STATISTICS WORKSHEET-1

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

a) True

b) False

SOLUTION: A

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

SOLUTION: A

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

SOLUTION: B

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

SOLUTION: D

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

SOLUTION: C

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

a) True

b) False

SOLUTION: B

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

SOLUTION: B

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

a) 0

b) 5

c) 1

d) 10

SOLUTION: A

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

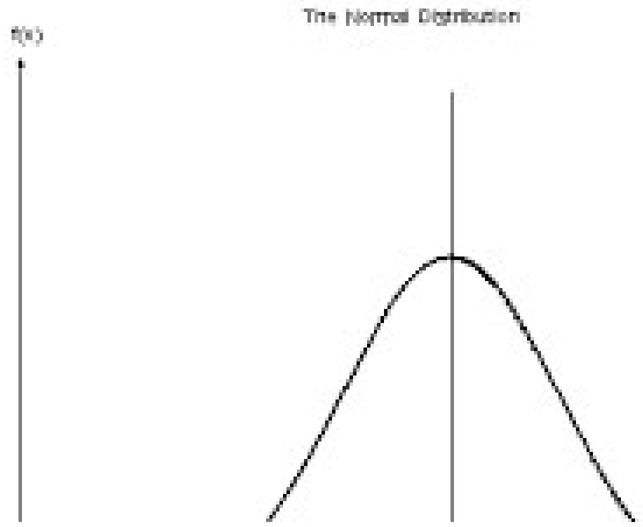
c) Outliers cannot conform to the regression relationship

d) None of the mentioned

SOLUTION: C

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

ANSWER:



Normal distribution, also known as the Gaussian distribution, is a continuous 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**. In simple terms, if a probability distribution forms a bell-shaped curve and mean, median and mode of the sample are equal then the variable has a normal distribution.**

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

ANSWER: When **dealing with missing data**, data scientists can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required. Before deciding which approach to employ, data scientists must understand why the data is missing.

Missing at Random (MAR): Missing at Random means the data is missing relative to the observed data. It is not related to the specific missing values. The data is not missing across all observations but only within sub-samples of the data. It is not known if the data should be there; instead, it is missing given the observed data. The missing data can be predicted based on the complete observed data.

Missing Completely at Random (MCAR): In the MCAR situation, the data is missing across all observations regardless of the expected value or other variables. Data scientists can compare two sets of data, one with missing observations and one without. Using a t-test, if there is no difference between the two data sets, the data is characterized as MCAR.Data may be missing due to test design, failure in the observations or failure in recording observations. This type of data is seen as MCAR because the reasons for its absence are external and not related to the value of the observation. It is typically safe to remove MCAR data because the results will be unbiased. The test may not be as powerful, but the results will be reliable.

Missing Not at Random (MNAR): The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing. In a survey, perhaps a specific group of people – say women ages 45 to 55 – did not answer a question. Like MAR, the data cannot be determined by the observed data, because the missing information is unknown. Data scientists must model the missing data to develop an unbiased estimate. Simply removing observations with missing data could result in a model with bias.

Deletion:

There are two primary methods for deleting data when dealing with missing data: list wise and dropping variables.

List wise:

In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, it may be the most efficient method to eliminate those cases from the analysis. However, in most cases, the data are not missing completely at random (MCAR). Deleting the instances with missing observations can result in biased parameters and estimates and reduce the statistical power of the analysis.

Pair wise:

Pair wise deletion assumes data are missing completely at random (MCAR), but all the cases with data, even those with missing data, are used in the analysis. Pair wise deletion allows data scientists to use more of the data. However, the resulting statistics may vary because they are based on different data sets. The results may be impossible to duplicate with a complete set of data.

Dropping Variables:

If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

Imputation techniques:

When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

Mean, Median and Mode

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

Time-Series Specific Methods

Another option is to use time-series specific methods when appropriate to impute data. There are four types of time-series data:

- No trend or seasonality.
- Trend, but no seasonality.
- Seasonality, but no trend.
- Both trend and seasonality.

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)

These options are used to analyze longitudinal repeated measures data, in which follow-up observations may be missing. In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

Linear Interpolation

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

When dealing with missing data, you should use this method in a time series that exhibits a trend line, but it's not appropriate for seasonal data.

Seasonal Adjustment with Linear Interpolation

When dealing with data that exhibits both trend and seasonality characteristics, use seasonal adjustment with linear interpolation. First you would perform the seasonal adjustment by computing a centered moving average or taking the average of multiple averages – say, two one-year averages – that are offset by one period relative to another. You can then complete data smoothing with linear interpolation as discussed above.

Multiple Imputations

Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

K Nearest Neighbors

In this method, data scientists choose a distance measure for k neighbors, and the average is used to impute an estimate. The data scientist must select the number of nearest neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

12. What is A/B testing?

ANSWER:

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools**.**

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

ANSWER:

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14. What is linear regression in statistics?

ANSWER:

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (Label or dependent variable) and one or more exploratory variables (Features or response or independent variables). The case of one explanatory variable is called a simple linear regression. For more than one explanatory variable or response, the process is called multiple linear regressions. 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. Most commonly, the conditional mean of the response given the values of the explanatory variables (response or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors.

15. What are the various branches of statistics?

ANSWER:

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.

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. Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through 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.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.