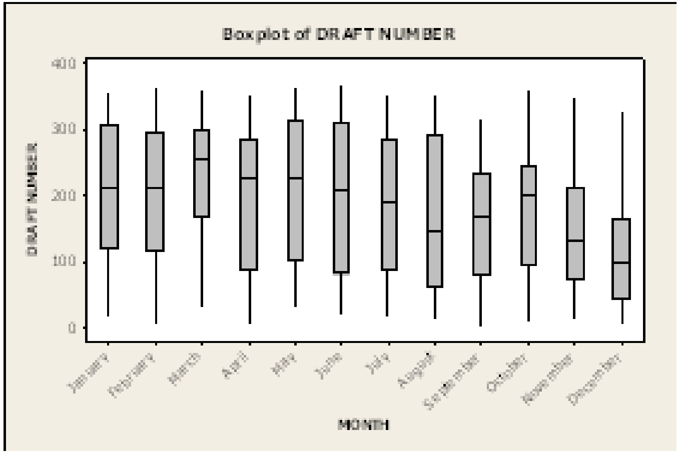
- **Interesting Example:** Side-by-side box plots can reveal interesting differences in data.

When the U.S. Government started drafting men into the Vietnam war, the draft order was determined by birth-dates that were randomly selected from a bin. There was some contention that the selection was not fair. Claims were made that those men with late birth-months were picked earlier in the draft. The scatterplot does not reveal such a pattern. The side-by-side box plots suggest this claim is valid.

Scatterplot
x-axis: Birth day (#1-365)
y-axis: Draft Number



Box Plots
x-axis: Birth Month
y-axis: Draft Number



- **Technology:** In practice, one would never create box plots by hand. Tips on how to use various software packages can be found at the textbook website: www.StevensStats.com

2.5 Weighted Averages & Simpson's Paradox

The weighted average (or weighted mean) works a lot like a regular (arithmetic) mean. In a weighted mean, not all values have the same importance. Some values carry a heavier weight than others. Sometimes the results of a weighted average are counter-intuitive such as in the case of Simpson's paradox.

- **Preliminary Example:** Calculating your grade point average (GPA)

  A GPA is calculated by first assigning each letter grade a numerical value (A = 4.0, B = 3.0, C = 2.0, D = 1.0, and F = 0.0). Then each grade is **weighted** by the number of credits before being averaged. Everyone knows that it is better for your GPA to get an `A` in a four credit course and an `F` in a one credit course than the other way around. But, do you know how big that difference really is? The two examples below illustrate the disparity.

Example:

4-credit A & 1-credit F

Credits (w)	Letter Grade	Numerical Grade (x)	$w \cdot x$
4	A	4.0	16.0
1	F	0.0	0.0
3	C	2.0	6.0
3	C	2.0	6.0
3	C	2.0	6.0
14			34

$$\text{GPA} = \frac{\sum(w \cdot x)}{\sum w} = \frac{34}{14} = 2.43$$

Your Turn:

4-credit F & 1-credit A

Credits (w)	Letter Grade	Numerical Grade (x)	$w \cdot x$
4	F		
1	A		
3	C		
3	C		
3	C		
14			

GPA =

- **Weighted Averages in General**

  The above example demonstrates the standard form for a weighted average.

$$\boxed{\text{weighted average:} \qquad \bar{x} = \frac{\sum(w_i \cdot x_i)}{\sum w_i} \qquad (2.5)}$$

This formula says to weight each data value (x_i) with the appropriate weight (w_i), add up the products of all of these and divide by the sum of all the weights. Your grade in a class is a weighted average of various items such as homework, quizzes, tests, final, and others. However, in this case the sum of the weights is usually one and so you don't see a division in there.

- **Averages of Averages:** In general, averaging averages is risky business. However, taking a weighted average allows you to do this accurately. For example, suppose a clinic employs two Registered Nurses (RN's) and four Licensed Practical Nurses (LPN's) with the average salaries given in the table below. What is the average salary for the nurses in this clinic?

Type of Nurse	Number Employed w	Average Salary x
RN	2	\$60,000
LPN	4	\$40,000

Wrong: Average of Averages $\dfrac{60,000+40,000}{2}$ = \$50,000.

Correct: Weighted Average

$$\bar{x} = \frac{\sum(w \cdot x)}{\sum w} = \frac{(2 \cdot 60,000) + (4 \cdot 40,000)}{2+4} = \mathbf{\$46,667}$$

- **Your Turn:** Professor Brown has two sections of Statistics, one in the morning and one in the afternoon. The morning section has 10 students and their average on Test #1 was 85. The afternoon section has 28 students and their average was 73. Calculate the average score on Test #1 for Professor Brown's Statistics Students.

- **Simpson's Paradox:** Sometimes, averages across categories directly contradict averages within categories. When this happens, it is called Simpson's Paradox.

Example: Professor Brown and Professor Sides both teach Statistics at the same college. They each have two sections of the course, one in the morning and one in the evening. The table below gives the class average on Test #1 with number of students in parentheses. Whose students did better?

	class average (*# students*)	
	AM Section	PM Section
Prof. Brown	**85** (*10*)	**73** (*28*)
Prof. Sides	**82** (*28*)	**70** (*10*)

AM: Brown's class did better than Sides'.
PM: Brown's class did better than Sides'

The Paradox:
Overall: Brown's students did worse.

Prof. Brown - Weighted Average: $\bar{x} = \dfrac{\sum(w \cdot x)}{\sum w} = \dfrac{(85 \cdot 10) + (73 \cdot 28)}{38} \approx \mathbf{76}$

Prof. Sides - Weighted Average: $\bar{x} = \dfrac{\sum(w \cdot x)}{\sum w} = \dfrac{(82 \cdot 28) + (70 \cdot 10)}{38} \approx \mathbf{79}$

Note: Prof. Sides' weighted average benefitted from the large number of AM students.

Chapter 2: Discussions

1. Is it possible that 70% of all students are *below average*?

2. In a Science article titled ”Gender Similarities Characterize Math Performance,” Hyde et al. (2008) reported their analysis of scores for over 7 million students in state NCLB math assessments. In their report they state that the average scores for males and females were nearly equal. However, the male data contained more variation as described by a variance that was between 1.11 and 1.2 times as great as for the females. Many headlines followed this report. Here are a few:

   (a) ”Math Scores Show No Gap for Girls, Study Finds” [Lewin (2008)]

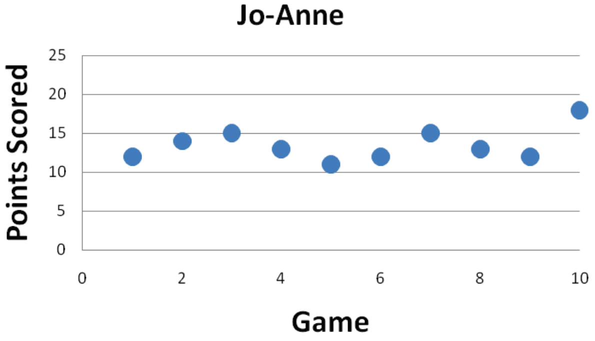
   (b) ”In math, girls and boys are equal” [Seattle Times News Service (2008)]

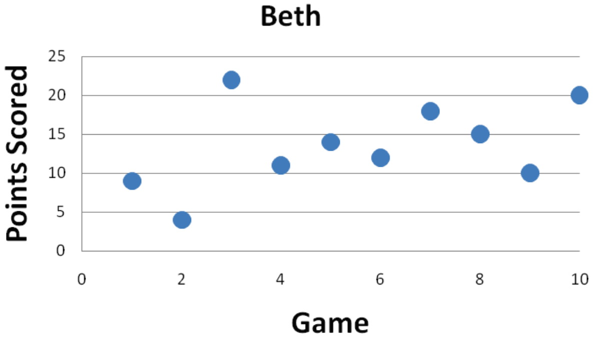
   (c) ”Math IS Harder for Girls” [Mac Donald (2008)]

   How could such varying headlines be justified by the same data set? What’s the missing headline?

