STATISTICS WORKSHEET-1

1. **Bernoulli random variables take (only) the values 1 and 0.**

Answer :-A

2. **Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?**

Answer :-A

3. **Which of the following is incorrect with respect to use of Poisson distribution?**

Answer :- B

4. **Point out the correct statement**.

Answer :-D

5. **______ random variables are used to model rates.**

Answer :-C

6. **10. Usually replacing the standard error by its estimated value does change the CLT.**

Answer :-A

7. **1. Which of the following testing is concerned with making decisions using data?**

Answer :-B

8. **4. Normalized data are centered at ______and have units equal to standard deviations of the original data.**

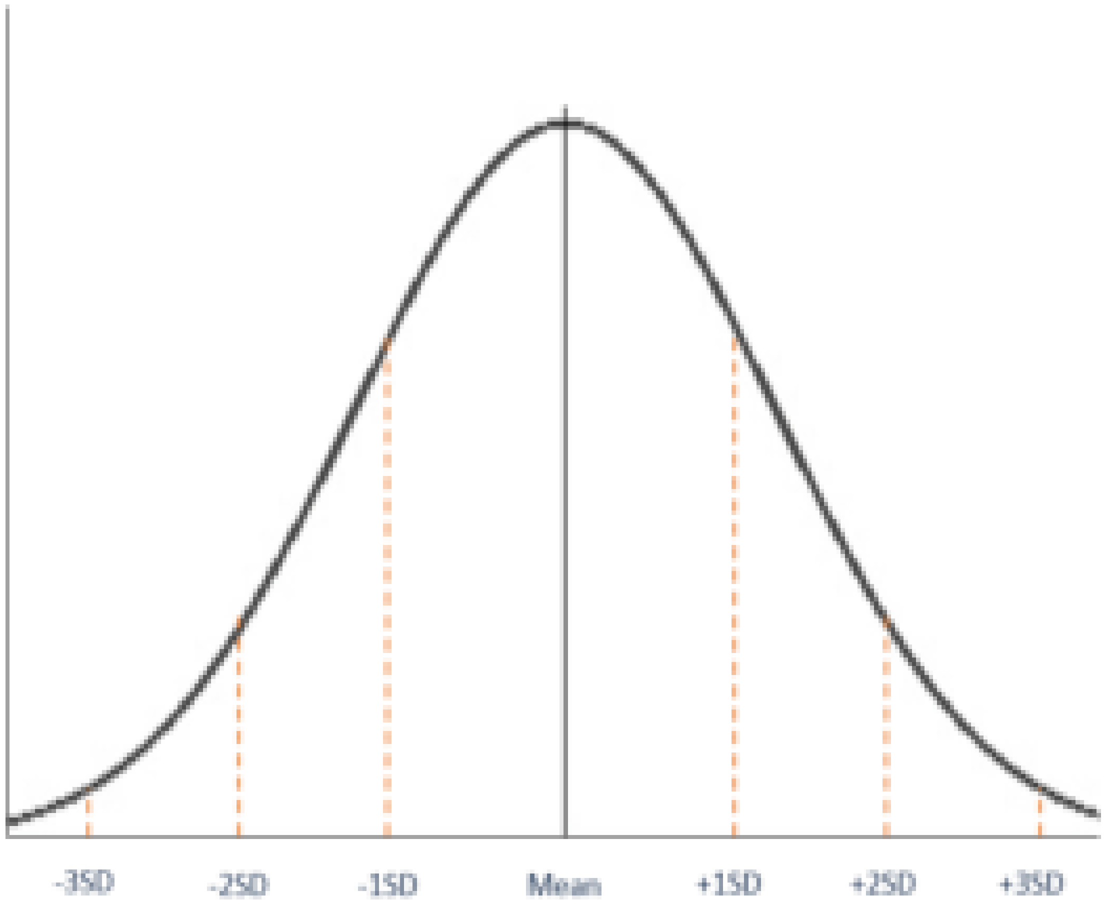
Answer :-A

9. **Which of the following statement is incorrect with respect to outliers?**

Answer :-D

10. What do you understand by the term Normal Distribution?

Answer :- The normal distribution is also referred to as Gaussian or Gauss distribution. The distribution is widely used in natural and social sciences. It is made relevant by the Central Limit Theorem, which states that the averages obtained from independent, identically distributed random variables tend to form normal distributions, regardless of the type of distributions they are sampled from.



Shape of Normal Distribution

A normal distribution is symmetric from the peak of the curve, where the mean is. This means that most of the observed data is clustered near the mean, while the data become less frequent when farther away from the mean. The resultant graph appears as bell-shaped where the mean, median, and mode are of the same values and appear at the peak of the curve.

The graph is a perfect symmetry, such that, if you fold it at the middle, you will get two equal halves since one-half of the observable data points fall on each side of the graph.

