STATISTICS WORKSHEET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?
   a) Total Variation = Residual Variation – Regression Variation
   b) Total Variation = Residual Variation + Regression Variation
   c) Total Variation = Residual Variation * Regression Variation
   d) All of the mentioned

   Ans: b) Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called__outcomes.
   a) random
   b) direct
   c) binomial
   d) none of the mentioned

   Ans: c) binomial

3. How many outcomes are possible with Bernoulli trial?
   a) 2
   b) 3
   c) 4
   d) None of the mentioned

   Ans: a) 2

4. If Ho is true and we reject it is called
   a) Type-I error
   b) Type-II error
   c) Standard error
   d) Sampling error

   Ans: a) Type-I error

5. Level of significance is also called:
   a) Power of the test
   b) Size of the test
   c) Level of confidence
   d) Confidence coefficient

   Ans:b) Size of the test

6. The chance of rejecting a true hypothesis decreases when sample size is:
   a) Decrease
   b) Increase
   c) Both of them
   d) None

   Ans:b) Increase

7. 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

8. What is the purpose of multiple testing in statistical inference?
   a) Minimize errors
   b) Minimize false positives
   c) Minimize false negatives
   d) All of the mentioned

   Ans : d) All of the mentioned

9. Normalized data are centred at_and have units equal to standard deviations of the original data
   a) 0
   b) 5
   c) 1
   d) 10

   Ans : a)0

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

10. # What Is Bayes' Theorem?

Bayes theorem is one of the most popular machine learning concepts that helps to calculate the probability of occurring one event with uncertain knowledge while other one has already occurred. It is one of the formulas that have their practical application in almost all the areas that derive their results from predictions and probability. One can think of weather predictions where the meteorologists use the Bayes' Theorem to predict the day-to-day weather forecast. The scientist can even forewarn the people about prevailing possibilities of bad weather so they can make the required arrangements. Similarly, there are many other fields where Bayes' Theorem is applied.

Bayes Theorem Formula:

If A and B are two independent events, then the probability of event A when B is true is calculated as the ratio of probability of event B such that event A is true and the individual probability of event A to the individual probability of event B. Bayes theorem Formula is written as:

P(A/B)=P(B/A).P(A)P(B)P(A/B)=P(B/A).P(A)P(B)

In the above mentioned Bayes Theorem Formula,

P (A | B) is the probability of event A being true when event B is true.

P (B | A) is the probability of event B being true when event A is true.

P (A) is the probability of event A being true.

P (B) is the probability of event B being true.

To State and Prove Bayes Theorem:

Conditional probability is the probability of one event when one or more other individual events are true. It can be better explained with the help of an example.

Suresh visits a library in which one of the book racks contains 3 rows. All the three rows are stacked with a mixture of reference books, journals and annual reports. Let us consider that Suresh picks a book from the second rack. The probability of whether the book picked by Suresh is a reference material depends on the other two events. (i.e. whether the book is a journal or an annual magazine). In general, conditional probability means the measure of probability of one event when the other event is true.

P(A/B)=P(A∩B)P(B)P(A/B)=P(A∩B)P(B)

where ,

A and B are two individual events and P (B) not equal to zero.

P (A | B) is the probability of event A being true when event B is true.

P (A ∩ B) is the probability of occurrence of both A and B.

P (B) is the probability of individual event B.

How to State and Prove Bayes Theorem:

Bayes theorem formula is stated as

P(A/B)=P(B/A).P(A)P(B)P(A/B)=P(B/A).P(A)P(B)

Bayes theorem proof can be derived using the concept of conditional probability. The probability of occurrence of both the events A and B is given in terms of their individual probabilities and conditional probability as:

P (A ∩ B) = P (A). P (B | A)P (A ∩ B) = P (A). P (B | A)

Similarly the occurrence of both the events simultaneously can also be given in terms of the probability of second event as:

P (A ∩ B) = P (B). P (A | B)P (A ∩ B) = P (B). P (A | B)

In both the equations, the left hand side is equal. So RHS can be equated.

P (B). P (A | B) = P (A). P (B | A)P (B). P (A | B) = P (A). P (B | A)

Further simplification gives the Bayes theorem formula as

$$P(A/B) = \frac{P(B/A).\ P(A)}{P(B)}$$

Example of Bayes' Theorem

Imagine you are a financial analyst at an investment bank. According to your research of publicly-traded companies, 60% of the companies that increased their share price by more than 5% in the last three years replaced their CEOs during the period.

At the same time, only 35% of the companies that did not increase their share price by more than 5% in the same period replaced their CEOs. Knowing that the probability that the stock prices grow by more than 5% is 4%, find the probability that the shares of a company that fires its CEO will increase by more than 5%.