3. You are the coach of a basketball team. You are making your play-off team and **have** to choose between Jo-Anne and Beth. Which one would you take? What if your team is the favorite? What if your team is the underdog?

Game	1	2	3	4	5	6	7	8	9	10	Statistics		
	Points Scored in Each Game										mean	standard deviation	median
Jo-Anne	12	14	15	13	11	12	15	13	12	18	13.5	2.1	13
Beth	9	4	22	11	14	12	18	15	10	20	13.5	5.5	13





Chapter 2: Summary Worksheet

1. Calculate the requested statistics for the given sample data.

   Sample Data: 68, 84, 93, 68, 70

   (a) mean:

   (b) median:

   (c) mode:

   (d) range:

   (e) sample variance:

   (f) sample standard deviation:

   (g) Suppose one of the 68's from this data set was switched to a 50. What would this do to the mean, median, mode, range, standard deviation, and variance?

2. Give the z-score (2 decimal places) for each test score with the given class mean and standard deviation. Assume the test scores are normally distributed.

   (a) For a 92 on a test with a class mean of 78 and a standard deviation of 12,

   $z =$

   (b) For a 75 on a test with a class mean of 60 and a standard deviation of 6,

   $z =$

   Which score is relatively higher, the 92 or the 75?

   Is either score unusual? If so, which one.

3. Give a 5-number summary and box plot for the 21 test scores indexed below.

index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
score	48	51	55	61	66	68	70	72	72	75	76	78	81	83	83	86	88	93	93	95	98

4. Calculate the GPA for a student with these grades:

	Letter
Credits	Grade
3	A
1	A
3	A
6	D
4	C

5. In Hockey Village, VT, the X-Ice Mites hockey team has 3 sub-teams, an A team, a B team, and a C team. The table gives the team size and average weight of the players on each team. Use a weighted average to calculate the mean weight of all the kids on the X-Ice Mites hockey team.

	# of	Average
Team	players	Weight (pounds)
A	10	78.5
B	13	64.1
C	7	55.3

Chapter 2: Problem Set

Numbers with an asterisk* have solutions in the back of the book.

Averages and Variation (2.1 & 2.2)

1.* **Sample Statistics:** For the following sample data, find the mean, median, mode, range, sample variance, and sample standard deviation. You should be able to do this by hand.

4, 8, 4, 6

2. **Sample Statistics:** For the following sample data, find the mean, median, mode, range, sample variance, and sample standard deviation. You should be able to do this by hand.

2, 3, 8, 1, 6

3.* **Sample Statistics:** For the following sample data, find the mean, median, mode, range, sample variance, and sample standard deviation. Check your answer with Software.

-1.5, 2.8, 3.4, -3.5, 7.6 -12.1

4. **Sample Statistics:** For the following sample data, find the mean, median, mode, range, sample variance, and sample standard deviation. Feel free to do this with software exclusively.

0.23, 0.75, 1.22, 0.53, 1.22 1.01 0.25

5.* **Simpson's Paradox, Wage Discrepancy:** Here is a fictitious example where an average across categories conflicts with the averages obtained within categories. This is called Simpson's Paradox.

Suppose you own a contracting company and employ 16 people (8 males and 8 females). Your employees are paid on an hourly basis and the wages (in dollars per hour) are given in the table below. You are accused of discriminatory pay practices because the average wage for the males ($29 per hour) is greater than the average wage for the females ($26 per hour). Using the same data, but refining your focus to include employee's experience, what is your best defense against such an accusation? What is the lurking variable that is really causing the difference in pay?

	less than 5 years experience						more than 5 years experience						mean
Male	20	26					29	29	31	31	33	33	29.0
Female	20	22	25	25	26	26	30	34					26.0

6. **Simpson's Paradox, Wage Discrepancy:** Here is another fictitious example where an average across categories conflicts with the averages obtained within categories. This is called Simpson's Paradox.

