iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 8 **Issue:** IX **Month of publication:** September 2020

**DOI:** https://doi.org/10.22214/ijraset.2020.31510

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 8 Issue IX Sep 2020- Available at www.ijraset.com*

# Elucidation and Dominance of Hypothesis Analogies in Data Science

Sahil Rahman
*School of Computer Science & Engineering, Galgotias University, Greater Noida,Uttar Pradesh, India.*

***Abstract:*** ***In the field of Data Science, Hypothesis Testing is an essential tool of statistical inference that plays a vital role in making an informed conclusion about the population using the sample data. It helps in making a decision as to which sample data best support mutually exclusive statement about the population. It is one of the essential concepts in statistics that form the foundation of Machine Learning because it is how you make decision if something happened, or if specific treatments have any side effects, or if groups differ from each other or if one variable forecast another and conclude about the characteristics of what you're comparing. However, while dealing with the Hypothesis Testing problems, the chances and probabilities for errors are uttermost. Mistakes created within the statement of the null and alternative hypothesis will have a severe impact within the interpretation of the result, which may lead to false or inaccurate decisions. One of the major concepts behind this faultiness is the confusion between the various testing or analogies and which test to choose for better results. Detailed analysis of these Hypothesis analogies has been done and presented. The target of this paper is to produce an outline of the issues faced by the students while learning the hypothesis testing and useful definitions and concepts for understanding the methods used, and their prospects and ease the confusions in merely following up the Hypothesis analogies and report current and future developments for it. There is no doubt in the fact that the field of Data Science is going to pick up a high pace in the upcoming years, so making the students understand this concept is highly needed because there has been an enormous growth in the number of students in these courses in the modern years.***

***Keywords: Hypothesis Analogies ● Statistical Methods ● Conceptual Understanding ● ALOHA ● Decision Making ● Sampling Distributions***

## I. INTRODUCTION

Hypothesis Testing has always happened to be a challenging concept encountered while studying statistics for Machine Learning. It forms the foundation of building and training the Machine Learning models for decision-making. Moreover, Hypothesis Testing is accepted to be a crucial topic in statistics for the majority of disciplines as it is used in a statistical analysis regarding population. It involves claiming the population and then test the claim by analysing the sampling data through a set of methods.

In the field of Machine Learning and Data Science, statistics plays a vital role. While heading towards these topics, students are introduced with a buzzy word "Hypothesis Testing". Researchers found out that it is the most challenging topic to teach, and the majority of students often find it more complicated and disorganized. Researchers have even found out that when students try out to perform the Hypothesis testing; they cannot perceive the logic of executing these steps or applying them in new contexts. It has also been noted that the majority of students are unable to identify what is the demand of the asked problem and which approach or test should be used for implementation, which led to causing the deficiency in understanding the concepts and failure in using appropriate formula among the different Hypothesis testing formulas.

Prior to the Hypothesis testing, students should be made to understand the underlying notions like Central Limit Theorem and the Sampling Distribution. Students should be made to have the skill to differentiate between the sample and population and to distinguish between the population proportion and population mean.

## II. HYPOTHESIS TESTING

In simple words, the terms "Hypothesis" means "making assumptions", and "Testing" means "a method to verify whether the end product is the same as the requirements", and collectively they form the term "Hypothesis Testing", that implies "the method to verify whether the hypothesis made is true or not". Hypothesis Testing is basically a belief that we propose about the population parameter that uses the experimental data in making statistical decisions. Let's see some few examples,

*1)* If a student studies for 8 hours, then he will surely be getting good score.
*2)* Consuming soft drinks in a regular habit may lead to obesity.
*3)* Eating apples everyday may lower the risk of diabetes.

For all the above-assumed examples, we require a mathematical conclusion that would prove helpful in stating whether the assumptions are valid. Let's put some lights on the characteristics of the Hypothesis,

*a)* It should be precise and clear.
*b)* It should be able to state the relationship between variables.
*c)* It should be specific and have the ability to conduct more tests.
*d)* The simplicity of the Hypothesis is inversely proportional to its significance.

Before diving deep into the hypothesis segment, let's quickly encounter with two most basic terms, "Population" and "Sample" that is having a substantial impact on this chapter. Let's suppose; there is an owner of a famous restaurant who wanted to know which of his dishes sell well, so he decided to survey customers to get a suitable review. Now, taking the survey from each customer would be a very strenuous task, so the owner decided to choose 200 customers randomly and surveyed them accordingly. Now, from the given example, it can be concluded that the total number of customers, the restaurant owner has, is the Population and the randomly chosen 200 customers are the Samples. A population is an area that includes all of the elements from a dataset, whereas a sample is a part of the Population,

Now, that we are aware of the terms "population" and "sample", let's summarize the steps involved while performing the Hypothesis Testing for a given problem. In the first step, the Null and Alternative hypothesis should be figured out and stated. Following up next, find out which suitable test-statistics or approach should be performed and detect the level of significance and level of confidence. This step is also known as the critical value, which is necessary to decide whether or not to reject the Null hypothesis. Next step is to form a probability statement and find out the corresponding p-value and finally, with the help of test statistics, arrive at a decision about the problem in concern.

