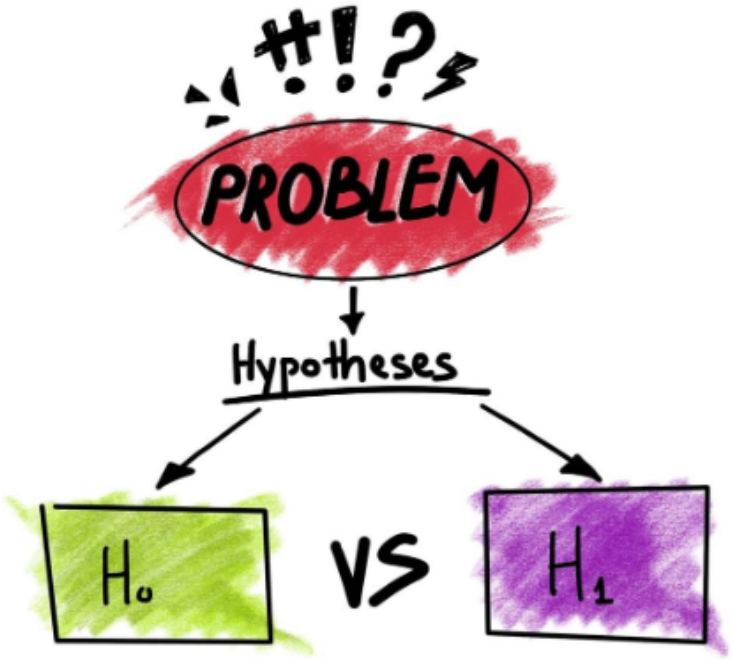
Hypothesis Testing

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about **population parameters**.



Null and Alternate Hypothesis

1. **Null hypothesis (H0)**:

In simple terms, the null hypothesis is a statement that assumes there is no significant effect or relationship between the variables being studied. It serves as the starting point for hypothesis testing and represents the **status quo** or the assumption of no effect until proven otherwise. The purpose of hypothesis testing is to gather evidence (data) to either reject or fail to reject the null hypothesis in favour of the alternative hypothesis, which claims there is a significant effect or relationship.

2. **Alternative hypothesis (H1 or Ha)**:

2. **Alternative hypothesis (H1 or Ha)**:

The alternative hypothesis, is a statement that contradicts the null hypothesis and claims there is a significant effect or relationship between the variables being studied. It represents the **research hypothesis** or the claim that the researcher wants to support through statistical analysis.

* Important Points

- How to decide what will be Null hypothesis and what will be Alternate Hypothesis(Typically the Null hypothesis says nothing new is happening)
- We try to gather evidence to reject the null hypothesis
- It's important to note that failing to reject the null hypothesis doesn't necessarily mean that the null hypothesis is true; it just means that there isn't enough evidence to support the alternative hypothesis.

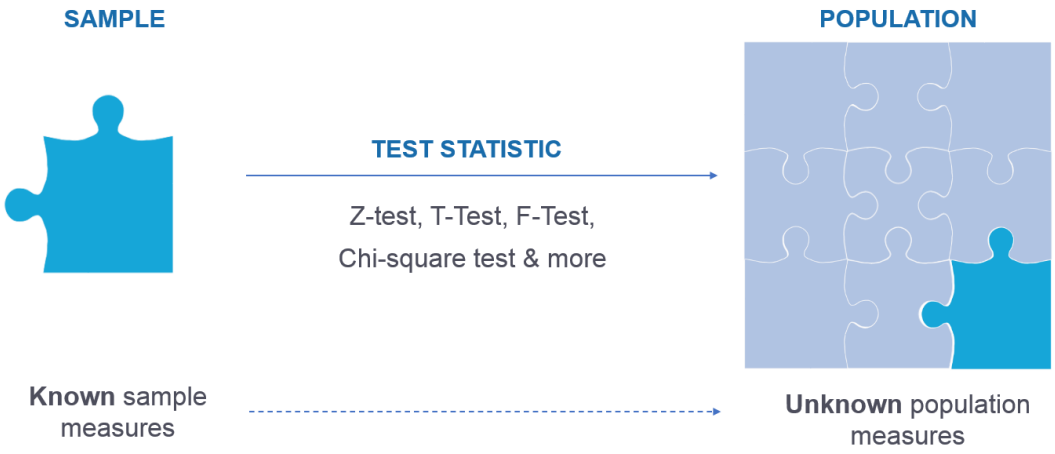
Example : Hypothesis tests are similar to jury trials, in a sense. In a jury trial, H0 is similar to the not-guilty verdict, and Ha is the guilty verdict. You assume in a jury trial that the defendant isn't guilty unless the prosecution can show beyond a reasonable doubt that he or she is guilty. If the jury says the evidence is beyond a reasonable doubt, they reject H0, not guilty, in favour of Ha , guilty.

1. Rejection Region Approach (Basic Approach)

A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected. i.e. if the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis.