   The manager at the GARP clothing branch in the mall is applauded for treating male and female sales representatives equally with respect to pay. The is demonstrated by the averages given in the table below. The average (mean) of the six males is \$14/hour which equals the average for the seven females. Using the same data, but refining your focus to include the employee's status (Assistant or Associate), can you conclude that the wages are biased one way or the other?

	Assistant Sales Rep				Associate Sales Rep					mean
Male	12	12	13	15	18	14	-	-	-	14.0
Female	10	12	-	-	14	14	16	16	16	14.0

7.* **Altering Data Sets:** Consider the following sample data

2, 4, 6, 6, 8 10

Now assume the following changes are made to this data. Comment on what would happen to the mean, median, mode, standard deviation and variance after the changes are made. You do not have to calculate all these values, just determine whether the statistic would increase, decrease, or stay the same.

(a) The 10 is replaced by a 20.

(b) The 2 becomes a 0 and the 10 becomes a 12.

(c) One of the 6's is replaced by a 0.

8. **Altering Data Sets:** Consider the following sample data

25, 28, 30, 35, 45 50

Now assume the following changes are made to this data. Comment on what would happen to the mean, median, mode, standard deviation and variance after the changes are made. You do not have to calculate all these values, just determine whether the statistic would increase, decrease, or stay the same.

(a) The 25 is replaced by a 10.

(b) The 25 becomes a 20 and the 50 becomes a 55.

(c) The 30 becomes a 35.

Measures of Relative Standing: z-scores (2.3)

Use the data found in this chart to answer the following questions.

Strata	Mean Height (inches)	Standard Deviation Height (inches)	Mean Weight (pounds)	Standard Deviation Weight (pounds)
U.S. Men	69.3	2.8	191	28
U.S. Women	64.0	2.8	145	32
NFL Quarterbacks	76.5	1.8	245	25
Top Female Models	70.0	2.2	115	18

9.* **Male Heights:** For the given heights of U.S. men, calculate the z-score, and comment on whether the height would be unusual for a U.S. man.

(a) 65.5 inches

(b) 70.2 inches

(c) 74.0 inches

(d) 78.0 inches

10. **Female Heights:** For the given heights of U.S. women, calculate the z-score, and comment on whether the height would be unusual for a U.S. woman.

(a) 57.8 inches

(b) 65.2 inches

(c) 68.2 inches

(d) 70.0 inches

11.* **Models:** Gisele Bundchen is a top female model. She is 71 inches tall and weighs 115 pounds.

(a) Is her height unusual with respect to top female models?

(b) Is her height unusual with respect to U.S. women?

(c) Is her weight unusual with respect to top female models?

(d) Is her weight unusual with respect to U.S. women?

12. **Quarterbacks:** Tom Brady is a quarterback in the NFL. He is 76.0 inches tall and weighs 225 pounds.

(a) Is his height unusual with respect to NFL quarterbacks?

(b) Is his height unusual with respect to U.S. men?

(c) Is his weight unusual with respect to NFL quarterbacks?

(d) Is his weight unusual with respect to U.S. men?

13.* With respect to their professional peers, who is taller: Gisele Bundchen or Tom Brady?

14. With respect to U.S. adults by gender, who is taller: Gisele Bundchen or Tom Brady?

15. If Gisele Bundchen and Tom Brady walk into a restaurant together, would you notice?

16.[*] **Wolf Spider:** I saw what I thought to be a Wolf spider on my garage. It was huge. Later there was an egg sack where the spider had been. Even later I saw at least 500 little baby spiders crawling out of the egg sack. I looked up in the internet that the average number of eggs laid by a Wolf spider is 302 with a standard deviation of 48.

(a) If my estimation and the internet information was accurate, is this an unusual number of spider eggs for a Wolf spider?

(b) What might be the explanation here?

