STATISTICS WORKSHEET-1

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

a) True

b) False

ANS : 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 :

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 : 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 : 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

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

ANS : Normal distribution, also known as the Gaussian distribution, 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.

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations

The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

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

ANS : **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. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.
- **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. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.
- **Not Missing At Random (NMAR):** When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Imputation Techniques

- Complete Case Analysis(CCA):- This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. ...
- Arbitrary Value Imputation. ...
- Frequent Category Imputation.

12. What is A/B testing?

ANS : better in a controlled environment. 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

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 bet.

13. Is mean imputation of missing data acceptable practice?

ANS: 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.

14. What is linear regression in statistics?

ANS : ***Linear regression*** quantifies the relationship between one or more *predictor variable(s)* 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 (the predictor variables) on height (the outcome variable). Linear regression is also known as *multiple regression*, *multivariate regression*, *ordinary least squares (OLS)*, and *regression*. This post will show you examples of linear regression, including an example of *simple linear regression* and an example of *multiple linear regression*.

Example of simple linear regression

The table below shows some data from the early days of the Italian clothing company Benetton. Each row in the table shows Benetton's sales for a year and the amount spent on advertising that year. In this case, our outcome of interest is sales—it is what we want to predict. If we use advertising as the predictor variable, linear regression estimates that Sales =168+23 Advertising **.** That is, if advertising expenditure is increased by one million Euro, then sales will be expected to increase by 23 million Euros, and if there was no advertising we would expect sales of 168 million Euros

Example of multiple linear regression

Linear regression with a single predictor variable is known as *simple regression.* In real-world applications, there is typically more than one predictor variable. Such regressions are called *multiple regression.* For more information, check out this post on why you should not use multiple linear regression for key Drive Analysis with example data for multiple linear regression examples.

Returning to the Benetton example, we can include year variable in the regression, which gives the result that sales= 323+14 Advertising +47 year. The interpretation of this equation is that every extra million Euro of advertising expenditure will lead to an extra 14 million Euro of sales and that sales will grow due to non-advertising factors by 47 million Euro per year.

15. What are the various branches of statistics?
ANS : The two branches of statistics are descriptive and inferential statistics.

Descriptive Statistics:

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

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.

.