Hypothesis Testing Steps for Rejection Region Approach

1. Formulate a Null and Alternate hypothesis
2. Select a significance level(This is the probability of rejecting the null hypothesis when it is actually true, usually set at 0.05 or 0.01)
3. Check assumptions (example distribution)
4. Decide which test is appropriate(Z-test, T-test, Chi-square test, ANOVA)
5. State the relevant test statistic
6. Conduct the test
7. Reject or not reject the Null



Example 1: Performing a Z test

Q: Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was **50 units** per day with a known population standard deviation of **5 units**. After implementing the training program, the company measures the productivity of a random sample of **30 employees**. The sample has an average productivity of **53 units** per day. The company wants to know if the new training program has significantly **increased productivity**.

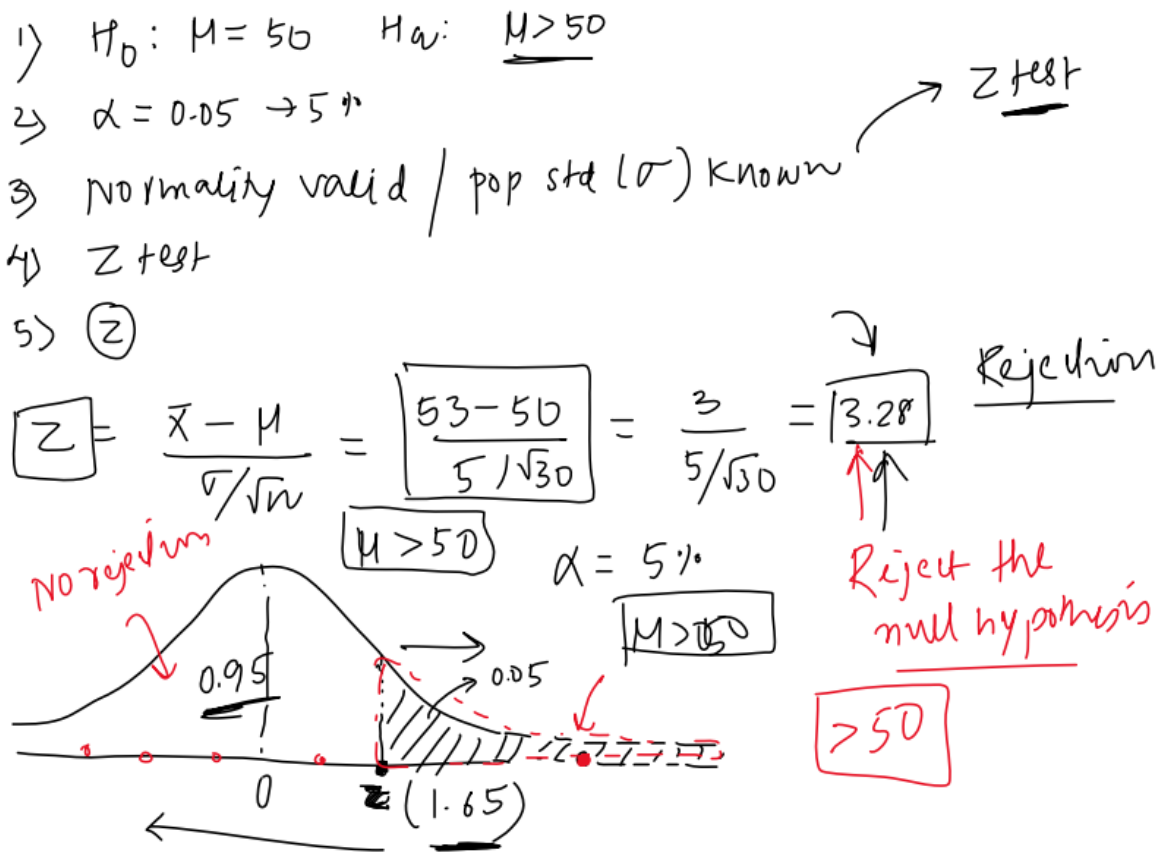
Solution " Here ,

μ = 50

σ(sigma) = 5

n =30

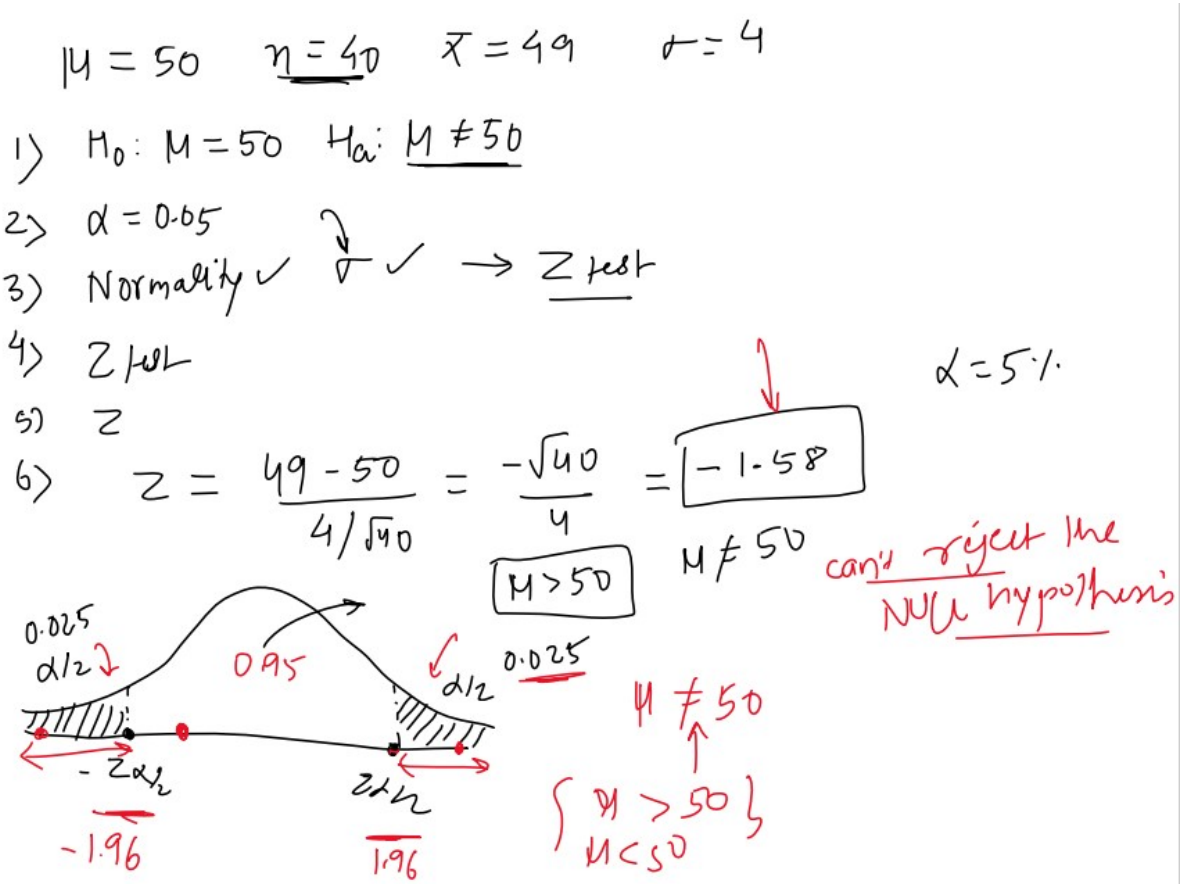
(x̄) = 53



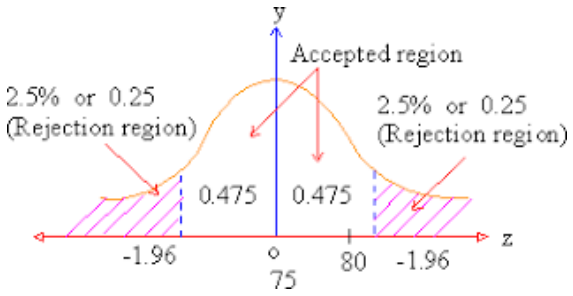
Example : 2

Suppose a snack food company claims that their Lays wafer packets contain an average weight of **50 grams per packet**. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed **50 grams**. The organization collects a random sample of **40 Lays** wafer packets and measures their weights. They find that the sample has an average weight of **49 grams**, with a known population **standard deviation of 4 grams**.

Solution :



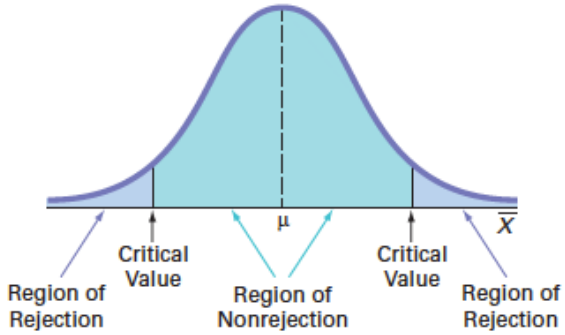
Rejection Region



Significance level

- it is denoted as α (alpha), is a predetermined threshold used in hypothesis testing to determine whether the null hypothesis should be rejected or not. It represents the probability of rejecting the null hypothesis when it is actually true, also known as Type 1 error.

- The critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.



Problem with Rejection Region Approach

The rejection region approach is a commonly used statistical method for hypothesis testing. It involves setting up a critical region (also known as the rejection region) based on a significance level, and if the test statistic falls within this region, the null hypothesis is rejected in favor of the alternative hypothesis. However, there can be certain issues or problems associated with the rejection region approach. Let's discuss a few of them:

1. **Subjectivity in choosing the significance level**: The rejection region approach requires selecting a significance level (usually denoted as α) before conducting the test. The significance level determines the probability of rejecting the null hypothesis when it is actually true. However, the choice of significance level is subjective and can vary among researchers or analysts. Different significance levels can lead to different conclusions, which may introduce inconsistency in the interpretation of results.

2. **Limited information about alternative hypotheses**: The rejection region approach primarily focuses on rejecting the null hypothesis in favor of the alternative hypothesis. However, it does not provide detailed information about the alternative hypothesis itself. It only determines whether the observed data falls within the rejection region or not. This limitation can make it difficult to understand the nature or magnitude of the effect being studied.

