Final: Stats 212

Pranav Tikkawar

August 15, 2024

Problem 1

Q: Recall the focus of methods such as z-tests and t-tests all focus on analyzing the population mean, μ, of some variable of interest in a population. Explain why many statistical methods are framed in terms of the population mean, μ. Discuss not only the practical reasons but also any theoretical reasons as well.

A: Many statistical methods are framed in terms of the population mean as it represents a “typical” value of a population. Primarily, the mean is a measure of central tendency: a single measure that “summarizes” the entire population. Since μ is based on all the potential candidates in a population, it does not exclude any individual in the determination of a typical individual in a population. Not only is estimating the population mean through the sample average a simple formula, but the Law of Large numbers can help determine a straightforward way to utilize sample averages to estimate a population mean.
Theoretically, the mean also provides many benefits when being a variable of interest. The mean is an unbiased estimator as it follows the property: $\mathbb{E}\left[\bar{X}\right] = \mu$. This result is true whether our data is an SRS from the population of interest or is modeled as an IID. This is super important as there are no conditions for this property to hold. The fact that $\bar{X}$ is unbiased for estimating μ is one of the main reasons why μ is chosen as the default measure of central tendency. However, unbiasedness is a property of an estimator that pertains to how it will ”tend” to behave across different samples; it does not guarantee that a value of the estimator for a particular sample will indeed be the parameter of interest.

Problem 2

Q: State the Central Limit Theorem and why it is important. Be sure to list all the assumptions that are needed for the result of the Central Limit Theorem.

A: The Central Limit Theorem states that if you have an independent and identically distributed sample of random variables $X_1, X_2, \ldots, X_n$ with mean

$\mu < \infty$ and standard deviation $\sigma < \infty$, then it follows that

$$\bar{X} \xrightarrow{d} \text{Norm}(\mu, \frac{\sigma^2}{n}) \text{ as } n \to \infty$$

The notation $\xrightarrow{d}$ denotes that provided the conditions above hold, it follows as the sample size grows, the distribution of $\bar{X}$ approaches that of $Norm(\mu, \frac{\sigma^2}{n})$. The implications of this are tremendous! For large enough n (usually about $n > 30$), the distribution of $\bar{X}$ will be approximately normal, regardless of the distribution of the original population. This is important as it allows us to make inferences about the population mean μ using the normal distribution. This is especially useful when we are interested in making inferences about the population mean μ through Confidence Intervals and Hypothesis Testing using the sample mean $\bar{X}$.

Problem 3

Q: Explain what the following terminologies means in the context of a hypothesis testing problem: Null Hypothesis, Alternative Hypothesis, Type I error, Type II error, α and β. It may help to make up an example to help you explain these concepts.

A:
Senario: Suppose a company claims that the average time it takes to complete a task is 10 minutes. You are skeptical of this claim and want to test it. Using this example I will go through each terminology and give an example.

Null Hypothesis (H_0): The null hypothesis is a statement that is assumed to be true unless there is sufficient evidence to the contrary. It is the hypothesis that is tested in a hypothesis test. It is the status quo, the default assumption.
Example: The null hypothesis would be that the average time it takes to complete a task is 10 minutes.

Alternative Hypothesis (H_1): The alternative hypothesis is the hypothesis that is accepted if the null hypothesis is rejected. It is the hypothesis that the researcher is trying to provide evidence for.
Example: The alternative hypothesis would be that the average time it takes to complete a task is not 10 minutes.

Type I error: A Type I error occurs when the null hypothesis is rejected when it is actually true. This is also known as a false positive. The probability of making a Type I error is denoted by α.
Example: A Type I error would occur if we reject the null hypothesis that the average time it takes to complete a task is 10 minutes when it is actually true.

Type II error: A Type II error occurs when the null hypothesis is not rejected when it is actually false. This is also known as a false negative. The probability of making a Type II error is denoted by β.
Example: A Type II error would occur if we do not reject the null hypothesis that the average time it takes to complete a task is 10 minutes when it is actually false.

α: The significance level, α, is the probability of making a Type I error. It is the probability of rejecting the null hypothesis when it is actually true.

β: The probability of making a Type II error. It is the probability of not rejecting the null hypothesis when it is actually false.

Problem 4

Q: Below are QQ Plots for two different data sets. Which of the datasets appears to be normally distributed, Dataset I or Dataset II? Explain your answer.

A: The QQ plot for Dataset I appears to be normally distributed. This is because the points on the QQ plot for Dataset I are very close to the line. This indicates that the data points are very close to the theoretical quantiles. On the other hand, the QQ plot for Dataset II is not normally distributed. This is because the points on the QQ plot for Dataset II create an S-shape. The data is similar to a fat-tailed distribution due to that fact that the quantiles of the empirical distribution are farther from the median than those of the normal.