STATISTICS – WORKSHEET 1

Q1 to Q9 have only one correct answer.

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 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) 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) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

   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 centred at _____ and have units equal to standard deviations of the original data.?

   Ans: a) 0

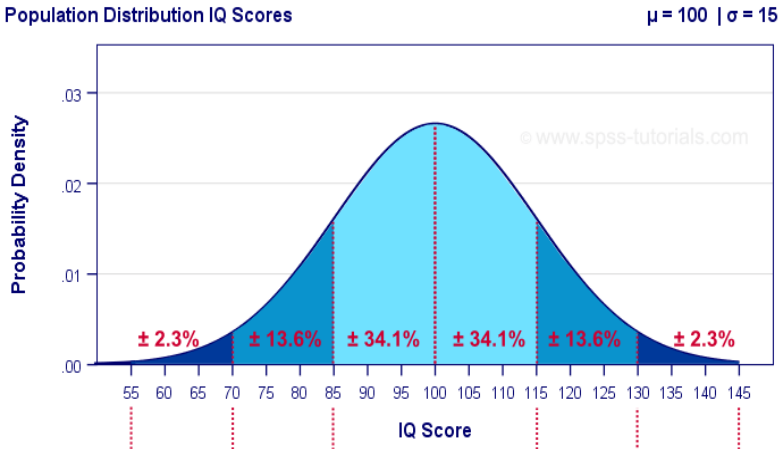
9. Which of the following statement is incorrect with respect to outliers?

   Ans: c) Outliers cannot conform to the regression relationship

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

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

Normal distribution is the most widely used of all distributions. Most of phenomena in real world are normally distributed. The mean of the normal distributed curve will be in the center and is symmetrical on either side from the mean forming a bell-shaped curve. The two tails of the curve on the either side never touch the horizontal line. Data beyond the normal distribution curve are the outliers. Threshold value of standard deviation is ±3, beyond are abnormal data(outliers).



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

Missing data can be handled in various ways. The most frequently way is imputing the missing data. Imputing is used because removing the data from the dataset every time is not feasible solution, which leads to reduction of dataset, this not only raises concerns for biasing the dataset but also leads to wrong insights of the data.

The most common way of replacing the missing data is to impute it with **mean/median/mode**. The other ways are replacing the missing data is with **Zero/Min/Max/adjacent cell value.**

12.What is A/B testing?

- 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.
- Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.
- A/B testing is also a form of statistical and two-sample hypothesis testing.
- Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.
- A/B testing also known as split testing.
- Example: If any company wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in color, headlines, and format. Here the objective is to check which newsletter brings higher traffic on the website.

13.Is mean imputation of missing data acceptable practice?

- Yes, if it is a numerical data

14.What is linear regression in statistics?

- Linear Regression in statistics is used to represent a relationship between dependent variable (y) and an independent variable (x).
- Independent variable (x) also called as predictor/input/ explanatory variable.
- Dependent variable (y) also called as response variable/output/outcome/measured.
- Simple Linear Regression: If the relationship between single 'x' and 'y' is linear it can be expressed by an equation:

$y = a + bx$, where a = intercept & b = slope (with increase in 'x', 'y' also increases by a unit).

- Multiple Linear regression: More than one independent variable expressed as

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + c$

- Even if the data points are scattered i.e. not in a straight line, the pattern can be obtained which gives an insight into the data.
- Example: With increase in age of a person, the level of glucose increases in the body.

15. What are the various branches of statistics?

Descriptive Statistics: These are the procedures used to summarize, organize, and make sense of a set of scores or observations. Descriptive statistics are further divided into Central tendency (mean, median, mode) and Dispersion of data (range, variance, standard deviation, percentile, skewness).

Inferential Statistics: These are procedures used that allow researchers to infer or generalize, observations made with samples to the larger population from which they were selected. Here we will work with hypothesis testing (z-score test(Z-test), t-test, chi-square test, ANOVA test, MANOVA test, ANCOVA test, MANCOVA test).