# Statistics– WORKSHEET 4

**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.
   1. True
   2. False

**Ans. A) True**

1. 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?
   1. Central Limit Theorem
   2. Central Mean Theorem
   3. Centroid Limit Theorem
   4. All of the mentioned

**Ans. A) Central Limit Theorem**

1. Which of the following is incorrect with respect to use of Poisson distribution?
   1. Modeling event/time data
   2. Modeling bounded count data
   3. Modeling contingency tables
   4. All of the mentioned

**Ans. B) Modeling bounded count data**

1. Point out the correct statement.
   1. The exponent of a normally distributed random variables follows what is called the log- normal distribution
   2. Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   3. The square of a standard normal random variable follows what is called chi-squared distribution
   4. All of the mentioned-
2. random variables are used to model rates.
   1. Empirical
   2. Binomial
   3. Poisson
   4. All of the mentioned

**Ans. C) Poisson**

1. Usually replacing the standard error by its estimated value does change the CLT.
   1. True
   2. False

**Ans. b) False**

1. Which of the following testing is concerned with making decisions using data?
   1. Probability
   2. Hypothesis
   3. Causal
   4. None of the mentioned

**Ans. b) Hypothesis**

1. Normalized data are centered at and have units equal to standard deviations of the original data.
   1. 0
   2. 5
   3. 1
   4. 10

**Ans. a) 0**

1. Which of the following statement is incorrect with respect to outliers?
   1. Outliers can have varying degrees of influence
   2. Outliers can be the result of spurious or real processes
   3. Outliers cannot conform to the regression relationship
   4. None of the mentioned

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

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

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

Ans. Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. In simple terms, if a probability distribution forms a bell-shaped curve and mean, median and mode of the sample are equal then the variable has a normal distribution.

The reason why Normal Distribution is so easy to explain because: -

1. Mean, median and mode are all equal.
2. Only mean and the standard deviation is required to explain the entire distribution.
3. How do you handle missing data? What imputation techniques do you recommend?

Ans.

* Missing data in a dataset are the result no response or non-availability of the required data. These are very common occurrence in dataset
* Higher amount of missing data adversely affects the model performance.
* However, there are several ways to treat the missing data.
* Either we can replace those missing data by a statistical value like the mean, median, mode or zero. Depending on the percentage of missing values w.r.t the dataset we can also choose to discard them.
* If the missing value percentage are lower than 10% of the dataset, we can opt to drop them.
* If the missing values is a categorical / ordinal value then we can impute those values by the mode.
* If the missing values is a continuous/discreet data then we can impute those missing values by their mean or median.

1. What is A/B testing?

Ans. 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.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

1. Is mean imputation of missing data acceptable practice?
2. What is linear regression in statistics?

Ans. When we see a relationship in a scatterplot, we can use a line to summarize the relationship in the data. We can also use that line to make predictions in the data. This process is called **linear regression**. There are more advanced ways to fit a line to data, but in general, we want the line to go through the "middle" of the points.

1. What are the various branches of statistics?

Ans.

**Types of Statistics:**

1. **Descriptive Statistics:** Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

**(a). Measure of central tendency – Mean, Median, & Mode**

**(b). Measure of Variability – Range, Variance, & Dispersion**

1. **Inferential Statistics**

Various types of inferential statistics are used widely nowadays and are very easy to interpret. These are given below:

* One sample test of difference/One sample hypothesis test
* Confidence Interval
* Contingency Tables and Chi-Square Statistic
* T-test or Anova
* Pearson Correlation
* Bi-variate Regression
* Multi-variate Regression

