Statistics Assignment 6 Solution

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

10. What is the difference between a boxplot and histogram?

Ans. A boxplot and histogram are both graphical representations of a set of data, but they differ in their focus and the information they convey.

A histogram is a graphical representation of the distribution of a continuous variable. It is a graph that displays the frequency or proportion of data points that fall within a series of intervals or "bins" of the variable's range. The x-axis shows the range of the variable, divided into the intervals or bins, while the y-axis shows the frequency or proportion of data points that fall within each bin. Histograms provide information about the shape, center, and spread of the distribution of the variable.

A boxplot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a continuous variable. It provides a summary of the minimum, maximum, median, quartiles, and potential outliers of the data. A boxplot displays the range of the variable using a box, where the top and bottom of the box represent the third and first quartiles, respectively, and the line inside the box represents the median. The whiskers extend from the box to the minimum and maximum values of the data, excluding potential outliers, which are shown as points or asterisks. Boxplots provide information about the shape, center, and spread of the distribution of the variable, as well as the presence of potential outliers.

The main difference between a histogram and a boxplot is that a histogram provides information about the frequency or proportion of data points that fall within intervals or bins of the variable, while a boxplot provides a summary of the minimum, maximum, median, quartiles, and potential outliers of the data. Histograms are useful for examining the shape of the distribution and identifying any patterns or clusters within the data, while boxplots are useful for summarizing the spread and identifying any potential outliers in the data.

11. How to select metrics?

Ans. Selecting the appropriate metrics is an important step in evaluating the performance of a machine learning model. The choice of metrics should be based on the specific problem and the objective of the model. Here are some common metrics used in machine learning and some guidelines for selecting them:

1. Classification problems: For binary classification problems (where the target variable has two possible values), common metrics include accuracy, precision, recall, F1 score, and

ROC-AUC. For multiclass classification problems (where the target variable has more than two possible values), common metrics include multiclass accuracy, macro- and micro-averaged precision, recall, and F1 score.

- Accuracy: measures the proportion of correctly classified instances out of the total instances. It is a good metric to use when the classes are balanced.
- Precision: measures the proportion of true positives among all positive predictions. It is a good metric to use when false positives are costly.
- Recall: measures the proportion of true positives among all actual positives. It is a good metric to use when false negatives are costly.
- F1 score: is the harmonic mean of precision and recall. It is a good metric to use when both false positives and false negatives are important.
- ROC-AUC: measures the area under the receiver operating characteristic curve, which plots the true positive rate against the false positive rate. It is a good metric to use when the cost of false positives and false negatives is different.

2. Regression problems: For regression problems, common metrics include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), R-squared, and adjusted R-squared.

- MSE: measures the average squared difference between the predicted and actual values. It is sensitive to outliers.
- RMSE: is the square root of the MSE. It is a good metric to use when you want to penalize larger errors more than smaller errors.
- MAE: measures the average absolute difference between the predicted and actual values. It is less sensitive to outliers.
- R-squared: measures the proportion of variance in the target variable that is explained by the model. It is a good metric to use when you want to compare the goodness of fit of different models.
- Adjusted R-squared: is a modification of R-squared that takes into account the number of predictors in the model. It is a good metric to use when you want to penalize the presence of unnecessary predictors in the model.

In summary, selecting the appropriate metrics in machine learning depends on the specific problem and the objective of the model. It is important to consider the trade-offs between different metrics and choose the ones that are most relevant to the problem at hand.

12. How do you assess the statistical significance of an insight?

Ans. To assess the statistical significance of an insight or a finding, you can use hypothesis testing, which involves comparing the observed data to a null hypothesis and determining the probability of observing the data given the null hypothesis.

