

CHAPTER 1: LOOKING AT DATA: DISTRIBUTIONS

1.1 DISPLAYING DISTRIBUTIONS

- **Statistics** is the science of collecting, organizing, and interpreting numerical facts which we call **data**. The goal of statistics is to gain information from data.
- There are two types of statistical goals:
 1. Describe a situation (**descriptive statistics**).
 2. Draw conclusions regarding hypothesized claims (**inferential statistics**).
- A **variable** is a characteristic or property of an individual that can be assigned a value.
- We will consider two types of variables:
 1. A **quantitative variable** assumes numerical values on which arithmetic operations can be performed sensibly.
 2. A **categorical variable** assigns an individual into one of several categories associated with that variable. The *values* of categorical variables are not numbers but labels. Arithmetic operations make no sense with categorical variables.
- The **distribution** of a variable is a description of the values it takes on and how often it takes on those values. The distribution describes the pattern of variation in the values of the variable.

The goal of Section 1.1 is to study the overall pattern of a set of data. We will use graphical tools to visualize the data. By displaying data, we can gain information on the variable that is measured. For a quantitative variable, we can see where the data are located, the spread of the data, and possible trends. For a categorical variable, we can observe patterns regarding how the data fall into the various categories.

NOTES:

We will cover four graphical methods for displaying data: bar graphs, stemplots, histograms, and time plots. Bar graphs are used with categorical variables whereas the others are all used with quantitative variables.

Bar Graphs: For each category of *one* variable, a bar is plotted. The heights of the bars correspond to the percentages in the marginal distribution of that variable. The horizontal axis contains the labels of the categories. The sum of the heights *across* the bars = 100%.

Stemplots and Histograms can help us

1. Locate the *center* of the distribution.
2. Summarize its *overall shape*. This includes checking if the distribution is
 - (i) **symmetric** (a mirror-image about the center)
 - (ii) **skewed to the left** (the left tail is longer than the right tail)
 - (iii) **skewed to the right** (the right tail is longer than the left tail).
3. Check for *obvious deviations* from the overall shape. Look for gaps or clustering in the distribution. Gaps are often formed by **outliers** which are observations not in accord with the other observations.

Graphical Method 1: **Stemplot** or **stem-and-leaf plot**

1. Create a **stem** and a **leaf** criterion. The stem is based on the leading digits while the leaf is based on the trailing digits.
2. List stem values in increasing order against a vertical bar.
3. Arrange leaves in increasing order from left to right against the bar along side of the appropriate stem value.

Graphical Method 2: Frequency or Relative Frequency **Histogram**

1. Divide the range of the data into non-overlapping classes of equal width.
2. Count the number of observations in each class. This is the *class frequency*. This is most easily done by making a table of all class frequencies, i.e. the number of individuals falling into each class. This is called a **frequency table**. For a **relative frequency table** divide the frequencies by the total number of observations in the data set. The relative frequency is the proportion of observations belonging to that class.
3. Draw the histogram. The horizontal scale corresponds to the classes and the vertical scale to the frequency (or relative frequency). A vertical bar is drawn above each class with the bar height being the class frequency (or relative frequency).

NOTES:

Graphical Method 3: Time Plot

1. Plot the values of the variable against the time order of the observations.

A **time series** is a set of measurements of a variable taken at regular intervals of time. An *index number*, such as the Consumer Price Index, is an example of a time series.

A *time plot* can show variation over time which can take on the form of

- (i) a **trend** (long-term change)
- (ii) **seasonal variation** (change occurring during specific time periods)
- (iii) **irregular fluctuations** (changes caused by unusual events)
- (iv) **cycles** (distinct up and down movement which is less regular than seasonal variation and not explained by seasonal effects).

1.2 DESCRIBING DISTRIBUTIONS

The center of a distribution: Mean and median.

- Let x_1, x_2, \dots, x_n be a set of n observations. The symbol Σ (called sigma) means to take a sum.
- Measure 1: The **mean** or **average** of a set of n observations, denoted \bar{x} , is defined as

$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum x_i.$$

- Measure 2: The **median** of a set of n observations, denoted M , is the midpoint of the distribution (the point at which half the observations fall above it and half below it).
- To find the median M :

1. Arrange the n observations in increasing order.
 2. If n is odd, the median M is the center observation.
 3. If n is even, the median M is the average of the two center observations.
- A measure is a **resistant measure** if it is insensitive to the influence of extreme observations.
 - For the center of a distribution, the mean is not a resistant measure while the median is a resistant measure.
 - The mean versus the median:
 - For perfectly symmetric distributions, the mean and the median are equal.
 - For right-skewed distributions, the mean is larger than the median.
 - For left-skewed distributions, the mean is smaller than the median.

Resistant measures of the spread of a distribution

- Goal: Develop strategies for dealing with outliers or other unusual data points.
 1. Detect outliers and investigate possible causes. Corrections can be obvious or, if justified, outliers can be deleted.
 2. When outliers cannot be deleted, resistant measures can be used so that outliers have little influence over conclusions.
- The p^{th} **percentile** of a distribution is the value such that p percent of the observations fall at or below it. Often, we are interested in
 - The 25th percentile, called the **first quartile** Q_1 .
 - The 75th percentile, called the **third quartile** Q_3 .
 - The 50th percentile which is the median M .
- Calculating quartiles:
 - Arrange the observations in increasing order.
 - If n is even, Q_1 is the median of the first half of the observations and Q_3 is the median of the second half of the observations.
 - If n is odd, there are equal numbers of observations below and above the median. Q_1 and Q_3 are the medians of the observations below and above the median, respectively.
- The **interquartile range** IQR is the range of values for the middle 50 percent of the distribution. Thus, $IQR = Q_3 - Q_1$. The IQR is a resistant measure of spread *about the median* and, therefore, is most useful when center of the distribution is measured by the median.
- The **Minimum** and **Maximum** are the smallest and largest observations.
- The **five-number summary** of a distribution consists (in increasing order) the Minimum, Q_1 , M , Q_3 , and the Maximum.

- As a rule of thumb for identifying outliers, treat any observation more than $1.5 \times IQR$ beyond the first or third quartile as a potential outlier.

NOTES:

Boxplots: Graphically displaying the 5-number summary.

1. Next to a vertical axis, draw a box so that the top of the box rests at Q_3 and the bottom rests at Q_1 . Thus, the box length is the IQR .
 2. Draw a horizontal line in the box at the value of the median M .
 3. Extend two lines (called whiskers) from the top and bottom of the box to the Maximum and Minimum observations.
- Some people prefer a **modified boxplot** which plots the whiskers at the Minimum and Maximum *only if* these values are less than $1.5 \times IQR$ beyond the quartiles. Otherwise, end the whiskers at the most extreme observations still within $1.5 \times IQR$ of the quartiles. In either case, always end the whiskers at observations.
 - The boxplot displays information regarding the spread and center of the distribution as well as skewness.
 - **Side-by-side boxplots** are used for displaying relations between a categorical variable and a quantitative variable.
 - Label the horizontal axis with the categories of the categorical variable. Above each category, draw the boxplot for the y values corresponding to that category.
 - Look for any overall pattern or unusual boxplots. We can make a side-by-side comparison of the distributions for each category.

NOTES:

Variance and Standard Deviation.

- The **deviation** of observation x_i about the mean is $x_i - \bar{x}$.
- The **variance** of a set of n observations, denoted s^2 , is

$$s^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] = \frac{1}{n-1} \sum (x_i - \bar{x})^2.$$

- The **standard deviation** s is the positive square root of the variance s^2 . The standard deviation is in the same units of measurement as the observations.
- The sum of all deviations is 0. That is, $\sum (x_i - \bar{x}) = 0$.
- The variance formula can be rewritten to make calculation easier and to reduce roundoff error:

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

- The standard deviation is a non-resistant measure of spread *about the mean*.
- If $s = 0$, then there is no spread and all observations are identical. If any two observations are different, then $s > 0$.
- A **linear transformation** changes the original variable x into a new variable x^* by an equation of the form $x^* = a + bx$, where $b \neq 0$.
- Linear transformations do not affect the general shape of a distribution. However, for the distribution of x^* ,

1. The mean, median, and quartiles are found by substituting the mean, median, and quartiles for x into $a + bx$:

$$\bar{x}^* = a + b\bar{x} \qquad M^* = a + bM \qquad Q_1^* = a + bQ_1 \qquad Q_3^* = a + bQ_3$$

2. The interquartile range and the standard deviation are found by multiplication by b :

$$IQR^* = b \times IQR \qquad s^* = b \times s$$

- Therefore, it is not necessary to calculate any measures for x^* from the original observations once the corresponding measures are calculated for x .

