**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

Answer: A

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

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: B

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

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

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: C

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

a) True

b) False

Answer: B

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

Answer: B

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

Answer: A

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

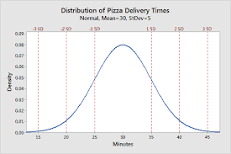
d) None of the mentioned

Answer: C

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

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

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.



The above diagram explains the bell curve:

* The average time lies between the bell curve.i.e Pizza delivery time i.e 25-30minutes which falls between the bell curve.
* A normal distribution is sometimes called the bell curve.
* The bell curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right.
* There are many things, such as intelligence, height, and blood pressure, that naturally follow a normal distribution.
* The center of a normal distribution is located at its peak, and 50% of the data lies above the mean, while 50% lies below. It follows that the mean, median, and mode are all equal in a normal distribution. Standard Deviation is also used in normal distribution data.

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

Missing data reduces the statistical power of the analysis, which can distort the validity of the results. If missing data is there in any dataset, it will be hard to predict or solve the problems.

There are types of missing data:-

* Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.
* Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data
* Not Missing At Random (NMAR):  When the missing data has a structure to it, we cannot treat it as missing at random.
* We can use deletion methods for such missing data but it will not be the best approach towards it. We might keep something important data through this method.
* So we impute missing data with some imputation techniques, so that we can give proper/appropriate results.

## Mean , Median & Mode Imputation:

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. A common technique is to use for mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low.

* 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 tree to estimate missing values and outputs OOB (out of bag) imputation error estimates.

## 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 uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables.

* Multiple imputation:

In the statistical literature, arguably the most advanced methodology for performing missing data imputation is multiple imputation. In multiple imputation we generate missing values from the dataset many times. The individual datasets are then pooled together into the final imputed dataset, with the values chosen to replace the missing data being drawn from the combined results in some way.

12. What is A/B testing?

* Like any type of scientific testing, A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.
* The A/B testing process can be simplified as follows:

1. You start the A/B testing process by making a claim (hypothesis).
2. You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.
3. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

* A/B tests, also known as split tests, allow you to compare 2 versions of something to learn which is more effective. Simply put, do our users like version A or version B?
* A/B testing can be used to determine the right price for the product, as this is perhaps one of the most difficult tasks when a new product or service is launched.

13. Is mean imputation of missing data acceptable practice?

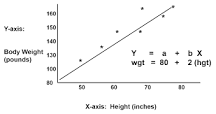
* The process of replacing null values in a data collection with the data’s mean is known as mean imputation.
* Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old 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.
* Mean imputation reduces the variance of the imputed variables.
* Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval

14. What is linear regression in statistics?

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). ... Linear regression has many practical uses.

Linear regression used for:

* Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.



* For example, suppose that height was the only determinant of body weight. In this example, if an individual was 70 inches tall, we would predict his weight to be: Weight = 80 + 2 x (70) = 220 lbs. In this simple linear regression, we are examining the impact of one independent variable on the outcome.

15. What are the various branches of statistics?

* Statistics is a study of presentation, analysis, collection, interpretation and organization of data
* There are two main branches of statistics  
  - Inferential Statistic.  
  - Descriptive Statistic.

Inferential Statistics:  
Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:  
Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.