3. **Lack of power analysis**: The rejection region approach does not explicitly consider the power of the statistical test. Power refers to the probability of correctly rejecting the null hypothesis when it is false. Without power analysis, it is challenging to determine the sample size required to detect a meaningful effect size or to evaluate the reliability of the test results.

4. **Ignoring effect sizes**: The rejection region approach is primarily concerned with the statistical significance of the test results rather than the practical significance or effect size. It may be possible to obtain statistically significant results with a large sample size, but the effect size may be too small to have any practical relevance. Focusing solely on statistical significance may lead to misleading conclusions.

5. **One-tailed vs. two-tailed tests**: The rejection region approach assumes a specific alternative hypothesis, either one-tailed (directional) or two-tailed (non-directional). However, selecting the appropriate type of test can be challenging, and making an incorrect choice can affect the validity of the conclusions. It requires careful consideration of the research question and prior knowledge.

To address some of these problems, alternative approaches such as confidence intervals, effect size estimation, and Bayesian methods have gained popularity. These methods provide a more comprehensive and informative analysis of the data, taking into account effect sizes, uncertainty intervals, and the strength of evidence for or against hypotheses.

Type 1 vs Type 2 Error

In hypothesis testing, there are two types of errors that can occur when making a decision about the null

- **Type-I** (False Positive) error occurs when the sample results, lead to the rejection of the null hypothesis when it is in fact true.

In other words, it's the mistake of finding a significant effect or relationship when there is none. The probability of committing a Type I error is denoted by **α (alpha)**, which is also known as the significance level. By choosing a significance level, researchers can control the risk of making a Type I error.

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

- **Type-II** (False Negative) error occurs when based on the sample results, the null hypothesis is not rejected when it is in fact false.

This means that the researcher fails to detect a significant effect or relationship when one actually exists. The probability of committing a Type II error is denoted by **β (beta)**.

Example: Type I vs Type II error

You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur:

- **Type I error (false positive)**: the test result says you have coronavirus, but you actually don't.
- **Type II error (false negative)**: the test result says you don't have coronavirus, but you actually do.

Trade-off between Type 1 and Type 2 errors

In summary, the trade-off between Type 1 and Type 2 errors can be summarized as follows:

- Type 1 error (false positive) occurs when the null hypothesis is wrongly rejected.
- Type 2 error (false negative) occurs when the null hypothesis is erroneously accepted.
- Decreasing the probability of Type 1 errors increases the probability of Type 2 errors, and vice versa.

- The trade-off involves finding a balance between the two error types based on the specific context, consequences, and costs associated with each error.
- The choice of significance level, sample size, and effect size can influence the trade-off.
- Researchers need to consider the practical implications and relative costs of each error type to make informed decisions in hypothesis testing.

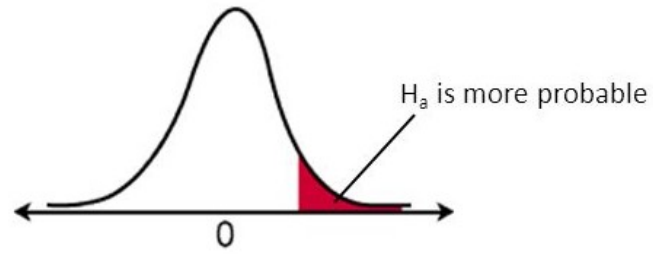
One sided vs two sided test

- **One-sided (one-tailed) test**: A one-sided test is used when the researcher is interested in testing the effect in a specific direction (either greater than or less than the value specified in the null hypothesis). The alternative hypothesis in a one-sided test contains an inequality (either ">" or "<").

Example: A researcher wants to test whether a new medication increases the average recovery rate compared to the existing medication.

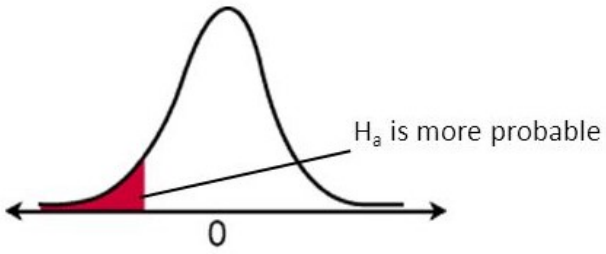
- **Two-sided (two-tailed) test**: A two-sided test is used when the researcher is interested in testing the effect in both directions (i.e., whether the value specified in the null hypothesis is different, either greater or lesser). The alternative hypothesis in a two-sided test contains a "not equal to" sign (≠).

Example: A researcher wants to test whether a new medication has a different average recovery rate compared to the existing medication.