NOTES:

1.3: THE NORMAL DISTRIBUTIONS

Density curves:

- A **mathematical model** provides a description of the overall shape of a distribution after omitting outliers and other deviations from a regular pattern.
- A mathematical model is often represented by a mathematical formula or is summarized in tables.
- Histograms can be approximated by a smooth curve. This curve displays the shape of the distribution.
- Because the sum of all relative frequencies = 1, the smooth curve used to describe the distribution will have a total area = 1 underneath it.
- For a **density curve**, the area under the curve for a range of values = the proportion of observations that fall within this range.
- The p^{th} **percentile** on a density curve is the point where p percent of the area lies to the left of that point and the remaining $(100 - p)$ percent lies to the right of that point.
- For many density curves, finding the mean, standard deviation, and percentiles for a density curve involves advanced mathematical methods (including calculus). To make things convenient, tables that include percentiles exist for commonly used density curves.

- Notation: to distinguish between the mean and standard deviation of observed data and the density curve, we use

	Observed	Density Curve
Mean	\bar{x}	μ
Standard Deviation	s	σ

NOTES:

Normal distributions:

- The **normal distribution** is a bell-shaped distribution that is symmetric about μ having the following property called the **68-95-99.7 Rule**:
 1. 68% of observations fall within σ of the mean μ .
 2. 95% of observations fall within 2σ of the mean μ .
 3. 99.7% of observations fall within 3σ of the mean μ .
- There are infinitely many different normal distributions. What uniquely defines a normal distribution is its mean μ and standard deviation σ . We denote a normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$.
- Any variable obtained from a linear transformation of a normal variable is also a normal variable. However, the mean and standard deviation of the transformed variable can change (following the rules described in Section 1.2).
- A normal distribution having mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution** and is denoted $N(0, 1)$.
- Important: If the variable X follows a normal distribution with mean μ and standard deviation σ , then the **standardized variable** $Z = \frac{X - \mu}{\sigma}$ follows the standard normal distribution. Or, in notational form,

$$\text{If } X \text{ is } N(\mu, \sigma), \text{ then } Z = \frac{X - \mu}{\sigma} \text{ is } N(0, 1).$$

- An observation is **standardized** by subtracting the mean and then dividing by the standard deviation.
- Interpretation: A standardized observation = the number of standard deviations the original observation is from the mean. Its sign (+ or -) indicates the direction from the mean (right or left).
- The **standard normal table** or **Z-table** gives the area under the standard normal curve to the left of Z . This table is accurate to two decimal places of accuracy for Z . The row refers to the integer and first decimal place of Z and the column refers to the second decimal place of Z . The table entry in that row and column
 - = the area under the standard normal curve to the left of Z
 - = the relative frequency of observations $\leq Z$.
- The relative frequencies associated with ANY normal distribution $N(\mu, \sigma)$ can be determined using the standard normal distribution $N(0, 1)$. The procedure is to standardize the observed variable X creating Z and then use the standard normal table to find the appropriate area.

NOTES:

Calculation rules for normal distributions:

If a variable X is $N(\mu, \sigma)$, then

1. the relative frequency (proportion) of observations to the left of some value x equals the area to the left of $z = \frac{x - \mu}{\sigma}$.
2. the relative frequency of observations to the right of some value x equals the area to the right of $z = \frac{x - \mu}{\sigma}$, or equivalently, it equals

$$\left[1 - \text{the area to the left of } z = \frac{x - \mu}{\sigma} \right].$$

3. the relative frequency of observations between two values x_1 and x_2 equals the area between $z_1 = \frac{x_1 - \mu}{\sigma}$ and $z_2 = \frac{x_2 - \mu}{\sigma}$. This equals

(Area to the left of z_2) - (Area to the left of z_1).

– General idea reiterated: Convert the X 's to Z 's by standardizing and then use the standard normal table.

NOTES:

CHAPTER 2: LOOKING AT DATA: RELATIONSHIPS

INTRODUCTION TO CHAPTER 2

- Goal: Study the relationships between two or more variables.
 - * Is a change in one variable associated with a change in another variable?
 - * If yes, what kind of change is it and how large is that change?
 - * Is the purpose to describe the nature of the relationship? Or, is it to show that variation in one of the variables can cause variation in the other?
- Recall the two types of variables to be studied:
 1. *Quantitative variables* take on numerical values for which numerical measures, such as means and standard deviations, are meaningful.
 2. *Categorical variables* record which of two or more categories an observation falls. Numerical measures, such as means and standard deviations, are **not** meaningful for categorical variables.
- In a statistical study,
 - * A **case** is an individual person, animal, or object for which values of variables are recorded.
 - * A **response variable** measures an outcome of the study.
 - * An **explanatory variable** is a variable chosen by the researcher with the intent of explaining the variability in the response variables.
- Response variables are often called *dependent variables*.
- Explanatory variables are often called *independent variables*.

NOTES:

2.1 SCATTERPLOTS

- Goal: To graphically display relationships between two variables. The graphs should give us some insight regarding the nature of the relationship.
- A **scatterplot** is a graphical representation of the relationship between two quantitative variables. For each case in the study, a pair of observations x and y corresponding to the two variables are plotted as points. The explanatory variable, x , is plotted on the horizontal scale of a scatterplot.
- Interpreting scatterplots:
 - * Look for an overall pattern that shows the form, direction, and strength of the relationship.
 - * Look for potential outliers or unusual deviations from the overall pattern.
- The form indicates the general shape of the pattern. In Chapter 2, the focus is studying forms that indicate linear relationships.
- The direction indicates whether there exists a clear increasing or decreasing pattern. We say two quantitative variables are
 - * **Positively associated** when increasing values of one variable tend to accompany *increasing* values of the other variable (and vice versa).
 - * **Negatively associated** when increasing values of one variable tend to accompany *decreasing* values of the other variable (and vice-versa).
 - * Because most categorical variables have no natural order from smallest to largest, we cannot speak of a negative or positive association between a categorical and a quantitative variable.
- The strength of the relationship is reflected by the amount of scatter about the pattern. For example, if the pattern is linear, are the points tightly fitted about the line (strong linear relationship) or are the points scattered widely about the line (weak linear relationship).

NOTES:

CHAPTER 2: LOOKING AT DATA: RELATIONSHIPS

2.2 CORRELATION

- **Situation:** The data consist of n observations. Each observation is an (x, y) pair:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

- **Goal:** Measure the strength of the *linear* relationship between the two variables x and y . We are not assuming as in the regression case that x is an explanatory variable and y is a response variable (although they may be).
- The statistic used is **the correlation coefficient** r , where

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right).$$

where s_x and s_y are the standard deviations of the x and y observations.

- To make computation easier, use the computing formula for r :

$$r = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{(n-1)s_x s_y}.$$

- Facts and Properties of the correlation coefficient r :
 1. $-1 \leq r \leq 1$
 2. If $r > 0$, there is a *positive* association between x and y .
 3. If $r < 0$, there is a *negative* association between x and y .
 4. The *stronger* the association, the closer r gets to 1 or -1 .
 5. The *weaker* the association, the closer r gets to 0, .
 6. If $r = 1$ or $r = -1$, all of the points (x_i, y_i) lie perfectly on a line.
 7. r measures only the strength of a linear relationship. If x and y follow a curved relationship, the strength of the association need not and often is not reflected in the value of r .
 8. If the scale of the measurement units for x and/or y is changed by a linear transformation, the value of r remains unchanged.
 9. The correlation coefficient r is *dimensionless*, that is, it has no unit of measurement.

NOTES:

2.3 LEAST SQUARES REGRESSION

- Goal 1: Fit a line through the data when there appears to be a linear pattern in the scatterplot, that is, there is a **linear dependence** of a quantitative response variable y on an explanatory variable x .
- The straight line that describes the dependence of one variable on another is called a **regression line**.
- We want the line to come as “close” as possible to the points. Because many possible lines seem “close” to the points, we need an objective method of finding a regression line. The most commonly used method is the **method of least squares**. The mathematical details will follow.

- Recall from algebra: The equation of a straight line has the form

$$\hat{y} = a + bx,$$

where b is the *slope*, a is the *intercept* for the explanatory variable x , and \hat{y} is the predicted response. The slope, b , is the change in \hat{y} corresponding to a unit increase in x . The intercept is the value of \hat{y} when $x = 0$.

- Goal 2: Once the line is fit, then we want to know how to use the regression line to gain information.
 - We can use the regression line to *predict* the response \hat{y} for a given value of the explanatory variable x .
 - The accuracy of a prediction from a regression line depends on the amount of scatter about the line – the less scatter, the more accurate the prediction.

- When we refer to scatter about the line, we are studying the deviations of the observed y values from the regression line. Formally, a **deviation** is the vertical distance between y and the regression line.
- A **residual** is the difference between the observed response y_i and the point \hat{y}_i on the regression line, that is,

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y_i - \hat{y}_i \\ &= y_i - (a + bx_i) \end{aligned}$$

- The **method of least squares** finds the regression line values of a and b that minimize the sum of the squared residuals, that is, the method of least squares minimizes

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2.$$

- The solution yielding the **least squares regression line of y on x** is

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \text{ and } a = \bar{y} - b\bar{x}.$$

- For easier computation, use the computing formula for b :

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}.$$

NOTES:

Relationship between Correlation and Regression

- The square of the correlation coefficient, r^2 , is the fraction (or proportion) of the variation in the y -values that is *explained* by the regression of y on x . Therefore, the larger the r^2 , the better the regression line fits the data.

