Unit 1 - Introduction to data

Reading: OpenIntro Statistics, Chapter 1

LO 1. Identify variables as numerical and categorical.

    (a) If variable is numerical, further classify as continuous or discrete based on whether or not the variable can take on an infinite number of values or only non-negative whole numbers, respectively.

    (b) If variable is categorical, determine if it is ordinal based on whether or not the levels have a natural ordering.

LO 2. Define dependent variables as variables that show some relationship with one another. Further categorize this relationship as positive or negative association, when possible.

LO 3. Define variables that are not associated as independent.

    ∗ *Reading: Section 1.1 and 1.2 of OpenIntro Statistics*

    ∗ *Video: Types of variables, Dr. Çetinkaya-Rundel (3:53)*

    ∗ *Test yourself: Give one example of each type of variable you have learned.*

LO 4. Classify a study as observational or experimental, and determine whether the study's results can be generalized to the population and whether they suggest correlation or causation.

LO 5. Question confounding variables and sources of bias in a given study.

LO 6. Distinguish between simple random, stratified, and cluster sampling, and recognize the benefits and drawbacks of choosing one sampling scheme over another.

LO 7. Identify the four principles of experimental design and recognize their purposes: control any possible confounders, randomize into treatment and control groups, replicate by using a sufficiently large sample or repeating the experiment, and block any variables that might influence the response.

LO 8. Identify if single or double blinding has been used in a study.

    ∗ *Reading: Sections 1.3 - 1.5 of OpenIntro Statistics*

    ∗ *Videos:*

        - *Random sampling vs. random assignment, Dr. Çetinkaya-Rundel (3:18)*

        - *Quick video on correlation vs. causation, YouTube (2:19)*

        - *Slower video on correlation vs. causation, KhanAcademy (10:45)*

    ∗ *Article: How Anecdotal Evidence Can Undermine Scientific Results, Scientific American, 2008*

    ∗ *Test yourself:*

LO 9. Use scatterplots for describing the relationship between two numerical variables making sure to note the direction (positive or negative), form (linear or non-linear) and the strength of the relationship as well as any unusual observations that stand out.

LO 10. When describing the distribution of a numerical variable, mention its shape, center, and spread, as well as any unusual observations.

LO 11. Note that there are three commonly used measures of center and spread:

- center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation).

- spread: spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution).

LO 12. Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodoal, multimodal, or uniform.

LO 13. Use histograms and box plots to visualize the shape, center, and spread of numerical distributions, and intensity maps for visualizing the spatial distribution of the data.

LO 14. Assess if robust statistics (e.g. median and IQR) are more appropriate measures of center and spread of numerical variables based on skewness and existence of extreme outliers.

LO 15. Recognize when transformations (e.g. log) can make the distribution of data more symmetric, and hence easier to model.

LO 16. Use frequency tables and bar plots to describe the distribution of one categorical variable.

LO 17. Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

LO 18. Use side-by-side box plots for assessing the relationship between a numerical and a categorical variable.

LO 19. Note that an observed difference in sample statistics suggesting dependence between variables may be due to random chance, and that we need to use hypothesis testing to determine if this difference is too large to be attributed to random chance.

LO 20. Set up null and alternative hypotheses for testing for independence between variables, and evaluate data's support for these hypotheses using a simulation technique.