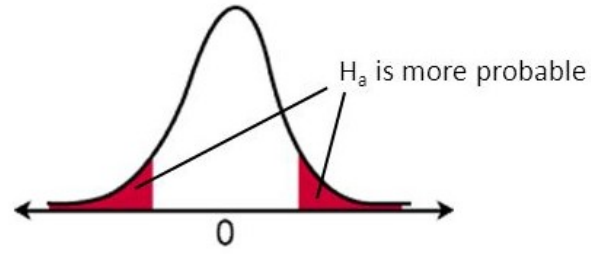
Right-tail test

$H_a: \mu > \text{value}$



Left-tail test

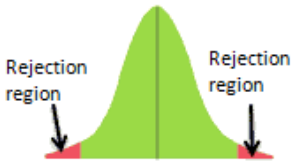
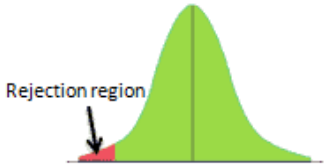
$H_a: \mu < \text{value}$



Two-tail test

$H_a: \mu \neq \text{value}$

The main difference between them lies in the directionality of the alternative hypothesis and how the significance level is distributed in the critical regions.

Z- Test	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)	Statistical conclusion
Two-tailed	$\mu=\mu_0$	$\mu\neq\mu_0$	
Left-tailed	$\mu\geq\mu_0$	$\mu<\mu_0$	
Right-tailed	$\mu\leq\mu_0$	$\mu>\mu_0$	

Advantages and Disadvantages?

Two-tailed test (two-sided):

- **Advantages**:

1. **Detects effects in both directions**: Two-tailed tests can detect effects in both directions, which makes them suitable for situations where the direction of the effect is uncertain or when researchers want to test for any difference between the groups or variables.
2. **More conservative**: Two-tailed tests are more conservative because the significance level (α) is split between both tails of the distribution. This reduces the risk of Type I errors in cases where the direction of the effect is uncertain.

- **Disadvantages**:

1. **Less powerful**: Two-tailed tests are generally less powerful than one-tailed tests because the significance level (α) is divided between both tails of the distribution. This means the test requires a larger effect size to reject the null hypothesis, which could lead to a higher risk of Type II errors (failing to reject the null hypothesis when it is false).
2. **Not appropriate for directional hypotheses**: Two-tailed tests are not ideal for cases where the research question or hypothesis is directional, as they test for differences in both directions, which may not be of interest or relevance.

One-tailed test (one-sided):

- **Advantages**:

1. **More powerful**: One-tailed tests are generally more powerful than two-tailed tests, as the entire significance level (α) is allocated to one tail of the distribution. This means that the test is more likely to detect an effect in the specified direction, assuming the effect exists.

2. **Directional hypothesis**: One-tailed tests are appropriate when there is a strong theoretical or practical reason to test for an effect in a specific direction.

- **Disadvantages**:

1. **Missed effects**: One-tailed tests can miss effects in the opposite direction of the specified alternative hypothesis. If an effect exists in the opposite direction, the test will not be able to detect it, which could lead to incorrect conclusions.

2. **Increased risk of Type I error**: One-tailed tests can be more prone to Type I errors if the effect is actually in the opposite direction than the one specified in the alternative hypothesis.

Where can be Hypothesis Testing Applied?

1. **Comparing means or proportions**: Hypothesis testing can be used to compare means or proportions between two or more groups to determine if there's a significant difference. This can be applied to compare average customer satisfaction scores, conversion rates, or employee performance across different groups.

2. **Analysing relationships between variables**: Hypothesis testing can be used to evaluate the association between variables, such as the correlation between age and income or the relationship between advertising spend and sales.
3. **Evaluating the goodness of fit**: Hypothesis testing can help assess if a particular theoretical distribution (e.g., normal, binomial, or Poisson) is a good fit for the observed data.

4. **Testing the independence of categorical variables**: Hypothesis testing can be used to determine if two categorical variables are independent or if there's a significant association between them. For example, it can be used to test if there's a relationship between the type of product and the likelihood of it being returned by a customer.

5. **A/B testing**: In marketing, product development, and website design, hypothesis testing is often used to compare the performance of two different versions (A and B) to determine which one is more effective in terms of conversion rates, user engagement, or other metrics.

Hypothesis Testing ML (Machine Learning) Applications

1. **Model comparison**: Hypothesis testing can be used to compare the performance of different machine learning models or algorithms on a given dataset. For example, you can use a paired t-test to compare the accuracy or error rate of two models on multiple cross- validation folds to determine if one model performs significantly better than the other.

2. **Feature selection**: Hypothesis testing can help identify which features are significantly related to the target variable or contribute meaningfully to the model's performance. For example, you can use a t-test, chi-square test, or ANOVA to test the relationship between individual features and the target variable. Features with significant relationships can be selected for building the model, while non-significant features may be excluded.
3. **Hyperparameter tuning**: Hypothesis testing can be used to evaluate the performance of a model trained with different hyperparameter settings. By comparing the performance of models with different hyperparameters, you can determine if one set of hyperparameters leads to significantly better performance.

4. **Assessing model assumptions**: In some cases, machine learning models rely on certain statistical assumptions, such as linearity or normality of residuals in linear regression. Hypothesis testing can help assess whether these assumptions are met, allowing you to determine if the model is appropriate for the data.