- Exchanging the roles of x and y (that is, regressing the x values on the y values) does not change the value of r^2 .
- By manipulation of the slope and correlation formulae, the slope of the least squares regression line, b , can be written in terms of the correlation coefficient r :

$$b = r \frac{s_y}{s_x}.$$

NOTES:

2.4 LIMITATIONS OF CORRELATION AND REGRESSION

- We are considering only linear relationships.
- r and least squares regression are *not resistant* to extreme values.
- There may be variables other than x which are not studied yet influence the response variable. The presence of these lurking variables can make results misleading.
- A strong correlation does not imply a cause and effect relationship.
- It is dangerous to extrapolate, that is, make a prediction at values outside the range of x .
- Consider the case where the y_i values used are actually *averages* of individual observations. Using averaged data in regression does not allow for *individual unit* predictions. Correlations based on averages usually are overestimates when applied to individuals.
- The correlation coefficient r is only one measure regarding the distributions of variables x and y . It is wise to also calculate other measures (from Chapter 1) to get a more complete description of the data.

- A **residual plot** is a plot of the $(y_i - \hat{y}_i, x_i)$ pairs. That is, a scatterplot plot of the residuals vs. the x values. The pattern of the residual plot indicates how well the regression line fits the data.
 - A ‘good’ fit has a residual plot whose points are centered about 0 and scattered randomly about 0. The closer the residuals are to 0, the better the fit.
 - A ‘bad’ fit will have large residuals (either positive or negative) OR a non-random pattern in the residual plot. A fan-shaped pattern indicates that the variation in y increases as x increases.
 - The residuals will have a mean of zero regardless of how well or how poorly the regression line fits the data.
- An observation (or x, y pair) is an **outlier** in a regression if it produces a large residual.
- An observation is **influential** if removing it would markedly affect the regression, that is, it would substantially change the position of the regression line.
- Least squares regression is not a resistant measure of linear association. Outliers as well as non-outliers can be influential observations.
- Although a residual plot will call attention to outliers, it will not necessarily indicate which observations are influential. That is, influential observations may or may not be outliers.
- An influential observation should be investigated to verify that it is correct. Even if it is correct, you should question whether it belongs to the population under study.
- A **lurking variable** is a variable that has an important effect on the response but is not included among the explanatory variables studied. Lurking variables sometimes vary systematically over time. Accordingly, by plotting y against the time order of the observation, evidence of the effect of lurking variables on y may appear.

NOTES:

2.5 THE QUESTION OF CAUSATION

- **Question:** When studying the relationship between two variables, does a change in the explanatory variable *cause* a change in the response variable? Or, in other words, is there a **cause-and-effect** relationship between the variables?
- **Important Note:** A strong association between two variables *does not imply* a causal link between the variables.
- There are three possible explanations (the three C's) for an association between x and y :
 1. **Causation:** A change in x causes a change in y .
 2. **Common Response:** Changes in both x and y are caused by changes in unobserved lurking variable(s).
 3. **Confounding:** The effect of the explanatory variable x on the response y cannot be determined because it is mixed with (cannot be separated from) the effects of other variables on y .
- To establish causation, a carefully designed experiment should be run in which the effects of possible lurking variables are controlled. In the experiment, the levels of x are changed by the researcher and y is observed.
- A designed experiment controls influences that are not of interest and changes the levels of the variables that are of interest. By doing so, we can determine which variables, when changed, will cause the response to change.
- When a designed experiment is not possible, causation should only be cautiously accepted, and the following should hold true:
 1. The association between x and y must be observed in many studies of different types among different groups. This would indicate the association is unlikely to be attributable to lurking variables.
 2. When effects of other variables are taken into account, the association between x and y still holds true.
 3. A causal relationship between x and y must have a plausible explanation.

NOTES:

CHAPTER 3: PRODUCING DATA

Goal: To study methods of producing data, and to determine which methods produce trustworthy data. This includes studying the basic ideas of statistical designs for

- Choosing a sample that represents a larger population.
- Laying out an experiment to study the effects of explanatory variables on a response.

3.1 FIRST STEPS

- An **exploratory data analysis** is an analysis which is usually performed on historical (or previously collected) data. Emphasis is placed on graphical methods to find data patterns showing the relationships among variables and to suggest further study. Conclusions cannot be generalized beyond the specific data observed. Conclusions regarding causality are not trustworthy.
- **Formal statistical inference** is the branch of statistics which attempts to answer specific questions with a known degree of confidence. The conclusions can usually be generalized beyond the specific data. One goal is to learn techniques of data collection that allow for successful statistical inference.
- A **statistical design** is an arrangement for collecting data from many individuals (objects). A good design takes the following questions into consideration:
 - (a) how many individuals are to be studied?
 - (b) how to select individuals for study?
 - (c) how to form groups of individuals when the study involves groups of individuals receiving different treatments?
- **Anecdotal evidence** is based on cases that were haphazardly selected and, therefore, may or may not be representative of any larger group of cases. Conclusions based on anecdotal evidence are not trustworthy.
- **Available data** are data that were collected in the past for other purposes (e.g. census data), but may prove useful in answering current questions.
- Good statistical designs for producing data rely on either **sampling** or **experiments**.
- Sampling *selects* a part of the population of interest to represent the whole population. The purpose of sampling is to collect information about some aspect(s) of that population.
- In an **experiment**, a treatment or a combination of treatments is actively *imposed* on the cases. The goal of an experiment is to gain information regarding the effects (if any) of the treatments on the response of interest.

- In an **observational study**, treatments are not actively imposed on the cases. Data are collected by observing cases under natural conditions. Conclusions regarding causality are tentative because lurking variables may be confounded with the explanatory variable. Conclusions regarding causality can be made with more confidence when a designed experiment is used.
- A **census** collects data on *every* case in the population of interest.

NOTES:

3.2 DESIGN OF EXPERIMENTS

- In an experiment, the **experimental units** are the objects subjected to specific experimental conditions. The specific experimental conditions are called **treatments**.
- If the experimental units are humans, they are called **subjects**.
- An explanatory variable in an experiment is often called a **factor**.
- A specific value or setting of a factor is called a **factor level**.

Benefits of Designed Experiments over Observational Studies

- (1) Well-designed can yield evidence for cause-effect relationships.

- (2) Allows for the study of effects of factors that are of particular interest.
- (3) Allows for the control of factors not of interest.
- (4) Allows for the study of combined effects of several factors simultaneously, and of interactions among the factors.

Two Common (but Weak) Design Formats

- (1) Apply treatment \rightarrow Observe response
- (2) Take an initial observation \rightarrow Apply treatment \rightarrow Take a final observation

Comparative Experiments

- In a **comparative experiment**, two or more treatments are applied to the set of experimental units. The responses of experimental units having different treatments are compared.
- The two design formats above are **not** comparative experiments because there is only one treatment applied to the experimental units.
- Comparative experiments
 1. Are more successful than uncontrolled (that is, no comparison group) or observational studies
 2. Reduce the possibility of *confounding* treatment effects.

Three Principles of Experimental Design

- Principle 1: Control the effects of factors that are not of interest (e.g. lurking variables).
 - A **placebo effect** is a response to a dummy treatment. A group of subjects given a placebo or no treatment is called a **control group**. The responses of units in the treatment group(s) are compared to the responses of units in the control group.
 - An experimental design or study is **biased** if it systematically favors certain outcomes.
- Principle 2: Random Assignment is the process of randomly assigning experimental units (subjects) to treatments to create treatment groups that are similar (except for chance variation) before treatments are applied.
 - Note: It is the experimental units that are randomly assigned.
 - Random assignment reduces the chance of systematic differences and reduces the chance of confounding the treatment effects.
 - When an observed difference in treatment effects is too large to reasonably have occurred purely by chance, we say that the difference is **statistically significant**.
 - If significant differences among treatments are found after running a comparative randomized experiment, we conclude that the differences are *caused* by the treatments.

- A mechanism is required to perform the process of randomization. Common mechanisms include (i) physical mechanisms, such as drawing a name from a hat (ii) a table of random digits, and (iii) a computerized random number generator.
- Principle 3: Replication is the process of repeating some or all treatments on additional experimental units.
- Replication can reduce the effects of chance variation. This will make the experiment more sensitive for detecting systematic effects of the treatments.

NOTES:

Cautions about Experimentation

- A **hidden bias** is a bias that occurs because of the way the experiment was conducted (despite the use of comparison and randomization).
- A **double-blind experiment** is an experiment where neither the subjects nor the evaluators know which treatment a subject received. This can prevent hidden bias. At least one member of the research team must know which treatment each subject received. Otherwise, the data cannot be analyzed for evidence of treatment effects.
- Another potential weakness of an experiment can be the **lack of realism**. This occurs when the experiment results do not reflect what happens in situations wider in scope than the experiment.

