Statistics

- **CHAPTER 1**
  - **Statistics:**
    - Numerical facts
    - The science of collecting, analyzing, presenting, and interpreting data
    - Making decisions
    - Science
  - Data is everywhere
  - **Inferences:** based on statistics obtained from the data
    - Be as good as the data
  - **Data can be obtained from:**
    - **Observational studies:** Can not control how subjects are assigned to groups and which treatments each group receives
    - **Experiments:** A controlled study in which the researcher attempts to understand cause-and-effect relationships
    - **Administering closed-ended questions:** Do not modify environment, simply administer a process closed-ended questions
  - Surveys
    - **Survey:** Process of collecting data
  - **Element:** A specific subject or object
  - **Population:** Collecting of all elements
    - N = population size
    - "All"
  - **Target population:** The population that is being studied
  - The main goal of statistic is to describe or characterize the population
  - **Census:** A survey that includes every element of the target population
  - **Sample:** A subset of the population that is selected for a study
    - "Selected"
    - "Each of the _____"
  - **Sampling:**
    - **With replacement:** A member of the population may be chosen more than once
    - **Without replacement:** A member of the population may be chosen only once
  - **Representative sample:** A sample that represents the characteristics of the population as closely as possible
  - **Process of data selection:**
    - Define the objectives of the survey or experiment
    - Define the variable and target population
    - Defining the sample-selection and data measuring or data-collecting schemes
  - **Sampling frame:** A list of the elects belonging to the population from which the sample will be drawn
  - **Sample design:** The process of selecting sample elements from the sampling frame
  - **Judgement samples:** Samples that are selected on the basis of being typical
  - **Probability samples:** Samples in which the elements to be selected are drawn on the basis of probability

Statistics

- **Probability:** A measure of the likelihood that a certain outcome will occur
  - Chance
  - Links descriptive and inferential statistics together
    - **Descriptive statistics:** Summarize data by construction tables drawing graphs, or calculating summary measures such as averages
      - Consists of methods for organizing, displaying, and describing data by using tables, graphs, and numbers
      - Provides ways to display data in a clear, understandable, and readable form
  - **Inferential statistics:** Consists of methods that use sample results to help make decisions or predictions about a population
    - Consists of methods that use sample results to help make decisions or predictions about a population
- **Biased sampling:** Method that produces data which systematically differs from the population
  - **Convenience sample:** Easily accessible
  - **Volunteer sample:** Chose to contribute the needed information
- **Simple random sample:** Every element in the population has a equal probability of being chosen
  - **Proportional sample:** Stratifying the sampling frame and then selecting some or all of the items
- **Systematic sample:** Every K*th* item of the sampling frame is selected
  - $N/n = K$
- **Variable:** The name given to what is being measured, counted, or observed from the elements of a population or a sample
  - x, y, or z
- **Observation:** Value of a variable for a specific element
- **Data set:** A collection of observations on one or more variable
- **Types of variables:**
  - **Quantitative:** A variable that measures a numerical quantity or amount
    - **Discrete:** Values are results of counting
      - **EX:** # of siblings, # of households with Netflix subscribers
    - **Continuous:** Values are results of measuring
      - **EX:** money, distance of house from campus, salaries
  - **Qualitative:** A variable that measures a quality or characteristic
    - Categorical variable
      - **Nominal:** Unordered categories
        - **EX:** color of a vehicle, blood type, gender
      - **Ordinal:** Order among the categories
        - **EX:** degree of injury: fatal, severe, moderate, minor
- **Classification of data:**
  - **Cross-section data:** Data collected on different elements at the same point in time or for the same period of time

- **Time-series data:** Data collected on the same element for the same variable at different periods of time

CHAPTER 2

- **Raw data:** Data on a single variable recorded in the sequence in which they are collected and before they are processed or ranked
  - Contains information on each element of a sample or population individually
  - Organized into: Tables, graphs, numerical measures
  - What values of the variable have been observed/measured
  - How often each value has occurred
- **Minimum:** Data value that is less than or equal to all other values in the data
  - =MIN( )
- **Maximum:** Data value that is greater than or equal to all other values in the data
  - MAX( )
- **Range:** A measure of dispersion that is defined as the difference between the largest and smallest observation
  - Range= MAX - MIN
- **Outliers:** Values that are very small or very large relative to the majority of the values in a data set
- **Stem-and-leaf Plot:** A graph for a raw cross-section quantitative data
  - Uses the actual numerical values of each data value as part of the graph
  - **Leaf:** Represents the values in the right-most decimal position in all data value
- **Line Plot:** The x-axis represents time measurements while the y-axis is a representative of measure or percentage of quantity
  - **Trend**: Long-run increase or decrease over time
    - Data taken over a long period of time
    - Upward or downward
    - Linear or nonlinear
  - **Seasonality:** Short-term regular wave-like patterns
    - Observed within 1 year
    - Often monthly or quarterly
- **Distribution table:** Exhibits how the collected data from a single variable are distributed over the unique values of the variable of interest
  - **Two columns:**
    - 1) Lists all the unique values of the variable
    - 2) Gives the number of elements that belong to each of the categories in the first column
- **Frequency:** The number of elements that belong to each of the category
- **Grouped data:** Data presented in the form of a data distribution table
- **Numerical measures:** Quantities that can be calculated from observed values of a variable
- **Parameter:** A numerical measure computed from a population data

  - Capital letter (N)
- **Statistic:** A numerical measure computed from a sample data
- **Peak:** A point in the graph that is higher than any other
- **Unimodel:** Has one peak
- **Bimodel:** Two peaks
- **Symmetric Distribution:** The left and right sides of the distribution, when divided at the middle value, form mirror images
- **Asymmetric:** Not symmetric
- **Skewed-to-the-right:** A greater proportion of the measurements lie on the right of the peak value
- **Skewed-to-the-left:** A greater proportion of the measurements lie on the left of the peak value
- **Line plot:** Time measurements while the y-axis is a representative of measure
- **Relative frequency (Proportion):** Shows the proportion of the total frequency that belongs to the corresponding category/class
  - Frequency/Sum of all frequencies
- **Percentage:** Shows how much percent of the total frequency belongs to the corresponding category/class
  - 100*(Relative frequency)
- **Proportion:** numerical measures obtained from a qualitative data set by taking the ratio of the number of elements with a specific characteristic to the number of elements in the data set
  - **Population proportion:** obtained by taking the ratios of the number of elements in a population with a specific characteristic to the total number of elements in a population
    - P = # OF ELEMENTS IN THE POPULATION WITH A SPECIFIC CHARACTERISTIC / N
  - **Sample proportion:** obtained by taking the ratios of the number of elements in a sample with a specific characteristic to the total number of elements in a sample
    - P = # OF ELEMENTS IN THE SAMPLE WITH A SPECIFIC CHARACTERISTIC / n
- **Percentage:** another way of expressing a proportion
  - 100%*(PROPORTION)
- **Contingency table:** a table in which frequencies correspond to two qualitative variables are presented
  - At least two rows and two columns
- **Rate:** used to refer to a percentage or probability
- **Bar graph:** a graph commonly used to display distribution tables for qualitative data and distribution tables of quantitative data that uses single values for classes
- **Type of distributions for quantitative data:**
  - 1. Determine the number of classes
  - 2. Determine the class width
  - 3. Determine the lower limit of the first class
  - 4. Form the intervals
  - 5. Determine frequencies

    - **Classes:** Identify all the unique values of the discrete quantitative variable of interest in the data set and list all the values as single-value classes; use the sort option to group the values and determine the unique values of the quantitative variable
        - **Single value classes:** used when the quantitative variable is discrete and assumes only a few distinct variables
        - **Interval classes:** used when the quantitative variable is continuous when the quantitative variable is discrete but assumes more than 10 unique values
            - **Approximate class width:** RANGE / NUMBER OF CLASSES
                - Round class width to whole number
            - **Class midpoint:** (LOWER LIMIT + UPPER LIMIT) / 2
- **Histogram:** used to display graphically a distribution table for quantitative data that uses intervals for classes
    - Bars drawn adjacent to each other (NO GAPS)
- **Polygon:** a graph formed by a line graph where the horizontal values are the midpoints of the distribution of the data and the vertical values are either frequency, relative frequency, or percentage
- **Distribution Table**
    - Determine the number of classes
    - Determine the class width
    - Determine the lower limit of the first class
    - Form the intervals required for each class
    - Determine frequencies
- **Shape**
    - Symmetric or asymmetric
    - Skewed to the right (long tail on the right) or skewed to the left

- **Numerical measures:** quantities that can be calculated from observed values of a variable that give a sense of the nature of the dataset with respect to center of the data, variability in the data, and relative position in the data
- **Parameter:** a numerical measure computed from a population data
- **Statistics:** a numerical measure computed from a sample data
- **Variance:** the mean of the squared deviations of the data values from the mean
    - **Population variance:** variance calculated from a population data
        - =VAR.P()
    - **Sample variance:** variance calculated from a sample data
        - VAR.S()
    - Unit: Unit^2
- **Standard deviation:** positive square root of the variance
    - Variance given: =SQRT()
    - Tells how closely the values of a data set are clustered around the mean

