ASSIGNMENT-3

STATISTICS WORKSHEET-3

1. D
2. C
3. 2
4. A
5. A
6. C
7. B
8. D
9. A

10. Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability

Bayes Theorem Statement

Let E_1, E_2,…, E_n be a set of events associated with a sample space S, where all the events E_1, E_2,…, E_n have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

| |�(��|�) = �(��)�(�|��)∑�=1��(��)�(�|��)
for any k = 1, 2, 3, …., n

Bayes Theorem Proof

According to the conditional probability formula,

|�(��|�) = �(��∩�)�(�)…(1)
Using the multiplication rule of probability,

|�(��∩�) = �(��)�(�|��)…(2)
Using total probability theorem,

�(�) = ∑�=1� �(��)�(�|��)…(3)
Putting the values from equations (2) and (3) in equation 1, we get

| |�(��|�) = �(��)�(�|��)∑�=1� �(��)�(�|��)

Note:

The following terminologies are also used when the Bayes theorem is applied:

Hypotheses: The events E_1, E_2,… E_n is called the hypotheses

Priori Probability: The probability P(E_i) is considered as the priori probability of hypothesis E_i

Posteriori Probability: The probability P(E_i|A) is considered as the posteriori probability of hypothesis E_i

Bayes' theorem is also called the formula for the probability of "causes". Since the E_i's are a partition of the sample space S, one and only one of the events E_i occurs (i.e. one of the events E_i must occur and the only one can occur). Hence, the above formula gives us the probability of a particular E_i (i.e. a "Cause"), given that the event A has occurred.

Bayes Theorem Formula

If A and B are two events, then the **formula for the Bayes theorem** is given by:

�(�|�)=�(�|�)�(�)�(�)�*h*����(�)≠0

Where P(A|B) is the probability of condition when event A is occurring while event B has already occurred.

Bayes Theorem Derivation

Bayes Theorem can be derived for events and random variables separately using the definition of conditional probability and density.

From the definition of conditional probability, Bayes theorem can be derived for events as given below:

$P(A|B) = P(A \cap B)/ P(B)$, where $P(B) \neq 0$

$P(B|A) = P(B \cap A)/ P(A)$, where $P(A) \neq 0$

Here, the joint probability $P(A \cap B)$ of both events A and B being true such that,

$P(B \cap A) = P(A \cap B)$

$P(A \cap B) = P(A | B) P(B) = P(B | A) P(A)$

$P(A|B) = [P(B|A) P(A)]/ P(B)$, where $P(B) \neq 0$

Similarly, from the definition of conditional density, Bayes theorem can be derived for two continuous random variables namely X and Y as given below:

��|�=�(�)=��,�(�,�)��(�)��|�=�(�)=��,�(�,�)��(�)
Therefore,

��|�=�(�)=��|�=�(�)��(�)��(�)

Bayes Theorem Applications
One of the many applications of Bayes' theorem is Bayesian inference, a particular approach to statistical inference. Bayesian inference has found application in various activities, including medicine, science, philosophy, engineering, sports, law, etc. For example, we can use Bayes' theorem to define the accuracy of medical test results by considering how likely any given person is to have a disease and the test's overall accuracy. Bayes' theorem relies on consolidating prior probability distributions to generate posterior probabilities. In Bayesian statistical inference, prior probability is the probability of an event before new data is collected.

11. A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A standard normal distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation (SD)

KEY TAKEAWAYS

- A Z-Score is a statistical measurement of a score's relationship to the mean in a group of scores.
- A Z-score can reveal to a trader if a value is typical for a specified data set or if it is atypical.
- In general, a Z-score of -3.0 to 3.0 suggests that a stock is trading within three standard deviations of its mean.
- Traders have developed many methods that use z-score to identify correlations between trades, trading positions, and evaluate trading strategies.

Z-Score Formula

The statistical formula for a value's z-score is calculated using the following formula:

$z = (x - \mu) / \sigma$

Where:

- z = Z-score
- x = the value being evaluated

- μ = the mean
- σ = the standard deviation

12. A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.

The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

KEY TAKEAWAYS

- A t-test is an inferential statistic used to determine if there is a statistically significant difference between the means of two variables.
- The t-test is a test used for hypothesis testing in statistics.
- Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.
- T-tests can be dependent or independent.

0 seconds of 1 minute, 38 secondsVolume 75%

T-Test

Understanding the T-Test

A t-test compares the average values of two data sets and determines if they came from the same population. In the above examples, a sample of students from class A and a sample of students from class B would not likely have the same mean and standard deviation. Similarly, samples taken from the placebo-fed control group and those taken from the drug prescribed group should have a slightly different mean and standard deviation.

Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement. It assumes a null hypothesis that the two means are equal.

Using the formulas, values are calculated and compared against the standard values. The assumed null hypothesis is accepted or rejected accordingly. If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and are probably not due to chance.

The t-test is just one of many tests used for this purpose. Statisticians use additional tests other than the t-test to examine more variables and larger sample sizes. For a large sample size, statisticians use a z-test. Other testing options include the chi-square test and the f-test.

Using a T-Test

Consider that a drug manufacturer tests a new medicine. Following standard procedure, the drug is given to one group of patients and a placebo to another group called the control group. The placebo is a substance with no therapeutic value and serves as a benchmark to measure how the other group, administered the actual drug, responds.

After the drug trial, the members of the placebo-fed control group reported an increase in average life expectancy of three years, while the members of the group who are prescribed the new drug reported an increase in average life expectancy of four years.

Initial observation indicates that the drug is working. However, it is also possible that the observation may be due to chance. A t-test can be used to determine if the results are correct and applicable to the entire population.

Four assumptions are made while using a t-test. The data collected must follow a continuous or ordinal scale, such as the scores for an IQ test, the data is collected from a randomly selected portion of the total population, the data will result in a normal distribution of a bell-shaped curve, and equal or homogenous variance exists when the standard variations are equal.

T-Test Formula

Calculating a t-test requires three fundamental data values. They include the difference between the mean values from each data set, or the mean difference, the standard deviation of each group, and the number of data values of each group.

This comparison helps to determine the effect of chance on the difference, and whether the difference is outside that chance range. The t-test questions whether the difference between the groups represents a true difference in the study or merely a random difference.

The t-test produces two values as its output: t-value and degrees of freedom. The t-value, or t-score, is a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets.

The numerator value is the difference between the mean of the two sample sets. The denominator is the variation that exists within the sample sets and is a measurement of the dispersion or variability.

This calculated t-value is then compared against a value obtained from a critical value table called the T-distribution table. Higher values of the t-score indicate that a large difference exists between the two sample sets. The smaller the t-value, the more similarity exists between the two sample sets.

T-Score

A large t-score, or t-value, indicates that the groups are different while a small t-score indicates that the groups are similar.
Degrees of freedom refer to the values in a study that has the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis. Computation of these values usually depends upon the number of data records available in the sample set.

13. In statistics, percentiles are used to understand and interpret data. The *n*th percentile of a set of data is the value at which *n* percent of the data is below it. In everyday life, percentiles are used to understand values such as test scores, health indicators, and other measurements. For example, an 18-year-old male who is six and a half feet tall is in the 99th percentile for his height. This means that of all the 18-year-old males, 99 percent have a height that is equal to or less than six and a half feet. An 18-year-old male who is only five and a half feet tall, on the other hand, is in the 16th percentile for his height, meaning only 16 percent of males his age are the same height or shorter.

- Percentiles are used to understand and interpret data. They indicate the values below which a certain percentage of the data in a data set is found.
- Percentiles can be calculated using the formula n = (P/100) x N, where P = percentile, N = number of values in a data set (sorted from smallest to largest), and n = ordinal rank of a given value.
- Percentiles are frequently used to understand test scores and biometric measurements.

Percentile Formula

Percentiles for the values in a given data set can be calculated using the formula:

n = (P/100) x N

where N = number of values in the data set, P = percentile, and n = ordinal rank of a given value (with the values in the data set sorted from smallest to largest). For example, take a class of 20 students that earned the following scores on their most recent test: 75, 77, 78, 78, 80, 81, 81, 82, 83, 84, 84, 84, 85, 87, 87, 88, 88, 88, 89, 90. These scores can be represented as a data set with 20 values: {75, 77, 78, 78, 80, 81, 81, 82, 83, 84, 84, 84, 85, 87, 87, 88, 88, 88, 89, 90}.

We can find the score that marks the 20th percentile by plugging in known values into the formula and solving for *n*:

n = (20/100) x 20

n = 4

The fourth value in the data set is the score 78. This means that 78 marks the 20th percentile; of the students in the class, 20 percent earned a score of 78 or lower.

Percentile scores have a variety of uses. Anytime that a set of data needs to be broken into digestible chunks, percentiles are helpful. They are often used to interpret test scores—such as SAT scores—so that test-takers can compare their performance to that of other students. For example, a student might earn a score of 90 percent on an exam. That sounds pretty impressive; however, it becomes less so when a score of 90 percent corresponds to the 20th percentile, meaning only 20 percent of the class earned a score of 90 percent or lower.

14. Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study

KEY TAKEAWAYS

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

The Formula for ANOVA is:

F=MSTMSE

where: F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

F=MSEMST**where:**F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

15. ANOVA is helpful for **testing three or more variables**. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.

The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables (such as the first example: age, sex, income). When you understand how each independent variable's mean is different from the others, you can begin to understand which of them has a connection to your dependent variable (landing page clicks), and begin to learn what is driving that behavior.

One-way ANOVA can help you know whether or not there are significant differences between the groups of your independent variables (such as USA vs Canada vs Mexico when testing a Location variable). You may want to test multiple independent variables (such as Location, employment status or education). When you understand how the groups within the independent variable differ (such as USA vs Canada vs Mexico, not location, employment status, or education), you can begin to understand which of them has a connection to your dependent variable (NPS score).