**WORKSHEET**

**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

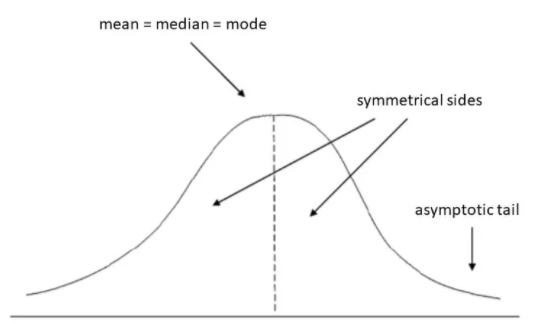
**Ans. C) Outliers cannot conform to the regression relationship**

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

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

**Ans: Normal distribution** is one of the most common [distribution function](https://www.britannica.com/science/distribution-function) for independent, randomly generated variables. It is a continuous probability distribution which is symmetrical on both sides of the mean, such that the right side of the center is the exact mirror image of the left side.The normal distribution curve represents an area where the probability and the total area under the curve is equal to 1.

In a normal distribution, most of the continuous data values tend to cluster around the mean, and the far away a value is away from the mean, the less likely it is to have occurrence. Here, the tails are also asymptotic, which denotes that they approach but cannot meet at the horizon. The normal distribution is also called the bell curve as the graph of its probability density takes the shape of a bell. It is also known as called Gaussian distribution, as the German mathematician Carl Gauss was the first theorised it. In a perfect normal distribution mean=median=mode which is at the peak of the bell curve.



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

**Ans:** First of all before proceeding lets understand types of missing data:

* a)Missing at random(MAR): It depicts that the propensity of a data point to be missing is independent of missing data, though it may be related to some observed data.
* b)Missing completely at random (MCAR):Here if the value goes missing then, it has no relation with it hypothetical value or values of other variables.
* c)Missing not at random (MNAR): Here it can go 2 ways. One case where the missing data depends on the hypothetical value. Second case where missing value is dependent on the value of few other variables.

In the first 2 cases as per the occurrences we can remove the mssing data. But in the 3rd case (MNAR) removing the data with missing values will lead to removal of essential data and creates a bias. So going further we can conclude that we can handle the missing data in 2 ways:

* Deletion: In this case the data with missing values are omitted as per requirement like listwise, pairwise or dropping variables. This method is suitable for first 2 types of missing values I’e MAR & MCAR.
* Imputation: In this case the missing data here unlike the previous method is filled in with the help of different suitable imputable techniques. MNAR type missing data are handled in this way.

**Few recommended imputation techniques are as follows:**

* K Nearest neighbors (KNN): This technique imputes missing values by finding the k most similar observations (on the basis of some distance measure) and taking the mean/median/mode of those neighbors. This method requires determination of the appropriate number of neighbors to consider, as well as how “distance” should best be measured.
* **Linear regression**: It can be used to impute missing values by using the existing variables to make a prediction about the missing value. Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is.
* **Last Observation Carried Forward (LOCF)/Next Observation Carried Backward (NOCB)** For time-series data, a straight-forward option is to use the last observed value for any missing data. Similarly, we could use the first observation after the missing value, which is referred to as Next Observation Carried Backward (NOCB).
* **Multiple Imputation by Chained Equations (MICE)**  
  With multiple imputation, the distribution of the observed data is taken into account, and several plausible estimates for the missing value are created. Multiple data sets are created and analyzed individually to obtain a set of parameter estimates. This method better accounts for uncertainty in the missing values, and is able to effectively handle continuous as well as categorical data.

**12. What is A/B testing?**

**Ans:** This is often termed as Split testing as it makes use of Control and variants in order to determine which version leaves the maximum impact and drives business metrics.

As mentioned earlier, there are 2 factors involved here control(A) or the original version and B as the variation of the original version.

In order to test the business metrics success, both of these versions are presented before website visitors to calculate the maximum impact. The one which shows the maximum impact on business metrics is termed as winner.

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

**Ans**: Mean imputation is not an optimal solution for the same. The reason being if data is missing at random, then estimates calculated by the help of mean shall remain unbiased . As most of the estimates are based upon the variable factors, therefore mean imputation may not be the correct solution to calculate.

**14. What is linear regression in statistics?**

**Ans:** Linear regression is the approach defined while connecting the modelling between the Scalar quantity and the variables which could be both dependent or independent.  During regression analysis this was the most common technique used to study the relation ship modelling, the reason being Models depended linearly upon their unknown parameters and also because the other statistical parameters are relatively easy to determine.

If the agenda deals with error reduction or predictions or forecasting then linear regression shall prove fruitful because it fits a predictive model which can be build based upon the data sets of values

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

**Ans: Statistics** deals with developing and studying different methods for collecting, analyzing and presenting the empirical data. It can be categorized into 2 types:

* **Descriptive statistics -** As the names suggests it deals with presentation and collection of data. While experimenting the executor needs to be aware of the data sets of values, biases and correct focus. The sets of values needs to be calculated on the average of the meta data or host data.
* **Inferential statistics -** From the output thus received from the descriptive statistics, this statisitics techniques involves drawing and pictorial representation of the data sets of values. The inferences which are drawn based on the above study is what's peculiar to the experiment conducted. While drawing one must be very careful to decide the biases as they are relatively easy to creep in.