MIDTERM REVIEW

Brian Cervantes Alvarez
November 6, 2023

Overview

1. We will delve into terms like the population of interest (the group we're studying) and the variable of interest (the attribute we're analyzing). We will also cover topics like population distribution, parameters, statistics, sampling distributions, population inference, and key principles of hypothesis testing, including null and alternative hypotheses, test statistics, Type I and Type II errors, significance levels, power, and the importance of consistency in test procedures, p-values, and confidence intervals.

2. We should understand the practical aspects of conducting various test procedures, including z-tests, t-tests, exact binomial tests, normal approximation binomial tests (z-tests), sign tests, signed-rank tests, chi-squared tests for variance, and the Kolmogorov-Smirnov test for distribution. We will gain a clear understanding of these tests, along with the calculation of confidence intervals and p-values.

3. We should be able to provide insights into answering critical questions related to p-value distributions under the null hypothesis for exact tests, the necessary assumptions for t-tests, normal approximation to binomial tests, and interpreting signed-rank tests as assessments of population mean/median. We will also examine the intricate relationships between p-values and hypothesis test decisions, as well as the connections between confidence intervals and hypothesis test outcomes. Moreover, our review page will clarify the distinctions between 95% and 99% confidence intervals (the latter being wider and encompassing the former), and it will highlight the consequences of incorrect variance usage in z-tests, chi-squared tests for variance on non-normally distributed data, and the implications of employing estimated parameter values in the Kolmogorov-Smirnov test.

Definitions

1. **Population of interest**:

   - The population of interest is the entire group or set of individuals or items that you want to study or draw conclusions about.
   - The population of interest, denoted as N, is the set of all registered voters in a particular country.

2. **Variable of interest**:

   - A variable of interest is a specific characteristic or attribute that you want to study within the population.
   - Let X be the variable of interest representing the age of the registered voters.

3. **Population distribution**:

   - The population distribution describes the way in which a variable of interest is distributed across the entire population.
   - The population distribution of the variable X follows a normal distribution with a mean μ and standard deviation σ.

4. **Parameter**:

   - A parameter is a numerical summary of a population distribution, often denoted by Greek letters.
   - The population mean is represented by μ and the population standard deviation is represented by σ.

5. **Statistic**:

   - A statistic is a numerical summary of a sample from the population, typically represented by Roman letters.
   - The sample mean is denoted as $\bar{x}$, and the sample standard deviation is denoted as s.

6. **Sampling distribution of a statistic**:

   - The sampling distribution of a statistic describes the distribution of that statistic across all possible samples of a given size.
   - The sampling distribution of the sample mean $\bar{x}$ for a simple random sample follows a normal distribution with a mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

7. **Population inference**:

   - Population inference is the process of making conclusions or inferences about the population based on information obtained from a sample.
   - We use sample data to make inferences about the population parameters, such as estimating the population mean.

8. **Representative sample, simple random sample**:

   - A representative sample is one that accurately reflects the characteristics of the population. A simple random sample is a sample selected in such a way that each individual in the population has an equal chance of being chosen.
   - A simple random sample is obtained when each registered voter has an equal probability of being selected, ensuring that it is a representative sample.

9. **Randomized experiment vs. observational study**:

   - A randomized experiment is a study where researchers intentionally manipulate an independent variable to observe its effect on a dependent variable. In contrast, an observational study observes and measures variables without intervention.

- In a randomized experiment, participants are randomly assigned to either a control group or a treatment group to assess the impact of a new drug. In an observational study, researchers observe the dietary habits of individuals and their cholesterol levels without any intervention.

