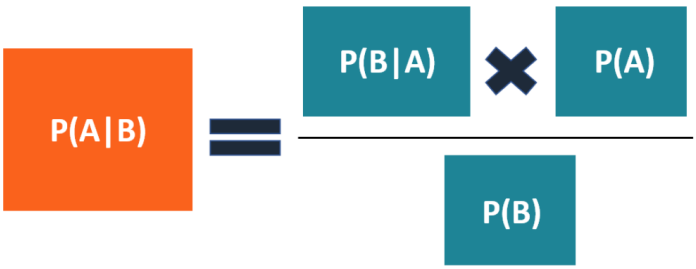
STATISTICS WORKSHEET-3

1. **Which of the following is the correct formula for total variation?**
   Answer- (b)
2. **Collection of exchangeable binary outcomes for the same covariate data are called outcomes**.
   Answer- (c)
3. **How many outcomes are possible with Bernoulli trial?**
   Answer-(a)
4. **If Ho is true and we reject it is called**
   Answer-(a)
5. **Level of significance is also called:**
   Answer-(c)
6. **The chance of rejecting a true hypothesis decreases when sample size is:**
   Answer-(b)
7. **Which of the following testing is concerned with making decisions using data?**
   Answer-(b)
8. **What is the purpose of multiple testing in statistical inference?**
   Answer-(a)
9. **Normalized data are centred at and have units equal to standard deviations of the original data**
   Answer-(a)
10. **What Is Bayes' Theorem?**

Answer- In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

The theorem is named after English statistician, Thomas Bayes, who discovered the formula in 1763. It is considered the foundation of the special statistical inference approach called the Bayes' inference.



Besides statistics, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology as the most notable examples. In addition, the theorem is commonly employed in different fields of finance. Some of the applications include but are not limited to, modeling the risk of lending money to borrowers or forecasting the probability of the success of an investment.

Formula for Bayes' Theorem

The Bayes' theorem is expressed in the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- P(A|B) – the probability of event A occurring, given event B has occurred
- P(B|A) – the probability of event B occurring, given event A has occurred
- P(A) – the probability of event A
- P(B) – the probability of event B

Note that events A and B are independent events (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).

A special case of the Bayes' theorem is when event A is a binary variable. In such a case, the theorem is expressed in the following way:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A^-)P(A^-) + P(B|A^+)P(A^+)}$$

Where:

- $P(B|A^-)$ – the probability of event B occurring given that event A^- has occurred
- $P(B|A^+)$ – the probability of event B occurring given that event A^+ has occurred

In the special case above, events A^- and A^+ are mutually exclusive outcomes of event A.

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%.

11. What is z-score?

Answer- **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).

The Formula for Z-Score
A z-score can be calculated using the following formula.

$$z = (X – \mu) / \sigma$$

where,
z = Z-Score,
X = The value of the element,
μ = The population mean, and
σ = The population standard deviation

How to Calculate Z-Score?
Usually, the population mean ((μ), the population standard deviation (σ), and the observed value (x) are provided in the problem statement, and substituting the same in the above Z-score equation yields us the Z-Score value. Depending upon whether the given Z-Score is positive or negative, one makes use of the respective positive Z-Table or negative Z-Table available online or on the back of your statistics textbook in the appendix.

12. What is t-test?

Answer- Imagine you are running an experiment where you want to compare two groups and quantify the difference between them.
For example:

- Compare if the people of one country are taller than people of another one.
- Compare if the brain of a person is more activated while watching happy movies than sad movies.

This comparison can be analyzed by conducting different statistical analysis, such as t-test, which is the one described in this article.

So, what is a t-test? It is a type of inferential statistic used to study if there is a statistical difference between two groups. Mathematically, it establishes the problem by assuming that the means of the two distributions are equal (H_o: $\mu_1=\mu_2$). If the t-test rejects the null hypothesis (H_o: $\mu_1=\mu_2$), it indicates that the groups are highly probably different.

This test should be implemented when the groups have 20–30 samples. If we want to examine more groups or larger sample sizes, there are other tests more accurate than t-tests such as z-test, chi-square test or f-test.

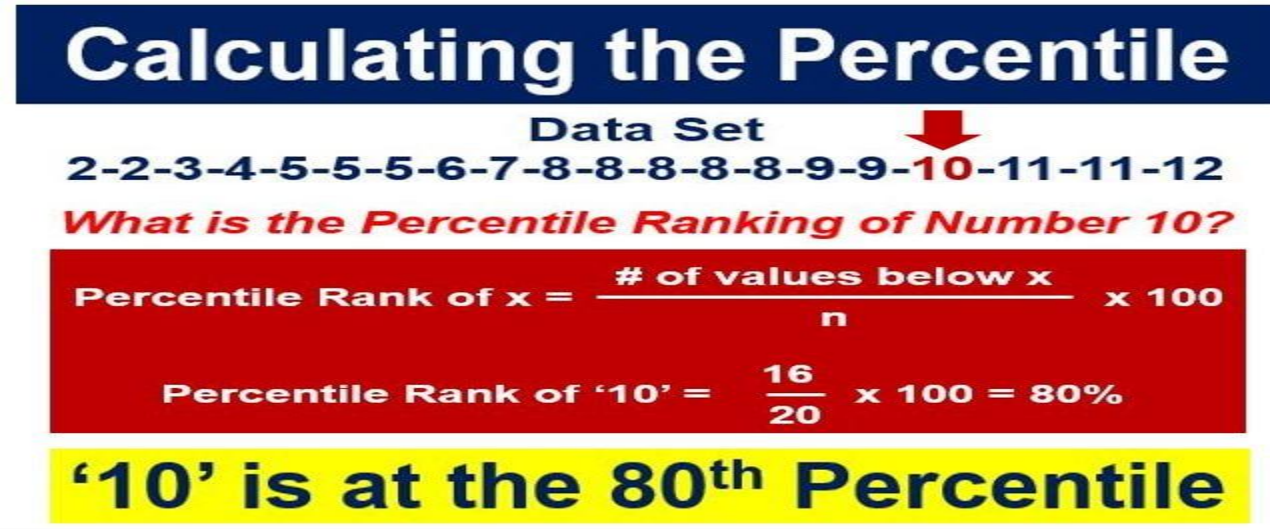
***Important**: The t-test rejects or fails to reject the null hypothesis, never accepts it.*

13. What is percentile?

Answer- Percentile or centile is a value or number that represents a percentage position on a range or list of data – the person or thing at that number of value is above that number in percentage. For example, if the exam results of schoolchildren at a specified school in a nationwide study is at the 65th percentile, it means that that school performed better than 65% of all the other schools in the country.

The richest 1% of the population is at the top percentile – 99% are poorer than them – while the poorest 1% is at the bottom percentile – 99% are richer than them.

The term is part of the family of words ending in 'ile', like 'quartile', that signpost positions on a scale of numbers.



The percentile ranking of any value equals the number of values below that value, divided by the total number of values in the set, times 100. In this image, there are 16 numbers below 10, and there

are 20 total numbers. Sixteen divided by twenty times 100 equals 80%. Ten is at the 80th percentile, because 80% of the numbers are below it.

Percentile, a measure used in statistics, always has a number next to it – it indicates that the person or thing being measured or evaluated is at the top of that number in percentage terms. For example, imagine a country has just 100 people, and Mr. Brown is at the 42nd percentile regarding physical strength. This means that there are 42 people physically weaker than him.

Percentiles are used extensively to report scores in academic exams and tests, such as LSAT, GRE and SAT. For example, in 2013, the 70th percentile for **GRE was 156 – so, if you scored 156, you did better than 70% of test takers.

14. **What is ANOVA?**
    Answer- The two fundamental concepts in inferential statistics are population and sample. The goal of the inferential statistics is to infer the properties of a population based on samples.

Population is all elements in a group whereas sample means a randomly selected subset of the population. It is not always feasible or possible to collect population data so we perform analysis using samples.

For instance, the college students in US is a population and randomly selected 1000 college students throughout US is a sample of this population.

It would not be reliable to directly apply the sample analysis results to the entire population. We need systematic ways to justify the sample results are applicable to the population. This is the reason why we need to statistical tests. They evaluate how likely the sample results are true representation of the population.

Consider we do research project on obesity. In the scope of our project, we want to compare the average weight of 20-year-old people in two different countries, A and B. Since we cannot collect the population data, we take samples and perform a statistical test.

We set the null and alternative hypothesis as below:

- Null hypothesis (H0) : The average weights of two groups are not different.
- Alternative hypothesis (H1): The average weights of two groups are different.

In case of comparing two groups, **t-test** is preferred over **ANOVA**. However, when we have more than two groups, t-test is not the optimal choice because a separate t-test needs to perform to compare each pair.

Assume we are comparing three countries, A, B, and C. We need to apply a t-test to A-B, A-C and B-C pairs. As the number of groups increase, this becomes harder to manage. Thus, we choose to go with ANOVA.

In the case of comparing three or more groups, ANOVA is preferred. There are two elements of ANOVA:

- Variation within each group
- Variation between groups

ANOVA result is based on the F ratio which is calculated as follows:

$$F = \frac{\text{Variation between groups}}{\text{Variation within groups}}$$

F ratio is a measure of the comparison between the variation between groups and variation withing groups.

Higher F ratio values indicate the variation between groups is larger than the individual variation of groups. In such cases, it is more likely that the mean of the groups are different.

By contrast, in case of lower F ratio values, the individual variation of groups are larger than the variations between groups. Thus, we can conclude that the elements in a group are highly different rather than the entire groups.

The larger the F ratio, the more likely that the groups have different means.

We have covered the intuition behind ANOVA and when it is typically used. The next step is do an example. We will use the R programming language to perform ANOVA test.

The rnorm function generates an array of numbers sampled from a normal distribution based on the given mean and standard variation values.

```
> rnorm(5, mean=10, sd=3)
[1]  8.624795  8.431731 10.570984  7.136710 11.801554
```

We will use the rnorm function to generate sample data and then stack groups in a data frame.

```
> A = rnorm(100, mean = 60, sd = 5)
> B = rnorm(100, mean = 71, sd = 10)
> C = rnorm(100, mean = 65, sd = 7)
> groups = stack(data.frame(cbind(A, B, C)))
```

	values	ind
1	67.30958	A
2	70.68381	A
3	68.64135	A
4	65.54049	A
5	59.89950	A

groups data frame

The values column contains the values and ind column shows which group it belongs to. The ANOVA test is done using the aov function:

```
> anovaResults = aov(values ~ ind, data = groups)
> summary(anovaResults)
             Df Sum Sq Mean Sq F value Pr(>F)
ind           2   8200    4100   58.56 <2e-16 ***
Residuals   297  20796      70
```

15.**How can ANOVA help?**

Answer- ANOVA can help to identify the sources of variation in a data set. This can help to improve the accuracy of data predictions and analyses. Additionally, ANOVA can help to identify relationships between different variables in a data set. This information can be used to improve data models and predictions. Overall, ANOVA can greatly improve the quality of data science research and results by allowing researchers to focus on the areas that need improvement.

How Can Businesses Use ANOVA?

There are a number of ways businesses can use ANOVA. One way is to use it to test the difference between two or more groups. This can be used to see if there is a difference in how a product is perceived by customers in different parts of the world or if there is a difference in how a product is perceived by men and women.

Another way businesses can use ANOVA is to test the difference between a control group and one or more experimental groups. This can be used to see if there is a difference in how a product is perceived by customers who have seen a marketing campaign versus those who have not, or to see if there is a difference in how a product is perceived by customers who have used it before and those who have not.

ANOVA is a valuable tool that can be used to improve the business and data science field. It can help data scientists and business owners better understand the data they are working with and identify any patterns that may exist.