Before commencing brief discussions on the steps and tests of the Hypothesis testing, let's quickly overview **Table 1** that shows the mistakes, an individual makes, while performing the Hypothesis Testing on a given problem.

Table 1. Categories of mistakes noted while dealing with the Hypothesis Analogies

| Steps involved in Hypothesis Testing | Classifications of mistakes occur |
|---|---|
| State the Null and the Alternative Hypothesis | 1. Stating incorrect population<br>2. Instead of the population parameter, stating the Sample statistics<br>3. Using incorrect signs or inequality symbol<br>4. False Hypothesis estimation<br>5. Senseless declaration |
| Calculate the Test statistic value | 1. Incorrect formula<br>2. Incorrect calculation<br>3. Senseless report |
| Determining the region of rejection from the level of significance | 1. Unable to differentiate between one-tailed test and two-tailed test<br>2. Wrong Z-test value<br>3. Senseless report |
| Producing the critical value and setting up the probabilities for the decision criteria | 1. Wrong critical value<br>2. Incorrect probability statement<br>3. Senseless report |
| Building a valid decision | 1. Wrong decision<br>2. Accurate decision but based on wrong calculations<br>3. Senseless declaration |
| Forming the decision in the conditions of the problem | 1. Incomplete communication<br>2. No link to the real problem<br>3. Senseless declaration |

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 8 Issue IX Sep 2020- Available at www.ijraset.com*

### *A. Null & Alternative Hypothesis*

In inferential statistics, the null hypothesis is an assertion that there is no statistically-significant connection between two measure phenomena or we can say that it is an assumption that states, "There should be no difference between the critical value and the p-value". With the help of the Null Hypothesis, we can check the likelihood of the statement being true in order to decide whether to accept or reject the alternative hypothesis. It is usually the hypothesis we have to try to disprove or discredit. It is denoted by the symbol $\boldsymbol{H_0}$ and can include signs such as =, ≥ or ≤. For example, if a population mean is equal to hypothesises mean, then the Null Hypothesis can be written as,

$$\boldsymbol{H_0}: \mu = \mu_0$$

The alternative hypothesis is an assumption that is contrary to or the negation of the null hypothesis. That implies that there is a statistically significant relationship between two measure phenomena, or we can say that it is an assumption that states, "There should be a difference between the p-value and the critical value". With the help of the Alternative Hypothesis, we can discover whether to accept or reject the statement based on the likelihood of the null hypothesis being true. It is denoted by the symbol $\boldsymbol{H_a}$ and can include signs such as ≠, > or <. From the earlier discussed example, if the Null Hypothesis is given as

$$\boldsymbol{H_0}: \mu = \mu_0$$

Then, the Alternative Hypothesis can be written as,

$$\boldsymbol{H_a}: \mu > \mu_0$$
$$\boldsymbol{H_a}: \mu < \mu_0$$
$$\boldsymbol{H_a}: \mu \neq \mu_0$$

The Null and the Alternative Hypothesis are called the mathematical opposites, that means only one among them can occur one at a time. They are inversely proportional to each other, i.e. if the null hypothesis is accepted, then the alternative hypothesis has to be rejected, or vice-versa.

### *B. Type I & Type II Errors*

No Hypothesis test is 100% practical. Since the test depends on probabilities, there is consistently an opportunity of making an off-base end. At the point when you do a hypothesis test, two kinds of mistakes are conceivable: Type I and Type II. The risk of these two mistakes are conversely related and dictated by the significance levels and the power for the test. Hence, you ought to figure out which blunder has progressively extreme ramifications for your circumstance before you characterize their risks. **Figure 1** shows the Graphical illustration of the connection between the Type I and Type II errors.

