Statistics worksheet - 1

Choose the correct option to answer your question.

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

Ans: a. True.

Explanation:

The Bernoulli distribution arises as the result of a binary outcome.

2. Which of the following thoerem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Ans: a. Central limit thoerem. (CLT)

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

Ans: b. Modelling bounded count data.

Explanation: Poisson distribution is used for modelling unbounded count data.

4. Point out the correct statement.

Ans: d. All of the mentioned.

Explanation: Many random variables, properly normalized, limit to a normal distribution.

5. _____ random variables are used to model rates.

Ans: c. Poisson.

Explanation: Poisson distribution is used to model counts.

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

Ans: b. False.

Explanation: Usually replacing the standard error by its estimated value doesnot changes the CLT.

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

Ans: b. Hypothesis.

Explanation: 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 orginal data.

Ans: a. 0

Explanation: 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?

Ans: c. Outliers cannot conform to the regression relationship.

Explanation: Outliers can conform to the regression relationship.

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 graph form, normal distribution will appear as a bell curve. That's why it's widely used in business, statistics, and in government bodies like the FDA:

- Heights of people.
- Measurement errors.
- Blood pressure.
- Points on a test.
- IQ scores.
- Salaries.

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

There are a lot of techniques to treat missing value. The best way to organize some of the most commonly used methods, if you use SAS to implement it -

- **Ignore the records with missing values.**

Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

- **Substitute a value such as mean.**

When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this

method could cause bias distribution and variance. That's where the following imputation methods come in.

- **Predict missing values.**

Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

Logistic Regression
Discriminant Regression
Markov Chain Monte Carlo (MCMC)

- **Predict missing values - Multiple Imputation.**

Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

In addition, there are a few required statistical assumptions for multiple imputation.

Whether the data is missing at random (MAR).

Multivariate normal distribution, for some of the modeling methods mentioned above (e.g. regression, MCMC).

At last, to report -

The type of imputation algorithm used.

Some justification for choosing a particular imputation method.

The proportion of missing observations.

The number of imputed datasets (m) created.

The variables used in the imputation model.

12. What is A/B testing?

A/B testing is an optimisation technique often used to understand how an altered variable affects audience or user engagement. It's a common method used in marketing, web design, product development, and user experience design to improve campaigns and goal conversion rates.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is the practice of replacing null values in a data set with the mean of the data.

Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.

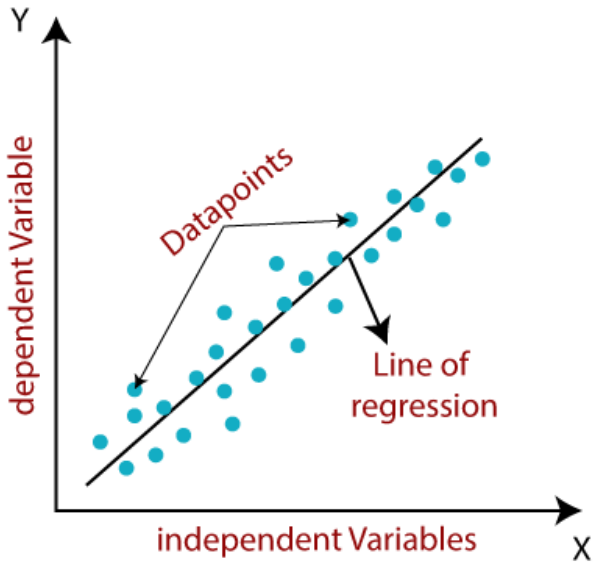
Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

14. What is linear regression in statistics?

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



15. What are the various branches of statistics?

Two branches, descriptive statistics and inferential statistics, comprise the field of statistics.

Descriptive Statistics

The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES:

The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

Inferential Statistics

The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE:

A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.