

STATISTICS WORKSHEET-1

Q1 to Q9 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
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
3. Which of the following is incorrect with respect to use of Poisson distribution?
 - a) Modeling event/time data
 - b) **Modeling bounded count data**
 - c) Modeling contingency tables
 - d) All of the mentioned
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**
5. _____ random variables are used to model rates.
 - a) Empirical
 - b) Binomial
 - c) **Poisson**
 - d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
 - a) True
 - 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
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
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

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Normal distribution, 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 graphical form, the normal distribution appears as a "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.

Formula for normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

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

One way of handling missing values is the **deletion of the rows or columns having null values**. If any columns have more than half of the values as null then you can drop the entire column. In the same way, rows can also be dropped if having one or more columns values as null.

Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

Most commonly used imputation Techniques are:

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.

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.

Old 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 utilising 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 imputation is 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 imputation 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.

12. What is A/B testing?

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 essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean. This method can lead into severely biased estimates even if data are MCAR. That is why, Mean imputation is typically considered terrible practice since it ignores feature correlation.

14. What is linear regression in statistics?

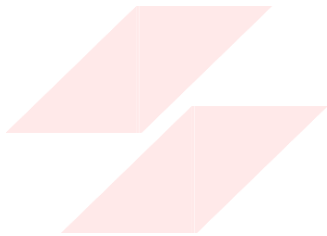
Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

- does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
- 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 $y = c + b \cdot x$, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15. What are the various branches of statistics?

There are three real branches of statistics: **data collection, descriptive statistics and inferential statistics.**



FLIP ROBO