Types of Experimental Designs

- A **completely randomized design** is a design in which the experimental units are randomly assigned to the treatments.
- A **block** is a group of experimental units or subjects that are known (or suspected) to be similar before the experiment is run, and therefore, are expected to respond similarly after receiving a treatment.
- In a **block design**, randomization occurs by randomly assigning experimental units to treatments within each block.
- A **matched pairs** design is a block design in which there are only two observations per block.
- Blocks are another form of *control* because effects of outside variables are controlled by bringing units with common outside characteristics into a block.

NOTES:

3.3 SAMPLING DESIGN

- A **population** is the *entire* group of cases or people about which information is wanted or needed.
- A **unit** is an individual case or person in the population.
- A **sample** is that portion of the population that is examined in order to gather information.
- Conclusions regarding the population are based on an analysis of the data collected in the sample.
- Samples based on *voluntary responses* tend to over-represent certain portions of the population. Voluntary responses are commonly from people with very strong opinions (usually negative opinions).
- A sampling scheme displays **bias** if the sample systematically favors certain parts of the population over others. If certain parts of the population are overrepresented in the sample, while other parts are underrepresented, then sampling bias exists.

Probability Sampling Designs

- A **probability sampling design** is a design which assigns each member of the population a known nonzero chance of being selected.
- Design 1: A **simple random sample (SRS)** of size n is a sample of n units selected from a population in such a way that every sample of size n has an equal chance of being selected.
- Design 2: A **stratified random sample** is a sample selected by first dividing the population into non-overlapping groups called **strata** and then taking a simple random sample within each stratum. Dividing the population into strata should be based on some criterion so that units are *similar within* a stratum but are *different between* strata.
- Stratified sampling *restricts* random selection at some level. A SRS does not restrict random selection.
- Samples can suffer from bias due to any of the following:
 - **Under-coverage**: The sample fails to represent the entire population because complete information about the population is inaccurate or unavailable.
 - **Nonresponse**: Selected individuals fail or refuse to respond.
 - **Response Bias**: The response given by the respondent may have been influenced by improperly or poorly worded survey questions or by the behavior of the interviewer. Response bias can occur because respondents may lie when presented with questions of a personal nature or questions regarding illegal activity.

NOTES:

Recall the following statistical principles:

- Sampling and experimental designs are based on the deliberate use of chance when collecting data.
- Random assignment is used to eliminate systematic favoritism or *bias* when assigning experimental units to treatments in an experiment.
- Random selection is used to eliminate systematic favoritism or *bias* when selecting units to be in a sample.
- The goal of sampling and experimentation is to draw conclusions about an underlying population of interest based on the data collected. This process of drawing conclusions is called *statistical inference*.

Toward Statistical Inference

- Statistical inference is often concerned with drawing conclusions about some numerical value associated with the population.
- A **parameter** is a value which describes some characteristic of a population (or possibly describes the entire population).
- A **statistic** is a value that can be computed from the observed (sample) data without making use of any unknown parameters.
- In general, the value of a parameter is unknown. Statistics computed from experimental or sampling data can provide information about the unknown parameter.
- Because only a part of the population is sampled in any sampling plan, the value of a statistic will vary in repeated random sampling. The inherent variability of a statistic associated with sampling is called **sampling variability**.
- The **sampling distribution** of a statistic is the distribution of values taken by the statistic over *all* possible samples of the same size selected from the same population.
- Frequently, the form of the sampling distribution of the statistic is known, and often its form is approximately normal.
- The **bias of a statistic** is the numerical difference between the center (mean) of the sampling distribution of the statistic and the true parameter value of interest.
- A statistic is **unbiased to estimate a parameter** if the center (mean) of the sampling distribution is equal to the true parameter value. In essence, the estimator is “on aim”. In practice, the mean of the sampling distribution will often equal the true parameter value.
- The **variability** of a statistic is described by the spread of its sampling distribution – the larger the spread in the sampling distribution, the larger the variability of the statistic.
- Effects of Random Selection vs. Effects of Sample Size

- Random selection produces a sampling distribution and eliminates bias. Therefore, properly chosen statistics calculated from data produced from designs using random selection will have no bias.
- Variability is controlled by the size of the sample. Increasing the size of the sample decreases the variability of the statistic.

NOTES:

CHAPTER 4: PROBABILITY: THE STUDY OF RANDOMNESS

4.1 RANDOMNESS

- A phenomenon (that which is observed or experienced) is **random** if the outcome of a single repetition is uncertain, but, when studying a large number of repetitions, there exists a regular distribution of relative frequencies.
- The mathematical study of randomness is called **probability theory**. Intuitively, a probability is a *long-term relative frequency*.
 - Probability theory enables us to study the orderly behavior of sample statistics which are used for gaining information about unknown population parameters.
 - When data come from a probability sample or from a randomized experiment, the values of the statistic are random.
 - Probability theory describes how the statistic will vary in repeated samples or repeated experiments when the underlying population remains unchanged.
- Because we cannot observe a random phenomenon indefinitely, we cannot observe a probability *exactly*. Mathematical probability is an *idealization* of what would happen to relative frequencies in an infinitely long series of trials.
- A second intuitive interpretation of probability is based on *personal opinion*, that is, probabilities are assigned to events based on a person's subjective opinion. Use of personal probabilities, however, leads us away from the mathematical study of long-term regularity of relative frequencies.
- Statistical sampling or experimental designs *deliberately* introduce randomness to gain information about the regular long-term pattern. This chapter will concentrate on the relative frequency interpretation of probability.

NOTES:

4.2 PROBABILITY MODELS

- Goal: Give a mathematical description called a *probability model* for randomness.
- A **probability model** consists of
 - (i) A **sample space** S which is the set of all possible outcomes of a random phenomenon.
 - (ii) An assignment of probabilities P to the outcomes in S .
- An **event** A is a set of outcomes of a random phenomenon (in other words, it is a subset of the sample space). We write the probability of the event A occurring as $P(A)$.
- Four Basic Probability Rules:

- **Rule 1:** Any probability $P(A)$ takes on a value between 0 and 1 inclusively, that is $0 \leq P(A) \leq 1$.
- **Rule 2:** $P(S) = 1$, that is, the probability of observing an outcome from the set of all possible outcomes is 1.
- **Rule 3 - Complement Rule:** For any event A , the event that A does not occur is called the **complement** of A . We denote the complement of A by A^c . The **complement rule** is

$$P(A^c) = 1 - P(A).$$

- **Rule 4 - Addition Rule for Disjoint Events:** Two events A and B are disjoint if they do not share any outcomes. If A and B are disjoint then the probability either A or B occurs is

$$P(A \text{ or } B) = P(A) + P(B).$$

- A **finite sample space** is a sample space with a finite number of possible outcomes.

General rules for probabilities when the sample space is finite:

- (1) Probabilities are assigned to each individual outcome such that
 - (a) each probability is between 0 and 1 inclusively and
 - (b) the sum of probabilities for all outcomes in the sample space = 1.
- (2) The probability of an event = the sum of the probabilities of the outcomes making up that event.

NOTES:

Equally Likely Outcomes

- Special Situation: A random phenomenon has k possible outcomes, and each outcome is *equally likely* to occur.
- Probabilities: Each individual outcome has probability $1/k$.
The probability of any event A is

$$\begin{aligned} P(A) &= \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} \\ &= \frac{\text{number of outcomes in } A}{k} \end{aligned}$$

- Warning: Most random phenomena *do not* have equally likely outcomes. In those cases, we will follow the general rules for finite sample spaces.

Independence and the Multiplication Rule

- Two events A and B are *independent* if the probability that one occurs does not change once you know that the other has occurred.
- **Rule 5 - Multiplication Rule for Independent Events:** The probability that both A and B occur is

$$P(A \text{ and } B) = P(A)P(B)$$

when A and B are independent events.

NOTES:

4.3 RANDOM VARIABLES

- A **random variable** is a variable whose value is a numerical outcome of random phenomenon.
- We use capital letters, such as X and Y , to denote a random variable.
- We use small letters, such as x and y , to denote a specific values that X and Y can take on.
- A statistic calculated from a random sample or a randomized comparative experiment is an example of a random variable.
- The assignment of probabilities for the values of a random variable X is called the **probability distribution** of X .
- Discrete Random Variables

- A **discrete random variable** X is a random variable which can assume only a countable number of values.
- Let x_1, x_2, \dots, x_k be the values X can assume. A *probability model* for X is given by assigning a probability p_i to each x_i , that is,

$$P(X = x_i) = p_i.$$

- The set of probabilities must satisfy each of the following:
 1. $0 \leq p_i \leq 1$ for each i
 2. $p_1 + p_2 + \dots + p_k = 1$
- The probability assigned to an event A , that is $P(X \text{ in } A)$, is found by summing the p_i for the outcomes x_i making up A .
- A **probability histogram** of a discrete random variable is a histogram of the probability distribution. In effect, the probability histogram is a relative frequency histogram for a very large number of trials.

