

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

Ans-a) True

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

Ans-a) Central Limit Theorem

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

Ans-b) Modeling bounded count data

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

Ans-d) All of the mentioned

5. _____ random variables are used to model rates.

- a) Empirical
- b) Binomial
- c) Poisson
- d) All of the mentioned

Ans-c) Poisson

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

- a) True
- b) False

Ans-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

Ans-b) Hypothesis

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

Ans-a) 0

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
Ans-c) Outliers cannot conform to the regression relationship

WORKSHEET

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans- The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal. For example, the Student's t, Cauchy, and logistic distributions are symmetric.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution

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

Ans-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.

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.

12. What is A/B testing?

Ans-A/B testing fosters data-driven decisions when planning changes on any website elements. By testing two versions of a web page in real-time, you'll be able to identify what will work best.

Without this data, you will find yourself in a trial-and-error situation. It may also take longer to find the best solution for your web design.

Overall, A/B testing on a [WordPress website](#) is ideal for improving the following aspects of a website:

- **Conversion rates** – when the site already receives a lot of traffic, but you want to convert more site visitors into customers, use A/B testing to see which design version helps them convert.
- **Bounce rates** – changing the placement of elements like navigation buttons or banners may help prevent users from leaving the website.
- **Users' pain points** – A/B testing will help you experiment with different solutions for the visitors' pain points.

13. Is mean imputation of missing data acceptable practice?

Ans- Mean imputation is **typically considered terrible practice** since it ignores feature correlation

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?

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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. What are the various branches of statistics?

Ans-There are three real branches of statistics: **data collection, descriptive statistics and inferential statistics**

Descriptive statistics:

This is a branch of statistics which deals with methods of collection of data, its presentation and organization in various forms, such as distribution tables, graphs (e.g., ogive, Lorenz curves, etc.), diagrams (e.g., pie charts) and finding measures of central tendency and measures of dispersion or spread which are used in the description of data. Managers, CEOs. etc. make use of descriptive statistics in presenting their annual reports, financial accounts and bank statements.

Descriptive statistics is used to present the data in an understandable way, so that a meaningful description can be made.

Inferential or predictive statistics:

This is a branch of statistics which deals with techniques used for analysis of data, making estimates that lead to predictions and drawing conclusions or inferences from

limited information taken on sample basis and testing the reliability of the estimates or predictions.

Inferential statistics is used to make comparisons or predictions about a larger group, known as population, using information gathered about a small part of that population called a sample.

Inferential statistics answers questions, such as "what is this data telling us about?" and "what should we do?" Techniques used are forecasting trends, hypothesis testing, kurtosis, skewness, etc.