Unit 1 - Introduction to data

Suggested reading: OpenIntro Statistics, Chapter 1

Suggested exercises:

* *Part 1 - Designing studies: 1.1, 1.3, 1.7, 1.15, 1.17, 1.21*
* *Part 2 - Exploratory data analysis: 1.27, 1.39, 1.41, 1.45, 1.49*
* *Part 3 - Introduction to inference via simulation: 1.51, 1.53*

---

* *Suggested reading: Section 1.1 and 1.2 of OpenIntro Statistics*

LO 1. Identify variables as numerical and categorical.

- If the 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.
- If the 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.

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

* *Suggested reading: Sections 1.3 - 1.5 of OpenIntro Statistics*
* *Relevant article:* ***How Anecdotal Evidence Can Undermine Scientific Results, Scientific American, 2008***

LO 4. Identify the explanatory variable in a pair of variables as the variable suspected of affecting the other, however note that labeling variables as explanatory and response does not guarantee that the relationship between the two is actually causal, even if there is an association identified between the two variables.

LO 5. Classify a study as observational or experimental, and determine and explain whether the study's results can be generalized to the population and whether the results suggest correlation or causation between the quantities studied.

- If random sampling has been employed in data collection, the results should be generalizable to the target population.
- If random assignment has been employed in study design, the results suggest causality.

LO 6. Question confounding variables and sources of bias in a given study.

LO 7. Distinguish between simple random, stratified, and cluster sampling, and recognize the benefits and drawbacks of choosing one sampling scheme over another.

- Simple random sampling: Each subject in the population is equally likely to be selected.
- Stratified sampling: First divide the population into homogenous strata (subjects within each stratum are similar, across strata are different), then randomly sample from within each strata.
- Cluster sampling: First divide the population into clusters (subjects within each cluster are non-homogenous, but clusters are similar to each other), then randomly sample a few clusters, and then randomly sample from within each cluster.

LO 8. 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 9. Identify if single or double blinding has been used in a study.

* *Relevant article: How Anecdotal Evidence Can Undermine Scientific Results, Scientific American, 2008*
* *Test yourself:*
  1. *Describe when a study's results can be generalized to the population at large and when causation can be inferred.*
  2. *Explain why random sampling allows for generalizability of results.*
  3. *Explain why random assignment allows for making causal conclusions.*
  4. *Describe a situation where cluster sampling is more efficient than simple random or stratified sampling.*
  5. *Explain how blinding can help eliminate the placebo effect and other biases.*

* *Suggested reading: Section 1.6 of OpenIntro Statistics*

LO 10. 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 11. When describing the distribution of a numerical variable, mention its shape, center, and spread, as well as any unusual observations.

LO 12. 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: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution).

LO 13. Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodoal, multimodal, or uniform.

LO 14. 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 15. Define a robust statistic (e.g. median, IQR) as a statistics that is not heavily affected by skewness and extreme outliers, and determine when such statistics are more appropriate measures of center and spread compared to other similar statistics.

LO 16. Recognize when transformations (e.g. log) can make the distribution of data more symmetric, and hence easier to model.

- *Test yourself:*
  1. *Describe what is meant by robust statistics and when they are used.*
  2. *Describe when and why we might want to apply a log transformation to a variable.*

- *Suggested reading: Section 1.7 of OpenIntro Statistics*

LO 17. Use frequency tables and bar plots to describe the distribution of one categorical variable.

LO 18. Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

LO 19. Use side-by-side box plots for assessing the relationship between a numerical and a categorical variable.

- *Test yourself:*
  1. *Interpret the plot in Figure 1.40 of the textbook (page 39).*
  2. *You collect data on 100 classmates, 70 females and 30 males. 10% of the class are smokers, and smoking is independent of gender. Calculate how many males and females would be expected to be smokers. Sketch a mosaic plot of this scenario.*

- *Suggested reading: Section 1.8 of OpenIntro Statistics*

LO 20. 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 observed difference is too large to be attributed to random chance.

LO 21. Set up null and alternative hypotheses for testing for independence between variables, and evaluate the data's support for these hypotheses using a simulation technique.

- *Test yourself: Explain why a difference in sample proportions across two groups does not necessarily indicate dependence between the two variables involved.*