Here are the general steps for hypothesis testing:

3. Formulate the null hypothesis and alternative hypothesis: The null hypothesis represents the status quo or the default assumption, while the alternative hypothesis represents the claim or the insight that you want to test.
4. Choose a significance level: The significance level, denoted by alpha, is the probability of rejecting the null hypothesis when it is actually true. Common values of alpha are 0.05 and 0.01.
5. Select a test statistic: The test statistic is a numerical value that summarizes the data and is used to determine the probability of observing the data given the null hypothesis.
6. Calculate the p-value: The p-value is the probability of observing a test statistic as extreme or more extreme than the observed one, assuming the null hypothesis is true.
7. Compare the p-value to the significance level: If the p-value is less than the significance level, then the observed data is considered statistically significant, and the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, the observed data is not statistically significant, and the null hypothesis is not rejected.
8. Interpret the results: If the null hypothesis is rejected, you can conclude that the insight or finding is statistically significant at the chosen significance level. If the null hypothesis is not rejected, you cannot conclude that the insight or finding is statistically significant, and you should investigate further or collect more data.

Note that the above steps are general and the specific test and procedures used for hypothesis testing can vary depending on the type of data, the research question, and the assumptions of the statistical test. It is also important to consider the practical significance and the effect size of the insight or finding, in addition to the statistical significance.

13. Give examples of data that doesnot have a Gaussian distribution, nor log normal.

Ans. There are many real-world datasets that do not have a Gaussian (normal) distribution or a log-normal distribution. Here are some examples:

9. Power law distributions: Power law distributions are characterized by a small number of observations with very large values and a large number of observations with small values. Examples include the distribution of wealth, city sizes, and the number of citations in scientific publications.
10. Exponential distributions: Exponential distributions are characterized by a constant probability of an event occurring at any point in time. Examples include the distribution of time between radioactive decay events or the distribution of time between consecutive customers arriving at a store.
11. Poisson distributions: Poisson distributions are characterized by the number of events occurring in a fixed interval of time or space, where the events are independent and rare. Examples include the number of phone calls received by a call center in an hour or the number of cars passing through a toll booth in a minute.

12. Bimodal distributions: Bimodal distributions are characterized by having two distinct peaks. Examples include the distribution of height in a population, where there may be two distinct groups of people (e.g., children and adults) with different average heights.
13. Uniform distributions: Uniform distributions are characterized by a constant probability of a value occurring within a range. Examples include the distribution of lottery numbers or the distribution of a coin flip.

Note that these are just a few examples, and there are many other types of distributions that can be found in real-world datasets. The choice of an appropriate probability distribution can depend on the characteristics of the data, the research question, and the assumptions of the statistical model.

14. Give an example where the median is a better measure than the mean.

Ans. The median is often a better measure than the mean when dealing with skewed data or outliers that can distort the average value. Here is an example:

Suppose you are looking at the salaries of employees in a company, and there is one executive who earns $10 million per year, while the rest of the employees earn an average of $50,000 per year. If you calculate the mean salary, it would be skewed upwards by the executive's salary, and it would not be a representative measure of the typical salary in the company. In this case, the median salary (the value that divides the dataset into two equal halves) would be a better measure of the typical salary, as it is not affected by extreme values or outliers. In this example, the median salary would be $50,000, which is a more accurate representation of the typical salary in the company than the mean of $1,000,000.

In general, the median is a more robust measure of central tendency than the mean, as it is not influenced by extreme values or outliers. The choice of an appropriate measure of central tendency depends on the nature of the data and the research question.

15. What is the Likelihood?

Ans. In statistics, likelihood is a term used to describe the probability of observing a set of data, given a certain model or parameter values. The likelihood function is a function of the parameters of a statistical model, which describes how likely the observed data are to have been generated by that model.

For example, if we have a set of data and a model that predicts the probability distribution of the data, we can calculate the likelihood of observing the data given the model parameters. The likelihood function can then be used to estimate the best values for the model parameters that fit the observed data.

The likelihood function is often denoted by the symbol $L(\theta|data)$, where θ represents the parameters of the model and data represents the observed data. The likelihood function is

typically maximized to estimate the best values of the model parameters, a process known as maximum likelihood estimation.

The likelihood function is closely related to the concept of probability, but there is a subtle difference. Probability is the likelihood of an event occurring, given the known values of the parameters. In contrast, likelihood is the likelihood of observing the data, given a particular set of parameter values. The likelihood function can be used to update our beliefs about the values of the parameters given the observed data, and it plays an important role in many statistical models and methods.