STATISTICS AND PROBABILITY THEORY ASSIGNMENT

Rajinder Singh

Q 1. Explain the difference between descriptive and inferential statistics. Provide examples of each.

Descriptive statistics refers to a branch of statistics that involves summarizing, organizing, and presenting data meaningfully and concisely. It focuses on describing and analyzing a dataset's main features and characteristics without making any generalizations or inferences to a larger population. It also involves a graphical representation of data through charts, graphs, and tables.

Examples of descriptive statistics include:

- Measures of central tendency: mean, median, mode
- Measures of variability: range, variance, standard deviation

On the other hand, **inferential statistics** makes the use of various analytical tools to draw inferences about the population data from sample data.

Examples include:

- Hypothesis testing
- Confidence intervals
- Regression analysis
- Correlation analysis

Inferential statistics help to draw conclusions or make predictions about the population while **descriptive statistics** summarizes the features of the data set.

Q 2. Define the Central Limit Theorem and discuss its significance in statistical inference.

Central Limit Theorem (CLT) states that the distribution of sample mean tends to normal distribution as the sample size increases.

The CLT is remarkable because it states that *distribution of sample mean ($\bar{x}$) tends to normal distribution regardless of the distribution of population from which the random sample is drawn*.

When sampling is done from a population with mean μ and standard deviation **σ**, the sampling distribution of sample mean $\bar{\boldsymbol{x}}$ will tend to normal distribution with mean **μ** and standard deviation $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ as the sample size **n** became large.

In general, a sample size of 30 or more elements is considered large enough for CLT to take effect.

There are 3 aspects of CLT

- When the sample size is large enough, the sampling distribution of $\bar{\boldsymbol{x}}$ is normal.
- The expected value of $\bar{\boldsymbol{x}}$ is μ
- The standard deviation of $\bar{\boldsymbol{x}}$ is σ/√n

Significance of CLT

Sampling and Estimation: The CLT is crucial for inferential statistics because it allows us to make inferences about population parameters based on samples. It provides a solid foundation for constructing confidence intervals and performing hypothesis tests, making statistical estimates more reliable and accurate.

Simplification of Complex Distributions: In many real-world scenarios, data may not follow a normal distribution, and their mathematical behavior can be quite complex. The CLT allows us to treat the sampling distribution of the mean as approximately normal, making statistical analysis much more manageable and feasible.

Basis for Hypothesis Testing: Hypothesis tests often rely on the assumption of normality, and the CLT allows us to apply these tests even when dealing with non-normally distributed populations, provided the sample size is large enough.

Modeling and Simulation: Many modeling and simulation techniques leverage the normal distribution due to its well-known properties. The CLT enables researchers to model and simulate complex phenomena by aggregating the effects of many random variables.

---

3. Discuss the concept of sampling and its role in statistical analysis.

Sampling is the process of selecting a subset from a population which is a true representative of entire population. Sample information is used to infer or conclude characteristics about the population.

There is an important role of sampling in statistic.

- Sampling can save money and time
- For given resources, sampling can broaden the scope of the dataset
- Sampling ensure greater accuracy
- If assessing the population is impossible, sampling is the only option.

It is important that sample should be drawn randomly form the entire population under the study. This increases the likelihood that our sample will be truly representative of population of interest and minimizes the chances of errors.

A random sample, also called as probability sampling, has a characteristic that every unit of the population has the same probability of being included in the sample. There is no bias in the selection process of an item in a sample.

---

4. Explain the process of hypothesis testing and the key components involved.

A hypothesis is something that has not yet been reported to be true. In other word, hypothesis (prediction) is your best guess about what you think will happen in the investigation based on some research or an experiment you have had.

Hypothesis testing is the process of determining whether or not a given hypothesis is true. It is an important application of statistic.

The first step in hypothesis test is to formalize it by specifying the **null hypothesis**. ***A null hypothesis*** *is an assertion about the value of a population parameter that we hold as true unless we have sufficient statistical evidence to conclude otherwise.*

- Null hypothesis is denoted by "H_0"

The **alternate hypothesis** is the negation of null hypothesis.

- Alternate hypothesis is denoted by "H_1"

For example, for the null hypothesis "**$\mu = 100$**", the alternate hypothesis is "**$\mu \neq 100$**".

Errors in hypothesis testing:

In the context of statistical hypothesis testing, rejecting a true null hypothesis is known as **type I error** and accepting a false null hypothesis is known as a **type II error**.

Instances of Type I and Type II errors:

	H_0 true	H_0 false
Accept H_0	No error	Type II error
Reject H_0	Type I error	No Error

P-Value

Given a null hypothesis and sample evidence with sample size "n", the "p-value" is the probability of getting a sample evidence that is equally or more unfavorable to the null hypothesis while the null hypothesis is actually true. The "p-value" is calculated giving the null hypothesis the maximum benefit of doubt.

The significance level

The most common policy in statistical hypothesis testing is establish a significance level. It is denoted by "α".

- When p-value falls below α, reject H_0.
- The standard values for α are 10%, 5% and 1%.
- If we set α = 5%, then (1- α) = 95% is the minimum confidence level that we set in order to reject H_0.

5. Describe the T-distribution and how it differs from the normal distribution.

t-distribution is a probability distribution which is used when we are working with small sample sizes or when the population variance is unknown.

When we draw samples from a normally distributed population and we don't know the population standard deviation, the distribution of sample means for some variable x drawn from this population can be described by the formula:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

A normal distribution, also known as a Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The t-distribution and the normal distribution are both probability distributions that are used to describe the behavior of data in different situations. Here are some of the key differences between them:

Shape:

- Normal Distribution: It has a bell-shaped curve and is symmetric around the mean. The shape of the normal distribution is the same regardless of the mean or standard deviation.
- T-distribution: Similar to the normal distribution in shape but has heavier tails. The t-distribution becomes wider and more variable as the sample size decreases, which is reflected in its degrees of freedom.

Degrees of Freedom:

- Normal Distribution: It does not depend on the degrees of freedom. The normal distribution is based on populations with a known variance or large sample sizes where the sample variance is a good approximation of the population variance.
- T-distribution: The shape of the t-distribution varies with the degrees of freedom. Degrees of freedom typically correlate with the sample size (n – 1 for a single sample).

Sample Size:

- Normal Distribution: It is used when dealing with large sample sizes (typically n > 30) or when the population variance is known.
- T-distribution: It is particularly useful for small sample sizes (typically n < 30) or when the population variance is unknown.

Tail Probability:

- Normal Distribution: Less probability in the tails; it assumes that extreme values are less likely to occur.
- T-distribution: More probability in the tails; it accounts for the greater variability expected with smaller samples and unknown population standard deviation, thus providing more conservative estimates.

Usage in Hypothesis Testing:

- Normal Distribution: Used in z-tests when the standard deviation of the population is known or the sample size is large.
- T-distribution: Used in t-tests which are applied when the standard deviation of the population is unknown and the sample size is small.

Convergence:

- T-distribution: As the sample size increases (and thus the degrees of freedom), the t-distribution approaches the normal distribution. In the limit, as degrees of freedom go to infinity, the t-distribution becomes identical to the normal distribution.

Standard Error:

- Normal Distribution: The standard error is based on the population standard deviation.
- T-distribution: The standard error is based on the sample standard deviation, which includes the correction factor of the square root of the degrees of freedom (n – 1).

---

6. Calculate the mean, median, and standard deviation for the following dataset: [10, 15, 20, 25, 30].

Sol.

Mean:

Mean is equal to sum of all the observations divided by number of all the observations.

Mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Mean = (10 + 15 + 20 + 25 + 30)/n

= 20

Median:

Median is an observation (or a point between two observations) in the centre of the data set.

In the above given observations, median = 20.

Standard Deviation

The standard deviation of a set of observations is the (positive) square root of the variance of the set.

Where, variance of a set of observations is the average squared deviation of the data points from their mean.

Sample standard deviation $s = \sqrt{s^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

$$= \frac{\sqrt{(10 - 20)^2 + (15 - 20)^2 + (20 - 20)^2 + (25 - 20)^2 + (30 - 20)^2}}{5-1}$$

$$= 7.90569415$$

---

7. A researcher wants to estimate the average height of students in a university. She samples 50 students and finds the mean height to be 65 inches with a standard deviation of 3 inches. Construct a 95% confidence interval for the population mean height.

Sol.

- Sample size (n) = 50
- Students mean height i.e sample mean ($\bar{x}$) = 65 inches
- Sample standard deviation (s) = 3

As we have a large sample size (>30), we will use z-distribution to constrict 95% CI.

- For normal distribution:

$$\bar{x} \pm Z\alpha_{/2} \frac{s}{\sqrt{n}}$$

For 95%CI: $Z\alpha_{/2}$ = 1.96

Substitute these values in above equation

$$= 65 \pm 1.96*3/\sqrt{50}$$

Calculate the upper and lower limits of 95%CI:

Lower limit:
$= 65 - 1.96*3/\sqrt{50}$
$= 64.169$

Upper limit:
$= 65 + 1.96*3/\sqrt{50}$
$= 65.831$

Therefore, based on the above data, we are 95% confident that the average height of students in the university lies between 64.169 and 65.831 inches.

---

8. A manufacturer claims that the average lifespan of its light bulbs is 1000 hours. A random sample of 50 light bulbs has a mean lifespan of 980 hours with a standard deviation of 50 hours.

Test the manufacturer's claim at a significance level of 0.05 using a right-tailed hypothesis test.

Sol.

We have population mean (μ) = 1000

Define the null and alternate hypothesis:

Null hypothesis (H_0): μ = 1000

As we will use a right tailed hypothesis test; therefore, **alternate hypothesis (H_1)** is μ > 1000

- Sample size (n) = 50
- Sample mean ($\bar{x}$) = 980
- Sample standard deviation (s) = 50
- Significance level (α) = 0.05

Calculate degree of freedom (df):

- df = n-1 = 50-1 = 49

Obtain critical value of p-value from t-table at df of 49 and significance level of 0.05.

- **p-value = 1.671**

As population SD is not known, we will use t-statistic.

For t-test

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Substitute the values

$$t = \frac{980-1000}{50/\sqrt{50}}$$

t = - 2.83

t-value (- 2.83) < p-value (1.671)

As calculated t-value is less than p-value; therefore, we failed to reject the null hypothesis at significance level of 0.05.

Therefore, the average lifespan of light bulbs is 1000 hours.

---

9. A pharmaceutical company is testing a new drug for lowering blood pressure. They want to determine if the drug is effective in reducing blood pressure levels. State the null and alternative

hypotheses for this study.

Sol.

Let's assume **population mean change in blood pressure** is "μ"

Based on this assumption we can state the null and alternate hypothesis as follows:

- H_0: μ = 0 (ie, no change in population mean blood pressure)
- H_1: μ < 0 (ie, reduction in population mean blood pressure)

10. A quality control manager at a factory wants to ensure that the average weight of products coming off the production line is 500 grams. She takes a random sample of 30 products and finds the mean weight to be 495 grams with a standard deviation of 10 grams. Test the manager's claim at a significance level of 0.01 using a left-tailed hypothesis test.

Sol.

Population mean weight (μ) = 500

Sample size (n) = 30

Sample mean weight (x^-) = 495

Sample standard deviation (s) = 10

Significance level (α) = 0.01

Define the null and alternate hypothesis:

- H_0: μ = 500
- H_1: μ < 500 (as we are using left tail test)

Degree of freedom = n-1

= 30-1

= 29

p-value (at df: 20 and significance level: 0.01) = - **2.462**

As population SD is not known, we will use t-statistic.

For t-test

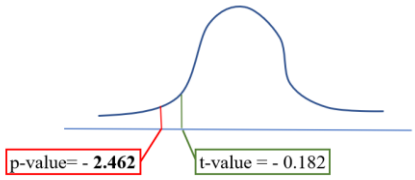
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Substitute the values

t = 495-500/(10/√30)

= -5/ 5.477

= - **0.182**



t-value (-0.182) > p-value (-2.462),

Hence, we failed to reject the null hypothesis. So, the average weight of products coming off the production line is 500 grams.