Statistics Worksheet

1. True
2. Central limit theorem
3. Modelling bounded count data
4. All of the mentioned
5. Poisson
6. False
7. Hypothesis
8. 0
9. Outliers cannot conform to the regression relationship

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

Ans) The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena.

For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

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 processing of the dataset as many machine learning algorithms do not support missing values.

Imputation Using (Mean/Median) Values:

This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

Imputation Using (Most Frequent) or (Zero/Constant) Values:

Most Frequent is another statistical strategy to impute missing values and YES!! It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column.

Imputation Using k-NN:

The k nearest neighbours is an algorithm that is used for simple classification. The algorithm uses ‘feature similarity’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This can be very useful in making predictions about the missing values by finding the k’s closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood.

12) What is A/B testing?

Ans) Like any type of scientific testing, A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference.

It is an analytical method for making decisions that estimates population parameters based on sample statistics.

The population refers to all the visitors coming to your website (or specific group of pages), while the sample refers to the number of visitors that participated in the test.

Let’s say, you make a decision to implement some change on your product pages based on A/B test results that tested a “sample” of the visitors to your website.

Ultimately, only a percentage of the visitors saw the challenger,

so that of course means not all the visitors. However, with A/B testing, you assume if the challenger (i.e. variation) in the test increased conversions for a group of visitors on product pages, it will thus have the same result for all the visitors of your product pages (we will delve into the accuracy of a variation’s validity later).

To recap, the A/B testing process can be simplified as follows:

You start the A/B testing process by making a claim (hypothesis).

You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.

The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

13) is mean imputation of missing data acceptable practice?

Ans) Perhaps that’s a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort.

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 mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

There are two problems

Problem 1: Mean imputation does not preserve the relationships among variables.

Problem 2: Mean Imputation Leads to An Underestimate of Standard Errors

14) What is linear regression in statistics?

Ans) 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 regress and. 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.

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.

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.

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 experiment 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.