The chapter summaries will briefly describe the most important concepts covered in the course.

Variables A typical data set is often represented with a matrix of information. Each row represents an individual or *unit*, while each column represents a *variable*. Variables may be *categorical*, where each individual is categorized into a discrete group, or *quantitative*, where each individual is measured on a numerical scale. Quantitative variables are either *discrete* (only take values from some discrete set of possible values) *continuous* (take values from a continuous range of possible values, although the recorded measurements are rounded).

Summation notation You should understand summation notation. For example,

$$\sum_{i=1}^{n} y_i$$

represents the sum of the values of the variable y.

Measures of Center The *mean* and the *median* are two common measures of center. The mean is calculated by summing the values and dividing by the number of values.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

It is the balancing point of the distribution of numbers.

The median is the middle number after they have been sorted. If there are an odd number of values, the median is the number at the unique middle position. If there are an even number of values, the median is average of the values at the two middle positions.

The median and the mean will be about the same for nearly *symmetric* distributions. If a distribution is *skewed to the right* (the right half is more spread out than the left half), the mean will be larger than the median. If a distribution is *skewed to the left* (the left half is more spread out than the right half), the mean will be smaller than the median.

The mean is more affected by extreme values. It may not be "typical" when there are extreme values present. The median is a more *robust* measure of center and is not affected by extreme values. The median may often be "typical" even when there are extreme values present.

Stem-and-leaf Diagrams *Stem-and-leaf diagrams* are a means to rewrite the numbers of a quantitative variable that shows the shape of the distribution using digits. They are not feasible for large data sets. Only one digit (often the last significant digit) is used as the *leaf* of each number with the remainder of the number a stem. If a stem-and-leaf diagram has too many rows, *rounding* and using a different place as the "leaf" can reduce the size. If a stem-and-leaf display has too few rows, *splitting* stems can make a better display. See the textbook for examples.

In a good stem-and-leaf diagram, digits are lined up to make it easier to compare relative frequencies. The list of stems should be regularly spaced, potentially including empty rows. This shows spacing. Digits within a row should be ordered if the desire is to use the stem-and-leaf diagram to find medians and quartiles.

Histograms *Histograms* are bar graphs of quantitative variables. A numerical interval that spans the values of the variable is divided into a number of smaller equally sized intervals and a bar is drawn covering each smaller interval where the height is proportional to the count of observations in the corresponding range.

Histograms with too few intervals over-summarize the shape of the distribution of numbers. Histograms with too many intervals look like broken combs and emphasize too many minor features of a distribution. A good histogram often has between 5 and 20 intervals, more when the sample size is larger.

The median of a distribution may be estimated by finding a vertical line that divides the area of the bars into two regions of equal area. The mean of a distribution may be estimated by finding the place where the bars would balance if it were made from a uniform solid material.

Histograms do a good job at displaying the shape, center, and spread of a distribution.

Measures of Dispersion *Spread* is a general statistical concept that describes the variablity in the distribution of a quantitative variable. One rough measure of spread is the *range* which is the maximum value minus the minimum value. This measure of spread is not very robust as it depends only on two extreme values.

The most commonly used measure of spread is the *standard deviation*. In words, it is (almost) the square root of the mean squared deviation from the mean. The formula is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

The standard deviation may be interpreted as the size of a typical deviation from the mean, provided that the distribution is not highly skewed.

You can roughly estimate the standard deviation from a histogram from this interpretation. Most observations will be closer to the mean than one standard deviation, but there will be a fair number of observations farther away.

For approximately fairly symmetric distributions with a bunch of observations in the middle that tail away in both directions, the standard deviation will be about one fourth the length of an interval that contains the middle 95% of the observations.

The common notation for standard deviation is s. At times I will use the notation SD as well.

The standard deviation is the square root of the *variance*.

Quantiles You should know what percentiles, quartiles, and quantiles are. The *first quartile*, Q_1, is another name for the *25th percentile*. It is the location that cuts off the smallest quarter of the data. The *third quartile* is another name for the *75th percentile*. It is the location that cuts off the largest quarter of the data. There are multiple definitions of quartiles, so quartiles measured from small data sets may be different.

A *five number summary* is the minimum, lower quartile, median, upper quartile, and maximum.

The *interquartile range* or *IQR* is the distance between the first and third quartiles, or $Q_3 - Q_1$. The IQR is a more robust measure of spread than the range.

Modified boxplots A *boxplot* is a graphical display of the five number summary. With the numerical axis drawn vertically, the box represents the middle half of the data from the first quartile to the third quartile. The box is split at the median.

In a regular boxplot, there is a whisker that extends from the top of the box to the maximum and another whisker from the bottom of the box to the minimum. Each whisker repesents one quarter of the data.

Modified boxplots will identify *outliers* (individual observations far from the overall pattern of the data) with individual points. In a modified boxplot, individual observations greater than the *upper fence* ($Q_3 + 1.5$IQR) or less than the *lower fence* ($Q_1 - 1.5$IQR) are noted as outliers. The whiskers would then extend from the box to the most extreme non-outliers.

Parallel boxplots are especially useful for comparing the distributions of two or more different quantitative variables.