Outline 1.1, Analyzing Categorical Data

- Categorical Variables
  - are labeled as categories
  - distribution is the count or percent of the data that falls under the category
- Bar Graphs and Pie Charts
  - Pie Charts
    - cannot easily be made by hand
    - assign each category a "slice" of the pie
    - are easier to make than bar graphs
  - Bar Graphs
    - show data more easily than pie charts
    - make it easier to see small data values
    - can also show more types of data
    - Bad Bar Graphs
      - use a disproportinate scale for quantity
      - use pictures
- Two-way Tables
  - Happen when two categorical variables are combined together in the same table
  - Can be analyzed by looking at each distribution separately
  - Distributions only containing one of a axis are called marginal distributions
  - The actual values in each marginal distribution mean nothing when compared to the other marginal distributions
  - In order to compare them the conditional distribution need to be calculated
  - The conditional distribution is the relative percent of each variable out of the whole of the marginal distribution
- Relationships between categorical tables
  - Side by Side bar graphs
  - Segmented bar graphs
  - All graphs need to provide an association between the variables in order to compare them

Outline 1.2, Displaying Quantitative Data with Graphs

- Dotplots
  - Each value is shown above a number line as a dot
  - For quantitative data
- Characteristics of Graphs
  - Shape
    - The shape of the graph
    - Skew
      - Includes the skew of the graph, if there is any
      - What direction the runoff of the data is oriented
      - Can be called "left-skewed", "negetive-skewed", and vise versa
    - Where peaks are
    - Where gaps are
    - Symmetry
      - How one side of the graph compares to the other
      - A lack of skew
  - Center
    - Where the graph is centered
    - Based on the mean, median, and mode
    - Is the "midpoint" of the data
  - Spread
    - How the data is distributed on the quantitative axis
  - Outliers
    - Data that does not fit inside the realm of the rest of the data
    - Is still regarded, just with a grain of salt
- Comparing Distributions
  - Important to make sense of the data
  - Stemplots
    - Uses a stem down the middle, with "leaves" coming off the edges to indicate data falling under the range of the stem
    - Not good for large datasets
    - Does not distribute the data as percisely as other formats

    - Rounding can fix some problems with too many digits
    - Stems can be split to make more acurate shapes
  - Histograms
    - Groups ranges of data
    - No gaps between data
    - Bar graphs have catagorical data, histograms have quantitive data

Outline 1.3, Describing Quantitative Data with Numbers

- The median is used to indicate the center of the distribution set, while the quartiles are used to determine spread from the center of the set, in conjunction with the minimum and maximum
- Outliers describe the parts of the data that do not fit into the expected range
- The mean and median combine to describe where the center of the distribution is
- A measure is resistant if changing/adding a single data point would not extremely change the measure
- The median is a resistant measure, as are the quartiles
- The mean and the standard deviation are both good descriptors for getting a rough idea about how a set would appear if normally distributed. The median and IQR are better for determining the shape of the distribution
- A numerical summary is always inferior to a graphical representation due to one conveying more data than the other
- The variance (s_x^2) and the standard deviation (s_x) are measures of the spread of data, with them directly correlating to how the data is spread

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

- A distribution is mainly characterized by the center, the spread, and the variability
- The five number summary can be incorporated into a box plot for overlay on a graph or a number line
- A five number summary is comprised of the median, quartiles, minimum, and maximum of the data set
- The quartiles are the medians of the halves of the data, separated by the median

Outline 2.1, Describing Location in a Distribution

- Percentiles
  - Are a measure of the percent of the data beneath the data point
  - Range from 0% to $99.\bar{9}$%
  - Used for telling how you relate to the rest of the data
- Z scores
  - Measure of how many standard deviations away a point is from the mean
  - Calculated as
$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$
  - Can also be used to compare across distributions
  - Better than a constant distance as the measure of spread is more universal
- Cumulative Relative Frequency
  - Like a histogram, but uses percents and includes all previous points
  - Makes better sense for certain applications where the change in value is cumulative
  - Also allows you to estimate backwards and between
- Data Transformations
  - Additive transformations
    - Shape does not change
    - Center does change
    - Spread does not change
    - Done by adding a constant to all of the points in the dataset
  - Multiplicative transformations
    - Shape does not change
    - Center usually changes
    - Spread does change
    - Done by multiplying all points in a dataset by a constant
  - Mixed transformations
    - Shape does not change
    - Center does changes
    - Spread does change
    - Done by applying two transformation to the same dataset

  - General
    * Transformations affect the 5 number summary directly
    * The shape of a distribution is always the same
- Conversion
  - Z scores can be converted into percentiles with a Z table
  - Can also go vice versa with more info