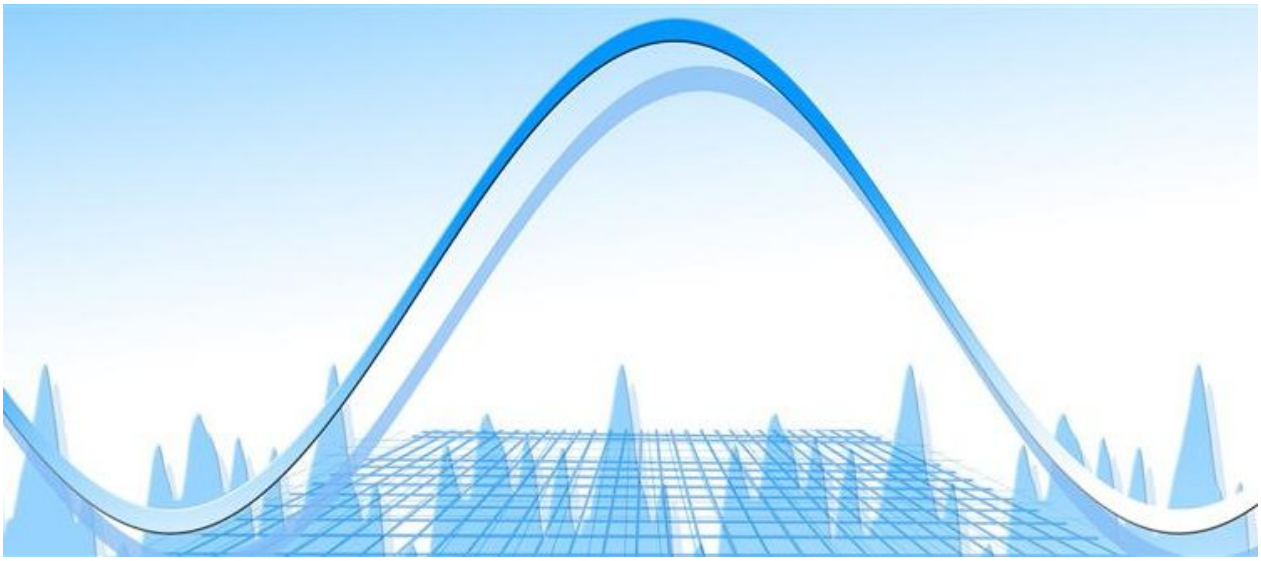
P-value

P-value is the probability of getting a sample as or more extreme(having more evidence against H0) than our own sample given the Null Hypothesis(H0) is true.

In statistics, the p-value is a measure of the strength of evidence against the null hypothesis. It is commonly used in hypothesis testing to determine whether the observed data is statistically significant or simply due to random chance.

- The null hypothesis is a statement that assumes there is no significant difference or relationship between variables in a population. The alternative hypothesis, on the other hand, suggests that there is a significant difference or relationship.
- The p-value represents the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed, assuming the null hypothesis is true. In other words, it quantifies the likelihood of observing the data or more extreme data, given that the null hypothesis is correct.
- Typically, in hypothesis testing, if the p-value is below a predetermined significance level (often denoted as α), which is typically set to 0.05 (5%), it is considered statistically significant. This means that the evidence suggests that the null hypothesis is unlikely, and the alternative hypothesis is more likely to be true. Conversely, if the p-value is above the significance level, we fail to reject the null hypothesis.

It's important to note that a statistically significant result does not guarantee practical or meaningful significance, and a non-significant result does not necessarily mean there is no effect. The p-value is just one factor to consider in the overall interpretation of statistical analysis.



In simple words p-value is a measure of the strength of the evidence against the Null Hypothesis that is provided by our sample data

