Course 1 - Introduction to Probability and Data

Week 1

LO 1. Identify variables as numerical and categorical.

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.

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.

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.

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.

Week 2

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.

LO 8. Use frequency tables and bar plots to describe the distribution of one categorical variable.

LO 9. Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

LO 10. Use side-by-side box plots for assessing the relationship between a numerical and a categorical variable.

Week 3

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.

LO 4. Distinguish between disjoint and independent events.

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.

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

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

LO 9. Distinguish between marginal and conditional probabilities.

LO 10. Construct tree diagrams to calculate conditional probabilities and probabilities of intersection of non-independent events using Bayes' theorem.

Week 4

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

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.

LO 5. Determine if a random variable is binomial using the four conditions.

LO 6. Calculate the number of possible scenarios for obtaining k successes in n trials using the choose function.

LO 7. Calculate probability of a given number of successes in a given number of trials using the binomial distribution.

LO 8. Calculate the expected number of successes in a given number of binomial trials and its standard deviation.

LO 9. When number of trials is sufficiently large ($np \geq 10$ and $n(1- p) \geq 10$), use the normal approximation to calculate binomial probabilities, and explain why this approach works.