***STATISTICS WORKSHEET***

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

a) True

b) False

Answer – 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

Answer – 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

Answer – c). Modelling contingency tables

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

Answer – d). All of the mentioned.

5. \_\_\_\_\_\_ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer – C). Poisson

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

a) True

b) False

Answer – B). False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer – B). Hypothesis

8. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

Answer – 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

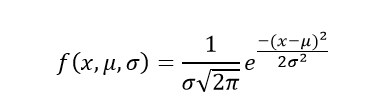
d) None of the mentioned

Answer – 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?

Answer - The **Normal Distribution**, also called the **Gaussian Distribution**, is the most significant continuous probability distribution. A large number of random variables are either nearly or exactly represented by the normal distribution. Furthermore, it can be used to approximate other [probability distributions](https://byjus.com/maths/probability-distribution/). The formula for Normal Distribution is mentioned below:



Where,

* x is the variable
* μ is the mean
* σ is the standard deviation

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

Answer – There are 3 common methods to handle missing data.

1. Mean or median imputation.
2. Multivariate imputation by Chained equations (MICE)
3. Random Forest
4. **Mean or median imputation** - When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. There are multiple reasons why this would not be most feasible option:

* There may not be enough observations with non-missing data to produce a reliable analysis.
* In predictive analytics, missing data can prevent the predictions for those observations which have missing data.
* External factors may require specific observations to be part of the analysis.

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low.

For large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2). Multivariate imputation by Chained equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE use predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, Bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

3). Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple [decision trees](https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods) to estimate missing values and outputs OOB (out of bag) imputation error estimates.

One thing random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data.

There are 3 imputation techniques I would like to discuss. The data can be missing in following ways:

* **Missing at Random (MAR):**Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
* **Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
* **Missing not at Random (MNAR):**Two possible reasons are that the missing value depends on the hypothetical value (e.g., People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable’s value.

12. What is A/B testing?

Answer - **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. A/B testing is one of the most prominent and widely used statistical tools. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics.

13. Is mean imputation of missing data acceptable practice?

Answer - Mean imputationis the practice of replacing null values in a data set with the mean of the data. Mean imputation reduces the variance of the data and increases bias in data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

Hence it is advisable not to follow mean imputation on missing data.

14. What is linear regression in statistics?

Answer - In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

15. What are the various branches of statistics?

Answer - Statistics examine the methodology for collecting, reviewing, analyzing, and making data conclusions.

Branches of Statistics: -

**1). Descriptive Statistics**

Descriptive statistics is the first part of statistics that deals with the collection of data. The statisticians need to be aware of the designing and experiments. They also need to choose the right focus group and avoid biases. In contrast, Descriptive statistics are used in use to do various kinds of analysis on different studies.

Descriptive statistics have two parts

* Central tendency measures.
* Variability measures.

**2). Inferential Statistics**

* The inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.
* Inference statistics often speak in terms of probability by using descriptive statistics. Besides, these techniques are used primarily by a statistician for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

Different types of inferential statistics include:

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