Interpreting p-value

With significance value

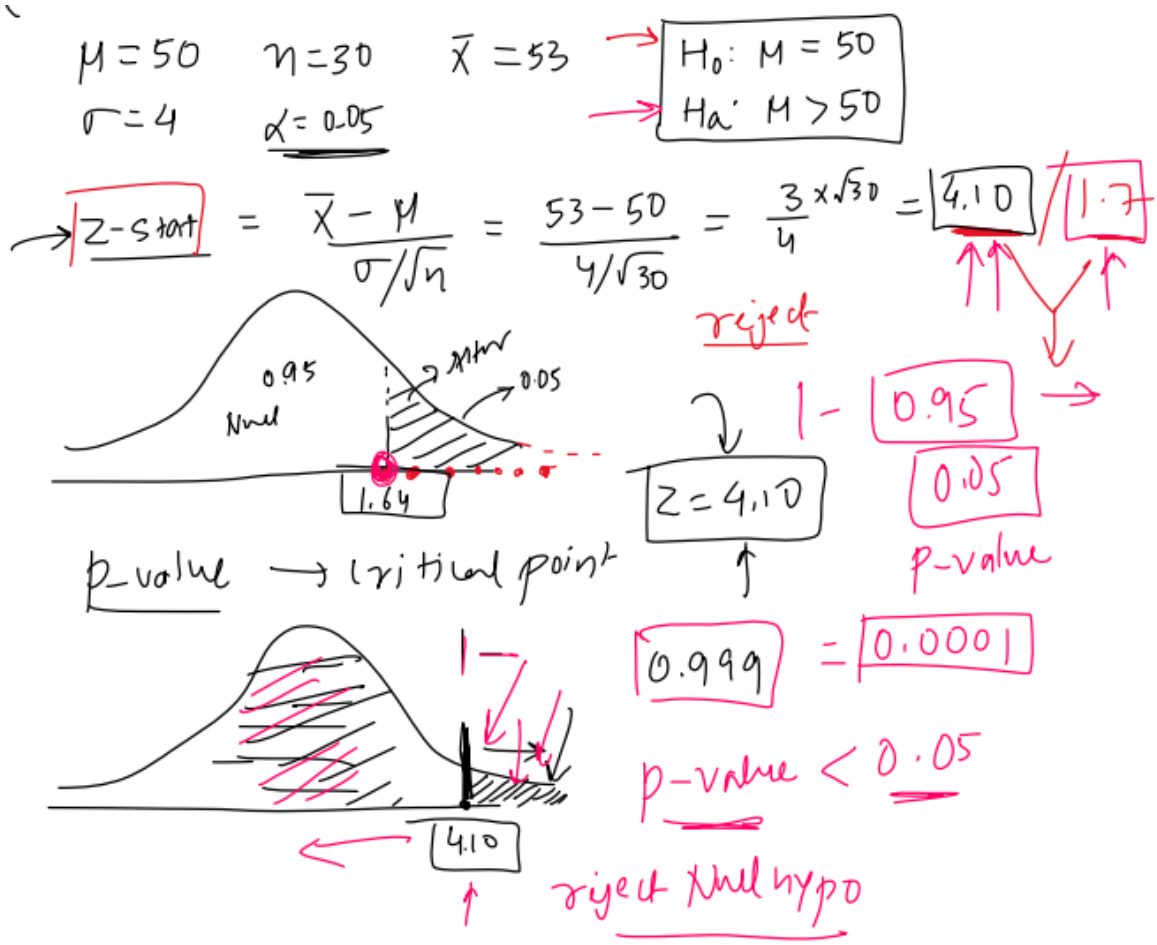
σ = 0.05 or 0.01

Without significance value

1. **Very small p-values** (e.g., p < 0.01) indicate strong evidence against the null hypothesis, suggesting that the observed effect or difference is unlikely to have occurred by chance alone.
2. **Small p-values** (e.g., 0.01 ≤ p < 0.05) indicate moderate evidence against the null hypothesis, suggesting that the observed effect or difference is less likely to have occurred by chance alone.
3. **Large p-values** (e.g., 0.05 ≤ p < 0.1) indicate weak evidence against the null hypothesis, suggesting that the observed effect or difference might have occurred by chance alone, but there is still some level of uncertainty.
4. **Very large p-values** (e.g., p ≥ 0.1) indicate weak or no evidence against the null hypothesis, suggesting that the observed effect or difference is likely to have occurred by chance alone.

