WORKSHEET SET 1 ASSIGNMENT

STATISTICS

QUESTION 1. 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

ANSWER: a) 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

ANSWER: a) Central Limit Theorem

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

a) Modelling event/time data

b) Modelling bounded count data

c) Modelling contingency tables

d) All of the mentioned

ANSWER: b) Modelling 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

ANSWER: d) All of the mentioned

5.] _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

ANSWER: c) Poisson

WORKSHEET SET 1 ASSIGNMENT

STATISTICS

6.] 10. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

ANSWER: b) 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

ANSWER: b) 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

ANSWER: a) 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 mentioned

ANSWER: c) Outliers cannot conform to the regression relationship

QUESTION 2. are subjective answer type questions, Answer them in your own words briefly.

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

ANSWER:

Normal distribution is also known as the Gaussian distribution, 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.

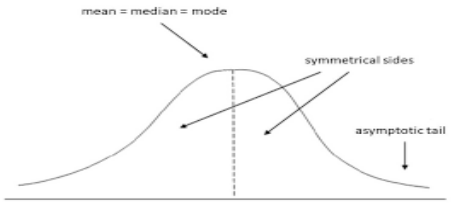
In graph form, normal distribution will appear as a bell curve.

WORKSHEET SET 1 ASSIGNMENT

STATISTICS

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the centre is a mirror image of the left side.

The area under the normal distribution curve represents probability and the total area under the curve sums to one.



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.

In reality, most pricing distributions are not perfectly normal.

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses.

The standard normal distribution has two parameters: the mean and the standard deviations For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

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 (provided it has finite variance).

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.

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

ANSWER:

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It’s most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

WORKSHEET SET 1 ASSIGNMENT STATISTICS

In Multiple imputation is another useful strategy for handling the missing data. In a multiple imputation, instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias.

Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Before deciding which approach to employ, data scientists must understand why the data is missing.

Best techniques to handle missing data

1. Use deletion methods to eliminate missing data. The deletion methods only work for certain datasets where participants have missing fields.
2. Use regression analysis to systematically eliminate data.
3. Data scientists can use data imputation techniques.

12.] What is A/B testing?

ANSWER:

A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

A/B testing is 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.

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 eCommerce, 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 optimizations (CRO), using which you can gather both qualitative and quantitative user insights.

WORKSHEET SET 1 ASSIGNMENT

STATISTICS

13.] Is mean imputation of missing data acceptable practice?

ANSWER:

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.

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

It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power.

But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.

On the other hand, there are many alternatives to the mean imputations that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

The mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

14.] What is linear regression in statistics?

ANSWER:

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple corelated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used.

The regression analysis, linear regression focuses on the conditional probability distributions 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 has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, forecasting or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables.

- After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

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 (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L^2-norm penalty) and lasso (L^1-norm penalty).

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.

15. What are the various branches of statistics?

ANSWER:

There are three real branches of statistics:

1. data collection,
2. descriptive statistics and
3. inferential statistics.

Data collection:

Data collection is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.

For data such as marks in a class test, this is fairly straightforward. Each student has a defined mark associated with them, so the marks are simply collected together to make the data set. Sometimes, data is harder to collect.

Counting the number of bees in a colony isn't easy, because they move and fly around; you may have to approximate in such cases. Also, if you are collecting data, you need to be careful where you get it from.

For example, suppose you want to conduct a poll on who people plan to vote for in an election. You can't realistically ask everyone in the whole country (the population), so you have to choose a representative sample of people. This isn't as easy as it sounds.

So, there are issues in the collection of the data; you need to make sure that the data has been collected fairly before you go on to deal with it, and try to present it and make conclusions.

WORKSHEET SET 1 ASSIGNMENT

STATISTICS

Descriptive statistics:

Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

Common visual techniques that include graphs, bar charts, pie charts and more, but we shall focus mainly on numerical techniques such as averages and spreads.

The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little to someone who sees it; it needs to be summarised.

In the 2010 General Election almost 30 million people voted. If each vote was simply written down and displayed, one after the other, you'd be totally lost; what happens is that a summary of votes is presented (for example as percentages: Conservative 36%, Labour 29%, Liberal Democrat 23%, Others 12%). This is an example of descriptive statistics – 'describing' or 'summarising' the overall data for people to understand.

Inferential statistics:

Inferential statistics is the aspect that deals with making conclusions about the data.

For example, a council might be considering altering the speed limit on a main road, after a number of accidents. They might do this by surveying the speeds of cars (data collection) and then arrive at a conclusion as to whether the speed limit needs to be lowered (if, for example, a number of cars are driving too fast).

Note, though, that this may not be the case; everyone might be driving at a perfectly acceptable speed, and the accidents are down to something other than speed (a blind spot or a pothole, for example). This is inferential statistics: take the data you have and make an 'inference' or 'conclusion' from it.