- Continuous Random Variables

- A **continuous random variable** X is a random variable which can take on *any* value in some interval of real numbers. This interval of real numbers may be finite or infinite in length.
 - A probability distribution for a continuous random variable X is determined by assigning for each event A (a sub-interval), the probability $P(A)$ which equals the area *above* A on the real number line and under a curve describing the distribution called a **density curve**.
 - A density curve $p(x)$ satisfies each of the following:
 1. $p(x) \geq 0$ for all x .
 2. The total area under $p(x)$ is 1.
 - Assigning probabilities for continuous distributions:
 1. Note: Any individual outcome c can be thought of as an interval of length 0. Therefore, a probability of 0 is assigned to individual outcomes, that is, $P(X = c) = 0$ for any outcome c .
 2. $P(X \geq x) = P(X > x)$.
 3. $P(X \leq x) = P(X < x)$.
 - A *normal random variable* is a continuous random variable. When calculating normal probabilities using the standard normal table, we are following the probability rules established for continuous random variables.
- The **standard normal table** or **Z-table** gives the area under the standard normal curve to the left of Z . This table is accurate to two decimal places of accuracy for Z . The row refers to the integer and first decimal place of Z and the column refers to the second decimal place of Z . The table entry in that row and column
= the probability that a randomly selected observation will be $\leq Z$.
 - The probabilities associated with ANY normal distribution $N(\mu, \sigma)$ can be determined using the standard normal distribution $N(0, 1)$. The procedure is to standardize the observed variable X creating Z and then use the standard normal table to find the appropriate area.

NOTES:

Calculation rules for normal distributions:

If X is $N(\mu, \sigma)$, then

1. $P(X \leq x)$ equals the area to the left of $z = \frac{x - \mu}{\sigma}$.
2. $P(X \geq x)$ equals the area to the right of $z = \frac{x - \mu}{\sigma}$, or equivalently, it equals

$$\left[1 - \text{the area to the left of } z = \frac{x - \mu}{\sigma} \right].$$

3. $P(x_1 \leq X \leq x_2)$ equals the area between $z_1 = \frac{x_1 - \mu}{\sigma}$ and $z_2 = \frac{x_2 - \mu}{\sigma}$. This equals

$$(\text{Area to the left of } z_2) - (\text{Area to the left of } z_1).$$

- General idea reiterated: Convert the X 's to Z 's by standardizing and then use the standard normal table.

NOTES:

4.4 MEANS AND VARIANCES OF RANDOM VARIABLES

- Goal 1: Expand the definitions of means and variances to distributions of random variables.
- Goal 2: Learn how to compute means and variances of distributions of random variables, and study the laws they obey.
- If all outcomes are not equally likely (this is true for most random variables), then the mean of the random variable is not a simple average. The mean is a *weighted* average of the outcomes, where the weights are the probabilities associated with the outcomes.
- The symbol μ_X is used to represent the mean of a probability distribution of a random variable X .
- If X is a discrete random variable which takes on the values x_1, x_2, \dots, x_k with probabilities p_1, p_2, \dots, p_k , respectively, then the **mean of the random variable X** is given by

$$\mu_X = x_1 p_1 + x_2 p_2 + \dots + x_k p_k = \sum x_i p_i.$$

The **variance of the random variable X** is given by

$$\sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \dots + (x_k - \mu_X)^2 p_k.$$

The standard deviation of X is the square root of the variance and is denoted by σ_X .

- Note: Because the probabilities sum to 1, a total weight of 1 is distributed among the outcomes when calculating the mean and variance.

- If X is a continuous random variable, then calculation of the mean and variance of X requires advanced mathematics. For *symmetric* continuous distributions, such as the normal distribution, the mean lies at the center of the distribution.
- The **law of large numbers** states that the observed sample mean, \bar{x} , *approaches* the mean μ_X of the distribution of outcomes for X as the number of independent trials increases.
 - Justification: The observed relative frequencies of each outcome will approach the probability of each outcome in the long run. In other words, after many trials, the relative frequencies provide good estimates of the probabilities used in the formula for calculating μ_X .
- In *the long run*, random fluctuations are averaged out when comparing the sample mean \bar{x} to the distribution mean μ_X . However, in *the short run*, the sample mean \bar{x} may not be near the distribution mean μ_X .

NOTES:

CHAPTER 5: FROM PROBABILITY TO INFERENCE

5.2: SAMPLE MEANS

- Recall: Statistics, such as the sample mean, percentiles, and standard deviation, are based on measured data. Statistical theory describes the sampling distribution of these statistics.
- Goal: Understand the sampling distribution of the sample mean \bar{x} . Two facts we will discuss are
 - (i) Averages are less variable than individual observations.
 - (ii) Averages are more normally distributed than individual observations.
- The sample mean \bar{x} from a sample or experiment is an *estimate* of the mean μ_X of the underlying population.
- Situation:
 - Select a SRS of size n from a population, and measure a variable X on each unit in the sample.
 - The data consist of observations on n random variables X_1, X_2, \dots, X_n .
 - A single X_i is a measurement on one unit selected at random from the population and follows the distribution of the population.
 - If the population is large relative to the sample, we consider X_1, X_2, \dots, X_n to be *independent* random variables each having the same distribution.
- The sample mean of a SRS of size n is

$$\bar{x} = \frac{1}{n} (X_1 + X_2 + \dots + X_n).$$

- If the population sampled has mean μ and variance σ^2 , then μ and σ^2 are also the mean and variance of each observation X_i . If we apply the addition rules for means and variances of random variables, we get

$$\mu_{\bar{x}} = \frac{1}{n} (\mu_{X_1} + \mu_{X_2} + \dots + \mu_{X_n}) \quad \text{and} \quad \sigma_{\bar{x}}^2 = \left(\frac{1}{n}\right)^2 (\sigma_{X_1}^2 + \sigma_{X_2}^2 + \dots + \sigma_{X_n}^2).$$

- After substitution of μ for each μ_{X_i} and σ^2 for each $\sigma_{X_i}^2$, we get

The Mean and Standard Deviation of a Sample Mean

$$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}, \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

- Note that as the sample size n increases, $\frac{\sigma^2}{n}$ decreases. Therefore, increasing the sample size decreases the variability of \bar{x} .

- Although we know the mean and variance of \bar{x} , the *shape* of the distribution of \bar{x} depends on the shape of the distribution of the population sampled.

NOTES:

- **Sampling Distribution of \bar{x} : The Normal Case**

If the population is normally distributed $N(\mu, \sigma)$, then the sample mean of n independent observations is normally distributed $N(\mu, \sigma/\sqrt{n})$.

- Any linear combination of independent normal random variables is normally distributed.
- For example: If X and Y are independent normal random variables such that X is $N(\mu_X, \sigma_X)$ and Y is $N(\mu_Y, \sigma_Y)$, then

$$X + Y \text{ is } N\left(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right) \quad \text{and} \quad X - Y \text{ is } N\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right).$$

- **Sampling Distribution of \bar{x} : The General Case**

- **The Central Limit Theorem:** For large n , the sampling distribution of \bar{x} is *approximately* $N(\mu, \sigma/\sqrt{n})$ for *any population* with finite standard deviation σ .
- Interpretation: The CLT states that no matter what distribution a single observation follows, the *distribution of the sample mean \bar{x} is approximately normal* for large enough n . The larger the sample size n , the closer \bar{x} is to being normally distributed.

NOTES:

CHAPTER 6: INTRODUCTION TO INFERENCE

PRELUDE

- Recall: Statistical inference
 - Draws conclusions about a population or process based on sample data.
 - Provides a statement, expressed in the language of probability, of how much confidence we can place in the conclusions.

In Chapter 6, we will

- Study two common types of statistical inference:
 1. *Confidence intervals* for estimating the value of a population parameter, and
 2. *Tests of significance* which assess the evidence for a claim.
- Both types of inferences are based on the sampling distributions of statistics and report probabilities that state *what would happen if we used the inference many times*.
- In Chapter 6, we consider the simplest setting: Inference about the mean of a normal population whose standard deviation *is known*. Understanding of the statistical reasoning for this case is stressed. This will aid in understanding inferences discussed in future chapters.
- When you use statistical inference, you are acting as if the data are a random sample or result from a randomized experiment. *If you do not use random sampling or random assignment, then your conclusions are open to challenge*.
- Although the details of formal statistical inference can be complex, the mathematics cannot remedy basic flaws in unreliable data, such as voluntary response samples and confounded experiments.

NOTES:

6.1 ESTIMATING WITH CONFIDENCE

- The sample mean \bar{x} is the natural estimator of the unknown population mean μ . Two important reasons:
 1. \bar{x} is an unbiased estimator of μ and
 2. The law of large numbers says that \bar{x} approaches μ as the sample size grows.
- Unbiasedness says only that there is no systematic tendency to underestimate or overestimate the truth. Unbiasedness is not enough for an estimate to be considered “good”. An estimate without an indication of its variability is of little value in statistical inference.
- We will use what we know about the normal distribution and the sampling distribution of \bar{x} to make inferences. The language of statistical inference uses this information about what would happen in the long run to express our confidence in the results of any one sample.
- The purpose of a **confidence interval** is to estimate an unknown parameter with an indication of *how accurate* the estimate is and *how confident* we are that the result is correct.
- Any confidence interval has two parts:
 1. A confidence interval computed from the data often has the following form:

$$\underline{\text{estimate} \pm \text{margin of error.}}$$

The *margin of error* shows how accurate we believe our guess is based on the variability of the estimate.