Percentiles, Quartiles, and Box Plots (2.4)

17. **AM -vs- PM Test Scores:** I have two sections of statistics, one in the morning (AM) with 22 students and one in the afternoon (PM) with 30 students. I gave each section the identical test. The results are ordered and indexed in the tables below. Answer the following questions regarding these data sets.

i	AM	PM
1	31	45
2	50	48
3	58	50
4	59	52
5	60	55
6	61	60
7	63	61
8	64	63
9	64	64
10	66	65
11	71	66
12	71	67
13	71	68
14	77	74
15	79	78
16	79	79
17	87	80
18	87	80
19	90	81
20	92	82
21	92	85
22	95	87
23		87
24		90
25		94
26		96
27		98
28		98
29		100
30		100

(a)[*] Consider the 22 scores from my AM section.

i. Calculate P_{90}

ii. Create the 5-number summary for my AM section.

iii. Create a box plot for the AM section.

(b) Consider the 30 scores from my PM section.

i. Calculate P_{90}

ii. Create the 5-number summary for my PM section.

iii. Create a box plot for the PM section.

(c)[*] Compare and contrast the two sections.

Weighted Averages & Simpson's Paradox (2.5)

18. **GPA:** Consider the report cards for Sam and Samantha given below.

Sam

Credits	Letter Grade
3	B
1	A
3	C
6	D
3	A

Samantha

Credits	Letter Grade
3	B
6	A
3	C
1	D
3	A

(a)* Calculate Sam's GPA.

(b) Calculate Samantha's GPA.

19. **Average Test Score:** Suppose there are three sections of a Statistics course taught by the same instructor. The class averages for each section on Test #1 are displayed in the table below. What is the average test score for all sections combined?

	Class Size	Class Average
Section 01	8	88
Section 02	16	74
Section 03	30	72

20. **Average Daily Balance:** Most credit cards charge interest based on the *average daily balance* per billing cycle. In this case each balance within the billing cycle is weighted by the number of days it exists. Suppose your credit card has a 30 day billing cycle and the balances over these 30 days are given in the table below.

Days	Transaction	balance (x)	# days (w)
1-6	remaining balance	\$1200	6
7-10	\$400 purchase	\$1600	4
11-20	\$300 purchase	\$1900	10
21-30	\$1000 payment	\$900	10

(a)* Calculate the average daily balance for these 30 days.

(b) What would have been your average daily balance if you paid \$1200 on day 13 instead of \$1000 on day 21?

21.* **Simpson's Paradox, Exercise -vs- Diet:** Below is a table for the mean weight lost (in pounds) by moderately (BMI < 40) and severely (BMI > 40) obese participants in a weight loss study over the course of 6 months. Some of these participants employed a diet only plan while others used an exercise only plan. The number in the parentheses gives the number of participants in each category.

	Mean weight loss (# participants)	
	Extremely Obese	Moderately Obese
Exercise Plan	**22** (5 participants)	**16** (25 participants)
Diet Plan	**19** (25 participants)	**13** (5 participants)

(a) Looking within each category of obesity, which plan seems to work best?

(b) Using weighted averages, calculate the mean weight loss across both categories of obesity for the exercise plan and then for the diet plan. Which plan seems to be more effective?

(c) Why did the *diet plan* do so well when using the weighted average?

22. **Simpson's Paradox, Jeter -vs- Justice**: In baseball, the batting average (BA) is defined as the quotient of hits divided by the times at bat. Below is a table for the 1995 and 1996 batting averages for Derek Jeter and David Justice. Who has the better batting average? Try a straight average of the two years and a weighted average based on the number of times at bat. You should get contradictory results.

Player	BA - 1995	BA - 1996
Derek Jeter	.250 (48 times at bat)	.314 (582 times at bat)
David Justice	.253 (411 times at bat)	.321 (140 times at bat)

(a) Looking within each category of year, which player had a better batting average?

(b) Using a weighted average across both years, calculate the mean batting average for each player. Which player had a better batting average?

(c) Why did Derek Jeter fair so well using the weighted average?