Chapter 3 Probability, Sampling, and Confidence Intervals

3.1 Probability

Principles of probability are essential to statistics. It is through probability that we understand how likely events are, which then allows us to make data-driven decisions. This course and textbook aren’t sufficient to gain an in-depth understanding of probability, but a few of the basics will be covered.

3.1.1 Probability Notation

Source: MATH 221 Textbook

You may already have a good understanding of the basics of probability. It is worth noting that there is a special notation used to denote probabilities. The probability that an event, x, will occur is written $P(x)$. As an example, the probability that you will roll a 6 on a die can be written as

P (Roll a 6 on a die)= $\frac{1}{6}$

3.1.2 Rules of Probability

Source: MATH 221 Textbook

Probabilities follow patterns, called **probability distributions,** or distributions, for short. There are three rules that a probability distribution must follow.

The three rules of probability are:

- **Rule 1**: The probability of an event X is a number between 0 and 1.

$$0 \leq P(X) \leq 1$$

- **Rule 2**: If you list all the outcomes of an experiment (such as rolling a die) the probability that one of these outcomes will occur is 1. In other words, the sum of the probabilities of all the possible outcomes of any experiment is 1.

$$\sum P(X) = 1$$

- **Rule 3**: (Complement Rule) The probability that an event X will not occur is 1 minus the probability that it will occur.

$$P(\text{not } X) = 1 - P(X)$$

You may have noticed that the Complement Rule is just a combination of the first two rules.

3.2 Sampling from a Population

Very rarely do we have access to an entire population for one reason or another (too large, not enough resources, etc.), so we are left with taking samples from the population that should be representative of that population. If we had access to the entire population then we wouldn’t need to do statistical tests or analysis, we would just make observations on the whole population. The goal of statistical analysis is to determine if what we see in a sample is likely to occur in the population. For example, if we observe a common trend among 100 BYU-Idaho can we assume that that trend will hold for all BYU-Idaho students? Or as a made-up larger scale example, assume that 1000 Toyota Camrys are tested and 1% of them are found to have a defect in the braking system. Can Toyota assume that 1% of all Toyta Camrys will have the same defect? Through statistical analysis we are able to obtain answers to these questions.

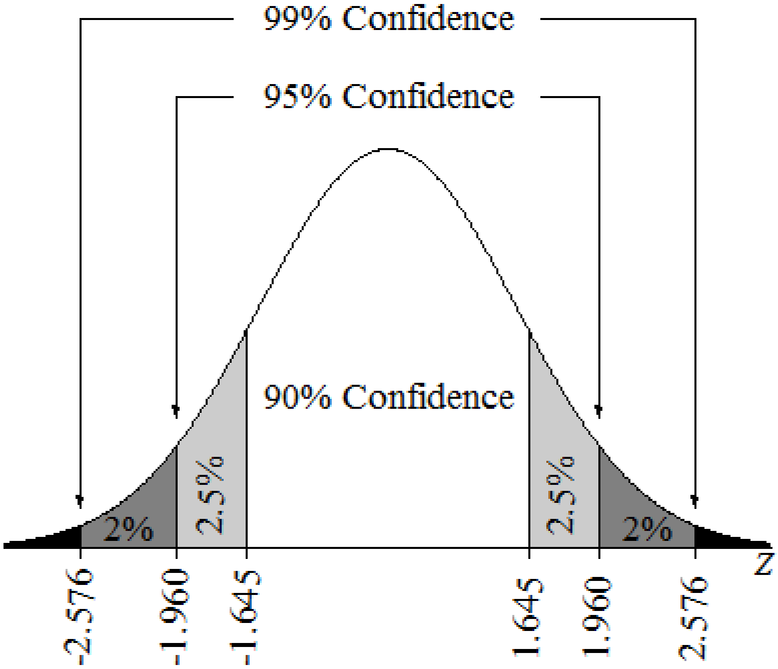
Hopefully this helps you see the importance of sampling. Even if we aren’t able to observe every member of a population, through proper sampling and statistical analysis we are able to gain insight into the population as a whole. For those insights to be valid, however, the sampling must be done in a statistically correct way. Not just any sample can be taken, so methods for sampling have been developed. Some of the more common methods are shown below.

Source: MATH 221 Textbook

- There are many sampling methods used to obtain a **sample** from a **population**:
    - A **simple random sample (SRS)** is a random selection taken from a population

- A **systematic sample** is every ***k***th item in the population, beginning at a random starting point
- A **cluster sample** is all items in one or more randomly selected clusters, or blocks
- A **stratified sample** divides data into similar groups and an **SRS** is taken from each group
- A **convenience sample** is one easily obtained in a less-than-systematic way and should be avoided whenever possible

3.3 Confidence Intervals



90%, 95%, and 99% Confidence Intervals

In statistics, we usually don’t know the exact values of the population parameters so we use statistical methods to approximate them. For example, we might take a random sample from the population, calculate the sample mean, and use that value as an estimate for the population mean. The sample mean is an example of a **point estimator** because it is a single value, or a point. There are also **interval estimators**, which instead offer a range, or interval, of values which are likely to contain the value we are trying to estimate. Arguably the most common interval estimator is the **confidence interval**.

3.3.1 Definition and Interpretation

First, a confidence interval is always associated with some percentage, or a probability. For example, a 95% confidence interval. Second, we find confidence intervals for values. Perhaps the most common confidence interval is a **95% confidence interval for the mean**. The correct interpretation of this 95% confidence interval would be: “We are 95% confident that the true mean lies within the lower and upper bounds of the confidence interval.”

Notice that with this interpretation we aren’t saying anything about the exact value of the population mean, we are simply giving a range of values that the mean is very likely to lie in.

Example

Source: MATH 221 Textbook

Consider the 95% confidence interval for the true mean of 25 rolls of a fair die. We find the 95% confidence interval to be: (2.37,3.71). When we interpret this confidence interval, we say, “We are 95% confident that the true mean is between 2.37 and 3.71.”

The word, “confident” implies that if we repeated this process many, many times, 95% of the confidence intervals we would get would contain the true mean μ. It does not imply anything about whether or not one specific confidence interval will contain the true mean.

We do not say that “there is a 95% probability (or chance) that the true mean is between 2.37 and 3.71.” The probability that the true mean μ is between 2.37 and 3.71 is either 1 or 0.

3.3.2 Finding a Confidence Interval

More often than not, confidence intervals will be 95% confidence intervals. Think back to the **68-95-99.7% Rule for Bell-Curves** from last chapter, especially note the 95. Assuming the data is approximately normally distributed, then approximately 95% of the data lies within two standard deviations of the mean, so by computing the values that are two standard deviations away from the mean on either side we compute the 95% confidence interval; with the upper bound of the CI being the mean plus two standard deviations ($\mu + 2\sigma$), and the lower bound being the mean minus two standard deviations ($\mu - 2\sigma$).

Another way to think of this is to say that if the data is approximately normally distributed then the true mean will lie within the 95% confidence interval approximately 95% of the time, or within two standard deviations of the sample mean 95% of the time.

For a step-by-step example of finding a confidence interval and a real-world example, see How to Determine the Confidence Interval for a Population Proportion.