2. A confidence level. The confidence level states the probability that the method will provide a correct answer. If we are finding a confidence interval for μ , the confidence level shows how confident we are that the interval will contain μ .
- For example, if you use 95% confidence intervals, in the long run 95% of your intervals will contain the true parameter value. Caution: *you can never know with absolute certainty* whether the result of applying a confidence interval to a particular set of data is correct.

General Definition of a Confidence Interval

- A level C confidence interval for a parameter θ is an interval computed from sample data by a method that has probability C of producing an interval containing the true value θ .
- Theory behind constructing a level C confidence interval for mean μ of a population based on data collected from a SRS of size n :
 - The sampling distribution of \bar{x} is $N(\mu, \sigma/\sqrt{n})$ when the population is $N(\mu, \sigma)$.
 - The sampling distribution of \bar{x} is approximately $N(\mu, \sigma/\sqrt{n})$ for any population when n is large.
 - In either case, for large samples, we can use the $N(\mu, \sigma/\sqrt{n})$ distribution probabilities when studying the sampling distribution of \bar{x} .

- The value z^* with probability p lying to its right under the standard normal density curve is called the **upper p critical value** of the standard normal distribution.
- Determining z^* for a level C confidence interval.
 - For a level C confidence interval, there is probability $1 - C$ outside the interval with equal probability to each side.
 - Or, in other words, z^* is the value such that the area to the right of z^* and to the left of $-z^*$ is $(1 - C)/2$.
 - The value of z^* can be determined from the standard normal probability table (Table A) or carefully using Table D.
- How to construct a level C confidence interval for mean μ of a population based on data collected from a SRS of size n :
 1. Find the number z^* such that any normal distribution has probability C within $\pm z^*$ standard deviations of its mean.
 2. Compute $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
- **SUMMARY:** Suppose that a SRS of size n is drawn from a population having unknown mean μ and known standard deviation σ . A level C confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where z^* is the upper $(1 - C)/2$ critical value for the standard normal distribution. This interval is *exact* when the population is normal and *approximately correct* for large n in other cases.

NOTES:

- Properties of Confidence Intervals

1. The user chooses the confidence level, and the margin of error follows from that choice.
2. High confidence is desirable because it says our method almost always gives correct answers.
3. A small margin of error is desirable because it says we have narrowed down the parameter to a smaller range of values. In other words, our interval estimate is more precise.
4. To obtain a higher confidence *for the same data*, you must be willing to accept a larger margin of error.
5. If the sample size n and the standard deviation σ are unchanged, then a larger z^* leads to a larger margin of error while a smaller z^* leads to a smaller margin of error.
6. If the standard deviation σ and the confidence level C remain unchanged, then increasing the sample size n decreases the margin of error whereas decreasing n increases the margin of error.

- Sample Size for Desired Margin of Error

The confidence interval for a population mean μ will have a specified margin of error m when the sample size is $n = \left(\frac{z^* \sigma}{m}\right)^2$.

- Note: The margin of error equals half the width of the confidence interval.

- Cautions: *Any formula for inference is correct only in specific circumstances.* Note the following warnings:

1. The data must be a SRS from the population. If the data can *plausibly* be thought of as independent observations from a population, then we are not in great danger by treating the data as a SRS.
 - The margin of error in a confidence interval only accounts for random sampling errors. The margin of error is obtained from the sampling distribution and indicates how much error can be expected because of chance variation in random samples of data.
 - Practical difficulties, such as nonresponse and under-coverage, can cause *additional* errors that may be larger than the random sampling error.
 - Therefore, the practical conduct of a survey influences the trustworthiness of its results in ways that are not accounted for by the margin of error.
2. The confidence interval formula is not correct for sampling designs more complex than a SRS (such as stratified or multistage). Correct methods for these designs exist but will not be discussed here.
3. There is no correct method for inference from data collected haphazardly. Note: \bar{x} computed from haphazardly collected data is biased.
4. Confidence intervals based on \bar{x} are not resistant to outliers because \bar{x} is not resistant to outliers.
5. If the sample size is *small* and the population is *not normal*, then the true confidence level will be different from the value C used in computing the interval.

6. The standard deviation σ of the population must be known. In general, this is unrealistic. Thus, the confidence interval $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ is of little use in practice. We will learn how to handle the situation when σ is *unknown* in Chapter 7.

NOTES:

6.2 TESTS OF SIGNIFICANCE

- Goal: Develop rules and principles for the second type of statistical inference: *tests of significance*.
- The first step in a test of significance is to state a claim that we will try to find evidence against.
- The statement being tested in a test of significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of “no effect” or “no difference”.
- The term “null hypothesis ” is abbreviated H_0 . A null hypothesis is a statement about a population, expressed in terms of some parameter or parameters.
- It is also convenient to give a name to the statement we hope or suspect is true instead of H_0 . This is called the **alternative hypothesis** and is abbreviated H_a .
- Hypotheses always refer to some population, *not to a particular sample outcome*. For this reason we state H_0 and H_a in terms of population parameters and not in terms of sample statistics.
- The alternative hypothesis states that a parameter differs from its null value in a specific direction (a **one-sided alternative**) or in either direction (a **two-sided alternative**).
- H_a is often more difficult to state than H_0 . For example, it may not always be clear whether H_a should be one-sided or two-sided.

- Two principles apply to most significance tests. These principles help in understanding the form of the tests:
 - The test is based on a statistic that estimates the parameter that appears in the hypotheses. Usually this is the same estimate we would use in the confidence interval for the parameter. When H_0 is true, we expect the estimate to be near the parameter value specified by H_0 .
 - Values of the estimate far from the parameter value specified by H_0 give evidence against H_0 . The alternative hypothesis determines which directions count *against* H_0 .
- A test of significance assesses the evidence against the null hypothesis in terms of probability. If the observed outcome is *unlikely under the supposition that the null hypothesis is true*, but is more probable if the alternative hypothesis is true, then the outcome provides evidence *against* H_0 and in *favor of* H_a . The less probable the outcome is, the stronger the evidence that H_0 is false.
- The alternative hypothesis determines what *kinds* of outcomes count as evidence against H_0 and in favor of H_a .
- In general, a test of significance finds the probability of getting an outcome *as extreme or more extreme than the actually observed outcome*. “Extreme” means “far from what we would expect if H_0 were true.” For a one-sided alternative hypothesis we are interested in “extreme” values in one direction while for a two-sided alternative hypothesis we are interested in “extreme” values in both directions.
- The probability, computed assuming H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P-value*** of the test.
- The *P-value* is a measure of the strength of the evidence against H_0 .
- The smaller the *P-value*, the stronger the evidence against H_0 provided by the data.

NOTES:

- To assess the evidence against H_0 , we can compare the P -value obtained to a fixed value that we regard as decisive. This amounts to stating *in advance* how much evidence against H_0 we require. The decisive value is called the **significance level** and is denoted by α .
 - If we choose $\alpha = .05$, we require that the data give evidence against H_0 so strong that it would happen *no more than 5%* of the time when H_0 is true.
 - If we choose $\alpha = .01$, we require that the data give evidence against H_0 so strong that it would happen *no more than 1%* of the time when H_0 is true.
- If the P -value is as small or smaller than α , then we say that the the data are **statistically significant** *at level α* , that is, the data yield a statistically significant result.
- Note: “Significant” in the statistical sense does not necessarily mean “important” in a practical sense. In statistics, the term is used to indicate only that the evidence against the null hypothesis reached the standard set by α .
- The P -value is *more informative* than a statement of significance because we can assess significance at *any* level we choose.
- **Tests of Significance:** A test of significance is a recipe or procedure for assessing the strength of the evidence provided by data against a null hypothesis. The steps common to all tests of significance are
 1. State the null hypothesis H_0 and the alternative hypothesis H_a .
 2. Specify the significance level α . This step is optional but is commonly used.
 3. Calculate the value of the test statistics on which the test will be based. This statistic will measure how well the data conform to H_0 .
 4. Find the P -value for the observed data. If the P -value is less than or equal to α , the test result is statistically significant at level α .
- The proper test statistic is determined by the hypotheses and the data collection design.

NOTES:

z-TESTS FOR A POPULATION MEAN

- Situation: We have a SRS of size n drawn from a normal population with unknown mean μ . We want to test the hypothesis that μ has a specified value. That is, the null hypothesis is

$$H_0: \mu = \mu_0.$$

- Motivation: Because \bar{x} is an unbiased estimator of μ , the test will be based on \bar{x} . Also, \bar{x} is normally distributed because the population is normal. As our test statistic we will use the standardized sample mean

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

The statistic z follows a standard normal distribution *when* H_0 is true. We can then determine the P -value based on standard normal distribution probabilities.