Outline 2.2, Density Curves and Normal Distributions

- Density Curve
  - The pattern of a distribution can be represented by a density curve
  - Allocates area to values based on frequency
  - Has total area 1
  - Relative to total value
  - Smooths out irregularities
  - We use the overall variables instead of the sample variables
  - Mean and median should be obvious on the graph
  - Tail pulls on mean, while less so on median
- Normal Distributions
  - Are the ideal distributions
  - Bell-shaped
  - Symmetric
  - Have a perfect curve
  - Follow the 68-95-99.7 standard deviation rule
  - Mean and standard deviation are constraining points
  - Are equivalent when standardized
  - z is a standard normal distribution centered on the y axis and with standard deviation 1
  - z scores can be converted into percentiles with precalculated tables
  - All distributions are some factor away from normality

Outline 3.1, Scatterplots and Correlation

- A scatterplot shows the correlation and relation between two quantitative distributions that are connected by the same source.
- If a relation is somewhat cause and effect, we call the independent variable an explanatory variable and the dependent variable is called the responsive variable.
- When we look at a scatterplot, we are looking for
  - Direction (Positive or negative correlation)
  - Strength (Line of best fit accuracy)
  - Form (type of line)
  - Outliers
- Correlation is show via the slope and the sign, with a positive slope increasing and a negative slope decreasing with the trend
- r is the correlation coefficient, and tells how strong the graph is and the distance of the points from it. Ranges from -1 to 1, with 0 being no correlation at all.

$$r = \frac{1}{n-1} \sum \frac{x_i - \bar{x}}{\sigma_x} \cdot \frac{y_i - \bar{y}}{\sigma_y}$$

Outline 3.2, Least Squares Regression

- Regression lines are a regression of a data set into a simpler, continuous model, which has a few distinct properties
  - Form is the shape of the regression line
  - Strength is how close the data is to the line
  - Regression lines can be used to predict the location of other points in the distribution
  - The slope is the best fitting change for the response variable in relation to the explanatory variable
  - The y intercept is the starting value of the explanatory set
  - The line can be represented with $\hat{y}$
  - Predicting data values outside the range is unsafe and not very reliable
- Most common way to create a regression line is to use the least squares regression
  - When creating one, you can examine the fit by looking at the residuals
  - If they have no correlation, it means that the line is a good fit

  - A residual plot can be made in order look look at the trend
  - The standard deviation of the residuals tells you how well the data fits the line
  - The squared regression coefficient r^2 tells you how much of the dataset can be explained by the line
  - The point $(\bar{x}, \bar{y})$ is always on the line

$$m = r\left(\frac{s_y}{s_x}\right)$$

$$b = \bar{y} - b\bar{x}$$

- Correlation and causation are two different things, and both have operate meanings in the same context
  - The line is not resistant, so a wild outlier can throw off the whole data
  - Just because data is strongly correlated, it does not mean a conclusion is sure

Outline 4.1, Sampling and Surveys

- A census is a survey for a dataset of a complete population
- A sample survey only collects data from a certain sample of the population
  - Samples are expanded to a population with inference, which is creating conclusions based on the sample's findings alone
  - Conveniences samples target those that are most willing and easy to accept
  - Voluntary samples are ones that choose to participate in the collection of data
  - A random set is simply a random subset of a population
    - A simple random sample gives every combination the same chance at being selected
    - A stratified random sample cherry picks people from different groups to be in the data to get a better fluctuation to find a better trend
    - A cluster sample is when areas are divided into clusters, and a SRS of these clusters are selected to participate in surveying
    - Random sampling decreases bias, but under-representation of some groups in the data can throw the data off if the RNG is bad
  - Humans are also a problem
    - They can just not respond
    - They can have bias on the survey and answer falsely
    - They can misinterpret questions and answer them improperly

Outline 4.2, Experiments

- Scientific questions can be addressed using data from observational experiments or studies
- It is important to only change one variable so that the difference between the results after testing is very likely to be in correlation to the variable that you are changing
- If the explanatory variable is influenced by other fields, it can become confounded and the relation between the response variable and the explanatory variable will be much lower than otherwise
- Observational studies are often misleading as the massive amount of factors that go into the result cloud the relation that is being studied
- In a study, treatments are imposed on subjects, or experimental units if they are not humans
- Every treatment is made up of changes in the environment, called factors
- Experimentation should include:
  - Comparison of treatments
  - Control treatment
  - Random assignment
  - Replicatability
- Completely random designs have everyone getting the exact same chance
- Placebos help experiments by eliminating the psychological aspect of testing and giving better and more accurate results
- Double blind experiments are when both the surveyor and the subject have no knowledge about the treatment
- If one know about the treatment, it is only single-blind
- If the result is very significant, then it is statistically significant
- Blocking together the treatments can create a more distributed and fair set in exchange for true randomness