Parameters of Normal Distribution

The two main parameters of a (normal) distribution are the mean and standard deviation. The parameters determine the shape and probabilities of the distribution. The shape of the distribution changes as the parameter values change.

1. Mean

The mean is used by researchers as a measure of central tendency. It can be used to describe the distribution of variables measured as ratios or intervals. In a normal distribution graph, the mean defines the location of the peak, and most of the data points are clustered around the mean. Any changes made to the value of the mean move the curve either to the left or right along the X-axis.

2. Standard Deviation

The standard deviation measures the dispersion of the data points relative to the mean. It determines how far away from the mean the data points are positioned and represents the distance between the mean and the observations.

On the graph, the standard deviation determines the width of the curve, and it tightens or expands the width of the distribution along the x-axis. Typically, a small standard deviation relative to the mean produces a steep curve, while a large standard deviation relative to the mean produces a flatter curve.

Properties

All forms of (normal) distribution share the following characteristics:

1. It is symmetric

A normal distribution comes with a perfectly symmetrical shape. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve.

2. The mean, median, and mode are equal

The middle point of a normal distribution is the point with the maximum frequency, which means that it possesses the most observations of the variable. The midpoint is also the point where these three measures fall. The measures are usually equal in a perfectly (normal) distribution.

3. Empirical rule

In normally distributed data, there is a constant proportion of distance lying under the curve between the mean and specific number of standard deviations from the mean. For example, 68.25% of all cases fall within +/- one standard deviation from the mean. 95% of all cases fall within +/- two standard deviations from the mean, while 99% of all cases fall within +/- three standard deviations from the mean.

4. Skewness and kurtosis

Skewness and kurtosis are coefficients that measure how different a distribution is from a normal distribution. Skewness measures the symmetry of a normal distribution while kurtosis measures the thickness of the tail ends relative to the tails of a normal distribution.

History of Normal Distribution

Most statisticians give credit to French scientist Abraham de Moivre for the discovery of normal distributions. In the second edition of "The Doctrine of Chances," Moivre noted that probabilities associated with discreetly generated random variables could be approximated by measuring the area under the graph of an exponential function.

Moivre's theory was expanded by another French scientist, Pierre-Simon Laplace, in "Analytic Theory of Probability." Laplace's work introduced the central limit theorem that proved that probabilities of independent random variables converge rapidly to the areas under an exponential function.

11. **How do you handle missing data? What imputation techniques do you recommend?**

Answer :- Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

And how would you choose that estimate? The following are some of the most prevalent methods:

Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people.It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution

Assume the value from a new person who was not included in the sample.To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables.To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10.Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

Cold deck imputation

A value picked deliberately from an individual with similar values on other variables.In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

Regression imputation

The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Stochastic regression imputation

The predicted value of a regression plus a random residual value.This has all of the benefits of regression imputation plus the random component's benefits.The majority of multiple imputation is based on stochastic regression imputation.

Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time.Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.
- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

Furthermore, standard errors are underestimated by all single imputation approaches.Because the imputed observations are estimates, their values have a random error associated with them. However, your programme is unaware of this when you enter that estimate as a data point. As a result, it ignores the additional source of error, resulting in too-small standard errors and p-values.

And, while imputation is straightforward in theory, it is difficult to master in reality. As a result, it isn't perfect, although it may suffice in some circumstances.

As a result of multiple imputation, numerous estimates are generated. In multiple imputation, two of the approaches indicated above–hot deck and stochastic regression–work as the imputation method.

The multiple estimates varied significantly because these two approaches contain a random component. This reintroduces some variance that your program can account for in order to provide reliable standard error estimates for your model.

About 20 years ago, multiple imputation was a big advance in statistics. It eliminates many (but not all) difficulties with missing data and, when done correctly, leads to unbiased parameter estimations and accurate standard errors.

12. What is A/B testing?

Answer **:-** A/B testing—also called split testing or bucket testing—compares the performance of two versions of content to see which one appeals more to visitors/viewers. It tests a control (A) version against a variant (B) version to measure which one is most successful based on your key metrics. As a digital marketing practitioner doing either B2B marketing or B2C marketing, your options for conducting A/B tests include:

- Website A/B testing (copy, images, colors designs, calls to action), which splits traffic between two versions—A and B. You monitor visitor actions to identify which version yields the highest number of 1) conversions or 2) visitors who performed the desired action.
- Email marketing A/B testing (subject line, images, calls to action), which splits recipients into two segments to determine which version generates a higher open rate.
- Content selected by editors or content selected by an algorithm based on user behavior to see which one results in more engagement.

Regardless of the focus, A/B testing helps you determine how to provide the best customer experience (CX).

13. Is mean imputation of missing data acceptable practice?

Answer :- The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Answer :- Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y =

estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, "how much additional sales income do I get for each additional $1000 spent on marketing?"

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, "what will the price of gold be in 6 months?"

15. **What are the various branches of statistics?**

Answer :- **Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.**

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.