STATISTICS

WORKSHEET-1

Q1 TO Q9

1)A

2)A

3)B

4)D

5)C

6)B

7)B

8)A

9)C

Q10 TO Q15

Q10) What do you understand by the term Normal Distribution?

Ans) Normal distribution is also known as the Gaussian distribution. It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. Normal distribution will appear as a bell curve in the graph.

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. Normal distributions are symmetrical, but not all symmetrical distributions are normal. The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analysis. The standard normal distribution has two parameters: mean and Standard deviation.

The normal distribution model is motivated by the Central limit theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled. Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data

could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

The assumption of a normal distribution is applied to asset prices as well as price action. Traders may plot price points over time to fit recent price action into a normal distribution. The further price action moves from the mean, in this case, the more likelihood that an asset is being over or undervalued. Traders can use the standard deviations to suggest potential trades. This type of trading is generally done on very short time frames as larger timescales make it much harder to pick entry and exit points.

Similarly, many statistical theories attempt to model asset prices under the assumption that they follow a normal distribution. In reality, price distributions tend to have fat tails and, therefore, have kurtosis greater than three. Such assets have had price movements greater than three standard deviations beyond the mean more often than would be expected under the assumption of a normal distribution. Even if an asset has went through a long period where it fits a normal distribution, there is no guarantee that the past performance truly informs the future prospects.

A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean. Bell curves (normal distributions) are used commonly in statistics, including in analysing economic and financial data.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans) Missing data is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research.

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null

hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions.

The manufacture and service of complex equipment, honey well and its customers compile vast amounts of maintenance data. For a number of reasons, this data is plagued with errors and lacunae. We discuss the type of data with which we are working in this section. Honeywell and its customers routinely compile maintenance information for plant and building equipment installed in various locations. Entry in these data bases is carried out by field personnel, and for various reasons is plagued by a high proportion of missing data fields. In addition, the entered data is sometimes erroneous, or is in a non-standard format and frequently has spelling errors.

Imputation techniques:

There are two types of imputation–single or multiple. Usually when people talk about imputation, they mean single. Single refers to the fact that you come up with a single estimate of the missing value, using one of the seven methods listed above. It is popular because it is conceptually simple and because the resulting sample has the same number of observations as the full data set.

Some imputation methods result in biased parameter estimates, such as means, correlations, and regression coefficients, unless the data are Missing Completely at Random. The bias is often *worse* than with listwise deletion, the default in most software.

The extent of the bias depends on many factors, including the imputation method, the missing data mechanism, the proportion of the data that is missing, and the information available in the data set. Moreover, all single imputation methods underestimate standard errors.

So multiple imputation comes up with multiple estimates. Two of the methods listed above work as the imputation method in multiple imputation–hot deck and stochastic regression.

Because these two methods have a random component, the multiple estimates are slightly different. This re-introduces some variation that your software can incorporate in order to give your model accurate estimates of standard error.

Multiple imputation was a huge breakthrough in statistics about 20 years ago. It solves a lot of problems with missing data and if done well, leads to unbiased parameter estimates and accurate standard errors.

Q 12. What is A/B testing?

Ans) The concept of A/B testing (also known as bucket testing, controlled experiment, etc.) applied to websites and the Internet. The different variations of your website to different people and measure which variation is the most effective at turning them into customers. If each visitor to your website is randomly shown one of these variations and you do this over the same period, then you have created a controlled experiment known as an A/B test.

A/B testing is a shorthand for a simple controlled experiment, in which two samples (a &b) of a single vector-variable are compared. These values are similar except for one variation which might affect a user behaviour. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows. A/B test are useful for understanding user and satisfaction of online features like a social media.

A/B tests are being used also for conducting complex experiments on subjects such as network effects when users are offline, how online services affect user actions, and how users influence one another. Many professions use the data from A/B tests.

This includes data engineers, marketers, designers and software engineers. Many positions rely on the data from A/B tests, as they allow companies to understand growth, increase revenue and optimize customer satisfaction.

A/B testing is claimed by some to be a change in philosophy and business-strategy in certain niches, though the approach is identical to a between-subject design, which is commonly used in a variety of research traditions. A/B testing as a philosophy of web development brings the field into line with a broader movement toward evidence-based practice. The benefits of A/B testing are considered to be that it can be performed continuously on almost anything, especially since most marketing automation software now typically comes with the ability to run A/B tests.

Q13. Is mean imputation of missing data acceptable practice?

Ans) In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit. Some items are more likely to generate a nonresponse than others: for items about private subjects such as income. Attrition is a type of missingness that can occur in longitudinal studies—for instance studying development where a measurement is repeated after a certain period of time. Missingness occurs when participants drop out before the test ends and one or more measurements are missing.

Data often are missing in research in economics, sociology, and political science because governments or private entities choose not to, or fail to, report critical statistics, or because the information is not available. Sometimes missing values are caused by the researcher—for example, when data collection is done improperly or mistakes are made in data entry.

These forms of missingness take different types, with different impacts on the validity of conclusions from research: Missing completely at random, missing at random, and missing not at random.

Missing data in medical research is a common problem that has long been recognised by statisticians and medical researchers alike. In general, if the effect of missing data is not taken into account the results of the statistical analyses will be biased and the amount of variability in the data will not be correctly estimated. There are three main types of missing data pattern: Missing Completely. At Random (MCAR), Missing at Random (MAR) and Not Missing At Random (NMAR). The type of missing data that a researcher has in their dataset determines the appropriate method to use in handling the missing data before a formal statistical analysis begins. The aim of this practice note is to describe these patterns of missing data and how they can occur, as well describing the methods of handling them. Simple and more complex methods are described, including the advantages and disadvantages of each method as well as their availability in routine software. It is good practice to perform a sensitivity analysis employing different missing data techniques in order to assess the robustness of the conclusions drawn from each approach.

Q14. What is linear regression in statistics?

Ans) In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression, the process is called multiple linear regression. This is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are model using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Mostly commonly, the conditional mean of the response given the values of the explanatory variables is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the

Joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm, or by minimizing a penalized version of the least squares cost function as in ridge regression and lasso. Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other, but

that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

Q15. What are the various branches of statistics?

Ans) Statistics is a branch of Mathematics, that deals with the collection, analysis, interpretation, and the presentation of the numerical data. In other words, it is defined as the collection of quantitative data. The main purpose of Statistics is to make an accurate conclusion using a limited sample about a greater population.

Types of statistics

Statistics can be classified into two different categories. The two different types of Statistics are: Descriptive Statistics, Inferential Statistics

In Statistics, descriptive Statistics describe the data, Inferential Statistics whereas help you make predictions from the data. In inferential statistics, the data are taken from the sample and allows you to generalize the population. In general, inference means “guess”, which means making inference about something. So, statistical inference means, making inference about the population. To take a conclusion about the population, it uses various statistical analysis techniques. In this article, one of the types of statistics called inferential statistics is explained in detail. Now, you are going to learn the proper definition of statistical inference, types, solutions, and examples.

Statistical inference is the process of analysing the result and making conclusions from data subject to random variation. It is also called inferential statistics. Hypothesis testing and confidence Intervals are the applications of the statistical inference. Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to assess the relationship between the dependent and independent variables. The purpose of statistical inference to estimate the uncertainty or sample to sample variation.

It allows us to provide a probable range of values for the true values of something in the population.