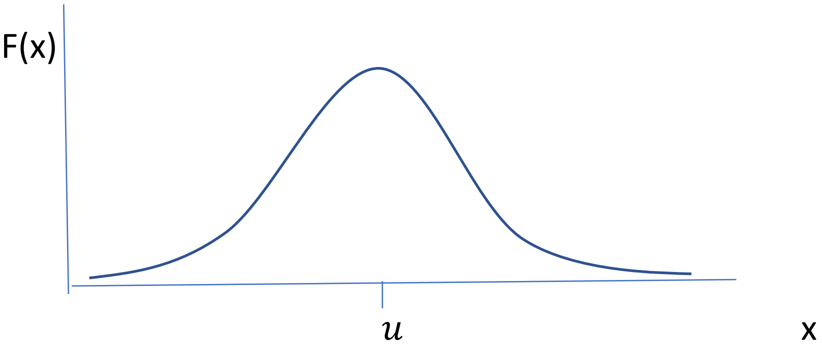
STATISTICS WORKSHEET-1

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

   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?
   Ans: a) Central Limit Theorem
3. Which of the following is incorrect with respect to use of Poisson distribution?
   Ans: b) Modeling bounded count data
4. Point out the correct statement.
   Ans: d) All of the mentioned
5. _______ random variables are used to model rates.
   Ans: c) Poisson
6. Usually replacing the standard error by its estimated value does change the CLT.
   Ans: b) False
7. Which of the following testing is concerned with making decisions using data?
   Ans: b) Hypothesis
8. Normalized data are centered at _______and have units equal to standard deviations of the original data.
   Ans: a) 0
9. Which of the following statement is incorrect with respect to outliers?
   Ans: c) Outliers cannot conform to the regression relationship.

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



The Normal Distribution , also known as the Gaussian distribution is a continuous probability distribution that is symmetrical around its mean , most of the observations cluster around the central peak and the probabilities for the values further away from the mean taper off equally in both directions .

The simplest case is , Normal Distribution over a scalar value x, in which case the PDF is:

$p(x|\mu, \sigma 2) = 1/ (\surd\ 2\pi\sigma 2)\ .\ e - (1/2\sigma\ 2)\ (x - \mu)\ 2$

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

Ans: There are a lot of techniques to handle missing data. The most commonly used methods are :

- **Ignore the data 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. I

Logistic Regression

Discriminant Regression

- **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.

12. 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 are shown to different segments . An A/B testing is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

To put this in more practical terms, a prediction is made that Page Variation B will perform better than Page Variation A. Then, data sets from both pages are observed and compared to determine if Page Variation B is a statistically significant improvement over Page Variation A.

This process is an example of statistical hypothesis testing.

13. Is mean imputation of missing data acceptable practice?

Ans:

The process of replacing missing values in a data collection with the mean of the data is known as mean imputation . Mean imputation is typically considered terrible practice because

- 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.
- Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

Ans: Linear regression describes linear trends in the relation between a response and an explanatory variable. Linear trends may be specified with the aid of linear equations.

A linear equation is an equation of the form:

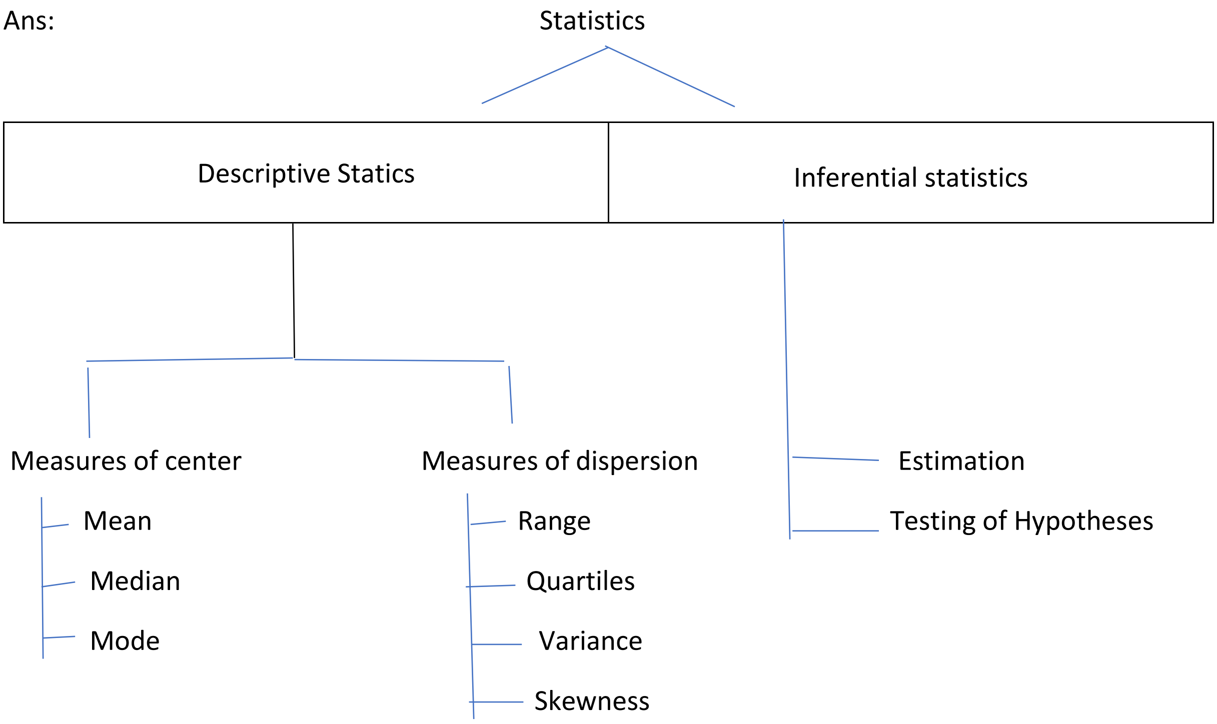
$$y = a + b \cdot x$$

where y and x are variables and a and b are the coefficients of the equation. The coefficient a is called the intercept and the coefficient b is called the slope.

A linear equation can be used in order to plot a line on a graph. With each value on the x-axis one may associate a value on the y-axis: the value that satisfies the linear equation. The collection of all such pairs of points, all possible x values and their associated y values, produces a straight line in the two-dimensional plane.

15. What are the various branches of statistics?

Ans:


Statistics
Descriptive Statics
Inferential statistics
Measures of center
Mean
Median
Mode
Measures of dispersion
Range
Quartiles
Variance
Skewness
Estimation
Testing of Hypotheses