STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
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?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
5. _______random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_______and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

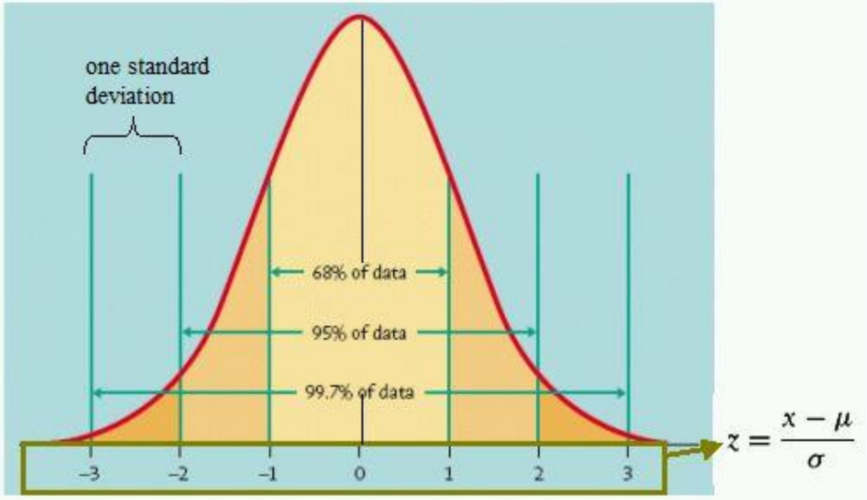
10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

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

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

Ans. A normal distribution is a type of continuous probability distribution in which most data points towards the middle of the range, while the rest taper off symmetrically towards either extreme. The middle of the range is also known as the mean of the distribution.

The normal distribution is also known as a GAUSSIAN DISTRIBUTION or probability bell curve. It is symmetric about the mean and indicates the value near the mean occur more frequently than the values that are farther away from the mean. c



IMPORTANCE of normal distribution:

It is the most important probability distribution for the independent random variables for some reasons.

1. It describe the distribution of values for many natural phenomena such as biology, physical science, mathematics, finance and economics.
2. It is important because, it can be used to approximate other probability curve such as, binomial, inverse, negative binomial.
3. It is the key idea behind the CENTRAL LIMIT THEOREM (CLT) which states that average calculated from independent and identical distribution random variables have approximately normal distributions. This is true regardless of the type of distribution from which the variables are sampled, as long as it has finite variance.

PARAMETERS OF NORMAL DISTRIBUTION

Two parameters are required to describe a normal distribution: MEAN and STANDARD DEVIATION

1. THE MEAN

It is central highest value of the whole curve. All the other values in the distribution either cluster around it or are at some distance away from it. Changing the mean on the graph, will shift the curve the entire curve along the x-axis.

2. THE STANDARD DEVIATION

In general, it is the measure of variability in the distribution. In a bell curve, it defines the width of the distribution and shows how far away from the mean the other values fall. In addition, it represents the typical distance between the average and the observations. Changing the deviation will change the distribution of values around the mean.

SKEWNESS IN NORMAL DISTRIBUTION

It represents the distribution degree of symmetry. Since the normal distribution is perfectly symmetric, it has a skewness of zero. In other distribution with the skewness less than or greater than zero, the left tail or the right tail will be longer, respectively.

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

Ans. Missing data can be handled in variety of ways and the most common is to ignore it, as the statistical programme make the self decision of ignoring data. It reduces the statistical power of the analysis, which can distort the validity of the result. However there are proven technique to deal with the missing data.

When dealing with missing data,we can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It’s most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

IMPUTATION

When data is missing, it may make sense to delete data, However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.

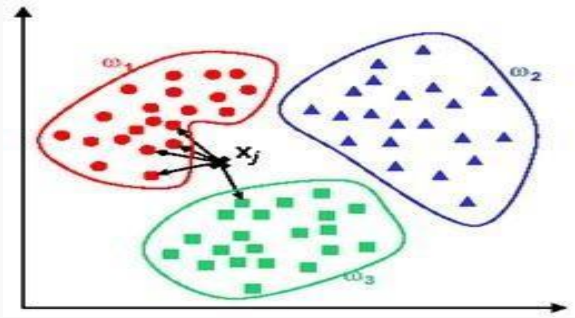
Instead of deletion, We must have multiple solutions to impute the value of missing data. Depending on the need of the model, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

MEAN MEDIAN AND MODE

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, we can calculate the mean or median of the existing observations However, when there are many missing variables, mean or median results can result in a loss of variation in the data.

K NEAREST NEIGHBOURS

kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.



Q.12 What is A/B testing?

Ans. A/B testing data science is a methodical way to evaluate the performance of two variants of a website, app, or campaign. It also goes by the name "**split testing."** By dividing traffic into two groups and serving one group the A/B version while serving the other group the control, A/B testing seeks to determine what works and doesn't work for us. This enables us to evaluate the impact of various versions on conversion rates and response rates.

How does it work:

Suppose a company launches a new cars and it hires two advertisement at the same time. These two companies make two advertisement for the promotion of the new car by taking the latest features of the car .

Here, we want to see which advertisement attracts the most visitors. To determine which advertisement performs better, we will gather data and analyze A/B testing results.

