STATISTICS WORKSHEET 1

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

Ans. (a) True

Que 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?

Ans. (a) Central Limit Therorem

Que 3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans. (b) Modeling bounded count data

Que 4. Point out the correct statement.

Ans. (c) The square of a standard normal random variable follows what is called chi-squared distribution

Que 5. __________ random variables are used to model rates.

Ans. (c) Poisson

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

Ans. (b) False

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

Ans. (b) Hypothesis

Que 8. Normalized data are centered at _________ and have units equal to standard deviations of the

original data.

Ans. (a) 0

Que 9. Which of the following statement is incorrect with respect to outliers?

Ans. (c) Outliers cannot conform to the regression relationship.

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

Ans. Normal Distribution:

The normal distribution is the most widely known and used of all distributions.

Because the normal distribution approximates many natural phenomena so will, it has developed into a standard of reference for many probability problems.

Why Normal Distribution?

Normal distributions are important in statistics. There importance is partly due to the central limit theorem. A normal distribution is sometimes informally called a bell curve. Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution.

The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.

There is a very strong connection between the size of a sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

It is a 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.

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

Ans. Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the Following ways :

Missing Completely At Random (MCAR) , Missing At Random(MAR) , Not Missing At Random (NMAR)

There are a lot of techniques to treat missing value. So here are few common methods

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the 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. However, 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 .

2. 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 to estimate missing values and output imputation error estimates.

One caveat is that 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.

Que 12. What is A/B testing?

Ans. A/B testing is a popular way to test your products and is gaining steam in the data science field.

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to 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.

In the above scenario, you may divide the products into two parts- A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.

A/B testing doesn’t work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

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

Ans. True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased.

Mean imputation replaces missing values of a certain variable by the mean of non-missing cases of that variable.

Advantages of the method:

- Missing values in your data do not reduce your sample size, as it would be the case with listwise deletion. Since mean imputation replaces all missing values, you can keep your whole database.
- Mean imputation is very simple to understand and to apply. You can explain the imputation method easily to your
  to your audience and everybody with basic knowledge in statistics will get what you've done.
- If the response mechanism is MCAR, the sample mean of your variable is not biased. Mean substitution might be a valid approach, in case that the univariate average of your variables is the only metric you are interested in.

There are a few advantages, but many serious drawbacks. On top of that, we can also benefit from the advantages With more advanced imputation methods. To make it short, there is basically no excuse for using mean imputation.

Que 14. What is linear regression in statistics?

Ans. In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

Simple linear regression uses one independent variable to explain or predict the outcome of the dependent Variable y, while multiple linear regression uses two or more independent variables to predict the outcome.

Regression can help finance and investment professionals as well as professionals in other businesses.

Linear regression quantifies the relationship between one or more predictor variables and one outcome variable.

Linear regression is commonly used for predictive analysis and modeling.

For example, it can be used to quantify the relative impacts of age, gender, and diet on height. Linear regression is also known as multiple regression, multivariate regression, ordinary least squares (OLS), and regression.

The principal advantage of linear regression is its simplicity, interpretability, scientific acceptance, and widespread availability. Linear regression is the first method to use for many problems. Analysts can use linear regression together with techniques such as variable recoding, transformation, or segmentation.

Que 15. What are the various branches of statistics?

Ans. Statistics plays a main role in the field of research. It helps us in the collection, analysis and presentation of data. Statistics is concerned with developing and studying different methods for collecting, analyzing and presenting the empirical data.

The field of statistic is composed of two broad categories - Descriptive and Inferential statistics. Both of them give us different insights about the data. One alone doesn't help us much to understand the complete picture of our data but using both of them together gives us a powerful tool for description and prediction.

Descriptive Statistics :

It describes the important characteristics/properties of the data using the measures the central tendency like mean/median/mode and the measures of dispersion like range, standard deviation, variance etc. Data can be summarized and represented in an accurate way using charts, tables and graphs.

Inferential Statistics :

It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.

Some differences to remember:

S No.	Descriptive Statistics	Inferential Statistics
1.	Concerned with the describing the target population.	Make inferences from the sample and generalize them to the population.
2.	Organize, analyze and present the data in a meaningful manner.	Compares, test and predicts future outcomes.
3.	Final results are shown in form of charts, tables and graphs.	Final result is the probability scores.
4.	Describe the data which is already known.	Tries to make conclusions about the population that is beyond the data available.
5.	Tools- Measures of central tendency (mean/median/mode), Spread of data (range, standard deviation etc.)	Tools- Hypothesis tests, Analysis of variance etc.