Ans 1: A

Ans 2: A

Ans 3: B

Ans 4: D

Ans 5: C

Ans 6: B

Ans 7: B

Ans 8: A

Ans 9: C

Ans 10:

Normal Distribution:-

A normal distribution of data is one in which the majority of data points are relatively similar, meaning they occur within a small range of values with fewer outliers on the high and low ends of the data range.

When data are normally distributed, plotting them on a graph results a bell-shaped and symmetrical image often called the bell curve. In such a distribution of data, mean, median, and mode are all the same value and coincide with the peak of the curve.

Properties of the Normal Distribution

1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

Ans 11:

There are many ways in which we can handle missing data.

Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step.

NORMAL IMPUTATION

If the data is numerical, we can use mean and median values to replace else if the data is categorical, we can use mode which is a frequently occurring value.

IMPUTATION BASED ON CLASS LABEL

Here, instead of taking the mean, median, or mode of all the values in the feature, we take values based on class.

MODEL-BASED IMPUTATION

We take missing value as the class and all the remaining columns as features. Then we train our data with any model and predict the missing values.

Ans 12:

A/B Testing is a tried-and-true method commonly performed using a traditional statistical inference approach grounded in a hypothesis test (e.g. t-test, z-score, chi-squared test). In simple words two tests are run in parallel:

1. **Treatment Group (Group A)** – This group is exposed to the new web page, popup form, etc.
2. **Control Group (Group B)** – This group experiences no change from the current setup.

The goal of the A/B is then to compare the conversion rates of the two groups using statistical inference.

The situation is vastly more complex and dynamic. Consider these situations:

- **Users have different characteristics**: Different ages, genders, new vs returning, etc
- **Users spend different amounts of time on the website**: Some hit the page right away, others spend more time on the site
- **Users find your website differently**: Some come from email or newsletters, others from web searches, others from social media
- **Users take different paths**: Users take actions on the website going to different pages prior to being confronted with the event and goal

Ans 13:

Mean imputation does not preserve the relationships among variables.

Most research studies are interested in the relationship among variables, mean imputation is not a good solution.

This solution that is so good at preserving unbiased estimates for the mean isn't so good for unbiased estimates of relationships.

Mean Imputation Leads to An Underestimate of Standard Errors

Any statistic that uses the imputed data will have a standard error that's too low. In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.

Ans 14:

Linear regression

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

1. Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
2. Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are

1. Determining the strength of predictors.
2. Forecasting an effect.
3. Trend forecasting.

Ans 15:

The Branches of Statistics

Two branches, *descriptive statistics* and *inferential statistics*, comprise the field of statistics.

Descriptive Statistics

CONCEPT The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

Descriptive statistics forms the basis for analysis and discussion in such diverse fields as securities trading, the social sciences, government, the health sciences, and professional sports. A general familiarity and widespread availability of descriptive methods in many calculating devices and business software can often make using this branch of statistics seem deceptively easy.

Inferential Statistics

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (*AARP*), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.

In this we start with a hypothesis and look to see whether the data are consistent with that hypothesis. Inferential statistical methods can be easily misapplied or misconstrued, and many inferential methods require the use of a calculator or computer.