Math 316

Section 2.1 Introduction

See book regarding types of variables: categorical, numerical, continuous, discrete. Each variable has observational units. Our book uses *upper case* for the *variable* and *lower case* for a *particular value*, i.e. an observation. See **Observational Units** and **Notation…** on page 28.

Section 2.2 Frequency Distribution

Frequency distribution: the possible outcomes and how many times each occurred. Relative frequency distribution: the possible outcomes and what *fraction* of the time each occurred. The relative frequency distribution is generally more useful information. **Example 2.2.1 and Example 2.2.5. Example 2.2.2.**

Dotplots tell us approximately (1) the center of the data and (2) how spread out the data are. **Example 2.2.3.**

Histograms give us similar information. **Example 2.2.4** (a discrete example) and **Example 2.2.6** (a continuous example).

For numerical observations with many possible values, we often group the data—that is, we split the entire range of values—into several smaller classes or intervals (usually of the same size). This is especially common when dealing with continuous values, like height or weight. Notice that how we specify the ranges/intervals of values can affect the histogram. **Example 2.2.6, Example 2.2.7.**

Of importance later: the *area* within a histogram corresponds to the *frequency* or *relative frequency* or *probability* of a particular range of values. **Figure 2.2.12.** The book states that this idea will be "indispensable" to us.

Read about shapes of distributions. **Online: Different shapes of data. Figure 2.2.14.**

Section 2.3 Measures of Center

A statistic is a numerical measure calculated from sample data. (The same value for an entire population is called a parameter.) Descriptive statistics describe the data. The two most useful descriptive statistics are (1) the center of the data and (2) how spread out the data are. A few examples would be for: salaries, home prices, exam scores, etc.

There are two common ways to measure/describe the center of the data:

Median $\tilde{y}$: middle value Mean $\bar{y}$**:** average value. (Notation: why the bar in $\bar{y}$?)

Formula in box on page 41. Examples 2.3.1, 2.3.2, 2.3.3. The Σ in the Sample Mean formula is a Greek "S" for "Sum." *We will usually work with the mean*. See book for more discussion on median vs. mean. **Example 2.3.6.**

Deviation: the difference between a particular value and what is normal; that is, deviance is the difference of a particular value from the mean. **Example 2.3.4**.

The mean can be quite affected by a single value—the median not very much. "**Robustness" on page 42**. **Online: Mean vs. Median** (compare to **Example 2.3.6**, which is also skewed to the right).

Section 2.4 Boxplots

Quartiles (and you can compute any percentile you want) are useful information. Boxplots are a bit like dotplots, but also include information about quartiles. They are a visual representation of the five-number summary. **Examples 2.4.1 – 2.4.3.**

The book states that outliers are the most interesting points in a data set. A formal definition of outliers and fences are given on Page 47. (To me, what is means to be an outlier is somewhat arbitrary, i.e. sort of made up.) Modified boxplots are sometimes used instead of boxplots to give better information about the main body of data and the outliers. **Example 2.4.5, Figures 2.4.2/3.**

Section 2.5 Relationships between Variables

Some types of things that we'll see later in the semester (don't stress about details now): **Examples/Figures 2.5.1/2. Figures 2.5.3/4. Table 2.5.3 and Figure 2.5.5.**

Section 2.6 Measures of dispersion

Range of data: **Example 2.6.1.**

The standard deviation (very important) is approximately the *average distance (deviation)* from the values in the data set to the mean. The standard deviation s is the square root of the variance s^2. So why the term "standard deviation"? *Standard* is synonymous with *typical* or *average*, and *deviation* is how *different* a particular observation is from what is normal (the mean). The standard deviation describes how spread out (i.e. how *dispersed*) the data are. **Example 2.6.2, Table 2.6.1, Figure 2.6.1.**

See **Figure 2.6.1**. The average (absolute value of the) deviation of the five observations from the mean of 73 is:

$$\frac{3+1+8+3+9}{5} = \frac{24}{5} = 4.8$$

It turns out that absolute values are not so useful to work with, so an alternative is to instead square each deviation, sum those squares, then square root the sum:

$$\sqrt{\frac{3^2+(-1)^2+(-8)^2+(-3)^2+9^2}{5-1}} \approx 6.4$$

See **Table 2.6.1**. The final version of this formula is given on Page 60 (like what we just did, but dividing by $n-1$ rather than n, where n is the number of observations). **See Page 62**, about 2/3 down the page, for a short discussion on why $n-1$ rather than n in that formula. (For now, don't worry about trying to understand exactly why we divide by $n-1$ rather than n.) **Example 2.6.4.**

How dispersed the data are (as measured by the standard deviation) is somewhat relative. The coefficient of variation is the standard deviation *relative to the mean*. Example: two groups of salaries (in thousands of $).

Group 1: 49, 55, 55, 75, 84, 90, 95, 140, 157, 200

Group 2: 949, 955, 955, 975, 984, 990, 995, 1040, 1057, 1100

For which group does how spread out (that is, the dispersion of, which is what the SD measures) the salaries are matter more? Use Excel.

Group	Mean	SD	Coefficient of variation
1	100	50	50/100 = 0.5
2	1000	50	50/1000 = 0.05

Pay attention to (and understand) the discussion about "Visualizing the Standard Deviation" on page 63. This will be important. The standard deviation tells us how dispersed (spread out) data are. **Example 2.6.6.** Very useful: **Page 64, Typical Percentages**. "Typical" doesn't mean that it is exactly this way for each set of data, it is approximate. **Examples 2.6.7 – 2.6.8.**

Section 2.7 Effect of Transformation of Variables

Linear transformations of data change their mean and standard deviation. Example:

	Original	Data $+ 5$	Data $* 3$	Data $* 3 + 5$
	1	6	3	8
	2	7	6	11
	3	8	9	14
	4	9	12	17
	7	12	21	26
Mean	3.4	8.4	10.2	15.2
Std dev.	2.302	2.302	6.906	6.906

Online: Histograms… Example 2.7.3. On your own: Nonlinear Transformations on Page 70. But don't worry about it too much for this course.

Section 2.8 Statistical Inference
Section 2.9 Perspective

Usually we do not have a complete set of data for the <u>population</u> in which we are interested. We usually have some subset, a <u>sample</u>, of that population. Examples:

Population	Statistic of interest	Sample
Babies born in United States	Average birth weight	Babies born in five randomly selected hospitals in 2017
Pepperdine students	Standard deviation of GPA	Students in Math 316
All moths of a certain type	Average wing span	Moths of that type caught in a certain trap

We generally find the mean $\bar{y}$ and standard deviation s of the sample, and use them as estimates for the mean μ and standard deviation σ of the entire population. **Table 2.9.1.** Greek letter μ "mu" is a Greek "m" which is pronounced as "myu" and σ is a Greek "s" for "standard deviation." A *sample* characteristic (e.g. *mean* or *standard deviation*) is a <u>statistic</u> while for the entire *population* is it called a <u>parameter</u>.

One of the most important issues is the reliability of the statistics of the sample as estimates for the parameters of the entire population. Of course it is critical that the sample accurately represents the entire population as well as possible. That is always a **huge** issue in creating and running studies of any type. The mathematics (the analysis of the data) is usually far easier than designing and running the study.