

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
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
3. Which of the following is incorrect with respect to use of Poisson distribution?  
Modeling bounded count data
4. Point out the correct statement.  
a) All of the mentioned
5. \_\_\_\_\_ random variables are used to model rates.  
Poisson
6. Usually replacing the standard error by its estimated value does change the CLT.  
False
7. Which of the following testing is concerned with making decisions using data?  
Hypothesis
8. Normalized data are centered at \_\_\_\_\_ and have units equal to standard deviations of the original data.  
a) 0
9. Which of the following statement is incorrect with respect to outliers?

Outliers cannot conform to the regression relationship

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?

The normal distribution, also known as the Gaussian distribution, it is the most important probability distribution in statistics for independent random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal. For example, the Student's t, Cauchy, and logistic distributions are symmetric.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, 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?

Use deletion methods to eliminate missing data

The deletion methods only work for certain datasets where participants have missing fields.

There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organisations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

Use regression analysis to systematically eliminate data

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

Average imputation and common-point imputation are recommendable techniques. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilizing this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

12. What is A/B testing?

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, 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?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The

---

variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. What are the various branches of statistics?

- There are three real branches of statistics:
1. Data collection
  2. Descriptive statistics and
  3. Inferential statistics.

**Data Collection**-Data collection is all about how the actual data is collected.

**Descriptive statistics** is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

**Inferential statistics** is the aspect that deals with making conclusions about the data.

