STATISTICS WORKSHEET-1

Multiple Choice Questions:

Q1) Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

Ans: A) True.

Q2) 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.

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

Ans: (b) Modeling bounded count data.

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
Ans: (d) All of the mentioned

Q5). _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
Ans: (c) Poisson

Q6). Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False
Ans: (b) False

Q7). 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.

Q8). 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

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
Ans: c) Outliers cannot conform to the regression relationship

Q 10 to Q 15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans: The normal distribution, also known as the Gaussian distribution, 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.

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

Ans: Best techniques to handle missing data:

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.

Data scientists can use data imputation techniques

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. 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 utilize 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).

Q12). What is A/B testing?

Ans: A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

Q13). Is mean imputation of missing data acceptable practice?

Ans: True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing. Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too. This is the original logic involved in mean imputation.

Q14). What is linear regression in statistics?

Ans: 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.

Q15). What are the various branches of statistics?

Ans: The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive Statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential Statistics

Inferential Statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.