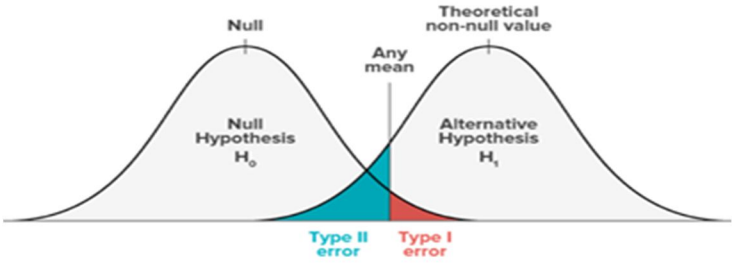



**Figure 1. Graphical depiction of the relation between Type I and Type II errors.**

According to the Type I error, if the null hypothesis is true, and you still reject it, you may get a Type I error. The probability of making this error is often represented by the Greek letter alpha **(α)**, where **α** implies the level of significance that has been put for the Hypothesis test. For example, **α** having the values 0.05, i.e. **α=0.05**, indicates that you are accepting a 5% chance that you are wrong when you reject the null hypothesis. For lowering this risk, a smaller value of alpha should be used. However, using a smaller level of alpha means that an individual or an observer will be unlikely to observe a true difference if the error really exists.

While the Type II error says if the null hypothesis is false, and you fail to reject it, you may get a Type II error. The probability of making this type of error is often represented by the Greek letter beta (β), where β implies the power of the test that has been set for the Hypothesis test. We can lower the risk of committing the Type II error by using a higher power value, which can be achieved by making the sample size large enough to determine a practical difference when the error truly exists. When the null hypothesis is false, the probability of rejecting it will be given as 1- β, and this value will be the power of the test. Table 2 shows the statistical errors committed while performing the hypothesis testing.

Table 2. Statistical errors in Hypothesis Testing

| Findings based on Sample | Reality about the Population | |
|---|---|---|
| | $H_0 \rightarrow$ True | $H_0 \rightarrow$ False |
| $\boldsymbol{H_0}$ Rejection | Type I Error (Probability = α) | Correct Decision (Probability = 1- β) |
| $\boldsymbol{H_0}$ Acceptance | Correct Decision (Probability = 1- α) | Type II Error (Probability = β) |

The Type I and Type II errors are inversely related to each other. If one among the errors increases, the other decreases. The rate of the Type I error is usually set in advance by the researcher, whereas the rate of the Type II error for a given test is difficult to determine because it needs to estimate the distribution of the alternative hypothesis, which is usually unknown.

*C. One-Tailed & Two-Tailed Test*

For a better understanding of this concept, the term "tailed" can be simplified as the term "sided". The alternative hypothesis can be classified into One-Tailed (One-sided) Test or Two-Tailed (Two-Sided) Test.

A one-tailed test can be defined as a statistical test in which one-sided critical area of a distribution is involved. It is also called as a "Directional Hypothesis" because it is used to specify the direction to be either greater than or less than the hypothesized value. A one-tailed test can be detected when there is a difference observed between the population parameter and the hypothesized value in a specific direction. However, it has a greater power than the two-tailed test but still cannot be identify whether the population parameter varies in the opposite direction. Furthermore, the one-tailed test is divided into two parts - Left-tailed test and Right-tailed test.

Let's suppose if we want to see whether the average marks in the Science subject of class 8$^{th}$ students of XYZ school are greater than 75, then a one-tailed alternative hypothesis will be performed because we are explicitly hypothesizing that the marks for students are greater than 75. Here, we can say that the null and the alternative hypothesis for the discussed example can be stated as,

$$\boldsymbol{H_0}: \mu = 75$$
$$\boldsymbol{H_a}: \mu > 75$$

A two-tailed test can be defined as a statistical test in which two-sided critical area of a distribution is involved. It is also called as a "Non-directional Hypothesis" because it is used to determine whether the population parameter is greater than or less than the hypothesized value. A two-tailed test can be detected when the population parameter differs in either direction but has less power than the one-tailed test. Let's look onto an example for a better comprehension of this concept.

The average marks in the Science subject of class 8$^{th}$ students of XYZ school are not equal to 85. Here, we can say that the null and the alternative hypothesis for the discussed example can be stated as,

$$\boldsymbol{H_0}: \mu = 85$$
$$\boldsymbol{H_a}: \mu \neq 85$$

Since, it is given that the $\mu \neq \mathbf{85}$, hence to prove the alternative hypothesis $\boldsymbol{H_a}$, the value of μ could be anything (greater than or less than) except 85, so from the previous knowledge, we can conclude that it is a Two-Tailed test.

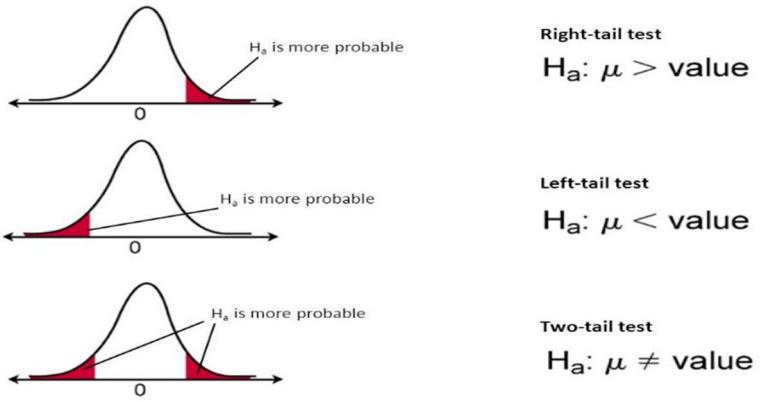



**Figure 2. Comparison of a One-tailed test (Right & Left) and a Two-tailed test**

With the help of Figure 2, we can have a comparison between a One-tailed right test, a One-tailed Left test and a Two-tailed test. The conclusion of whether to utilize a one-tailed or a two-tailed test is essential because the value of a test statistic that falls in the rejection region in a one-tailed test may not work out in a two-tailed test, even though both the tests uses the same probability level. Hence, we should go for the one-tailed test only if we have a positive vibe to expect that the difference will be in a specific direction. In contrast, a two-tailed test is more stable than a one-tailed test because it takes a more extreme test statistic for the rejection of the null hypothesis.

*D. Significance Level & Confidence Level*

The significance level is the threshold or the probability by which we will reject a null hypothesis when it is true, or we can say that it is the probability of the occurrence of Type I error. It means that if the probability of getting the statistics for any sample is lower than its significance level, then we can reject the null hypothesis and claim that we have enough evidence for the alternative. However, if the probability of getting the statistics for any sample is at or higher than its significance level, then we can't afford to reject the null hypothesis and claim that we don't have enough evidence for the alternative. It is denoted by the Greek word alpha **(α)**. For a one-tailed test, each tail = **α**, where alpha signifies the area in the tail, i.e. the rejection area, and for a two-tailed test, each tail = **α /2**, where alpha signifies the total area in the two tails. **Table 3** represents the relation of the level of significance along with the corresponding confidence interval in terms of Z value.

Table 3. Relation of significance level and confidence level

| Level of Significance α | Corresponding confidence interval in terms of Z value |
|---|---|
| 0.01 | -1.65σ to +1.65σ |
| 0.05 | -1.96σ to +1.96σ |
| 0.1 | -2.58σ to +2.58σ |

The motive of proceeding with a random sample formula for production and calculating statistics, i.e. the mean from the data, is to estimate the population mean. However, the sample statistics approximation of the primary population value has always been an issue. So, a confidence interval, i.e. the percentage of all the possible samples, addresses this issue because it come up with a range of values which is expected to contain the population parameter of interest.

Now, at a confidence level, such as 95%, these confidence intervals are constructed. It means that it is the probability that the confidence interval contains the population parameter in approximately 95% of the cases or in simple words, we can say that the level of confidence is the probability with which we will accept a null hypothesis when it is true. So, there is 1 in 20 chance, i.e. 5% chance that our confidence interval does not include the true mean.

The confidence level stated as a proportion, rather than as a percentage is called as a confidence coefficient. For example, if you have a confidence level of 95%, the confidence coefficient would be 0.95. In contrast, if the confidence coefficient is high, the confident you are that your results are accurate. The confidence coefficient can be stated by the equation,

$\boldsymbol{c = 1 - \alpha}$, or we can say that,

$$Level\ of\ Confidence = 1 - Level\ of\ Significance$$

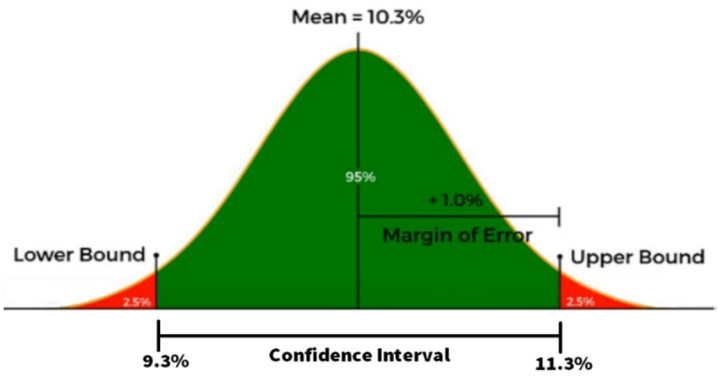



**Figure 3. Level of Confidence**

The striking example in Figure 3 shows that the 10.3% conversion rate is the mean. ±1.0% is the margin for error, and this in turn provides a confidence interval varying from 9.3% to 11.3%. Let's say that 10.3% ±1.0% at 95% confidence is our actual conversion rate. Here, 95% is the confidence level and 2.5%(left) + 2.5% (right) = 5% is the significance level indicating that if we take a total 20 samples, we can know with the complete certainty that the sample conversion rate would fall between 9.3% and 11.3% for atleast 19 of those samples.