Performing the z -test:

- To test the hypothesis $H_0: \mu = \mu_0$ based on a SRS of size n from a population with unknown mean μ and known standard deviation σ , compute the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

In terms of a standard normal random variable Z , the P -value for a test of H_0 against

$$\begin{aligned} H_a: \mu > \mu_0 & \text{ is } P(Z \geq z), \\ H_a: \mu < \mu_0 & \text{ is } P(Z \leq z), \text{ and} \\ H_a: \mu \neq \mu_0 & \text{ is } 2P(Z \geq |z|). \end{aligned}$$

These P -values are exact if the population distribution is normal and approximately correct for large n in other cases.

- Sometimes we require a specific degree of evidence, stated as a significance level α , in order to reject the null hypothesis. In terms of the P -value, the outcome of a test is significant at level α if $P\text{-value} \leq \alpha$.
- When statistical software is not used, the P -value can be difficult to calculate. Fortunately, we can decide if a result is statistically significant *without* calculating the P -value by using the same table of critical values used to obtain confidence intervals (Table D) or the standard normal z -table (Table A).
- The number z^* with probability p falling to the right of it under the standard normal density curve is the **upper p critical value** for the standard normal distribution.

Fixed significance level z -tests for a population mean:

- To test the hypothesis $H_0: \mu = \mu_0$ based on a SRS of size n from a population with unknown mean μ and known standard deviation σ , compute the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Reject $H_0: \mu = \mu_0$ at significance level α against a **one-sided** alternative

$$H_a: \mu > \mu_0 \quad \text{if} \quad z \geq z^*$$

or

$$H_a: \mu < \mu_0 \quad \text{if} \quad z \leq -z^*,$$

where z^* is the upper α critical value.

Reject $H_0: \mu = \mu_0$ at significance level α against a **two-sided** alternative

$$H_a: \mu \neq \mu_0 \quad \text{if} \quad |z| \geq z^*,$$

where z^* is the upper $\alpha/2$ critical value.

- Relationship between significance tests and confidence intervals:
 - A level α two-sided significance test rejects a hypothesis $H_0: \mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ .
- P -values versus fixed level α :
 - The P -value is the *smallest* level α for which the data yield a statistically significant test for rejecting H_0 .
 - Knowing the P -value allows us to assess significance at any level α . A P -value is more informative than a reject-or-not finding at a fixed significance level.
 - Assessing significance at level α is easier because it is based on a critical value with no required probability calculation.
 - In practice, the use of statistical tables is becoming outdated because computer software is used to perform most statistical tests. The software automatically calculates the P -value. By initially learning how to perform statistical tests using tables, the principles behind statistical testing are learned enabling a researcher to correctly interpret computer output.

NOTES:

6.3 USE AND ABUSE OF TESTS

Statistical tests are only valid under certain circumstances. Important comments regarding the proper use and unfortunate abuse of significance tests are discussed in this section.

Using Significance Tests

1. Choosing a level of significance

- Sometimes a researcher will make a decision or take some action if the evidence supplied by the data reaches a certain standard (which is commonly the significance level α).
- Choosing a level α in advance makes sense *if you must make a decision*, but not if you only want to describe the strength of your evidence. This is when the P -value should be used.
- If you do use a fixed α significance test to make a decision, α should be chosen based on how much evidence is required to reject H_0 . This depends on
 - (a) How plausible is H_0 ? If H_0 represents an established assumption in your field, then strong evidence (small α) will be needed to reject it.
 - (b) What are the consequences of rejecting H_0 ? For example, if the potential cost of rejecting H_0 in favor of some alternative is large, strong evidence (small α) is required.
- Users of statistics have over-emphasized certain standard α levels, in particular, $\alpha = .01, .05, .10$. Nonetheless, there is *no practical distinction* between the P -values $.049$ and $.051$. It makes no sense to treat $\alpha = .05$ as a universal rule for what is significant.
- There is no sharp border between *significant* and *insignificant*, only increasingly strong evidence against H_0 as the P -value decreases.

2. What statistical significance doesn't mean

- When a null hypothesis ('no effect' or 'no difference') is rejected at the standard levels, $\alpha = .05$ or $\alpha = .01$, there is evidence that an effect is present. Warning: Despite rejecting H_0 , the effect may be extremely small (possibly small enough not to have any practical value).
- When large samples are used, even tiny deviations from the null hypothesis will be significant.
- Moral: Statistical significance is not the same thing as (does not imply) practical significance.
- Before attaching too much importance on statistical significance, the researcher should determine the P -value and investigate the experimental results. This involves plotting the data, looking for outliers which could strongly influence results, and verifying the result claimed to be significant is visible in the plots. In general, use exploratory data analysis methods to verify the results of a hypothesis test.

3. Don't ignore lack of significance

- It is an unfortunate fact that research in some fields is rarely published unless significance at some level, usually $\alpha = .05$, is attained. Such a publication policy impedes the spread of knowledge.
- If a researcher has good reason to suspect that an effect is present and then fails to find significant evidence of it, this may be just as important or even more important than finding evidence in favor of it.
- Keeping silent about negative results may lead other researchers to waste time and money by repeating the same experiment in an attempt to find an effect that isn't there.
- An important aspect of planning a study is to verify that the proposed test has a high probability of detecting an effect of the size you hope to find. This probability is called the *power* of the test. Calculation of power is discussed in the text, but will not be covered in this course.

NOTES:

Abuse of Significance Tests

1. Statistical inference is not valid for all sets of data

- Badly designed surveys or experiments often produce invalid results. Formal statistical inference cannot correct basic flaws in data collection.
- Tests of significance and confidence intervals are based on the laws of probability. Random sampling or random assignment ensures that these laws apply.
- Often data that did not result from randomized samples or experiments are analyzed. To apply statistical inference to such data, we must have confidence in a probability model for the data, for example, assuming the data are independent and normally distributed. In such cases, the probability model can be and should be checked by examining the data.
- Moral: Always ask how the data were produced. Do not be impressed by *P*-values until you are confident that the data deserve a formal statistical analysis.

2. Beware of searching for significance

- The reasoning behind statistical significance works well if you first, decide what effect you are seeking, then design an experiment or sample to search for it, and, second, use a test of significance to weigh the evidence you get.
- Too often, a statistically significant result is taken as a sign of successful research. Thus, it becomes tempting to make statistical significance the object of the research.
- A common (but unproductive) tactic is to perform many tests on the same data until one or more so-called 'significant' results are found.
- It is not convincing to search for any effect or pattern *after collecting the data*, find one, and claim it is significant.
- It is more convincing to hypothesize *prior to collecting the data* that an effect or pattern will be present, design a study to look for it, and find it at a small significance level.
- Patterns or effects suggested by exploring the data may be important, but you cannot legitimately test a hypothesis on the same data that *suggested* the hypothesis.
- Moral: Once you have a hypothesis, design a proper study to search specifically for the effect you suspect. If this study results in a statistically significant result, you have real evidence regarding the effect of interest.

NOTES:

CHAPTER 7: INFERENCE FOR DISTRIBUTIONS

PRELUDE

- In Chapter 7, we will develop confidence intervals and significance tests for inference about a population mean μ and for comparing the means of two populations.
- In Chapter 6, the emphasis was placed on statistical *reasoning*. In Chapter 7 the emphasis is on statistical *practice* so we no longer need to assume σ is known.
- You will be introduced to the commonly used t procedures for inference which are based on the t distributions.

7.1 INFERENCE FOR THE MEAN OF A POPULATION

- Recall: Both confidence intervals and tests of significance for mean μ of a normal population are based on the sampling distribution of \bar{x} , that is, \bar{x} is $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
- However, σ is generally *unknown*. In such cases, we estimate $\frac{\sigma}{\sqrt{n}}$ by $\frac{s}{\sqrt{n}}$ where s is the sample standard deviation.
- $\frac{s}{\sqrt{n}}$ is called the **standard error of the sample mean** \bar{x} .
- In general, when the standard deviation of a statistic is estimated from the data, the result is called the **standard error of the statistic**.
- Recall: The standardized sample mean $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ has a standard normal distribution $N(0, 1)$. However, when we substitute s/\sqrt{n} for σ/\sqrt{n} , the statistic *does not* have a normal distribution. It *does* have what is called a t distribution.

- Suppose that a SRS of size n is drawn from a $N(\mu_0, \sigma)$ population. Then the **one-sample t statistic**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

has the t distribution with $n - 1$ degrees of freedom.

- There is a different t distribution for each sample size. A particular t distribution is specified by giving the **degrees of freedom**.
- The degrees of freedom for this t statistic come from s which has $n - 1$ degrees of freedom. (See Chapter 1).
- Notation: The t distribution with k degrees of freedom is denoted by $t(k)$.
- The density curve for a t distribution is similar in shape to the standard normal curve, that is, it is bell-shaped and symmetric about 0.

