STATISTICS

WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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: Central Limit Theorem

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

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

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Ans: 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: All of the mentiond

For explanation: Many random variables, properly normalized, limit to a normal distribution.

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans: 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: 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 mention

Ans: outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans: A normal distribution is the proper term for a probability bell curve. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal.The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

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

Ans: The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of

missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

Following ways to handle the missing values in the datasets

1.Deleting Rows with missing values

2.Impute missing values for continuous variable

3.Impute missing values for categorical variable

4.Other Imputation Methods

5.Using Algorithms that support missing values

6.Prediction of missing values

7.Imputation using Deep Learning Library — Datawig

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. ...

2.Multivariate Imputation by Chained Equations (MICE) MICE assumes that the missing data are Missing at Random (MAR). ...

3.Mean imputation. Simply calculate the mean of the observed values for that variable for all individuals who are non-missing. ...

Hot deck imputation. ...

Cold deck imputation. ...

Regression imputation. ...

Stochastic regression imputation. ...

Interpolation and extrapolation.

12. What is A/B testing?

Ans: A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. ... A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.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. A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc. If you're not A/B testing your website, you're surely losing out on a lot of potential business revenue

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. That's a good thing. ... Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.Bad practice in general If just estimating means, mean imputation preserves the mean of the observed data Leads to an underestimate of the standard deviation.Distorts relationships between variables by "pulling" estimates of the correlation toward zero

14. What is linear regression in statistics?

Ans:The linear regression model describes the dependent variable with a straight line that is defined by the equation Y = a + b × X, where a is the y-intersect of the line, and b is its slope. ... The regression line enables one to predict the value of the dependent variable Y from that of the independent variable X. Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line.Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

15. What are the various branches of statistics?

Ans: The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

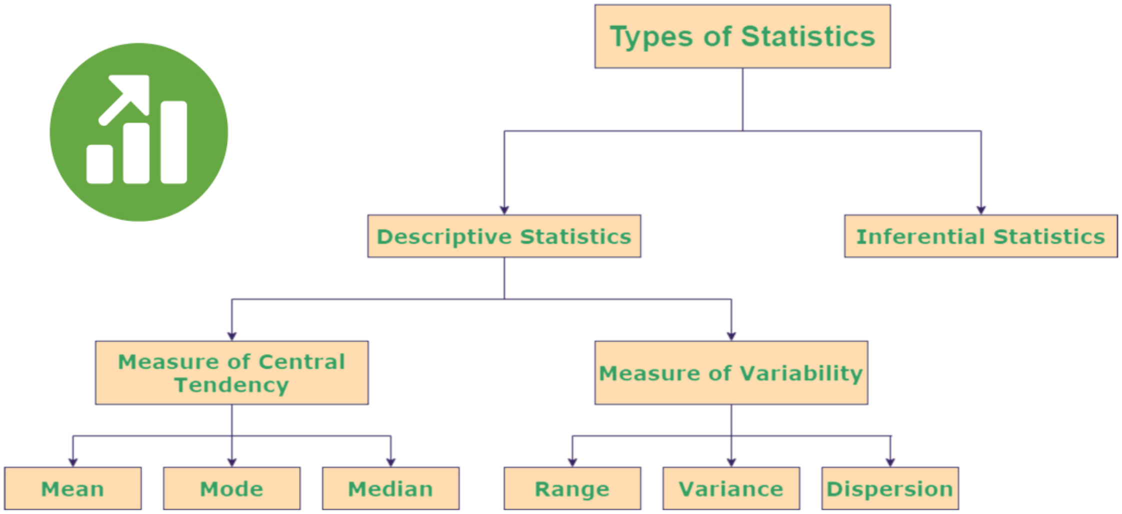
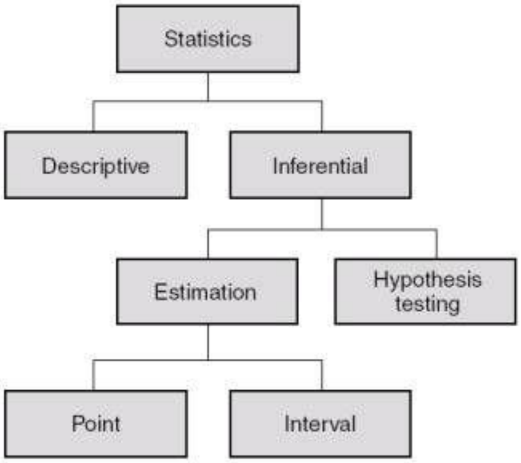


FIGURE 15.1 Major divisions in the field of statistics


Statistics
Descriptive
Inferential
Estimation
Hypothesis testing
Point
Interval