Introduction to statistics

Assignment no. 02

Rahee sutrave

Descriptive Statistics:

- gives information about raw data regarding its description or features.
- Describes situation, visible characteristics of database Help to organise, analyse, present data in meaningful manner.
- Explains already known data related to particular sample of small size, and simple to perform.
- Charts, graph, bar chart for descriptive statistics.
- It measures of Central Tendency - These help to describe the central position of the data by using measures such as mean, median, and mode
- Measures of Dispersion - These measures help to see how spread out the data is in a distribution with respect to a central point. range, standard deviation, variance, quartiles, and absolute deviation are the measures of dispersion.

Inferential statistics:

- draw inferences about the population or sample by using data extracted from the population.
- Explains probability or prediction of occurrence of event, and quite complex in performing.
- Draw inferences or conclusion about whole population.
- Probability models use for inferential statistics
- Hypothesis Testing - This technique involves the use of hypothesis tests such as the z test, f test, t test, etc. to make inferences about the population data. It requires setting up the null hypothesis, alternative hypothesis, and testing the decision criteria.
- Regression Analysis - Such a technique is used to check the relationship between dependent and independent variables. The most commonly used type of regression is linear regression.

Population:

- it is entire set of data drawn for a statistical study.
- Population can be a group of individuals, a set of items, etc. makes up the data pool for a study.
- It is data on your study of interest, use to draw conclusions.
- It can be a group of individuals, objects, events, organizations, etc.
- Example: residents of a country would constitute the Population.

Sample:

- It is smaller and more manageable representation of a larger group.
- It is subset of a larger population that contains characteristics of the population.
- sample is an unbiased subset of the population that best represents the whole data.
- It uses in statistical testing when the population size is too large.
- Example: All residents who live above the poverty line would be the Sample.

Hypothesis:

- approximate explanation that relates to the set of facts that can be tested by certain further investigations.
- It uses to estimate relation between two statistical variables.
- Hypothesis testing provides a way to verify whether the results of an experiment are valid.
- It is claimed that brisk walking for half an hour every day reverses diabetes. In order to accept this in your lifestyle, you may need evidence that supports this claim or hypothesis

Types of hypothesis

Null hypothesis :

- it is an assumption as there is no effect or no difference.
- suggests that there is no statistical significance exist between populations.
- Abbreviation use as H0
- H0: There is no difference in the mean return from A and B, or the difference between A and B is zero.
- Example: people living in slum areas has lower employability opportunities. The general opinion of public is that they have less accessibility of education, employment skills. This is null hypothesis

Alternative hypothesis:

- it is an statement in which there is some statistical significance between two measured phenomenon.
- suggests that there is some statistical significance exist between populations.
- Abbreviation use as H1.
- H1: There is some difference in the mean return from A and B, or the difference between A and B is non zero.
- Example: people living in slum areas has lower employability opportunities. The general opinion of public is that they have less accessibility of education, employment skills. Schemes like MGNREGA, SKILL INDIA, sarv shikshan abhiyan widers employment opportunities. This is alternative hypothesis.

Errors in hypothesis:
errors made due to incorrect evaluation of the outcome of hypothesis testing.

Type I error:

- When doing hypothesis testing, one ends up incorrectly rejecting the null hypothesis (default state of being) when in reality it holds true. The probability of rejecting a null hypothesis when it actually holds good is called as Type I error.
- It essentially means that unexpected outcomes or alternate hypotheses can be true.
- It should keep as minimum as possible. It known as false positive.
- Example: The claim made or the hypothesis is that the person has committed a crime or is guilty. The null hypothesis will be that the person is not guilty or innocent. Based on the evidence gathered, the null hypothesis as that the person is not guilty gets rejected. And alternate hypothesis is that the person is held guilty.

However, the rejection of null hypothesis is false. This means that the person is held guilty although he/she was not guilty. This is Type I error.

Type II error:

- When doing a hypothesis testing, when we fail to reject null hypothesis and he should actually have rejected it, this error or mistake is termed as Type II error.
- It should keep as minimum as possible. It known as false positive.
- Type II errors can turn out to be very fatal and expensive
- Type II error are also termed as False Negatives.
- Example: Person committing a crime, the null hypothesis is that the person is not guilty. In other words, the person is innocent. Rejecting the null hypothesis would mean that the person is guilty.

Failing to reject the null hypothesis is that the person is not guilty or the null hypothesis holds good.
In case, the failing to reject the null hypothesis is a mistake, this means that the person is held not guilty when he/she really was guilty.

	Null hypothesis is TRUE	Null hypothesis is FALSE
Null hypothesis rejected	TYPE I ERROR	CORRECT SOLUTION
Fail to reject null hypothesis	CORRECT SOLUTION	TYPE II ERROR

central limit theorem:

- The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of sample taken from a population.
- The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.
- A normal distribution is a symmetrical, bell-shaped distribution, with increasingly fewer observations the further from the center of the distribution. i. e. calculation of mean, median, mode.

- The central limit theorem states that the sampling distribution of the mean will always follow a normal distribution under the following conditions:

1.The sample size is sufficiently large. This condition is usually met if the sample size is $n \geq 30$.

2.The samples are independent and identically distributed (i. e.) random variables. This condition is usually met if the sampling is random.

- The population's distribution has finite variance. Central limit theorem doesn't apply to distributions with infinite variance, such as the Cauchy distribution. Most distributions have finite variance.
- *t* tests, z test, f test, ANOVAs, and linear regression, have their statistical power comes from assumptions about populations' distributions that are based on the central limit theorem.

Linear regression:

- Linear regression analysis is used to predict the value of a variable based on the value of another variable.
- The variable you want to predict is called the dependent variable.
- The variable you are using to predict the other variable's value is called the independent variable
- Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions.

Types of regression

- Simple linear regression:

Simple linear regression is used to estimate the relationship between two quantitative variables.
You can use simple linear regression when:

1. How strong the relationship is between two variables
2. The value of the dependent variable at a certain value of the independent variable.

Assumption :

1.Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn’t change significantly across the values of the independent variable.
2.Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3.The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).
4.Normality: The data follows a normal distribution

- Multiple linear regression:

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:

1.How strong the relationship is between two or more independent variables and one dependent variable.

2.The value of the dependent variable at a certain value of the independent variables.

Assumption:

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables.
3. In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.
4. Normality: The data follows a normal distribution.