STATISTICS_Worksheet_1

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

a) True b) False

Answer1-: True (A Bernoulli random variable is the simplest kind of random variable. It can take on two values, 1 and 0. It takes on a 1 if an experiment with probability resulted in success and a 0 otherwise.)

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

Answer2-: Central Limit Theorem , (the CENTRAL LIMIT THEOREM (CLT) - one of the most important theorems in all of statistics. It states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases.)

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

Answer3-: Modeling bounded count data. (Poisson distribution is used for modeling unbounded 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

Answer4-: The correct choice is (d) All of the mentioned (Many random variables, properly normalized, limit to a normal distribution.)

5. __ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer5-: C, Poisson (Poisson distribution is used to model counts.)

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

a) True

b) False

Answer6-: False (Usually replacing the standard error by its estimated value doesn't change the CLT)

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer7-: B, Hypothesis (The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis.)

8. Normalized data are centered at____ and have units equal to standard deviations of the original data.

a) 0   b)   c) 1   d) 10

Answer8-: A, 0 (In statistics and applications of statistics, normalization can have a range of meanings)

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

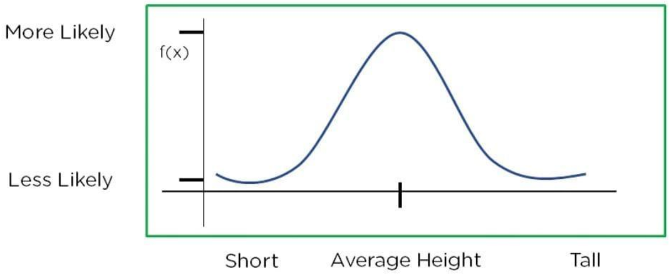
c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Answer9-: C, Outliers can conform to the regression relationship.

10. What do you understand by normal distribution?

Answer10-: A normal distribution is the continuous probability distribution with a probability density function that gives you a symmetrical bell curve. Simply put, it is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends.



Normal Distribution

From the above graph, the distribution is mostly about the mean or the average of all heights. Apart from this, most data is around the mean. On moving away, the probability density decreases too. This kind of curve is called a Bell Curve, and it is a common feature of a normal distribution.

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

Answer11-: Missing data can be dealt with in a variety of ways. The most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that our statistical programme will make the decision for us. The application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy is imputation. Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values.

The following are some of the recommended methods:

Mean imputation

Calculating 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.

Substitution

Assuming 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.

12. What is A/B testing?

Answer12-: 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.

13. Is mean imputation of missing data acceptable practice?

Answer13-: The process of replacing null values in a data collection with the data’s mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Answer14-: In statistics, linear regression is a linear approach for modeling 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.

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

15. What are the various branches of statistics?

Answer15-: Descriptive statistics and inferential statistics are the two main branches of statistics. Descriptive statistics mainly involves the collection and presentation of data. Inferential statistics deals with inferring the right conclusions from the analysis performed using descriptive statistics.