*E. Critical Value Approach & P-Value Approach*

A critical value approach can be stated as an approach, including a position on the test statistic distribution under the null hypothesis that explains a set of values that entitled for rejecting the null hypothesis. This set is called the critical or rejection region. Usually, the one-tailed test has one critical value, while the two-tailed test has two critical values. If the test statistic is not as intense as the critical value, then the null hypothesis is not rejected, and if the test statistic is more intense than the critical value, then the null hypothesis is rejected in favour of the alternative hypothesis.

On the other hand, the p-value approach stands for the probability-value approach, which is the probability of getting a statistic atleast this far away from the mean if we were to assume that the null hypothesis is true. It is a measure for the power of the evidence in the data against the null hypothesis. Usually, the smaller is the p-value, the more substantial would be the sample evidence if for rejecting the null hypothesis. More specifically, the p-value is the lowest value of **σ** that results in the rejection of the null hypothesis. If the p-value is less than the significance level (**σ**), then the null hypothesis must be rejected. In contrast, if the p-value is greater than or equal to the significance level (**σ**), then the null hypothesis cannot be rejected.

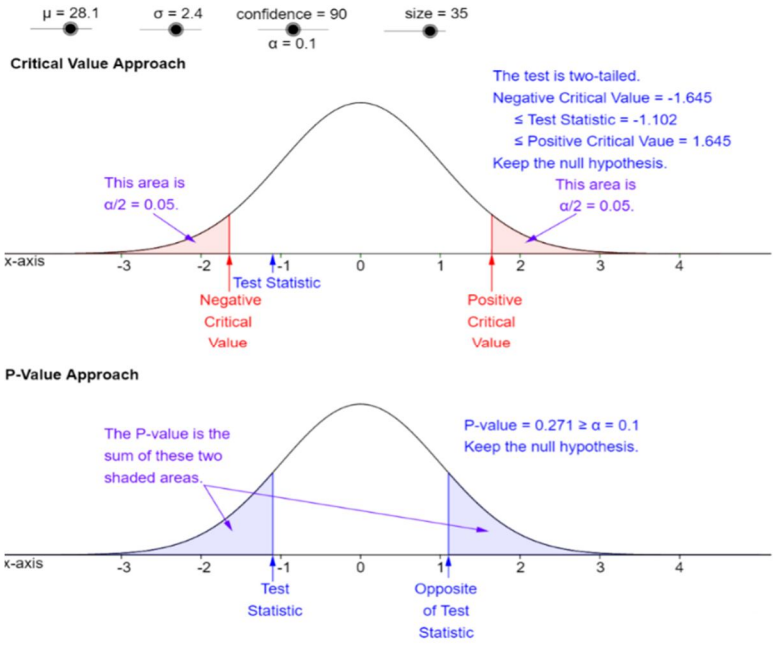



**Figure 4. Distribution overview of Critical Value Approach and P-Value Approach**

Figure 4. depicts the distribution overview of the critical value approach and p-value approach and how both of them holds the situation when mean, standard deviation, confidence level or significance level, and size are given out to be 28.1, 2.4, 90% or 0.10, and 35, respectively.

## III. TEST STATISTIC

In simple words, a Test Statistic is a random variable that is deliberated from the sample data and used in a hypothesis testing. The need for this concept arises in detecting whether to reject the null hypothesis. It allows us to quantify how close things are to our expectations or theory. It shows the difference between the observed data and what we expect if the null hypothesis is true. It compares the data with what is expected, under the null hypothesis. It can also be used to calculate the p-value. Based on the probability model assumed in the null hypothesis, different hypothesis tests use various test statistics; that is why it is essential to choose a suitable test statistic wisely. In further topics, we are going to understand when and how to use the test statistic. Some standard test and test statistics are shown in Table 4.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 8 Issue IX Sep 2020- Available at www.ijraset.com*

Table 4. Hypothesis Tests and their Test Statistics

| Hypothesis Test | Test Statistic |
|---|---|
| Z-Test | Z-statistic |
| t-Test | t-statistic |
| ANOVA | F-statistic |
| Chi-squared Tests | Chi-squared statistic |

### *A. Z-Test Statistic*

In Z-Test statistic, we can do the hypothesis test about means, the difference between means, proportion, or even difference between proportions. It tells how many fluctuations are above or below the mean score. At the centre, the mean is always 0, and the standard deviation is 1. The limit ranges from $-3\sigma$ to $+3\sigma$. Let's determine when to use a Z-statistics. We use the Z-statistic when,

*1)* The population of the standard deviation **(σ)** is mentioned.
*2)* The sample size **(n)** is greater than or equal to 30.
*3)* The null hypothesis $(H_0)$ may be one-tailed or two-tailed.
*4)* The data points are independent of each other.
*5)* The data should be normally distributed.