  - **Population standard deviation:** the standard deviation calculated from a population data
    - =STDEV.P()
  - **Sample standard deviation:** standard deviation calculated from a sample data
    - =STDEV.S()
- **Coefficient of variance:** a numerical measure that relates the standard deviation of a dataset to its mean
  - **CV:** (STANDARD DEVIATION / MEAN) * 100
  - The higher the CV, the greater the dispersion in the variable

- **Calculating grouped data:**
  - M= midpoint
  - f= frequency
  - n= sum of frequency
- **Z-scores:** determine the distance between an observation and the mean; unitless
  - How many standard deviations an observation is away from the mean
  - Z-SCORE: (OBSERVATION - MEAN) / STANDARD DEVIATION
  - **Not unusual:** Z-score between $-2 < Z < 2$
  - **Somewhat unusual:** Z-score between $-3 < Z < -2$ or $2 < Z < 3$
  - **Possible outlier:** $Z < 3$ or $Z > 3$
- **Probability:** measurement of the likelihood that a certain outcome will occur
  - $0 <$ PROBABILITY < 1
- **Random variable**: a numerical description of the outcome of an experiment
- **Normal probability distribution:** graphically represented by a unimodal and bell shaped curve satisfying the following properties
  - Total area under the curve is 1.0
  - Population standard deviation > 0 and population mean less than Infiniti and greater than negative infinity
  - Curve is symmetric to the mean
  - Two tails of curve extend indefinitely
  - **Normal probabilities:** =NORM.DIST(a, mean, standard_deviation, TRUE)
    - **Area to left:** =NORM.DIST(a, mean, standard_deviation, TRUE)
    - **Area to right:** 1 - =NORM.DIST(a, mean, standard_deviation, TRUE)
- **Empirical rule:**
  - Approximately **68%** of the observations lie within one standard deviation of the mean
  - Approximately **95%** of the observations lie within two standard deviations of the mean
  - Approximately **99.7%** of observations lie within three standard deviations of the mean
- **Normal random variable when probability is known:**
  - =NORM.INV(p/100, mean, standard_deviation)
- **Standard normal distribution:**
  - population mean = o

- population standard deviation = 1
- As you increase the standard deviation with the same mean, the normal curve will be shorter and wider
- As you increase the mean with the same standard deviation, the normal curve will shift to the right

- **kth percentile:** (K*N/100)
  - k% of the observation are less than the kth percentile, and 100-k% of the observations are greater than the kth percentile
- **Percentile rank of xp:** (# OF OBSERVATIONS LESS THAN XP / TOTAL # OF OBSERVATIONS)* 100
  - q% of the observations are less than the xp and 100-q% of the observations are greater than or equal to xp

- **5-Number Summary**
  - Lower Adjacent value
  - First Quartile
  - Median
  - Third Quartile
  - Upper Adjacent value
- **Median**: a numerical measure that represents value of the middle term in a data set that has been ranked in increasing order
  - Mean: balance point
  - Median: halfway point
- **Quartiles:** three summary measures that divide a ranked data set into four equal parts
  - **Q1:** value of the middle term among observations that are less than the median
  - **Q2:** same as the median of a data set
  - **Q3**: the value of the middle term among the observations that are greater than the median
- **Interquartile Range:** the range of the middle 50% of the data
  - Difference between third quartile and first quartile
    - Q3-Q1
- **Lower adjacent value:** smallest observation greater than or equal to the lower inner fence
  - **Lower inner fence:** Q1 - (1.5 * IQR)
- **Upper adjacent value**: largest observation less than or equal to the upper inner fence
  - **Upper inner fence:** Q3 + (1.5 * IQR)
- **Outlier:**
  - If x < lower inner fence
  - If x > upper inner fence
- **Box-and-Whisker plot:**
  - A- (Lower adjacent value)
  - Q1
  - Median
  - Q3
  - A+ (Upper adjacent value)

Statistics

- **Interpreting:**
  - **Symmetric:** median line is approximately equal distance to Q1 and Q3, whiskers are of equal length
  - **Skewed-to-right:** median line close to Q1, top whisker longer
  - **Skewed-to-left:** median line is closer to the Q3, bottom whisker longer
- **Measures of central tendency:** numerical measures that represent the center of the distribution
  - Mean, Median, Mode
- **Measures of dispersion:** numerical measures that provide information about the variability among the data values
  - Range, IQR, Variance, Standard Deviation, Coefficient of Variance
- **Measure of position:** numerical measure that determine the position of a single value in relation to two other values in the data set
  - Z-Scores, Quartiles, Percentiles
- **Creating Box-and-Whisker Plot:**

