Midterm Review

- Key concepts: variable, observation, population, sample, etc.
- Numerical vs. Categorical data
- Graphics for numerical and categorical data
- Describe the distribution from the graphics
- Measures of center and spread
- Five-number summary and boxplot
- Apply empirical rule and z-score
- Interpret scatter plots
- Find and interpret the correlation coefficient
- Find and interpret a regression model
- Assess the regression model with the coefficient of determination and residual plot
- Association vs. Causation
- Sampling and potential bias
- Observational studies vs. experimental studies
- Key features in controlled experiments
- Basic probability rules for calculating theoretical probabilities.

Examining the distribution of a variable

1. Categorical variable:

-- Takes category or label values. Beware of variable coding! It does not make sense to perform numerical operations.

- Graphs:
  - pie chart: display each category as a slice of the pie
  - bar chart: display each category as a bar
- Numerical summaries:
  - Mode: category of the highest frequency
  - Variability/diversity: if many observations spread across many different categories, then the variability is high, if many observations fall into the same categories, then the variability is low

2. Numerical variable:

-- take numerical values and represent some kind of measurement.

- Graphs:
  - Dot plot: plots each individual value as a dot on the number line
  - Histogram: plots the number (count) of observations that fall in intervals of values.

When examining the distribution of a quantitative variable, one should describe:

- The overall pattern of the data: Shape: symmetry/skewness; peakedness/modality; Center; Spread
- Any deviations from the pattern (outliers)

3. Numerical measures of the center

1) Mode: the most frequently occurring value.
    - May not be unique
    - May not exist
2) Mean: the arithmetic average
    - Sensitive to extreme values (as it factors in their magnitude).
    - Appropriate only for symmetric distributions with no outliers
3) Median: the middle value in sorted data with 50% of data below it and 50% above it
    - Resistant to extreme values (as it depends on the order).

4. Numerical measures of the spread

1) Range = Maximum – Minimum
    - Pro: easy to calculate, useful for a quick measurement of variability
    - Con: sensitive to extreme values
2) Inter-Quartile Range: range covered by the middle 50% of the data.

$$IQR = Q3 - Q1$$

    - The first quartile (Q1) is the median of the lower half of the data, about one quarter (25%) of the data points fall below it.
    - The third quartile (Q3) the median of the upper half of the data, about three quarters (75%) of the data points fall below it.

    How to find the IQR?
    i. Sort the data in numerical order
    ii. Split the data into 2 halves at the median (do not include the median in the lower/upper half)
    iii. Find the median of the lower half of the sorted data (Q1)
    iv. Find the median of the upper half of the sorted data (Q3)
    v. The IQR is the distance between Q3 and Q1 (Q3-Q1)
3) Standard deviation: The average distance between the data points and the mean.
    i. Find the mean.
    ii. For each data point, find the difference between it and the mean, and square the difference.
    iii. Find variance by summing up the squared differences and dividing by (n-1), where n is the number of observations.
    iv. Take the square root of the variance found in step iii to get the standard deviation.

Note: It is essential to choose the appropriate measure of center and spread based on the nature of the data and the research question.
Since the mean and standard deviations are highly influenced by extreme values, they should be used as numerical descriptions of the center and spread only for distributions that are roughly symmetric and have no outliers. The median and IQR are useful when extreme values are present, but they do not take into account all data points like the mean and standard deviation do.

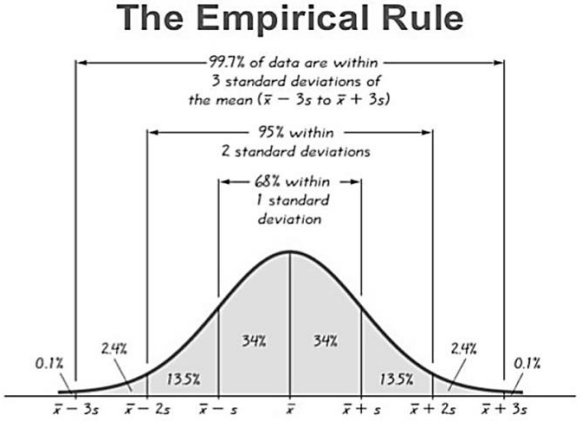
5. Five-number summary and Boxplot

- The five-number summary: Min, Q1, Q2, Q3, Max
    - The five-number summary provides a complete numerical description of a distribution: the **center** (Q2), and the **extremes** (min and max, which give the range) and the **quartiles** (Q1, Q3, which give the IQR) describe the **spread**.
    - The 1.5(IQR) criterion can be used to detect outliers: Observations that fall below Q1-1.5(IQR) or above Q3+1.5(IQR).
- The boxplot visually displays the five number summary and any observation that was classified as a potential outlier using the 1.5 (IQR) criterion.
    - The main box goes from Q1 to Q3, with a line indicating the median (Q2).
    - The lower whisker extends to the smallest observation $\geqslant$ Q1-1.5(IQR).
    - The upper whisker extends to the largest observation $\leqslant$Q3+1.5(IQR).
    - Any potential outliers are marked individually.

Note that the boxplot is particularly useful for comparing multiple distributions at once, as the boxes and whiskers allow for easy visual comparison of the center, spread, and range of each distribution. However, it is important to keep in mind that the boxplot may not reveal certain characteristics of the distribution, such as its shape.

6. Empirical rule and Z-scores

- For **symmetric unimodal** distributions, the **Empirical Rule** tells us what percentage of the observations falls within 1, 2, and 3 standard deviations of the mean
  The Empirical Rule states that in unimodal and symmetric distributions,
    - The middle 68% of the data is within 1 standard deviation
    - The middle 95% of the data is within 2 standard deviations
    - The middle 99.7% of the data is within 3 standard deviations of the mean



- **Z-score** measures the distance of an observation to the mean in standard deviations.
    - **z = (x-xbar)/s**
    - Z-score is **unitless**, makes different groups comparable.
    - The **magnitude** of z-score indicates the degree of deviation from the mean.
    - The **sign** of the z-score indicates relative position from the mean.
    - Data values with z-scores greater than 2 or less than -2 are considered “unusual” or “rare”.

Examining the relationship between variables

1. **Case Categorical → Categorical**
   - Graph: grouped bar chart
   - Numerical summaries: Two-way table, relative frequency within each group

2. **Case Categorical → Numerical**
   - Graph: side-by-side boxplots.
   - Numerical summaries: five-number summary, mean, std. dev. within each group, depending on the distribution of the data.

3. **Case Numerical → Numerical**
   - Graph: scatterplot
     - Overall pattern → direction, form, strength.
     - Deviations from the pattern → outliers.
   - Numerical summaries: Correlation coefficient, regression model.

4. **Linear relationship**
   - When the relationship between numerical variables is approximately linear, the correlation coefficient (r) measures the direction and the strength.
     - The sign of r indicates the direction.
     - The magnitude of r indicates the strength.
     - r is unitless and heavily influenced by inconsistent outliers
   - Least squares regression line
     - Line of best fit found using the least squares criterion.
     - The equation can be calculated using five statistics: the mean and sd of X, the mean and sd of Y, the correlation coefficient between two variables.
     - The slope represents the average change in Y that's associated with a 1-unit increase in X.
     - The intercept represents the predicted value of Y when X takes the value of zero. Interpret the intercept only when it is meaningful for X to be 0.
   - We evaluate the regression model with:
     - The coefficient of determination (r2, r-squared)
     - The residual plot. A good fit is when the residuals are randomly scattered around 0 without any obvious pattern
   - Caution:
     - Always check the scatterplot for linearity
     - Beware of extrapolation, which means predicting values beyond the range of observed data.
     - Association does not imply causation.
   - Confounding variables: hidden variables that are related to both the explanatory and response variable and cause them to be false correlated.

Producing Data

1. **Potential bias in data production**
   - **Selection bias:**
     - **Undercoverage**: excluding individuals with or without certain characteristics
     - **Volunteer bias**: self-selected individuals who have strong opinions about the subject
     - **Nonresponse bias**: individuals who do not respond differ significantly
   - Measurement bias
     - Systematic error in measurement process, leading to inaccurate data
     - Can arise from faulty instruments, poor experimental design, or human error
   - Estimator bias
     - Occurs when the estimation method used consistently overestimates or underestimates the true value

2. **Sampling**
   - The goal of sampling is to obtain a representative sample that accurately reflects the population of interest
   - Simple random sampling (SRS) is a common method of sampling in which each member of the population has an equal chance of being selected.
     - This requires a sampling frame, which is a list of all the members of the population
     - SRS can help control for undercoverage bias

3. **Study design**
   - Observational study
     - Researchers do not interfere with the explanatory/treatment variable.
     - The values of the treatment variable occur naturally.
     - It is difficult to establish causation because of confounding variables..
   - Experimental study
     - Individuals are assigned to a treatment group or a control group
     - Key features of a well-designed experiment
       - Large sample size
       - Random assignment: controls for confounding variables
       - Blinding
       - Placebo

Probability

- Theoretical probability v.s. Empirical probability
  - Difference and connection
- Probability of an event A: denoted by P(A), it is a number between 0 and 1 that represents how likely for an event A to occur.
- Complement rule: $P(A^c) = 1-P(A)$
- Equally likely outcomes: P(A) = (number of outcomes in A) / (number of outcomes in S)