1. **Formulate a hypothesis:**

A hypothesis states how the change of a test variable impacts a performance metric on a population. An example of a hypothesis is the following:

Changing the color of the add-to-cart button from blue to red (the test variable) will increase the conversion rate.

2. Create control and treatment versions of your test variable:

The term “A/B” in A/B testing refers to the two versions of the thing you’re testing.

3. Determine the sample size for statistical significance:

Depending on the use case and the number of users a service has, it can be impossible to run an A/B test on all the population. The next best alternative is to run the A/B test on a subset or sample of users. To do this, practitioners usually determine a statistically significant sample of users that is large enough for them to make conclusions about the population.

4. Run the test, and analyze the results :

To analyze the results, we calculate the difference in the test metric–conversion rate–between the treatment and control groups. If the difference is significant enough, we can confidently conclude that one version is indeed better than the other.

Q.13 Is mean imputation of missing data acceptable practice?

Ans. 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. 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.

Advantages of Mean Substitution

1. Missing values in our data **do not reduce our sample size**, as it would be the case with listwise deletion . Since mean imputation replaces all missing values, you can keep your whole database.
2. Mean imputation is very **simple to understand and to apply** . we can explain the imputation method easily to our audience and everybody with basic knowledge in statistics will get what we've done.
3. If the response mechanism is MCAR, the **sample mean of your variable is not biased**. Mean substitution might be a valid approach, in case that the univariate average of your variables is the only metric we are interested in.

Drawbacks of Mean Substitution

1. Mean substitution leads to **bias in multivariate estimates** such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2. **Standard errors and variance** of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.
3. If the response mechanism is MAR or MNAR, even the **sample mean of your variable is bias**. Assume that we want to estimate the mean of a population's income and people with high income are less likely to respond; our estimate of the mean income would be biased downwards.

Q.14 What is linear regression in statistics?

Ans. 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 has two primary purposes—understanding the relationships between variables and forecasting.

1. The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.
2. A linear regression equation allows you to predict the mean value of the dependent variable given values of the independent variables that you specify.

Different types of linear regression models

There are two different types of linear regression models. They are the following:

- **Simple linear regression**: The following represents the simple linear regression where there is just one independent variable, X, which is used to predict the dependent variable Y.

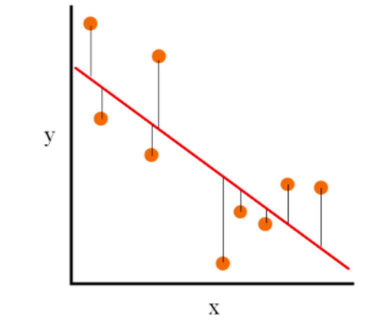


Fig 1. Simple linear regression

- **Multiple linear regression**: The following represents the multiple linear regression where there are two or more independent variables (X1, X2) that are used for predicting the dependent variable Y.

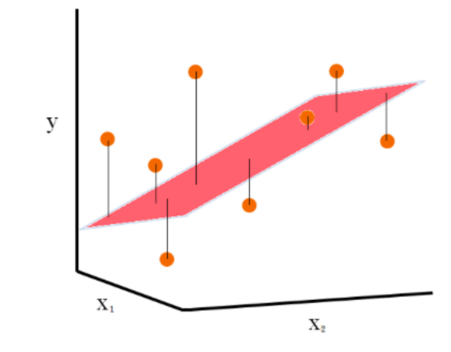


Fig 2. Multiple linear regression

Key Terminologies for Regression Models

The following are some key terminologies in relation to measuring the residuals and performance of the linear regression models:

- **Root Mean Square Error (RMSE)**: Root Mean Square Error (RMSE) is a measure of how well a linear regression model fits the data. It is calculated by taking the squared difference between the predicted values and the observed values, finding an average of all the differences, and then taking the square root of this number. The lower the RMSE value, the better fit of our model to the data.

- **R-Squared**: R-Squared is a statistical measure that represents how well the linear regression model fits the data. It is also known as the coefficient of determination. It is a statistic that measures the proportion of variation in the dependent variable (the y-axis) that can be explained by the independent variables (the x-axis). The R-squared value ranges between 0 and 1, with 1 indicating a perfect fit. A higher R-squared value indicates that more of the variance in the dependent variable can be explained by the independent variables.

- **Residual standard error (RSE)**: Residual Standard Error (RSE) is a measure of how well linear regression models fit the data. It is calculated by taking the sum of the squared residuals (the differences between the observed values and predicted values), dividing it by the degrees of freedom, then taking its square root. The RSE provides an indication of how much variation there is in the data that cannot be explained by the regression model. Lower RSE values indicate better fits to the data.

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

Ans. The two basic branches of statistics are **Descriptive and Inferential.** These both are employed in the scientific analysis of data. Let us now study these two simultaneously.

Firstly, Descriptive statistics: In this type of statistics, the data is summarized through the given observations. The summarization is one method for using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorized into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Secondly, Inferential statistics: This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.

Importance of Statistics

- Statistics executes the work simply and gives a transparent picture of the work we do regularly.
- The statistical methods help us to examine different areas such as medicine, business, economic, social science and others.
- Statistics equips us with different kinds of organized data with the help of graphs, tables, diagrams and charts.
- Statistics makes us understand the bulk of data in a simple way.
- Statistics is the way to collecting accurate quantitative data.