10. **Causal inference**:
    - Causal inference is the process of drawing conclusions about cause-and-effect relationships between variables.
    - To make a causal inference, one must establish a causal relationship between the independent variable (e.g., a drug) and the dependent variable (e.g., health outcomes) by considering factors such as randomization, study design, and confounding variables.
11. **Confounding**:
    - Confounding occurs when an extraneous variable influences both the independent and dependent variables, leading to a false or misleading interpretation of their relationship.
    - In a study on the effect of exercise on heart health, age can be a confounding variable if it affects both exercise habits and heart health, making it important to control for age in the analysis.
12. **Statistical vs. practical significance**:
    - Statistical significance refers to the probability of obtaining a result as extreme as, or more extreme than, the observed result under the null hypothesis. It is typically measured using a significance level (e.g., α).
    - Practical significance refers to whether the observed effect or result is meaningful or relevant in a real-world context.
13. **Empirical distribution function**:
    - The empirical distribution function (EDF) is a non-parametric estimate of the cumulative distribution function (CDF) based on observed data. It assigns a probability to each data point, showing the proportion of data points less than or equal to a given value.
14. **Weak Law of Large Numbers**:
    - The Weak Law of Large Numbers states that as the sample size increases, the sample mean converges in probability to the population mean. In other words, as you take larger samples, the sample mean becomes a better estimate of the population mean.
15. **Central Limit Theorem**:
    - The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean (or other sample statistics) approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution. This is a fundamental concept in inferential statistics.
16. **Hypothesis test: null and alternative hypotheses, test statistic, reference distribution, rejection region**:
    - In a hypothesis test, the null hypothesis (H_0) is a statement of no effect or no difference, while the alternative hypothesis (H_1 or H_a) is a statement of an effect or difference.
    - The test statistic is a value computed from sample data that is used to assess the evidence against the null hypothesis.
    - The reference distribution is the probability distribution used to determine the expected behavior of the test statistic under the null hypothesis.
    - The rejection region is a critical region in the reference distribution where, if the test statistic falls into this region, the null hypothesis is rejected.

17. **Type I and type II error**:

    - A Type I error occurs when the null hypothesis is rejected when it is actually true, indicating a false positive.
    - A Type II error occurs when the null hypothesis is not rejected when it is actually false, indicating a false negative.

18. **Significance level and power**:

    - Significance level (α) is the probability of making a Type I error, which is set in advance as the threshold for rejecting the null hypothesis.
    - Power ($1 - \beta$) is the probability of correctly rejecting a false null hypothesis (i.e., correctly detecting an effect or difference).

19. **Exactness and consistency of a test procedure**:

    - Exactness of a test procedure means that it controls the Type I error rate precisely at the specified significance level (α).
    - Consistency means that as the sample size increases, the probability of making a Type II error approaches zero.

20. **p-value**:

    - The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the one computed from the sample data, assuming that the null hypothesis is true. A smaller p-value provides stronger evidence against the null hypothesis.

21. **Confidence interval**:

    - A confidence interval is a range of values that is constructed from sample data and provides a plausible range for the population parameter. It is often used for estimating the true value of a parameter with a certain level of confidence (e.g., 95% confidence interval).

Test Procedures

1. **z-test/confidence interval/p-value**:

   - A z-test is used when you have a large sample size (typically $n \geq 30$) and you know the population standard deviation. It assesses whether a sample mean is significantly different from a known population mean.
   - A confidence interval for the mean can be constructed to estimate the range within which the true population mean is likely to fall.
   - A p-value is calculated to determine the significance of the results. A small p-value (e.g., < 0.05) suggests that the sample mean is significantly different from the population mean.
   - Example: Conducting a z-test to determine if the average weight of apples in a shipment of 100 apples is different from the known population mean weight of apples.

2. **t-test/confidence interval/p-value**:

   - A t-test is used when you have a small sample size (typically $n < 30$) and you don't know the population standard deviation. It assesses whether a sample mean is significantly different from a hypothesized population mean.
   - A confidence interval for the mean can be constructed to estimate the range within which the true population mean is likely to fall.
   - A p-value is calculated to determine the significance of the results. A small p-value (e.g., < 0.05) suggests that the sample mean is significantly different from the hypothesized population mean.
   - Example: Conducting a t-test to determine if the average test scores of a group of students differ from the expected mean score.

3. **Exact binomial test/p-value**:

   - The exact binomial test is used to assess whether an observed binomial proportion (e.g., success rate) differs from a hypothesized proportion.
   - A p-value is calculated to determine the significance of the results. A small p-value suggests that the observed proportion is significantly different from the hypothesized proportion.
   - Example: Conducting an exact binomial test to determine if the success rate of a manufacturing process is significantly different from a target success rate.

4. **Normal approximation binomial test (z-test)/confidence interval/p-value**:

   - In cases where the sample size is sufficiently large, you can use a normal approximation to conduct a binomial test.
   - A confidence interval can be constructed to estimate the range within which the true binomial proportion is likely to fall.
   - A p-value is calculated to determine the significance of the results. A small p-value suggests that the observed proportion is significantly different from the hypothesized proportion.
   - Example: Using a normal approximation binomial test to determine if the click-through rate of an online ad campaign differs from the expected rate.

5. **Sign test/confidence interval/p-value**:

   - The sign test is a non-parametric test used to determine if there is a significant difference between two related measurements, especially when the data is not normally distributed.
   - A confidence interval is not typically used in the sign test.
   - A p-value is calculated to determine the significance of the results. A small p-value suggests a significant difference between the measurements.

- Example: Using a sign test to assess whether there is a significant change in patient blood pressure before and after a treatment.

6. **Signed-rank test/p-value**:

   - The signed-rank test, also known as the Wilcoxon signed-rank test, is used to assess whether there is a significant difference between two related measurements.
   - A confidence interval is not typically used in the signed-rank test.
   - A p-value is calculated to determine the significance of the results. A small p-value suggests a significant difference between the measurements.
   - Example: Conducting a signed-rank test to determine if there is a significant change in the reaction times of individuals before and after a training program.

7. **Chi-squared test for variance/confidence interval/p-value**:

   - The chi-squared test for variance is used to assess whether the variance of a sample is significantly different from a hypothesized population variance.
   - A confidence interval for the variance can be constructed to estimate the likely range of the true population variance.
   - A p-value is calculated to determine the significance of the results. A small p-value suggests a significant difference in variance.
   - Example: Using a chi-squared test for variance to determine if the variance in test scores of two different schools is significantly different.

8. **Alternative t-test for variance**:

   - An alternative t-test is used to assess whether two samples have significantly different variances.
   - This test does not typically involve confidence intervals.
   - Example: Conducting an alternative t-test to determine if the variance in the heights of male and female participants is significantly different.

9. **Kolmogorov-Smirnov test for distribution**:

   - The Kolmogorov-Smirnov test assesses whether a sample comes from a specific distribution (e.g., normal, uniform).
   - A p-value is calculated to determine the significance of the results. A small p-value suggests that the sample distribution is significantly different from the hypothesized distribution.
   - Example: Using the Kolmogorov-Smirnov test to determine if a dataset of exam scores follows a normal distribution.

Important Concepts

1. **What is the distribution of p-values under the null hypothesis for an exact test?**
    - Under the null hypothesis, the distribution of p-values for an exact test follows a uniform distribution between 0 and 1. In other words, if the null hypothesis is true, p-values are equally likely to take any value between 0 and 1.
2. **What assumptions are necessary when performing the t-test?**
    - The data follows a roughly normal distribution (approximately symmetric).
    - The samples are independent.
    - Homogeneity of variances (variances are roughly equal for the compared groups).
3. **What assumptions are necessary when performing the normal approximation to the binomial test?**
    - The sample size is sufficiently large (typically $n \geq 30$).
    - The success-failure condition is met, ensuring that np and n(1-p) are both greater than 5, where p is the probability of success.
4. **What assumptions are necessary when interpreting the signed-rank test as a test of population mean/median?**
    - The assumption is that the distribution of differences between paired observations is symmetric and continuous.
5. **What is the relationship between p-values and hypothesis test decisions?**
    - A smaller p-value indicates stronger evidence against the null hypothesis. Typically, if the p-value is less than the chosen significance level (e.g., 0.05), the null hypothesis is rejected in favor of the alternative hypothesis. Conversely, a larger p-value suggests weaker evidence against the null hypothesis.
6. **What is the relationship between confidence intervals and hypothesis test decisions?**
    - A confidence interval provides a range of plausible values for a parameter (e.g., mean). If a null hypothesis value falls outside the confidence interval, it suggests that the null hypothesis may be inconsistent with the data, which corresponds to rejecting the null hypothesis in a hypothesis test.
7. **How is a 95% confidence interval related to a 99% confidence interval (wider, narrower, subset, superset)?**
    - A 95% confidence interval is narrower than a 99% confidence interval. In general, as the confidence level increases (e.g., from 95% to 99%), the width of the interval increases. A 99% confidence interval is a superset of a 95% confidence interval, meaning it contains a larger range of values.
8. **What happens if you use an incorrect variance in performing a z-test (too large? too small?)?**
    - If you use an incorrect, too large variance in a z-test, it can lead to underestimating the significance of the results, potentially resulting in a Type II error (failing to detect a true effect).
    - If you use an incorrect, too small variance in a z-test, it can lead to overestimating the significance of the results, potentially resulting in a Type I error (false positive).

9. **What happens if you perform the chi-squared test for variance on non-normal data?**

    - The chi-squared test for variance assumes that the data follows a normal distribution. If you apply this test to non-normal data, the results may be invalid and not reflect the true variance of the population.

10. **What happens if you perform the Kolmogorov-Smirnov test using estimated parameter values (rather than pre-specified parameter values)?**

    - The Kolmogorov-Smirnov test is used to compare a sample distribution to a known theoretical distribution. If estimated parameter values are used instead of pre-specified values, the test may still provide useful information, but the results may not be as accurate, and the test's power to detect differences could be reduced.