Statistics with R Specialization
Course 1: Introduction to Probability and Data with R

Designing Studies

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 associated 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.*

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.

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.*

Exploring Numerical Data

LO 1. 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 2. When describing the distribution of a numerical variable, mention its shape, center, and spread, as well as any unusual observations.

LO 3. 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 4. Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodal, multimodal, or uniform.

LO 5. 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 6. Define a robust statistic (e.g. median, IQR) as a statistic 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 7. 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.*

Exploring Categorical Data

LO 1. Use frequency tables and bar plots to describe the distribution of one categorical variable.

LO 2. Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

LO 3. 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 1.30 of OpenIntro Statistics (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.*

Defining Probability

LO 1. Define the probability of an outcome as the proportion of times the outcome would occur if we observed the random process that gives rise to it an infinite number of times.

LO 2. Explain why the long-run relative frequency of repeated independent events settles down to the true probability as the number of trials increases, i.e. why the law of large numbers holds.

LO 3. Define disjoint (mutually exclusive) events as events that cannot both happen at the same time:

- If A and B are disjoint, P(A and B) = 0

LO 4. Distinguish between disjoint and independent events.

- If A and B are independent, then having information on A does not tell us anything about B (and vice versa).
- If A and B are disjoint, then knowing that A occurs tells us that B cannot occur (and vice versa).
- Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

LO 5. Draw Venn diagrams representing events and their probabilities.

LO 6. Define a probability distribution as a list of the possible outcomes with corresponding probabilities that satisfies three rules:

- The outcomes listed must be disjoint.
- Each probability must be between 0 and 1.
- The probabilities must total 1.

LO 7. Define complementary outcomes as mutually exclusive outcomes of the same random process whose probabilities add up to 1.

- If A and B are complementary, P(A) + P(B) = 1

LO 8. Distinguish between union of events (A or B) and intersection of events (A and B).

- Calculate the probability of union of events using the (general) addition rule:

If A and B are not mutually exclusive, P(A or B) = P(A) + P(B) − P(A and B)

If A and B are mutually exclusive, P (A or B) = P (A) + P (B), since for mutually exclusive events P(A and B) = 0

- Calculate the probability of intersection of independent events using the multiplication rule:

  If A and B are independent, P(A and B) = P(A) × P(B)

  If A and B are dependent, P(A and B) = P(A|B) × P(B)

Test yourself:

1. What is the probability of getting a head on the 6th coin flip if in the first 5 flips the coin landed on a head each time?

2. True / False: Being right handed and having blue eyes are mutually exclusive events.

3. P(A) = 0.5, P(B) = 0.6, and there are no other possible outcomes in the sample space. What is P(A and B)?

Conditional Probability

LO 1. Distinguish between marginal and conditional probabilities.

LO 2. Construct tree diagrams to calculate conditional probabilities and probabilities of intersection of non-independent events using Bayes' theorem: $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$

Test yourself: *50% of students in a class are social science majors and the rest are not. 70% of the social science students and 40% of the non-social science students are in a relationship. Create a contingency table and a tree diagram summarizing these probabilities. Calculate the percentage of students in this class who are in a relationship.*

The Normal Distribution

LO 1. Define the standardized (Z) score of a data point as the number of standard deviations it is away from the mean: $Z=(x-\mu)/\sigma$.

LO 2. Use the Z score

- if the distribution is normal: to determine the percentile score of a data point (using technology or normal probability tables)
- regardless of the shape of the distribution: to assess whether or not the particular observation is considered to be unusual (more than 2 standard deviations away from the mean)

LO 3. Depending on the shape of the distribution determine whether the median would have a negative, positive, or 0 Z score keeping in mind that the mean always has a Z score of 0.

LO 4. Assess whether or not a distribution is nearly normal using the 68-95-99.7% rule or graphical methods such as a normal probability plot.

Test yourself: *True/False: In a right skewed distribution the Z score of the median is positive.*

Binomial Distribution

LO 1. Determine if a random variable is binomial using the four conditions.

- The trials are independent.
- The number of trials, n, is fixed.
- Each trial outcome can be classified as a success or failure.
- The probability of a success, p, is the same for each trial.

LO 2. Calculate the number of possible scenarios for obtaining k successes in n trials using the choose function: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

LO 3. Calculate probability of a given number of successes in a given number of trials using the binomial distribution: $P(k = K) = \binom{n}{k} p^k (1-p)^{(n-k)}$.

LO 4. Calculate the expected number of successes in a given number of binomial trials ($\mu = np$) and its standard deviation ($\sigma = \sqrt{np(1-p)}$).

LO 5. When number of trials is sufficiently large (np ≥ 10 and n(1−p) ≥ 10), use the normal approximation to calculate binomial probabilities, and explain why this approach works.

Test yourself:

1. *True/False: We can use the binomial distribution to determine the probability that in 10 rolls of a die the first 6 occurs on the 8th roll.*
2. *True / False: If a family has 3 kids, there are 8 possible combinations of gender order.*
3. *True/ False: When n = 100 and p = 0.92 we can use the normal approximation to the binomial to calculate the probability of 90 or more successes.*