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

Ans- 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

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans- a) Central Limit Theorem

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

a) Modelling event/time data

b) Modelling bounded count data

c) Modelling contingency tables

d) All of the mentioned

Ans- b) Modelling bounded 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

Ans- d) All of the mentioned

5. _______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans- c) Poisson

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

a) True

b) False

Ans- 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

Ans- b) Hypothesis

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

Ans- a) 0

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

Ans- c) Outliers cannot conform to the regression relationship

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

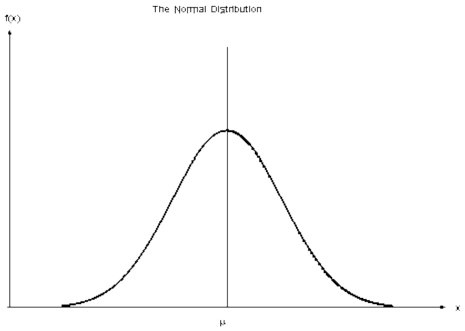
Ans- The normal distribution is also known as Gaussian Distribution, as it is the most important probability distribution. Because of the shape of its graph, it is also called Bell-shaped curve.

Normal distribution is a continuous probability distribution that is symmetrical around its mean. From the mean, the probability values are equally taper off in both directions.

Properties of a normal distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center.
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

Basic diagram-



Formula: $p(x) = e^{-(x - \mu)2/2\sigma2}/\sigma$Square root of$\sqrt{}2\pi$.

F(x)= Probability density functions

σ = Standard Deviation

μ = Mean

Examples = heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

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

Ans- As per the name, its clearly stated that missing data is the data which appear when no value is available in one or more variables of an individual. Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.

Missing data can be handled in many ways. The first and most common way it to ignore it. Statistical programme will automatically takes the decision on your behalf if we make no decisions.

There are some common ways to handle the missing data if we chose not to ignore it. These methods are-

Listwise or case deletion- The most common approach to the missing data is to simply omit those cases with the missing data and analyze the remaining data. This approach is known as the listwise deletion.

Pairwise deletion- Pairwise deletion eliminates information only when the particular data-point needed to test a particular assumption is missing. If there is missing data elsewhere in the data set, the existing values are used in the statistical testing. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion, which may delete the case with any missing data.

Mean substitution- The mean value of a variable is used in place of the missing data value for that same variable. This allows the researchers to utilize the collected data in an incomplete dataset.

However, with missing values that are not strictly random, especially in the presence of a great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias.

Regression imputation- In regression imputation, the existing variables are used to make a prediction, and then the predicted value is substituted as if an actual obtained value. This approach has a number of advantages, because the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution.

Last observation carried forward- This method replaces every missing value with the last observed value from the same subject. Whenever a value is missing, it is replaced with the last observed value.

Maximum likelihood- In these, the assumption that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand. After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.

When there are missing but relatively complete data, the statistics explaining the relationships among the variables may be computed using the maximum likelihood method. That is, the missing data may be estimated by using the conditional distribution of the other variables.

Expectation-Maximization- Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods .

This approach begins with the expectation step, during which the parameters (e.g., variances, covariances, and means) are estimated, perhaps using the listwise deletion. Those estimates are then used to create a regression equation to predict the missing data. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to "fill in" the missing data. The expectation and maximization steps are repeated until the system stabilizes, when the covariance matrix for the subsequent iteration is virtually the same as that for the preceding iteration.

Multiple imputation- Multiple imputation is another useful strategy for handling the missing data. In a multiple imputation, instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values.

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.

There are 7 imputation techniques-

- Mean imputation. Simply calculate the mean of the observed values for that variable for all individuals who are non-missing. ...
- Substitution. ...
- Hot deck imputation. ...
- Cold deck imputation. ...
- Regression imputation. ...
- Stochastic regression imputation. ...
- Interpolation and extrapolation.

12. What is A/B testing?

Ans- A/B testing, also known as split testing, is the process of comparing two versions of a web page, email, or other marketing asset and measuring the difference in performance.

It is a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behaviour. A/B tests are widely considered the simplest form of controlled experiment.

You do this giving one version to one group and the other version to another group. Then you can see how each variation performs. Think of it like a competition. You’re pitting two versions of your asset against one another to see which comes out on top. Knowing which marketing asset works better can help inform future decisions when it comes to web pages, email copy, or anything else.

13. Is mean imputation of missing data acceptable practice?

Ans- Mean imputation of missing data is not as much acceptable practice.

Since mean imputation ignores correlation, its consider as typical practice.

1- Mean substitution leads to bias in multivariate estimates such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2- Standard errors and variance of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate.
3- If the response mechanism is MAR or MNAR, even the sample mean of your variable is biased . Assume that you want to estimate the mean of a population's income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.

14. What is linear regression in statistics?

Ans- Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data.

Linear regression is an important tool in analytics. The technique uses statistical calculations to plot a trend line in a set of data points. The trend line could be anything from the number of people diagnosed with skin cancer to the financial performance of a company. Linear regression shows a relationship between an independent variable and a dependent variable being studied.

There are a number of ways to calculate linear regression. One of the most common is the ordinary least-squares method, which estimates unknown variables in the data, which visually turns into the sum of the vertical distances between the data points and the trend line.

The calculations to perform linear regressions can be quite complex. Fortunately, linear regression models are included in most major calculations packages, such as Excel, R, MATLAB and Mathematica.

Linear regression is a way to model the relationship between two variables. You might also recognize the equation as the slope formula. The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

15. What are the various branches of statistics.

Ans- There are two real branches of statistics: **Descriptive statistics and Inferential statistics.**

Descriptive Statistics- Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis.

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.

After collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable or the relation between two variables (e.g., age and creativity).

Inferential Statistics - Inferential statistics helps to conclusions and make predictions based on your data.

When you have collected data from a sample, you can use inferential statistics to understand the larger population from which the sample is taken.

Inferential statistics have two main uses:

- making estimates about populations (for example, the mean SAT score of all 11th graders in the US).
- testing hypotheses to draw conclusions about populations (for example, the relationship between SAT scores and family income).