The formula for the Z-statistic test can be given as,

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where,

$\bar{x} \rightarrow$ Sample Mean
$\mu \rightarrow$ Population Mean
$\sigma \rightarrow$ Standard Deviation of the Population
$n \rightarrow$ Sample Size

### *B. Programming Approach*

Let's see an example and find the Z value using the Z-test statistic formula with the help of a programming approach,
The principal of the XYZ school claims that the students of his school are having an intelligence above average. 40 students are chosen randomly, and the mean IQ scores is noted to be 112.5. Now, we must find out whether the evidence is sufficient to support the principal's claim. Given that the population mean of student's IQ is 95 with a standard deviation of 15. Let the significance level be 0.05.
Follow the Hypothesis Testing steps.

*1)* Stating the Null and Alternative hypothesis. Here, the variables NH and AH represents the Null Hypothesis and Alternative Hypothesis, respectively.

```
NH = "PM = 100"
AH = "PM > 100"
```

**OR**

```
NH: μ = 75
AH: μ > 75
```

*2)* State the significance level. Here, the significance level is given as 0.05, which is equal to a Z-score of 1.645, so store the value of 1.645 in the variable APLHA,

```
ALPHA = 1.645
```

*3)* Now, let the variable SM be the sample mean, PM be the population mean, SDP be the standard deviation of the population and N be the total number of observations and put the values of each variable, respectively.

```
SM, PM, SDP, N = 112.5, 95, 15, 40
```

*4)* Now, form a Z-test statistic formula and calculate the Z value.

```
import math
Z = ((SM - PM)/(SDP/math.sqrt(N)))
print(Z)
```

```
7.378647873726218
```

*5)* Now compare the Z-value of 7.378 with the ALPHA value of 1.645. If the Z-value would be greater than the ALPHA value, then the Null Hypothesis can be rejected and if the Z-value would be less than the ALPHA value, then the Null Hypothesis cannot be rejected.

```
if Z > ALPHA:
  print("Null Hypothesis can be Rejected")
elif Z < ALPHA:
  print("Null Hypothesis cannot be rejected")
```

```
Null Hypothesis can be Rejected
```

*6)* It can be clearly seen that the Z-value is greater than the ALPHA value so the Null Hypothesis can be rejected, and we can state that the Principal's claim is right.

*Note: For performing the above Z-test statistic example, the Python programming language is used along with one of its libraries, i.e. math. The IDLE used is the Google COLAB

*C. Student t-test Statistic*

In t-test statistic, we can do the hypothesis test about the difference between the means of two groups which may be corelated in certain aspects. It allows testing of an assumption applicable to a population. It removes the problem of negative scores and decimal value and presents an excellent discriminating range. At the centre, the mean is always 50, and the standard deviation is 10. The limit ranges from **0** to **100.** Hence, in the Student t-test, the population's standard deviation is replaced by the sample's standard deviation in the usual Z-test formula. This kind of test statistic requires the number of independent observations in a dataset, i.e. the degree of freedom, and can be denoted as $\boldsymbol{df = 1 - n}$, where $\boldsymbol{n}$ is the sample size. Let's determine when to use a t-statistics. We use the t-statistic when,

*1)* The population of the sample **(S)** is mentioned.
*2)* The sample size **(n)** is less than 30.
*3)* The null hypothesis $\boldsymbol{(H_0)}$ may be one-tailed or two-tailed.

The formula for the t-statistic test can be given as,

$$\boldsymbol{t = \frac{\bar{x} - \mu}{S/\sqrt{n}}}$$

where,

$\boldsymbol{\bar{x}} \rightarrow$ Sample Mean
$\boldsymbol{\mu} \rightarrow$ Population Mean
$\boldsymbol{S} \rightarrow$ Standard Deviation of the Sample
$\boldsymbol{n} \rightarrow$ Sample Size

The t-distribution resembles the Z-distribution but with thicker tails. The tails are thicker because we are estimating the true population standard deviation. Estimation adds a little more uncertainty which means thicker tails since the extreme values are little more familiar. However, as we get more and more data, the t-distribution converges to the Z-distribution, so with large samples, the Z-tests and t-test should give a similar p-value.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 8 Issue IX Sep 2020- Available at www.ijraset.com*

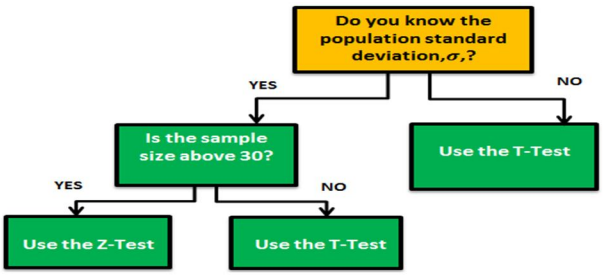



**Figure 5. Z-test Statistic Vs t-test statistic**

For a simple explanation, the flowchart in Figure 5 shows when the Z-test statistics or t-test statistics should be used.

