WORKSHEET-1

***Q.1: ans**. – a.*

***Q.2: ans**. – a.*

***Q.3: ans**. – b.*

***Q.4: ans**. – b.*

***Q.5: ans**. – c.*

***Q.6: ans**. – b.*

***Q.7: ans**. – b.*

***Q.8: ans**. – a.*

***Q.9: ans**. – c.*

***Q.10: ans**. - A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution. The normal distribution is often called the bell curve because the graph of its probability density looks like a bell. It is also known as **Gaussian distribution**, after the German mathematician **Carl Gauss** who first described it. We say the distribution of data is normal when it is symmetric around the mean, when +-68% of the data lies within one standard deviation of the mean, 95% within two standard deviation of the mean and 99.8% of it within three standard deviations of the mean.*

***Q.11: ans**. - Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical program will make the decision for you. Your application will remove things in a list wise sequence most of the time. Depending on why and how much data is gone, list wise deletion may or may not be a good idea. Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values. He following is some of the most prevalent methods:*

- ***MEAN IMPUTATION:***

Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

- ***SUBSTITUTION:***

Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

WORKSHEET-1

- ***HOT DECK IMPUTATION:***

A value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10.Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

- ***COLD DECK IMPUTATION:***

A value picked deliberately from an individual with similar values on other variables. In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

- ***REGRESSION IMPUTATION:***

The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilizing the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

- ***STOCHASTIC REGRESSION IMPUTATION:***

The predicted value of a regression plus a random residual value. This has all of the benefits of regression imputation plus the random component's benefits. The majority of multiple imputations are based on stochastic regression imputation.

- ***INTERPOLATION AND EXTRAPOLATION:***

  An estimate based on other observations made by the same person. It generally only works with data that is collected over time. Proceed with caution, though. For a variable like height in children– one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

- ***SINGLE OR MULTIPLE IMPUTATION:***

  - *Single and multiple imputations are the two forms of imputation. When people say imputation, they usually mean single.*
  - *The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.*
  - *It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.*

WORKSHEET-1

- *When list wise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.*
- *Unless the data is missing completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with list wise deletion, which is most software's default.*
- *The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.*

***Q.12: ans**. - 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 drives 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.*

The version that moves your business metric(s) in the positive direction is known as the 'winner.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

The metrics for conversion are unique to each website. For instance, in the case of e-commerce, it may be the sale of the products. Meanwhile, for B2B, it may be the generation of qualified leads.

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.

***Q.13: 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.

***Q.14: ans**. - Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:*

WORKSHEET-1

(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, and

(3) trend forecasting.

Types of Linear Regression:

I) Simple linear regression.
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

ii) Multiple linear regression.
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

iii) Logistic regression*.*
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

iv) Ordinal regression*.*
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

v) Multinomial regression*.*
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

vi) Discriminant analysis*.*
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability. Occam's razor describes the problem extremely well – a simple model is usually preferable to a more complex model. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

WORKSHEET-1

Q.15: 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.*

Descriptive Statistics:

Descriptive statistics *deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.*

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics:

Inferential statistics, *as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.*

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.