- The spread of a t distribution, however, is a bit greater than the spread of a standard normal distribution. This is due to the extra variability caused by substituting the random variable s for the fixed parameter σ .
- As the degrees of freedom increase, the $t(k)$ density curves approaches then $N(0, 1)$ curve. This is because s approaches σ as n increases. Therefore, the t statistic approaches the standard normal z statistic.
- Table E in the text gives the upper p critical values for many of the t distributions and for common values of p .
- Recall: p is the upper tail probability needed for significance tests and by the confidence level C for confidence intervals.
- With the t distributions, samples from normal populations *with unknown* σ can be analyzed. This involves
 1. Replacing σ/\sqrt{n} by s/\sqrt{n} in the confidence interval and significance test procedures in Chapter 6.
 2. Using the upper p critical values of the appropriate t distribution instead of the standard normal distribution.

NOTES:

The One-sample t Procedures

- Suppose that a SRS of size n is drawn from population with unknown mean μ . A level C confidence interval for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}},$$

where t^* is the upper $(1 - C)/2$ critical value for the $t(n - 1)$ distribution.

- The interval is exact when the population distribution is normal and is approximately correct for large n in other cases.
- To test $H_0: \mu = \mu_0$ based on a SRS of size n , compute the one-sample t statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

In terms of a random variable T having a $t(n - 1)$ distribution, the P -value for a test of H_0 against

$$H_a: \mu > \mu_0 \text{ is } P(T \geq t),$$

$$H_a: \mu < \mu_0 \text{ is } P(T \leq t), \text{ and}$$

$$H_a: \mu \neq \mu_0 \text{ is } 2P(T \geq |t|).$$

- The P -values are exact when the population distribution is normal and are approximately correct for large n in other cases.

For a fixed α test:

- For one-sided alternatives, we reject the null hypothesis $H_0: \mu = \mu_0$ in favor of

$$H_a: \mu > \mu_0 \text{ if } t \geq t^*$$

or

$$H_a: \mu < \mu_0 \text{ if } t \leq -t^*,$$

where t^* is the upper α critical value of the $t(n - 1)$ distribution.

- For a two-sided alternative, we reject the null hypothesis $H_0: \mu = \mu_0$ in favor of

$$H_a: \mu \neq \mu_0 \text{ if } |t| \geq t^*,$$

where t^* is the upper $\alpha/2$ critical value of the $t(n - 1)$ distribution.

NOTES:

The Matched Pairs t Procedures

- In a *matched pairs* study, subjects are matched in pairs and the outcomes are compared within each matched pair.
- A common situation calling for matched pairs is before-and-after observations on the same subject.
- Matched pairs data are restated as single sample data by taking differences within each pair.
- Because the matched pairs are restated in the form of a single sample, we will make inferences about a single population, that is, the *population of all differences* within matched pairs.
- It is incorrect to ignore the pairs and analyze the data as if we had two samples because inference procedures comparing two samples (Section 7.2) assume the samples are selected independently of each other.
- With matched pairs, the data are not independent *within pairs* but are independent *across pairs*.
- In a matched pairs analysis, we assume that the population of *differences* has a normal distribution because the t procedures are applied to the differences.

NOTES:

Robustness of t Procedures

- A statistical inference procedure is called **robust** if the probability calculations required are insensitive to violations of the assumptions made.
- Recall: The results of one-sample t procedures are exactly correct only when the population is normal. However, real populations are never exactly normal. Therefore, the usefulness of the t procedures depends on how strongly they are affected by non-normality.
 1. t procedures *are not robust* when non-normality is caused by outliers, because \bar{x} and s are not resistant to outliers.

2. t procedures *are not robust* when non-normality is caused by *strong* skewness.
 3. If n is large, t procedures *are robust* when the population distribution is nonnormal (except for the two cases above). Two reasons why this is true:
 - Because of the Central Limit Theorem, when n is large the sampling distribution of \bar{x} is close to normal.
 - As n increases the sample standard deviation s approaches the population standard deviation σ . For large samples s will be an accurate estimate of σ whether or not the population has a normal distribution.
- Practical sample size guidelines for inference on a single mean:
 - *Sample size* < 15 : Use t procedures if the data are close to normal. If the data are clearly nonnormal or if outliers are present, do not use t .
 - *Sample size* ≥ 15 : The t procedures can be used except in the presence of outliers or strong skewness.
 - *Large samples*: The t procedures can be used even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.

NOTES:

CHAPTER 10: INFERENCE FOR REGRESSION

PRELUDE

- In Chapter 2 regression was motivated as a descriptive method. The straight line equation

$$y = a + bx$$

summarized the linear relationship between 2 quantitative variables, a response variable y and an explanatory variable x .

- In Chapter 10 we will learn how to do statistical inference for regression. We will be assuming that there is a true population regression line which we will estimate using data collected from the population.
- Parameters have been indicated with Greek letters and we will continue with that convention. Accordingly the notation will change. The population regression line (a *parameter*) will be denoted by

$$\beta_0 + \beta_1 x$$

and the estimated line (a *statistic*) will be denoted by

$$b_0 + b_1 x.$$

The sample intercept b_0 is an estimate of the population intercept β_0 and the sample slope b_1 is an estimate of the population intercept β_1 .

- We will learn how to construct confidence intervals for these 2 parameters and also how to carry out hypothesis tests.

10.1 SIMPLE LINEAR REGRESSION

- Statistical inference requires a probability model. For each value of x we assume that we have subpopulation of y values with each subpopulation being normally distributed with mean

$$\mu_y = \beta_0 + \beta_1 x.$$

In other words we assume that the means of the subpopulations of y values vary linearly with x .

- Although the means vary linearly with x we will assume that the subpopulations all have the same standard deviation σ .
- Each observation of a response variable y is thought of as being the sum of 2 pieces: the mean μ_y and a random noise piece we will call ϵ .

- Specifically the model says (see page 661):

Given n observations on the explanatory variable x and the response variable y ,

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

the **statistical model for simple linear regression** states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Here $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The random deviations ϵ_i are assumed to be independent and normally distributed with mean 0 and standard deviation σ . The parameters in the model are β_0, β_1 , and σ .

- *Interpretation of β_0 and β_1* : The intercept β_0 is the mean of the subpopulation of y values with $x = 0$. This may be a nonsense value because $x = 0$ may not even be a possible value depending on the context. The slope β_1 represents the change in the **mean** response when the explanatory variable increases by 1 unit.
- *Estimating the parameters*:

- Slope β_1 : The estimate of the true slope is

$$b_1 = r \frac{s_y}{s_x}$$

where r is the correlation between y and x , s_x is the standard deviation of x , and s_y is the standard deviation of y .

- Intercept β_0 : The estimate of the true intercept is

$$b_0 = \bar{y} - b_1 \bar{x}$$

where \bar{x} and \bar{y} are the sample means of the x and y values, respectively.

- Mean Response when $x = x^*$: The estimate of the mean of the subpopulation of y values when $x = x^*$ is

$$\hat{y} = b_0 + b_1 x^*.$$

- Standard deviation σ : Estimation of the standard deviation requires the residuals

$$e_i = \text{observed response} - \text{predicted response} \quad (1)$$

$$= y_i - \hat{y}_i \quad (2)$$

$$= y_i - b_0 - b_1 x_i \quad (3)$$

The residuals can be thought of as a kind of estimate of the noise terms ϵ_i in the model. The residuals sum to 0. The estimate of the standard deviation is

$$s = \sqrt{\frac{\sum e_i^2}{n - 2}} \quad (4)$$

$$= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}. \quad (5)$$

Note that we divide by $n - 2$ which is called the **degrees of freedom** associated with s .

- Inference for β_0 and β_1

- If the assumptions are met then b_0 and b_1 are normally distributed unbiased estimators of β_0 and β_1 . Even if the normality assumption for the ϵ_i terms in the model is suspect the Central Limit Theorem comes into play for large n and b_0 and b_1 will be approximately normally distributed. We can use this result to construct confidence intervals and derive tests for β_0 and β_1 .

- A level C confidence interval for the intercept β_0 is

$$b_0 \pm t^* SE_{b_0}$$

where t^* is the value for the $t(n - 2)$ density curve with area C between $-t^*$ and t^* and SE_{b_0} is the standard error of b_0 . You do not need to know how to compute the standard error. It will always be given to you.

- A level C confidence interval for the slope β_1 is

$$b_1 \pm t^* SE_{b_1}$$

where t^* is the value for the $t(n - 2)$ density curve with area C between $-t^*$ and t^* and SE_{b_1} is the standard error of b_1 . You do not need to know how to compute the standard error. It will always be given to you.

- To test the hypothesis $H_0 : \beta_1 = 0$ we compute the test statistic

$$t = \frac{b_1}{SE_{b_1}}$$

If the null hypothesis is true this test statistic has a t distribution with $n - 2$ degrees of freedom. In terms of a random variable T having this distribution the P -value for a test of H_0 against

1. $H_a : \beta_1 > 0$ is $P(T \geq t)$
2. $H_a : \beta_1 < 0$ is $P(T \leq t)$
3. $H_a : \beta_1 \neq 0$ is $2P(T \geq |t|)$

- We will not be concerned with testing hypotheses about β_0 . The test of $H_0 : \beta_1 = 0$ is a test of whether or not a linear relationship exists between y and x . If we fail to reject the null hypothesis then we do not have strong evidence of a straight line relationship between the 2 variables.

NOTES: