Elementary Statistics Chapters 1-4 Exam Prep

JJC Stat128

A Document By:
Nathan Warner

July 08,2023
Computer Science
Northern Illinois University
United States

Contents

1 Vocab

- **Population:** The entire group to be studied is called the population.
- **Sample:** In statistics, it is often impractical or impossible to get access to the entire **population**, which is why we only look at a **sample.** A sample is a **subset** of the population being studied.
- **Individual:** An individual is a person or object that is a member of the population being studied.
- **Statistic:** A statistic is a numerical summary of a sample.
- **Descriptive Statistics:** Descriptive statistics consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.
- **Inferential Statistics:** inferential Statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.
- **Parameter:** A parameter is a numerical summary of a population.
- **Variables:** The characteristics of the individuals in a study. Variables vary, which means they can take on different values.
- **Constants:** Variables that do not vary. Inferential statistics is not necessary with constants.
- **Qualitative, or categorical variables** allow for the classification of individuals base on some attribute or characteristic.
- **Quantitative variables** provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.
- A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. A discrete variable cannot take on every possible value between any two possible values.
- A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values. Continuous variables typically result from measurement. Continuous variables are often rounded. If a certain make of car gets 24 miles per gallon (mpg) of gasoline, its miles per gallon must be greater than or equal to 23.5 and less than 24.5, or $23.5 \leqslant mpg \leqslant 24.5$
- The list of observed values for a variable is **data.**
- **Qualitative data** are observations corresponding to a **qualitative variable.**
- **Quantitative data** are observations corresponding to a quantitative variable.
- **Discrete data** are observations corresponding to a discrete variable.
- **Continuous data** are observations corresponding to a continuous variable.
- **Explanatory Variable:** An explanatory variable, also known as an independent variable or predictor variable, is a variable that is manipulated or controlled by researchers in an experiment or study. It is the variable that is hypothesized to have an impact on the outcome or dependent variable.
- **Lurking variable**: An explanatory variable that was not considered in a study, but that affects the value of the response variable.
- **Response Variable**: The response variable, also known as the dependent variable or outcome variable, is the variable that is measured or observed to determine the effect or response of the explanatory variable(s). It is the variable that researchers are interested in studying or predicting.
- **Confounding:** Occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.

- **Census:** List of individuals in a population along with certain characteristics of each individual.
- **Random Sampling:** The process of using chance to select individuals from a population to be included in the sample.
- **Simple Random Sampling:** A sample of size n from a population of size N is obtained through simple random sampling if every possible sample of size n has an equal chance of occurring. The sample is then called a simple random sample.
  - $n < N$
- **frame:** a list of all the individuals within the population.
  - **Stratified sample**: is obtained by dividing the population into nonoverlapping groups called strata and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogenous (similar) in some way.
    - Within Stratified samples, the number of individuals sampled from each stratum should be proportional to the size of the strata in the population.
  - **Systematic sample** is obtained by selecting every kth individual from the population. The first individual selected corresponds to a number between 1 and k
  - **Cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.
  - **Convenience sample:** the individuals are easily obtained and not based on randomness.
- **Bias:** If the results of the sample are not representative of the population. Sampling bias means that the technique used to obtain the sample's individuals tends to favor one part of the population over another. Any convenience sample has sampling bias because the individuals are not chosen through a random sample.
- **Undercoverage:** Occurs when the proportion of one segment of the population is lower in a sample than it is in the population. This can result if the frame used to obtain the sample is incomplete or not representative of the population.
- **Sampling bias:** sampling bias is a bias in which a sample is collected in such a way that some members of the intended population have a lower or higher sampling probability than others. It results in a biased sample of a population in which all individuals, or instances, were not equally likely to have been selected
- **Nonresponse bias:** exists when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do
  - This can be controlled with **callbacks**.
  - This can also be controlled with **rewards or incentives**
- **Response bias:** Exists when the answers on a survey do not reflect the true feelings of the respondent.
- **Open Question:** Allows the respondent to choose his or her response
- **Closed Question:** requires the respondent to choose from a list of predetermined responses
- **Nonsampling errors:** result from undercoverage, nonresponse bias, response bias, or data-entry error. Such errors could also be present in a census.
- **Sampling error:** results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.
- **Experiment:** is a controlled study conducted to determine the effect of varying one or more explanatory variables or **factors** has on a response variable.
- **Factor:** A variable whose effect on the response variable is to be assessed by the experimenter
- **Treatment:** Any combination of the values of the factors is called a treatment

- **Experimental Unit (or subject)** is a person, object or some other well-defined item upon which a treatment is applied
- **Control Group:** Serves as a baseline treatment that can be used to compare to other treatments.
- **Placebo:** is an innocuous medication, such as a sugar tablet, that looks, tastes, and smells like the experimental medication.
- **Blinding:** refers to nondisclosure of the treatment an experimental unit is receiving.
- **Single-blind** experiment is one in which the experimental unit (or subject) does not know which treatment he or she is receiving.
- **Double-blind** experiment is one in which neither the experimental unit nor the researcher in contact with the experimental unit knows which treatment the experimental unit is receiving.
- **Design:** To design an experiment means to describe the overall plan in conducting the experiment. Conducting an experiment requires a series of steps.
- **completely randomized design:** is one in which each experimental unit is randomly assigned to a treatment.
- **matched-pairs design:** is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way (that is, the same person before and after a treatment, twins, husband and wife, same geographical location, and so on). There are only two levels of treatment in a matched-pairs design.
- **A frequency distribution** lists each category of data and the number of occurrences for each category of data.
- **The relative frequency** is the proportion (or percent) of observations within a category and is found using the formula

$$Relative\ frequency = \frac{Frequency}{Sum\ of\ all\ frequency}.$$

- **A relative frequency distribution** lists each category of data together with the relative frequency.
- **A bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represents the category's frequency or relative frequency.
- **A Pareto chart** is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.
- **A pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.
- A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same, and the rectangles touch each other.
- **Classes:** The Categories in which data is grouped
- **lower class limit:** the smallest value within the class
- **upper class limit:** the largest value within the class
- **Class Width:** is the difference between consecutive lower class limits.
- A table is **open ended** if the first class has no lower class limit or the last class has no upper class limit
- We draw a **dot plot** by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.

- **uniform distribution:** frequency of each value of the variable is evenly spread across the values of the variable.
- **bell-shaped distribution:** highest frequency occurs in the middle and frequencies tail off to the left and right of the middle.
- **skewed right:** the tail to the right of the peak is longer than the tail to the left of the peak
- **skewed left:** tail to the left of the peak is longer than the tail to the right of the peak.
- **The arithmetic mean** of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations.
- **The population arithmetic mean**, μ, (pronounced "mew"), is a parameter that is computed using data from all the individuals in a population.

$$\mu = \frac{x_1 + x_2 + x_N}{N} = \frac{\sum x_i}{N}.$$

- **The sample arithmetic mean**, $\bar{x}$ (pronounced x-bar"), is a statistic that is computed using data from individuals in a sample.

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

- **The median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.
  - For odd n:

$$M = \frac{n+1}{2}.$$

  - For even n:

$$M = Average\ of\ \frac{n}{2}, \frac{n}{2} + 1.$$

- A numerical summary of data is said to be **resistant** if observations that are extreme (very large or small) relative to the data do not affect its value substantially.
  - So the median is resistant, but the mean is not resistant.
- **The mode** of a variable is the observation of the variable that occurs most frequently in the data set.
  - If no observation occurs more than once, we say that the data have **no mode**.
- **Bimodal:** If the data has two modes
- **Multimodal:** If the data has more than two modes
- A numerical summary of data is said to be **resistant** if observations that are extreme (very large or small) relative to the data do not affect its value substantially.
  - So the median is resistant, but the mean is not resistant.
- **Dispersion:** Degree to which the data are spread out.
- **Range:** The range, r, of a variable is the difference between the largest and smallest data value. That is,

$$range = R = Largest\ data\ value - smallest\ data\ value.$$

  Note: Range is **not** resistant
- **Deviation:** a deviation refers to the difference between an individual data point and a central value, such as the mean or median. It represents how much a particular data point varies or deviates from the average or typical value in a data set. When can comopute a deviation with:

$$Individual\ data\ point - mean.$$

  We call this calculation, "deviation about the mean"

Note: The sum of the deviations about the mean always equals zero

- **The population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. The population standard deviation is symbolically represented by σ (lowercase Greek sigma). The formula is given by:

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+...+(x_N-\mu)^2}{N}} = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}.$$

Note: Standard Deviation is **not** resistant

- **The sample standard deviation**, s, of a variable is the square root of the sum of squared deviations about the sample mean divided by $n-1$, where n is the sample size. The formula is given as

$$s = \sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+...+(x_n-\overline{x})^2}{n-1}} = \sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}.$$

Note: Standard Deviation is **not** resistant

- we call $n-1$ the **degrees of freedom** because the first $n-1$ observations have freedom to be any value, but the n^{th} observation has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.
- The variance of a variable is the square of the standard deviation.
  - **The population variance** is σ^2
  - **The Sample Variance** is s^2

Note: The units of measure in variance are squared values. So if the variable is measured in dollars, the variance is measured in dollars squared. This makes interpreting the variance difficult.

- **Class Midpoint:** The class midpoint is the sum of consecutive lower class limits divided by 2
- **Approximate Population Mean (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\mu = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n}.$$

where: x_i is the midpoint or value of the i th class
f_i is the frequency of the i th class
n is the number of classes

- **Approximate Sample Mean (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1f_1+x_2f_2+...+x_nf_n}{f_1+f_2+...+f_n}.$$

where: x_i is the midpoint or value of the i th class
f_i is the frequency of the i th class
n is the number of classes

- **The weighted mean**, $\overline{x}_w$, of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

$$\overline{x}_w = \frac{\sum w_i x_i}{w_i} = \frac{w_1 + x_1 + w_2 + x_2 + ... + w_n + x_n}{w_1 + w_2 + ... + w_n}.$$

  Where: w_i is the weight of the i^{th} observation
  x_i is the value of the i^{th} observation.

- **Approximate Population Standard Deviation (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}}.$$

  Where: x_i is the midpoint or value of the ith class
  f_i is the frequency of the i^{th} class

- **Approximate Sample Standard Deviation (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2 f_i}{\sum f_i - 1}}.$$

  Where: x_i is the midpoint or value of the ith class
  f_i is the frequency of the i^{th} class

- **The z-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation.

  - **Population Z-score**

$$z = \frac{x - \mu}{\sigma}.$$

  - **Sample Z-score**

$$z = \frac{x - \overline{x}}{s}.$$

  Note: The Z-score is unitless. It has mean 0 and a standard deviation of 1
  Round z-scores to the nearest hundredth

- The median is a special case of a general concept called the **percentile.**

- **the k^{th} percentile,** denoted P_k, of a set of data is a value such that k percent of the observations are less than or equal to the value.

- The most common percentiles are **quartiles**, which divide data sets into fourths, or four equal parts.

- The **interquartile range, IQR,** is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the first and third quartiles and is found using this formula

$$IQR = Q_3 - Q_1.$$

- **Outliers:** When analyzing data, we must check for extreme observations, called outliers. Outliers can occur by chance, because of errors in the measurement of a variable, during data entry, or from errors in sampling.

- **Fences** serve as cutoff points for determining outliers.

$$Lower\ Fence\ = Q_1 - 1.5 \cdot IQR$$
$$Upper\ Fence\ = Q_3 + 1.5 \cdot IQR.$$

- The **five-number summary** of a set of data consists of the smallest data value, Q_1 the median, Q_3 and the largest data value. We use the five-number summary to learn information about the extremes of the data set. The summary is organized as follows:

$$\text{Minimum } Q_1 \ M \ Q_3 \text{ Maximum}$$

- **bivariate data:** data in which two variables are measured on an individual. For example, we might want to know whether the amount of cola consumed per week is related to a person's bone density. The individuals would be the people in the study, and the two variables would be the amount of cola consumed weekly and bone density.

- **The response (dependent) variable** is the variable whose value can be explained by the value of the explanatory (or predictor or independent) variable.

- **A scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

- Two variables that are linearly related are **positively associated** when above-average values of one variable are associated with above-average values of the other variable (or below-average values of one variable are associated with below-average values of the other variable). That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.

- Two variables that are linearly related are **negatively associated** when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

- The **linear correlation coefficient**, or Pearson product moment correlation coefficient, is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formula for the sample correlation coefficient.

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

  Where:

  x_i is the i th observation of the explanatory variable
  $\overline{x}$ is the sample mean of the explanatory variable
  $s_x$is the sample standard deviation of the explanatory variable
  y_i is the i th observation of the response variable
  $\overline{y}$ is the sample mean of the response variable
  $\mathbf{S}_y$ is the sample standard deviation of the response variable
  n is the number of individuals in the sample

- **The least-squares regression line** minimizes the sum of the squared errors (or residuals). This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$ (read "y-hat"). We represent this as $\sum residuals^2$

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

  Where:

$$b_1 = r \cdot \frac{s_y}{s_x} \text{ is the slope of the least-squares regression line.}$$

And:

$$b_0 = \overline{y} - b_1\overline{x} \text{ is the y-Intercept of the least-squares regression line.}$$

- The observed distance for this club-head speed is 274 yards. The difference between the observed and predicted values of y is the error, or **residual**.

$$Residual \ = observed - predicted.$$

- **The coefficient of determination**, R^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

$$R^2 = r^2.$$

- **An influential observation** significantly affects the least-squares regression line's slope and/or y-intercept. (It also affects the value of the correlation coefficient.) Methods exist for determining whether a particular observation is influential; however, they are beyond the scope of this course. Nonetheless, we can still get a sense as to whether a particular observation is influential right now.
- the difference in our predicted value, and our actual value, is due to factors (variables) other than the club-head speed (wind speed and position of the ball on the club face, for example) and to random error. The differences just discussed are called **deviations**.
- **Total Deviation:** The deviation between the observed value, y, and mean value, $\overline{y}$, of the response variable.

$$y - \overline{y}$$
$$Or: \text{Explained Deviation} + \text{Unexplained Deviation.}$$

- **Explained Deviation:** The deviation between the predicted value, $\hat{y}$, and mean value, $\overline{y}$, of the response variable.

$$\hat{y} - \overline{y}.$$

- **Unexplained Deviation:** The deviation between the observed value, y, and predicted value, $\hat{y}$, of the response variable

$$y - \hat{y}.$$

- If a plot of the residuals against the explanatory variable shows the spread of the residuals increasing or decreasing as the explanatory variable increases, then a strict requirement of the linear model is violated. This requirement is called **constant error variance**. The statistical term for constant error variance is **homoscedasticity**
- **A marginal distribution** of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.
- **A conditional distribution** lists the relative frequency of each category of the response variable, given a specific value of the explanatory variable in the contingency table.
- **Simpson's Paradox**, which describes a situation in which an association between two variables inverts or goes away when a third variable is introduced to the analysis.

2 Formulas

- **The relative frequency** is the proportion (or percent) of observations within a category and is found using the formula

$$Relative\ frequency = \frac{Frequency}{Sum\ of\ all\ frequency}.$$

- **The population arithmetic mean**, μ, (pronounced "mew"), is a parameter that is computed using data from all the individuals in a population.

$$\mu = \frac{x_1 + x_2 + x_N}{N} = \frac{\sum x_i}{N}.$$

- **The sample arithmetic mean**, $\bar{x}$ (pronounced x-bar"), is a statistic that is computed using data from individuals in a sample.

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

- **The median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.
  - For odd n:

$$M = \frac{n+1}{2}.$$

  - For even n:

$$M = Average\ of\ \frac{n}{2}, \frac{n}{2} + 1.$$

- **Range:** The range, r, of a variable is the difference between the largest and smallest data value. That is,

$$range = R = Largest\ data\ value - smallest\ data\ value.$$

- **Deviation:** a deviation refers to the difference between an individual data point and a central value, such as the mean or median. It represents how much a particular data point varies or deviates from the average or typical value in a data set. When can comopute a deviation with:

$$Individual\ data\ point - mean.$$

- **The population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. The population standard deviation is symbolically represented by σ (lowercase Greek sigma). The formula is given by:

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}.$$

  Note: Standard Deviation is **not** resistant

- **The sample standard deviation**, s, of a variable is the square root of the sum of squared deviations about the sample mean divided by $n-1$, where n is the sample size. The formula is given as

$$s = \sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+...+(x_n-\overline{x})^2}{n-1}}$$
$$= \sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}.$$

  Note: Standard Deviation is **not** resistant

- **Approximate Population Mean (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\mu = \frac{\sum x_i f_i}{\sum f_i}$$
$$= \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n}.$$

  where: x_i is the midpoint or value of the i th class
  f_i is the frequency of the i th class
  n is the number of classes

- **Approximate Sample Mean (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i}$$
$$= \frac{x_1 f_1 + x_2 f_2 + ... + x_n f_n}{f_1 + f_2 + ... + f_n}.$$

  where: x_i is the midpoint or value of the i th class
  f_i is the frequency of the i th class
  n is the number of classes

- **The weighted mean**, $\overline{x}_w$, of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

$$\overline{x}_w = \frac{\sum w_i x_i}{w_i} = \frac{w_1 + x_1 + w_2 + x_2 + ... + w_n + x_n}{w_1 + w_2 + ... + w_n}.$$

  Where: w_i is the weight of the i^{th} observation
  x_i is the value of the i^{th} observation.

- **Approximate Population Standard Deviation (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2 f_i}{\sum f_i}}.$$

  Where: x_i is the midpoint or value of the ith class
  f_i is the frequency of the i^{th} class

- **Approximate Sample Standard Deviation (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$s = \sqrt{\frac{\sum(x_i-\overline{x})^2 f_i}{\sum f_i - 1}}.$$

  Where: x_i is the midpoint or value of the ith class
  f_i is the frequency of the i^{th} class

- **The z-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation.
  - **Population Z-score**

$$z = \frac{x - \mu}{\sigma}.$$

  - **Sample Z-score**

$$z = \frac{x - \overline{x}}{s}.$$

  Note: The Z-score is unitless. It has mean 0 and a standard deviation of 1
  Round z-scores to the nearest hundredth

- The **interquartile range, IQR,** is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the first and third quartiles and is found using this formula

$$IQR = Q_3 - Q_1.$$

- **Fences** serve as cutoff points for determining outliers.

$$\begin{aligned} Lower\ Fence\ &= Q_1 - 1.5 \cdot IQR \\ Upper\ Fence\ &= Q_3 + 1.5 \cdot IQR. \end{aligned}$$

- The **five-number summary** of a set of data consists of the smallest data value, Q_1 the median, Q_3 and the largest data value. We use the five-number summary to learn information about the extremes of the data set. The summary is organized as follows:

$$\text{Minimum } Q_1 \ M \ Q_3 \text{ Maximum}$$

- The **linear correlation coefficient**, or Pearson product moment correlation coefficient, is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formula for the sample correlation coefficient.

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

  Where:

  x_i is the i th observation of the explanatory variable
  $\overline{x}$ is the sample mean of the explanatory variable
  $s_x$is the sample standard deviation of the explanatory variable
  y_i is the i th observation of the response variable
  $\overline{y}$ is the sample mean of the response variable
  $\mathbf{S}_y$ is the sample standard deviation of the response variable
  n is the number of individuals in the sample

- **The least-squares regression line** minimizes the sum of the squared errors (or residuals). This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$ (read "y-hat"). We represent this as $\sum residuals^2$

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

  Where:

$$b_1 = r \cdot \frac{s_y}{s_x} \text{ is the slope of the least-squares regression line.}$$

  And:

$$b_0 = \overline{y} - b_1 \overline{x} \text{ is the y-Intercept of the least-squares regression line.}$$

- The observed distance for this club-head speed is 274 yards. The difference between the observed and predicted values of y is the error, or **residual**.

$$Residual \ = observed - predicted.$$

- **The coefficient of determination**, R^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

$$R^2 = r^2.$$

- **Total Deviation:** The deviation between the observed value, y, and mean value, $\overline{y}$, of the response variable.

$$y - \overline{y}$$
$$Or : \text{Explained Deviation} + \text{Unexplained Deviation}.$$

- **Explained Deviation:** The deviation between the predicted value, $\hat{y}$, and mean value, $\overline{y}$, of the response variable.

$$\hat{y} - \overline{y}.$$

- **Unexplained Deviation:** The deviation between the observed value, y, and predicted value, $\hat{y}$, of the response variable

$$y - \hat{y}.$$

- If a plot of the residuals against the explanatory variable shows the spread of the residuals increasing or decreasing as the explanatory variable increases, then a strict requirement of the linear model is violated. This requirement is called **constant error variance**. The statistical term for constant error variance is **homoscedasticity**

3 Concepts To Know:

- Statistics and Statistical Thinking.
  - Describe Variability
  - Understand Sources of variability
  - The process of statistics
    1. Identify the problem to be solved. It is important to clearly lay out the questions that the researcher wants answered, along with clearly specifying which population the study applies.
    2. Collect the data.
    3. Describe the data.
    4. Perform inference.
- Inferential/Descriptive Statistics
  - Inferential Statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.
  - Descriptive statistics consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.
- Variables
  - Qualitative (Categorical) / Quantitative
  - Discrete (count) / Continuous (measure)
- Data
  - Qualitative (Categorical) / Quantitative
  - Discrete (count) / Continuous (measure)
- Distinguish between an Observational Study and a Designed Experiment
- Explain the Various Types of Observational Studies
  - Cross-sectional Studies: Observational studies that collect information about individuals at a specific point in time, or over a very short period of time.
  - Case-control Studies: These studies are retrospective, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals that have certain characteristics are matched with those that do not.
    - Positive: Control group allows for a comparison
    - Negative: Individuals must remember details
    - Negative: Records might not exist
  - Cohort Studies: A cohort study first identifies a group of individuals to participate in the study (cohort). The cohort is then observed over a period of time. Over this time period, characteristics about the individuals are recorded. Because the data is collected over time, cohort studies are prospective.
    - Advantage: Researcher does not need to rely on the memory of study participants or existing records.
    - Disadvantage: Requires a lot of time.
    - Disadvantage: Could be expensive.
- Effective Sampling Techniques:
  1. Simple random sampling
  2. Stratified sampling
  3. Systematic sampling
  4. Cluster sampling

- Obtaining a Simple Random Sample Using Stat-Crunch or calculator
- Bonus:
  - nCk formula
  - Probability formula
- Obtain a Stratified Sample
- Obtain a Systematic Sample
  - Choosing a value of k if N is not known
  - Mathematically deducing a value of k if N is known
  - Steps to deduce k and compute sequence:
    1. If possible, approximate the population size N.
    2. Determine the sample size desired, n.
    3. Divide N by n and round down to the nearest integer. This value is k.
    4. Randomly select a number between 1 and k, call this number a (starting point).
    5. The sample will consist of the following individuals:
- Obtain a Cluster Sample
  - We conclude that if the clusters have homogeneous individuals, it is better to have more clusters with fewer individuals in each cluster.
  - When each cluster is heterogeneous, fewer clusters with more individuals in each cluster are appropriate.
- 3 sources of bias
- Steps in designing an experiment
- Explain a matched-pairs design
- Make Frequency Distribution Chart
- Make Relative Frequency Distribution Chart
- Make bar graph
- Dual bar graph
- Pie chart
- Creating Tables with classes (using bin function $\rightarrow$ data $>$ bin)
- Histogram
- Determining the lower limit of the first class and class width formula
- Drawing a dot plot
- Identify Shape of a Distribution
- Most common graphical misrepresentations of data:
  - Scale
  - Inconsistent Scale
  - Misplaced Origin
- Finding population mean (formula)
- Finding a simple random sample from a population

- Finding sample mean (formula)
- Steps in finding median, finding median if n is even and if n is odd (formula)
- How to find mode
- When to use mean, median, or mode.
- Relation among the Mean, Median, and Distribution Shape
- Finding Range (formula)
- What is deviation about the mean? What should the sum of all these values be?
- Population Standard Deviation (formula)
  - The sum of the deviations about the mean always equals zero
- Sample Standard Deviation (formula)
- What does a higher standard deviation mean?
- Is the standard deviation resistant?
- What is variance?
- The empirical rule
- Approximate population mean (formula)
- Approximate sample mean (formula)
- Weighted mean (formula)
- Approximate population standard deviation (formula)
- Approximate sample standard deviation (formula)
- What is a z-score?
- Population z-score (formula)
- Sample z-score (formula)
- A z-score is unitless (mean=0, std.dev=1)
- Negative z-score
- Interpret percentiles
- Find and interpret quartiles
  - Quartiles are resistant
- Find IQR (formula)
- Deciding Which Measure of Central Tendency and Dispersion to Report
- Fences (finding fences) (formula)
- How fences find Outliers
- Five Number Summary
- Boxplots (Create them)
- Scatter plot axes

- Positive and negative association
- sample linear correlation (formula)
- Properties of linear correlation cofficient
- Chart
- correlation causation
- equation of least-squares regression line (formula)
- use it to make predictions
- what is a residual (formula)
- sum of least squares residual
- Coefficient of determination
- Three tyes of deviation and how to calculate (total deviation = explained vs unexplained)
- Finding the Coefficient of Determination
- is a linear model appropriate?
- is the variance of the residuals constant?
- are their any outliers?
- Marginal Distribution
- Conditional Distribution
- Simpson's Paradox
- Drawing bar graph of conditional distribution