*D. Anova (F-Test) Statistic*

The t-test statistic works perfectly fine while dealing with two groups but takes a lot of time when to compare more than two groups at the same time. The Analysis of variance or ANOVA is a statistical method used in analysing the means of two or more groups at the same time. It is often called as the F-test statistic. This concept is put by Sir Ronald Fisher. Unlike the Z and t-distributions, the ANOVA distribution does not have any negative values. Let's see the steps to find out the F ratio,

*1)* First find $\Sigma X$ & $\Sigma X^2$ for all the groups

*2)* *Correction Term:* Here, the correction term is denoted by $C_x$ and $N$ is the total number of samples.

$$C_x = \frac{(\Sigma X)^2}{N}$$

*3)* *Sum of Squares of Total:* It is the difference between the sum of all the squared of all the observations and the correction term.

$$SS_T = \Sigma X^2 - C_x$$

*4)* *Sum of Squares Among Groups:* It is the difference between the squared submission of all the observation divided by the small $n$ and the correction term. Here, $n$ be the number of samples in each category.

$$SS_A = \frac{(\Sigma X)^2}{n} - C_x$$

*5)* *Sum of Squares Within Groups:* It is the difference between the Sum of squares of total and the Sum of squares among groups.

$$SS_w = SS_T - SS_A$$

*6)* Mean of Sum of Squares among groups: It is the Sum of squares among group divided by $k-1$ , where $k$ be the number of categories. $k-1$ be the degree of freedom.

$$MSS_A = \frac{SS_A}{k-1}$$

*7)* *Mean of Sum of Squares within Groups:* It is the Sum of squares within group divided by $N-k$, where $N$ be the total number of samples and $k$ be the number of categories. $N-k$ is actually the degree of freedom.

$$MSS_w = \frac{SS_w}{N-k}$$

*8)* *F Ratio:* It is the fraction between the Mean of Sum of Squares among groups and Mean of Sum of Squares within groups.

$$F_{ratio} = \frac{MSS_A}{MSS_w}$$

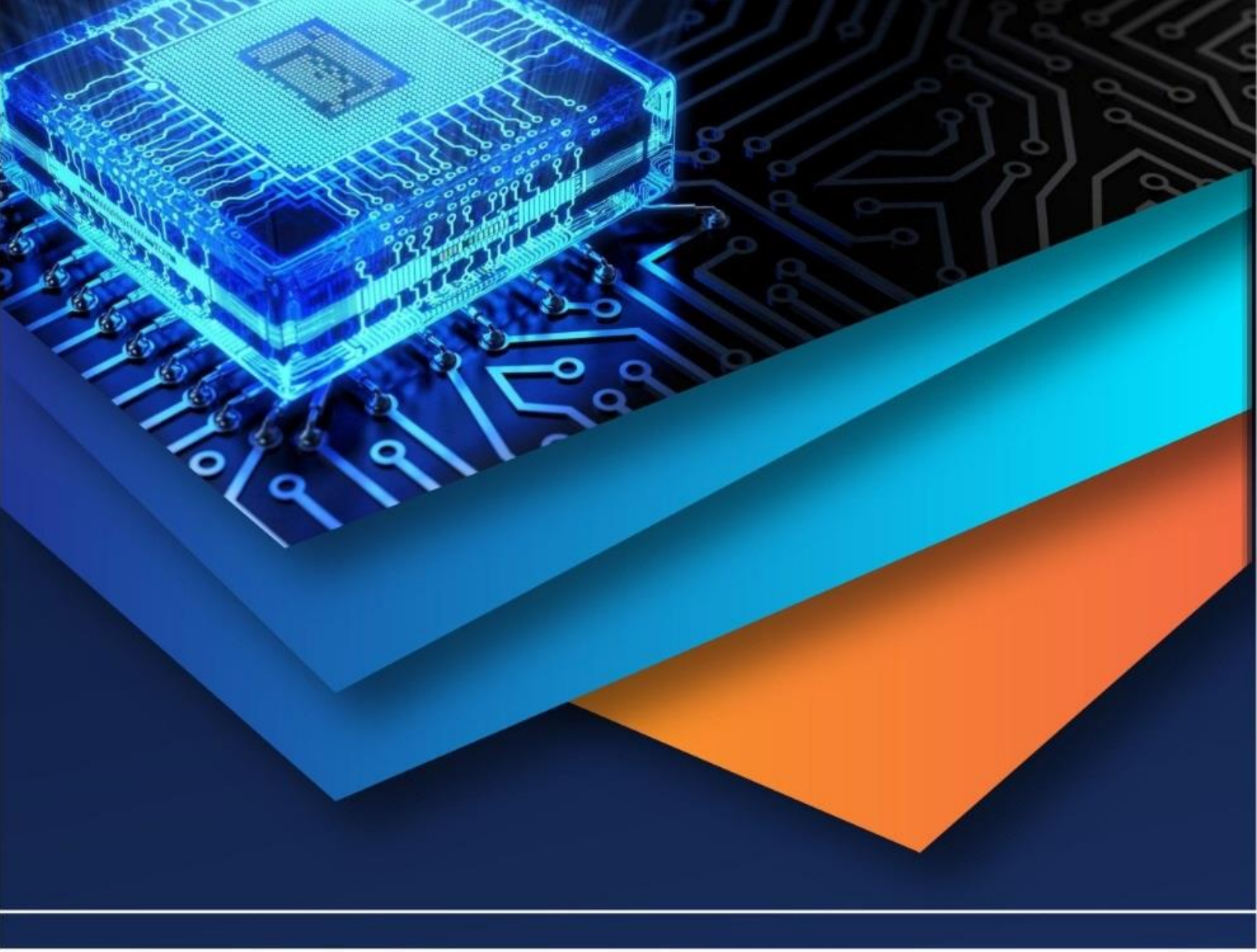
After following up the steps, we would get the value of the F-test statistic or the F ratio. Now, we could further use this F value to compute the hypothesis testing.

