STATISTICS WORKSHEET-1

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

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?**

Central Limit Theorem

3. **Which of the following is incorrect with respect to use of Poisson**

distribution?

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

5. **___ Poisson distribution___ random variables are used to model rate.**

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

False

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

Hypothesis

8. **4. Normalized data are centered at__0____and have units equal to standard deviations of the original data.**

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

Outliers cannot conform to the regression relationship

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

A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

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

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

Imputation techniques which I would recommend are :-

There are different ways of replacing the missing values. You can use the python libraries Pandas and Sci-kit learn as follows:

Replacing With Arbitrary Value

If you can make an educated guess about the missing value then you can replace it with some arbitrary value using the following code.

Replacing With Mean

This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.

Replacing With Mode

Mode is the most frequently occurring value. It is used in the case of categorical features.

Replacing With Median

Median is the middlemost value. It’s better to use the median value for imputation in the case of outliers.

12. What is A/B testing?

A/B Testing is 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 drives business metrics.

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.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is **typically considered terrible practice** since it ignores feature correlation.

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).

15. What are the various branches of statistics?

Statistics

___ _ __|______

Descriptive **Inferential**

□ Central Tendency	Dispersion	Z Score
	Of data	Hypothesis Testing
Mean	Range	T Test
Median	Percentile	z Test(Z Score)
Mode	Skew	Co-Relation Test
	Variance	Chi-Square Test
	Standard Deviation	Anova Test