Chapter highlights

Chapter 1 focused on various ways that researchers collect data. The key concepts are the difference between a sample and an experiment and the role that randomization plays in each.

- Researchers take a **random sample** in order to draw an **inference** to the larger population from which they sampled. When examining observational data, even if the individuals were randomly sampled, a correlation does not imply a causal link.
- In an **experiment**, researchers impose a treatment and use **random assignment** in order to draw **causal conclusions** about the effects of the treatment. While often implied, inferences to a larger population may not be valid if the subjects were not also *randomly sampled* from that population.

Related to this are some important distinctions regarding terminology. The terms stratifying and blocking cannot be used interchangeably. Likewise, taking a simple random sample is different than randomly assigning individuals to treatment groups.

- **Stratifying vs Blocking**. Stratifying is used when sampling, where the purpose is to *sample* a subgroup from each stratum in order to arrive at a better *estimate* for the parameter of interest. Blocking is used in an experiment to *separate* subjects into blocks and then *compare* responses within those blocks. All subjects in a block are used in the experiment, not just a sample of them.
- **Random sampling vs Random assignment**. Random sampling refers to sampling a subset of a population for the purpose of inference to that population. Random assignment is used in an experiment to separate subjects into groups for the purpose of comparison between those groups.

When randomization is not employed, as in an **observational study**, neither inferences nor causal conclusions can be drawn. Always be mindful of possible **confounding factors** when interpreting the results of observation studies.

Chapter highlights

A raw data matrix/table may have thousands of rows. The data need to be summarized in order to makes sense of all the information. In this chapter, we looked at ways to summarize data **graphically**, **numerically**, and **verbally**.

Categorical data

- A single **categorical variable** is summarized with **counts** or **proportions** in a **one-way table**. A **bar chart** is used to show the frequency or relative frequency of the categories that the variable takes on.
- Two categorical variables can be summarized in a **two-way table** and with a **side-by-side bar chart** or a **segmented bar chart**.

Numerical data

- When looking at a single **numerical variable**, we try to understand the **distribution** of the variable. The distribution of a variable can be represented with a frequency table and with a graph, such as a **stem-and-leaf plot** or **dot plot** for small data sets, or a **histogram** for larger data sets. If only a summary is desired, a **box plot** may be used.
- The **distribution** of a variable can be described and summarized with **center** (mean or median), **spread** (SD or IQR), and **shape** (right skewed, left skewed, approximately symmetric).
- **Z-scores** and **percentiles** are useful for identifying a data point's relative position within a data set.
- **Outliers** are values that appear extreme relative to the rest of the data. Investigating outliers can provide insight into properties of the data or may reveal data collection/entry errors.
- When **comparing the distribution** of two variables, use two dot plots, two histograms, a back-to-back stem-and-leaf, or parallel box plots.
- To look at the **association** between two numerical variables, use a **scatterplot**.

Graphs and numbers can summarize data, but they alone are insufficient. It is the role of the researcher or data scientist to ask questions, to use these tools to identify patterns and departure from patterns, and to make sense of this in the context of the data. Strong writing skills are critical for being able to communicate the results to a wider audience.

Chapter highlights

This chapter focused on understanding likelihood and chance variation, first by solving individual probability questions and then by investigating probability distributions.

The main probability techniques covered in this chapter are as follows:

- The **General Multiplication Rule** for **and** probabilities (intersection), along with the special case when events are **independent**.
- The **General Addition Rule** for **or** probabilities (union), along with the special case when events are **mutually exclusive**.
- The **Conditional Probability Rule**.
- Tree diagrams and **Bayes' Theorem** to solve more complex conditional problems.
- The **Binomial Formula** for finding the probability of exactly x successes in n independent trials.
- **Simulations** and the use of random digits to estimate probabilities.

Fundamental to all of these problems is understanding when events are independent and when they are mutually exclusive. Two events are **independent** when the outcome of one does not affect the outcome of the other, i.e. $P(A|B) = P(A)$. Two events are **mutually exclusive** when they cannot both happen together, i.e. $P(A \text{ and } B) = 0$.

Moving from solving individual probability questions to studying probability distributions helps us better understand chance processes and quantify expected chance variation.

- For a **discrete probability distribution**, the **sum** of the probabilities must equal 1. For a **continuous probability distribution**, the **area under the curve** represents a probability and the total area under the curve must equal 1.
- As with any distribution, one can calculate the mean and standard deviation of a probability distribution. In the context of a probability distribution, the **mean** and **standard deviation** describe the average and the typical deviation from the average, respectively, after many, many repetitions of the chance process.
- A probability distribution can be summarized by its **center** (mean, median), **spread** (SD, IQR), and **shape** (right skewed, left skewed, approximately symmetric).
- Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.
- The mean of the sum of two random variables equals the sum of the means. However, this is not true for standard deviations. Instead, when finding the standard deviation of a sum or difference of random variables, take the square root of the sum of each of the standard deviations squared.

The study of probability is useful for measuring uncertainty and assessing risk. In addition, probability serves as the foundation for inference, providing a framework for evaluating when an outcome falls outside of the range of what would be expected by chance alone.

Chapter highlights

This chapter began by introducing the normal distribution. A common thread that ran through this chapter is the use of the **normal approximation** in various contexts. The key steps are included for each of the normal approximation scenarios below.

1. Normal approximation for **data**:
   - Verify that population is approximately normal.
   - Use the given mean μ and SD σ to find the Z-score for the given x value.

2. Normal approximation for a **sample mean/sum**:
   Verify that population is approximately normal or that $n \geq 30$.
   Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to find the Z-score for the given/calculated sample mean.

3. Normal approximation for the **number of successes** (binomial):
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$ to find the Z-score for the given number of successes.

4. Normal approximation for a **sample proportion**:
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ to find the Z-score for the given sample proportion.

5. Normal approximation for the **sum of two independent random variables**:
   - Verify that each random variable is approximately normal.
   - Use $E(X+Y) = E(X) + E(Y)$ and $SD(X+Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$ to find the Z-score for the given sum.

Cases 1 and 2 apply to **numerical** variables, while cases 3 and 4 are for **categorical** yes/no variables. Case 5 applies to both numerical and categorical variables.

Note that in the binomial case and in the case of proportions, we never look to see if the *population* is normal. That would not make sense because the "population" is simply a bunch of no/yes, 0/1 values and could not possibly be normal.

The **Central Limit Theorem** is the mathematical rule that ensures that when the sample size is sufficiently large, the sample mean/sum and sample proportion/count will be approximately normal.