## IV. CONCLUSION

Hypothesis Testing has a vivid background that stands as a strong backbone for the required fields, such as Machine Learning and Data Science. Before jumping to the concepts of Machine Learning, students should be introduced and securely taught the Hypothesis Analogies so that they could think and observe the problems in a uniquely productive way. It has been predicted that the upcoming future holds a lot of potential for the Data Science field, so to have a grip on this field, statistics should be treated as an essential aspect. Knowing the same would help to build and train a better model with high accuracy, which would prove to be a tremendous achievement.

## REFERENCES

[1] Evangelista, F. & Hemenway, C. 2002. The use of the Jigsaw in hypothesis testing. 2nd International Conference on the Teaching of Mathematics at the Graduate Level. Hersonissos, Crete, Greece.
[2] Garfield, J.B. & Ben-Zvi, D. 2008. Developing students' statistical reasoning: Connecting research and teaching practice. Emeryville, CA: Springer.
[3] Glaser, R. 2003. Assessing expert knowledge representations of introductory statistics. CSE Tech. Rep. No. 600. Los Angeles. University of California, Center for Research on Evaluation, Standards, and Student Testing.
[4] Gliner, J.A., Leech, N.L. & Morgan, G.A. 2002. Problems with null hypothesis significance testing (NHST): What do the textbooks say? The Journal of Experimental Education 71(1): 83-92.
[5] Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? Methods of Psychological Research 7(1): 1-20.
[6] Link, W.C. 2002. An examination of student mistakes in setting up hypothesis testing problems. Louisiana-Mississippi Section of the Mathematical Association of America.
[7] Lipson, K., Kokonis, S. & Francis, G. 2006. Developing a computer interaction to enhance student understanding in statistical inference. Proceedings of the 7th International Conference on Teaching Statistics, ICOTS-7.
[8] Lipson, K., Kokonis, S. & Francis, G. 2003. Investigation of students' experiences with a web-based computer simulation. Proceedings of the 2003 IASE Satellite Conference on Statistics and the Internet. Berlin.
[9] Rossman, A.J. & Chance, B. 2004. Anticipating and addressing student misconceptions. ARTIST Conference on assessment in Statistics, Lawrence University. August, 1-4.
[10] Smith, T.M. 2008. An investigation into student understanding of statistical hypothesis testing. Doctoral Dissertation. Retrieved from Digital Repository at the University of Maryland. (umi-umd-5658.pdf)
[11] Sotos, C., Vanhoof, S., Noortgate, W. & Onghena, P. 2007. Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. Educational Research Review 2(2): 98-113.
[12] Sotos, C., Vanhoof, S., Noortgate, W. & Onghena, P. 2009. How confident are students in their misconceptions about hypothesis tests? Journal of Statistics Education 17(2).
[13] Weinberg, A., Wiesner, E. & Pfaff, T.J. 2010. Using informal inferential reasoning to develop formal concepts: Analyzing an activity. Journal of Statistics Education 18(2).
[14] Zieffler, A., Garfiled, J., DelMas, R. & Reading, C. 2008. A Framework to support research on Informal Inferential Reasoning. Statistics Education Research Journal 7(2), 40-58.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY