STATISTICS WORKSHEET-1

Answers:

1. Bernoulli random variables take (only) the values 1 and 0.

   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

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

   b) Modelling bounded count data

4. Point out the correct statement.

   d) All of the mentioned

5. ______ random variables are used to model rates.

   c) Poisson

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

   b) False

7. Which of the following testing is concerned with making decisions using data?

   b) Hypothesis

8. Normalized data are centred at ______and have units equal to standard deviations of the original data.

   a) 0

9. Which of the following statement is incorrect with respect to outliers?

   c) Outliers cannot conform to the regression relationship

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

    The Normal distribution is also called Gaussian distribution.
    The normal distribution is the most commonly seen continuous distribution in nature. Just as Binomial distribution, every event is independent from one another. In the normal distribution the mean, median and mode all line up such that the centre of the distribution is the mean. Because of this, exactly half of the results fall either side of the mean. The Normal distribution is also identifiable by its bell shape and may sometimes be referred as a bell curve.

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

    Real-world data often has missing values. Data can have missing values for a number of reasons such as observations that were not recorded and data corruption. Handling missing data is important as many machine learning algorithms do not support data with missing values.
    The simplest imputation technique we would use is replacing missing values with the mean or median values of the dataset at large, or some similar summary statistic. This has the advantage of being the simplest possible approach, and one that doesn't introduce any undue bias into the dataset.

    The followings are the other imputation techniques we can apply to handle missing data:

Substitution

Impute the value from a new individual who was not selected to be in the sample. In other words, go find a new subject and use their value instead.

Hot deck imputation

A randomly chosen value from an individual in the sample who has similar values on other variables. In other words, find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable. One advantage is you are constrained to only possible values. In other words, if Age in our study is restricted to being between 5 and 10, we will always get a value between 5 and 10 this way. Another is the random component, which adds in some variability. This is important for accurate standard errors.

Cold deck imputation

A systematically chosen value from an individual which has similar values on other variables. This is similar to Hot Deck in most ways, but removes the random variation. So for example, we may always choose the third individual in the same experimental condition and block.

Regression imputation

The predicted value obtained by regressing the missing variable on other variables. So instead of just taking the mean, we're taking the predicted value, based on other variables. This preserves relationships among variables involved in the imputation model, but not variability around predicted values.

Stochastic regression imputation

The predicted value from a regression plus a random residual value. This has all the advantages of regression imputation but adds in the advantages of the random component. Most multiple imputations are based off of some form of stochastic regression imputation.

Interpolation and extrapolation

An estimated value from other observations from the same individual. It usually only works in longitudinal data. Interpolation, for example, might make more sense for a variable like height in children that cannot go back down over time. Extrapolation means we are estimating beyond the actual range of the data and that requires making more assumptions that we should.

12. What is A/B testing?

A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

13. Is mean imputation of missing data acceptable practice?

Yes, 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. That's a good thing, since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

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.

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?

The two main branches of statistics are descriptive statistics and inferential statistics.

1. Descriptive Statistics: If data can be described without any statistical tools then it is called descriptive statistics. Ex: marks in class, height of student.
2. Inferential Statistics: If data is too big then we use inferential statistics,

   We take a few samples from different data and we find the average. This is called inferential statistics. The average is then applicable to all the data from where we have selected our samples.