STATISTICS WORKSHEET-1

Ques1) Bernoulli random variables take (only) the values 1 and 0?

a) True

b) False

ANS :-**a)** True

Ques 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

Ques 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

Ques 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

Ques 5) ______ random variables are used to model rates.

a) Empirical b) Binomial

c) Poisson d) All of the mentioned

ANS :-**c)** Poisson

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

a) True b) False

ANS :- **b)** False

Ques 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

Ques 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

Ques 9) 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

Ques 10) what do you understand by the term Normal Distribution?

ANS A probability function that specifies how the values of a variable are distributed is called the normal distribution .It is symmetric since most of the observations assemble around the central peak of the curve. The probabilities for values of the distribution are distant from the mean narrow off evenly in both directions. It is also known as Gaussian distribution. It is used in the natural sciences and social arts to describe real valued random variables whose distributions are unknown.

Characteristics of normal distribution:-

1) It is symmetric, unimodal (i.e., one mode), and asymptotic.
2) The values of mean, median, and mode are all equal.
3) A normal distribution is quite symmetrical about its center. That means the left side of the center of the peak is a mirror image of the right side. There is also only one peak (i.e., one mode) in a normal distribution.

Ques11) How do you handle missing data? What imputation techniques do you recommend?

ANS There are a lot of techniques to treat missing value. I am trying to think what is the best way to organize some of the most commonly used methods, if you use SAS to implement it -

- **Ignore the records with missing values.**

Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

- **Substitute a value such as mean**.

When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this method could cause bias distribution and variance. That's where the following imputation methods come in.

- **Predict missing values.**

Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

  - Logistic Regression
  - Discriminant Regression
  - Markov Chain Monte Carlo (MCMC)
  - ...
- **Predict missing values - Multiple Imputation**. Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

Ques12). What is A/B testing?

Ans A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics. Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

Ques13) Is mean imputation of missing data acceptable practice?

Ans Mean imputation is simple but is equally dangerous as well. It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power. But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.

Different problems of mean imputation are:-

1) Mean imputation does not preserve the relationships among variables.

2) Mean Imputation Leads to an Underestimate of Standard Errors.

Ques 14) What is linear regression in statistics?

Ans Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

Ques 15) What are the various branches of statistics?

Ans Statistics have majorly splits into two types:

1. Descriptive statistics
2. Inferential statistics

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorise into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.