1. Rewrite the Five-number summary in the following order (Q1,A-,Median,A+,Q3)
2. Make line plot with markers - right click on point - format data series
    1. Marker option:
        1. Built in
        2. Type: long line
        3. Size: 15
    2. Fill: solid fill
    3. Border: No line
    4. Line: No line
3. Chart tools - Design - Add chart elements
        1. Lines: High-low lines
        2. Up/down bars: Up/down bars
4. Right click axis - format axis to include outliers
    1. Insert - illustrations - shapes - add outlier

- **Scatter Plot:** provides a convenient way to describe whether a linear relationship exists between two quantitative variables
- **Patterns:**
  - Straight line/linear: either upward or downward slope
  - Curved patter/nonlinear
  - Clusters
  - Outliers
- **X=** independent variable
- **Y=** dependent variable
- **Linear correlation coefficient:** describes strength and direction of the linear relationship between x and y
  - -1 to 1
  - Unitless
  - **Excel:**
    - =CORREL(x=xdata,y=ydata)
  - **Interpreting linear correlation coefficient:**

  - r= 0: no linear relationship
  - r= 1,-1: perfect linear relationship
  - -1 < r < -.5: strong negative
  - -.5 < r < 0: weak negative
  - 0 < r < .5: weak positive
  - .5 < r < 1: stromg positive
- **Simple regression:** describes the relationship between two or more numerical variables
  - One independent variable and one dependent variable: independent variable is being used to explain the variation in the dependent variable
- **Models:**
  - Simple: 1 explanatory variable
  - Multiple: 2+ explanatory variables
- **Least squares regression line:** y = a + bx
- How to formulate equation of best fit regression line
  - **B =** r * (Sy/Sx)
    - **Sy:** standard deviation of y
    - **Sx:** standard deviation of x
  - **A =** ybar - (b* xbar)
    - **ybar:** mean of y
    - **xbar:** mean of x
  - **Yhat =** predicted value of the response variable
- Interpreting
  - **For a:** The predicted value of (y variable) is (value of a) (unit of y variable) when (x variable) is 0 (unit of x variable)
  - **For b:** The average change in (y variable) is (value of b) (unit of y variable) for each (unit of x variable) of change in (x variable)
- **Coefficient of determination (r^2):** measure of how good the regression model fits the data
  - Proportion of variability in the dependent variable explained by the regression model
  - The closer r^2 is to 1, the better the best-fit regression model
  - r^2:
    - **Excel:** =CORREL()^2
- **Residual:** the difference between the observed value of the response and the predicted value of the response
  - **Residual:** Observed - Predicted
  - Standard deviation of residuals
    - **Excel:** =STEYX(ydata, xdata)
- **SSE:** sum of squares of error
  - **Summation of (y-yhat)^2**
- **Estimation:** the assignment of value to a population parameter based on a value of the corresponding sample statistic
- **Estimator:** sample statistic used to estimate a population parameter
- Estimation:
  - **Point Estimation:** the value of an unknown population parameter is estimated by a single number
    - **Phat =** x / n (# of units sampled with desired characteristic / sample size)

  - **Interval Estimation:** value of an unknown population parameter is estimated by an interval constructed around the point estimate that is believed with some percentage contains the unknown value of the population parameter
- **Confidence level:** (1-alpha)*100%
- **Margin of Error (E):** E = Z(1-alpha/2) * (standard deviation/ square root of n)
  - Z(1-alpha/2): **Excel =**NORM.INV(1-alpha/2, 0, 1)
- **Interpretation of confidence interval:** we are (1-alpha)*100% confident that the true value of the unknown population mean is between lower limit of confidence interval and upper limit of confidence interval.
- Confidence interval:
  - **Population mean:** xbar - (Z(1-alpha/2)*(standard deviation/square root of n) <= mean <= xbar + (Z(1-alpha/2)*(standard deviation/square root of n)
  - Sample mean: xbar
- **Sample size:** n = (Z(1-alpha/2)*standard deviation / E)^2
  - Sample size large if:
    - n * phat > 5
    - n * (1-phat) > 5
- **Population proportion:** phat - (Z(1-alpha/2)* square root of (phat * (1-phat)/n) <= p <= phat + (Z(1-alpha/2)* square root of (phat * (1-phat)/n)