STATISTICS WORKSHEET-1

ANSWERS

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

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?

ANS: **(a) Central Limit Theorem**

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

ANS: **(b) Modelling bounded count data**

4. Point out the correct statement.

ANS: **(d) All of the mentioned**

5. _____ random variables are used to model rates.

ANS: **(c) Poisson**

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

ANS: **(b) False**

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

ANS: **(b) Hypothesis**

8. Normalized data are centered at___ ___and have units equal to standard deviations of the original data. ANS: **(a) 0**

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

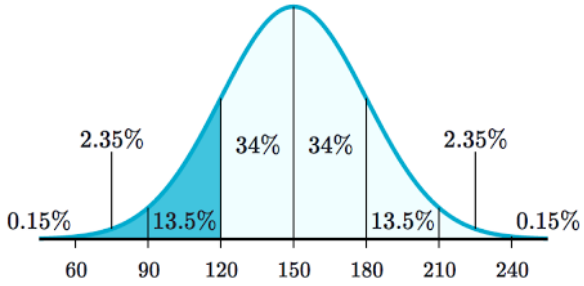
ANS: **(c) Outliers cannot conform to the regression relationship**

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

ANS: The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.

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. approximately normally distributed; measurement errors also often have a normal distribution. The normal distribution is easy to work with mathematically. In many practical

cases, the methods developed using normal theory work quite well even when the distribution is not normal.



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

ANS: Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.

Below is a sample of the missing data from the Titanic dataset. You can see the columns ‘Age’ and ‘Cabin’ have some missing values.

There can be multiple reasons why certain values are missing from the data.

Reasons for the missing data from the dataset affect the approach of handling missing data. So it’s necessary to understand why the data could be missing.

Some of the reasons are listed below:

- Past data might get corrupted due to improper maintenance.
- Observations are not recorded for certain fields due to some reasons. There might be a failure in recording the values due to human error.

Types Of Missing Value

Missing Completely At Random (MCAR)

In MCAR, the probability of data being missing is the same for all the observations. In this case, there is no relationship between the missing data and any other values observed or unobserved (the data which is not recorded) within the given dataset. That is, missing values are completely independent of other data. There is no pattern. In the case of MCAR, the data could be missing due to human error, some system/equipment failure, loss of sample, or some unsatisfactory technicalities while recording the values.

Missing At Random (MAR)

Missing at random (MAR) means that the reason for missing values can be explained by variables on which you have complete information as there is some relationship between the missing data and other values/data. In this case, the data is not missing for all the observations. It is missing only within sub-samples of the data and there is some pattern in the missing values.

Missing Not At Random (MNAR)

Missing values depend on the unobserved data. If there is some structure/pattern in missing data and other observed data can not explain it, then it is Missing Not At Random (MNAR). If the missing data does not fall under the MCAR or MAR then it can be categorized as MNAR.

It can happen due to the reluctance of people in providing the required information. A specific group of people may not answer some questions in a survey.

12. What is A/B testing?

ANS: A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

A/B testing, also known as split testing, is a marketing experiment wherein you split your audience to test a number of variations of a campaign and determine which performs better. In other words, you can show version A of a piece of marketing content to one half of your audience, and version B to another

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. 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: If we want to use a variable x to draw conclusions concerning a variable y: y is called dependent or response variable. x is called independent, predictor, or explanatory variable. If the relationship between two variables is linear is can be summarized by a straight line. A straight line can be described by an equation:

y= a + bx

a is called the intercept and b the slope of the equation. The slope is the amount by which y increases when x increases by 1 unit.

15.What are the various branches of statistics?

ANS: There are basically four branches into which statistics is divided.

1. **Mathematical or theoretical statistics**

   It helps in forming the experimental and statistical distribution.

2. **Statistical methods or functions**

   It helps in the collection, tabulation and interpretation of the data. It helps in analyzing the data and returns insight from the data

3. **Descriptive statistics**

   It helps in summarizing and organizing any data set characteristics. It also helps in the representation of data in both classification and diagrammatic way.

4. **Inferential Statistics**

   Inferential statistics are often used to compare the differences between the treatment groups. Inferential statistics use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects.