STATISTICS WORKSHEET-1

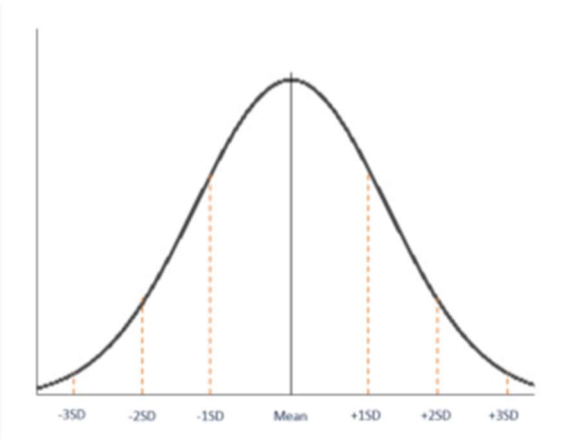
ANSWER

1. **A**
2. **A**
3. **B**
4. **D**
5. **C**
6. **B**
7. **B**
8. **A**
9. **C**

SUBJECTIVE ANSWER

ANSWER10:

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 changes.

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.

ANSWER-11

7 ways to handle missing values in the dataset:

1. Delete Rows with Missing Values:

Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.

Pros:

- A model trained with the removal of all missing values creates a robust model.

Cons:

- Loss of a lot of information.
- Works poorly if the percentage of missing values is excessive in comparison to the complete dataset

2. Impute missing values with Mean/Median:

Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values. The missing values are replaced by the mean value in the above example, in the same way, it can be replaced by the median value.

Pros:

- Prevent data loss which results in deletion of rows or columns
- Works well with a small dataset and easy to implement.

Cons:

- Works only with numerical continuous variables.
- Can cause data leakage
- Does not factor the covariance between features.

3. Imputation method for categorical columns:

When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category.

Pros:

- Prevent data loss which results in deletion of rows or columns
- Works well with a small dataset and easy to implement.
- Negates the loss of data by adding a unique category

Cons:

- Works only with categorical variables.
- Addition of new features to the model while encoding, which may result in poor performance

4. Other Imputation Methods:

Depending on the nature of the data or data type, some other imputation methods may be more appropriate to impute missing values.

For example, for the data variable having longitudinal behaviours, it might make sense to use the last valid observation to fill the missing value. This is known as the Last observation carried forward (LOCF) method.

For the time-series dataset variable, it makes sense to use the interpolation of the variable before and after a timestamp for a missing value.

5. Using Algorithms that support missing values:

All the machine learning algorithms don't support missing values but some ML algorithms are robust to missing values in the dataset. The k-NN algorithm can ignore a column from a distance measure when a value is missing. Naive Bayes can also support missing values when making a prediction. These algorithms can be used when the dataset contains null or missing values.

The sklearn implementations of naive Bayes and k-Nearest Neighbors in Python does not support the presence of the missing values.

Another algorithm that can be used here is Random Forest that works well on non-linear and the categorical data. It adapts to the data structure taking into consideration the high variance or the bias, producing better results on large datasets.

Pros:

- No need to handle missing values in each column as ML algorithms will handle it efficiently

Cons:

- No implementation of these ML algorithms in the scikit-learn library.

6. Prediction of missing values:

In the earlier methods to handle missing values, we do not use correlation advantage of the variable containing the missing value and other variables. Using the other features which don't have nulls can be used to predict missing values.

The regression or classification model can be used for the prediction of missing values depending on nature (categorical or continuous) of the feature having missing value.

Pros:

- Gives a better result than earlier methods
- Takes into account the covariance between missing value column and other columns.

Cons:

- Considered only as a proxy for the true values

7. Imputation using Deep Learning Library — Datawig

This method works very well with categorical, continuous, and non-numerical features. Datawig is a library that learns ML models using Deep Neural Networks to impute missing values in the datagram.

Datawig can take a data frame and fit an imputation model for each column with missing values, with all other columns as inputs.

Below is the code to impute missing values in the Age column

Pros:

- Quite accurate compared to other methods.
- It supports CPUs and GPUs.

Cons:

- Can be quite slow with large datasets.

I recommend best imputation techniques is ***Impute missing values with Mean/Median*.**

ANSWER-12

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

ANSWER-13

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing. ... Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

ANSWER-14

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
- 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

- Determining the strength of predictors,
- Forecasting an effect
- Trend forecasting.

ANSWER-15

Two main branch branches of statistics:

A. Descriptive Statistics
B. Inferential Statistics

A. Descriptive Statistics

Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations. Descriptive statistics can be categorized into

- Measures of central tendency
- Measures of variability

To easily understand the analyzed data, both measures of tendency and measures of variability use tables, general discussions, and graphs.

B. Inferential Statistics

Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.

The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis