Jane Chinwuko
Project: Batch DS2307
PFA Worksheet Set 1

Statistics.

Q1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

Answer is A

The simplest type of random variable is a Bernoulli random variable. It has two possible values: 1 and 0

Q2. Which of the following theorem states that the distribution of averages of iid variables, properly normalised, 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 is A

The sampling distribution of the mean will always have a normal distribution if the sample size is large enough, according to the central limit theorem.

Q3. 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

Answer is B

Unbounded count data is modelled using the Poisson distribution.

Q4. 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 is D

Q5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer is C

The Poisson random variable "x" determines the experiment's success rate.

Q6. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

Answer is B

Normaly, the CLT is unaffected by changing the standard error's predicted value.

Q7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer is B

Hypothesis testing uses data from a sample to make inferences about a population parameter or population probability distribution.

Q8. Normalised data are centered at_____and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

Answer is A

Q9. 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 is C

Outliers does conform to the regression relationship.

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

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

Normal distribution is a probability distribution for independent, random variables in statistics. Its graph appears as a symmetrical bell curve. The mean, median and mode of a normal distribution are equal. Normal distribution is also known as Gaussian distribution.

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

Missing data is when there are no records for an important variable in a dataset. Missing data can affect the results of any data analysis or the reliability of machine learning models, depending on their volume.

Missing data can be handled using different methods, such as:

1) **By deleting these missing values. This is not recommended because one can delete important data from the dataset. The entire row or column is deleted from the dataset.**
2) **Imputing the missing values. Here the missing values are replaced with some arbitrary value.**

There are different imputation techniques for handling missing data.

1) **By replacing with the mean-Using the mean of the respective column.**
2) **By replacing with the mode- Mostly used for categorical features.**
3) **By replacing with the median- Better used in the case of outliers.**
4) **By replacing with the previous value- Also known as forward fill; used in time series data.**
5) **By replacing with the next value- Also known as backward fill.**
6) **Using Interpolation-Uses interpolate method from pandas such as linear, polynomial and quadratic.**

I recommend using either mean, median or mode techniques for missing values. These are the most widely used methods of imputing missing data when there are a few missing observations.

Q12. What is A/B testing?

An A/B testing is a type of statistical hypothesis test in which a relationship between two sets of data is hypothesised, and the two sets are then compared to see if the association is statistically significant or not.

Q13. Is mean imputation of missing data acceptable practice?

Mean imputation is not considered the best practice because it reduces the variance of the imputed variable. It also ignores the distribution and correlation of data and creates unrealistic values.

Q14. What is linear regression in statistics?

Linear regression is a regression model that uses a straight line to evaluate the relationship between one independent variable value (x) and a dependent variable value (y). The line that fits our model the best is the regression line.

Q15. What are the various branches of statistics?

There are two main branches of statistics: descriptive and inferential.

Descriptive statistics handles the presentation and collection of data.

Inferential statistics involves making the appropriate inferences from a statistical analysis that has been carried out using descriptive statistics.