Before finding the probabilities, you must first define the notation of the probabilities.

- P(A) – the probability that the stock price increases by 5%
- P(B) – the probability that the CEO is replaced
- P(A|B) – the probability of the stock price increases by 5% given that the CEO has been replaced
- P(B|A) – the probability of the CEO replacement given the stock price has increased by 5%.

Using the Bayes' theorem, we can find the required probability:

$$P(A|B) = \frac{0.60 \times 0.04}{0.60 \times 0.04 + 0.35 \times (1 - 0.04)} = 0.067 \text{ or } 6.67\%$$

Thus, the probability that the shares of a company that replaces its CEO will grow by more than 5% is 6.67%.

Applications of Bayes' Theorem

There are plenty of applications of the Bayes' Theorem in the real world. Don't worry if you do not understand all the mathematics involved right away. Just getting a sense of how it works is good enough to start off.

Bayesian Decision Theory is a statistical approach to the problem of pattern classification. Under this theory, it is assumed that the underlying probability distribution for the categories is known. Thus, we obtain an ideal Bayes Classifier against which all other classifiers are judged for performance.

We will discuss the three main applications of Bayes' Theorem:

- Naive Bayes' Classifiers
- Discriminant Functions and Decision Surfaces
- Bayesian Parameter Estimation

11. What is z-score?

Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the population mean (μ) and also the population standard deviation (σ).

Z score can be statistically defined as the standard deviation of a raw score from its mean. It is one of the most important numerical measurements used in statistics. Z score is also most of the times regarded as the standard score. A Z score can be placed on the normal distribution curve. Z scores may be either positive or negative or null. Positive value of Z - score is the indication that the Z score is above the mean whereas the negative value indicates that the Z score is below the mean. To calculate Z score, we subtract the population mean from an individual raw score followed by the division of the difference obtained by the standard deviation of the population. This process involving the conversion is referred to as standardizing or normalizing.

Formula

The equation is given by $z = (x – \mu)/ \sigma$.

μ = mean

σ = standard deviation

x = test value

When we have multiple samples and want to describe the standard deviation of those sample means, we use the following formula:

$z = (x – \mu)/ (\sigma/\sqrt{n})$

Interpretation

1. If a z-score is equal to -1, then it denotes an element, which is 1 standard deviation less than the mean.

2. If a z score is less than 0, then it denotes an element less than the mean.

3. If a z score is greater than 0, then it denotes an element greater than the mean.

4. If the z score is equal to 0, then it denotes an element equal to the mean.

5. If the z score is equal to 1, it denotes an element, which is 1 standard deviation greater than the mean; a z score equal to 2 signifies 2 standard deviations greater than the mean; etc.

Example 1

The test score is 190. The test has a mean of 130 and a standard deviation of 30. Find the z score. (Assume it is a normal distribution)

Solution:

Given test score $x = 190$

Mean, $\mu = 130$

Standard deviation, $\sigma = 30$

So $z = (x – \mu)/ \sigma$

$= (190 – 130)/ 30$

$= 60/30$

$= 2$

Hence, the required z score is 2.

Example 2

The test scores of students in a class test has a mean of 70 and with a standard deviation of 12. What is the probable percentage of students scored more than 85?

Solution: The z score for the given data is,

z= (85-70)/12=1.25

From the z score table, the fraction of the data within this score is 0.8944.

This means 89.44 % of the students are within the test scores of 85 and hence the percentage of students who are above the test scores of 85 = (100-89.44)% = 10.56 %.

12. What is t-test?

The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.

You want to know whether the mean petal length of iris flowers differs according to their species. You find two different species of irises growing in a garden and measure 25 petals of each species. You can test the difference between these two groups using a t-test and null and alterative hypotheses.

The null hypothesis (H0) is that the true difference between these group means is zero.
The alternate hypothesis (Ha) is that the true difference is different from zero.

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. 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.

A t-test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test or a post-hoc test.

The t-test is a parametric test of difference, meaning that it makes the same assumptions about your data as other parametric tests. The t-test assumes your data:

- are independent

- are (approximately) normally distributed.
- have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)
- If your data do not fit these assumptions, you can try a nonparametric alternative to the t-test, such as the Wilcoxon Signed-Rank test for data with unequal variances.

Type of t-test

When choosing a t-test, you will need to consider two things: whether the groups being compared come from a single population or two different populations, and whether you want to test the difference in a specific direction.

i) One-sample, two-sample, or paired t-test

- If the groups come from a single population (e.g. measuring before and after an experimental treatment), perform a **paired t-test**.
- If the groups come from two different populations (e.g. two different species, or people from two separate cities), perform a **two-sample t-test** (a.k.a. **independent t-test**).
- If there is one group being compared against a standard value (e.g. comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample t-test**.

ii) One-tailed or two-tailed t-test

- If you only care whether the two populations are different from one another, perform a **two-tailed t-test**.
- If you want to know whether one population mean is greater than or less than the other, perform a **one-tailed t-test.**

T-test formula

The formula for the two-sample t-test (a.k.a. the Student's t-test) is shown below.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

In this formula, t is the t-value, x_1 and x_2 are the means of the two groups being compared, s_2 is the pooled standard error of the two groups, and n_1 and n_2 are the number of observations in each of the groups.

A larger t-value shows that the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups. You can compare your calculated t-value against the values in a critical value chart to determine whether your t-value is greater than what would be expected by chance. If so, you can reject the null hypothesis and conclude that the two groups are in fact different.

13. What is percentile?

Percentiles are used in statistics to give you a number that describes the value that a given percent of the values are lower than.
A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is at the 86th percentile, where 86 is the percentile rank, it is equal to the value below which 86% of the observations may be found. In contrast, if it is in the 86th percentile, the score is at or below the value of which 86% of the observations may be found. Every score is in the 100th percentile.
The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3). In general, percentiles and quartiles are specific types of quantiles.
The range of values containing the central half of the observations is called the interquartile range: that is, the range between the 25th and 75th percentiles (the range including the values that are up to 25% higher or down to 25% lower than the median).
It is used with the median value to report data that are markedly non-normally distributed.

Example

i) Use the NumPy **percentile()** method to find the percentiles:

```
import numpy

age= [5,31,43,48,50,41,7,11,15,39,80,82,32,2,8,6,25,36,27,61,31]

x=numpy.percentile(ages, 75)

print(x)
```

43.0

ii) What is the age that 90% of the people are younger than?

```
import numpy

ages= [5,31,43,48,50,41,7,11,15,39,80,82,32,2,8,6,25,36,27,61,31]

x=numpy.percentile(ages, 90)

print(x)
```

61.0

14. What is ANOVA?

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.

The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method. ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers." It was employed in experimental psychology and later expanded to subjects that were more complex.

The Formula for ANOVA is:

$$F = MSE/MST$$

where:F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

An ANOVA conducted on a design in which there is only one factor is called a **one-way ANOVA**. If an experiment has two factors, then the ANOVA is called a two-way ANOVA. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels. ANOVAs can also be used for within-group/repeated and between subjects designs. For this chapter we will focus on between subject one-way ANOVA.

In a One-Way ANOVA we compare two types of variance: the variance between groups and the variance within groups.

Types of ANOVA

one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same.

The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

Limitations

During an analysis of variance test, it is important to consider the independent and dependent variables. The independent variable, referred to as a factor in ANOVA, is the variable that is manipulated in an experiment or study. The dependent variable refers to what is affected by the independent variable; it is the variable being measured. For example, in a study of how exercise affects mood, the amount of exercise is the independent variable and the mood of the participants is the dependent variable.

15. How can ANOVA help?

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

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. It is employed with subjects, test groups, between groups and within groups.

A common approach to figure out a reliable treatment method would be to analyse the days it took the patients to be cured. We can use a statistical technique which can compare these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples on the basis of their means, is called ANOVA.

- You might use ANOVA to compare three different treatments against each other, to compare two different diets against each other, or compare two different exercise programs against each other. For example, let's say you want to know if there's a difference between the average heights of four different types of trees in a forest. Instead of calculating whether each pair is statistically different from one another, you could run one ANOVA test to find out whether any of them are significantly different from one another.
- ANOVA is also used as a method of testing how well different groups of data fit together. Let's say you have a group of dogs, and you want to know whether they are all the same size or if some dogs are bigger than others. You can use ANOVA to test whether the groups differ from each other. You can also use ANOVA to compare more than two groups at once. You could test whether German Shepherds are the same size as Poodles, teacup Poodles, teacup Chihuahuas, and regular-sized Chihuahuas. This way, you could see if all of these dog breeds are the same size, or if one breed is larger than another breed.
- You can use ANOVA to test for statistical differences between two or more groups to see if there is a significant difference between the means of those groups. ANOVA determines whether a test is valid by looking at the variation between and within groups.
- If a test shows a large standard deviation between groups, then the differences are likely due to random chance; however, if the standard deviation within groups is large, then it may be due to real differences between groups.

- An important thing to know about ANOVA tests is that they assume all groups are sampled from populations with equal variances. If the variances between groups are not equal, you'll need to use Welch's ANOVA instead