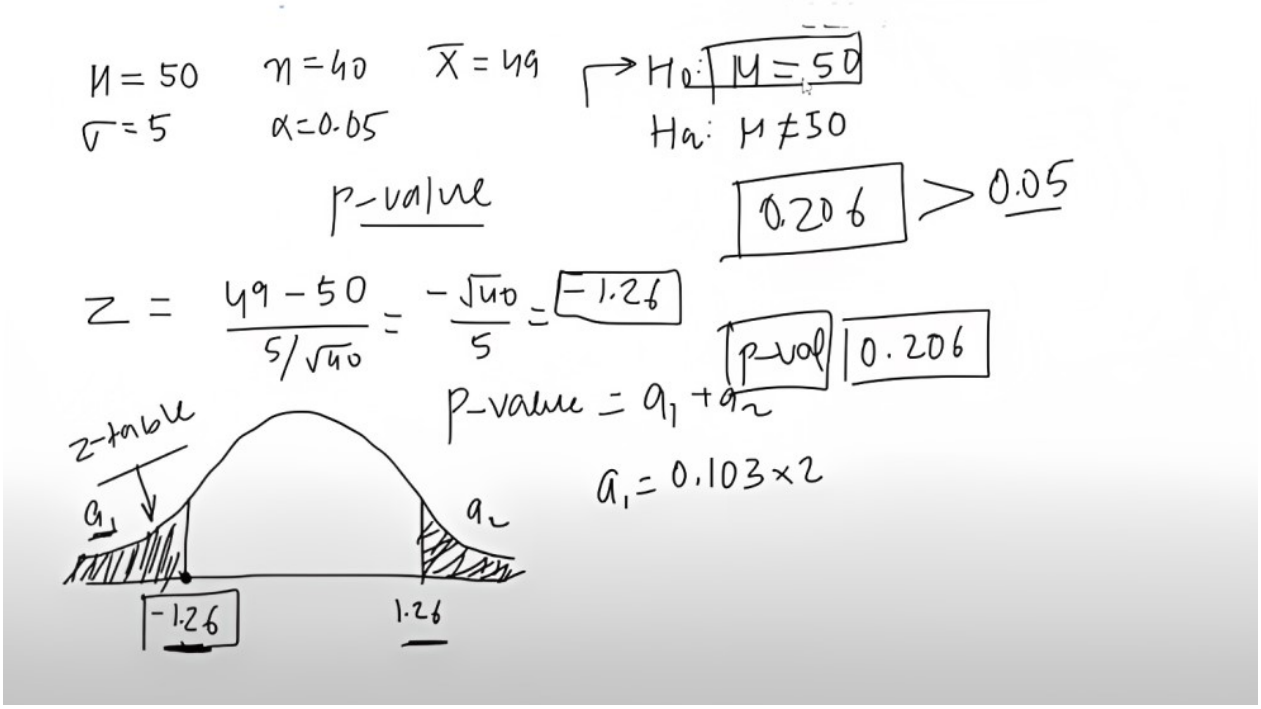
P-value in context of Z-test

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was **50** units per day. After implementing the training program, the company measures the productivity of a random sample of **30** employees. The sample has an average productivity of **53** units per day and the pop std is **4**. The company wants to know if the new training program has significantly increased productivity.



Example : 2 (Z -Test with 2 tail test)

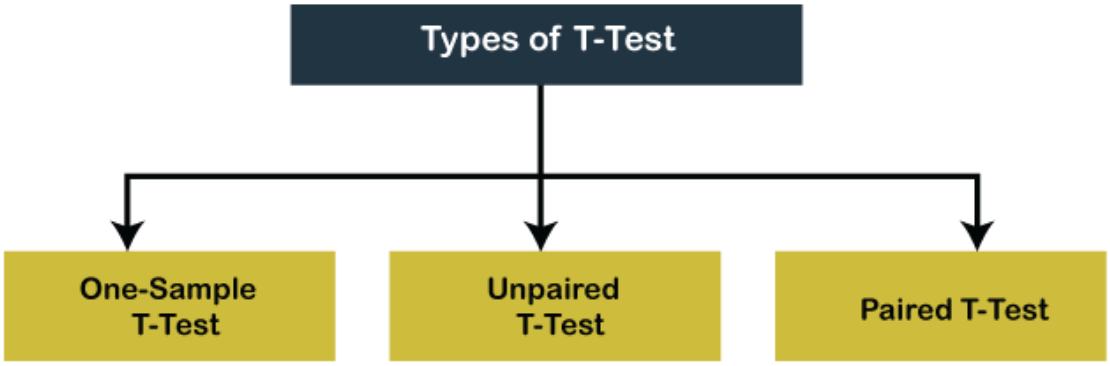
Suppose a snack food company claims that their Lays wafer packets contain an average weight of **50** grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed **50** grams. The organization collects a random sample of **40** Lays wafer packets and measures their weights. They find that the sample has an average weight of **49** grams, with a pop standard deviation of **5** grams.



T -test

A t-test is a statistical test used in hypothesis testing to compare the means of two samples or to compare a sample mean to a known population mean. The t-test is based on the t□distribution, which is used when the population standard deviation is unknown and the sample size is small.

There are three main types of t-tests:



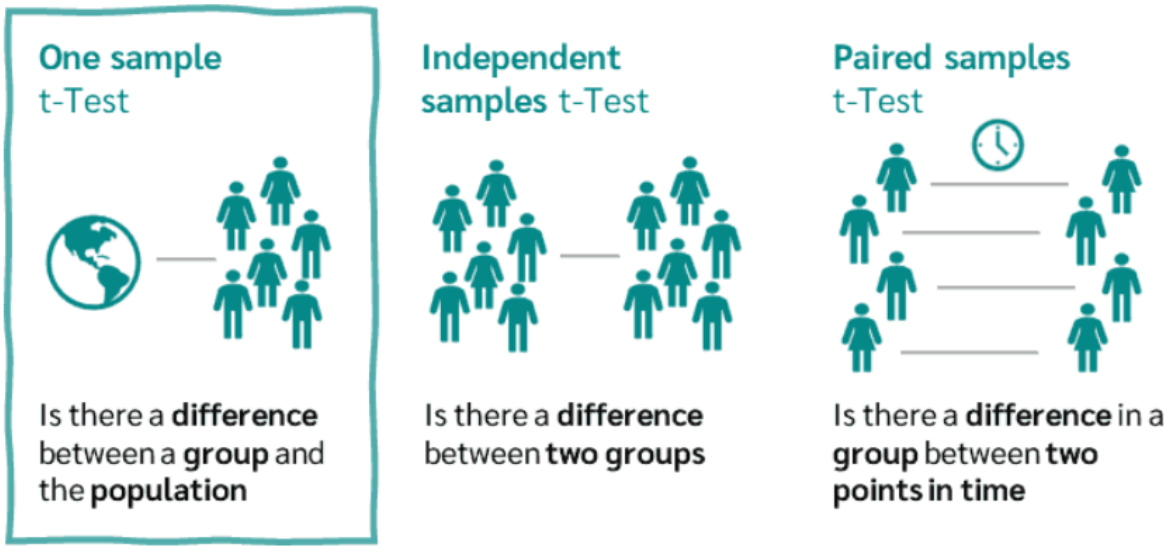
- **One-sample t-test**: The one-sample t-test is used to compare the mean of a single sample to a known population mean. The null hypothesis states that there is no significant difference between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.
- **Independent two-sample t-test**: The independent two-sample t-test is used to compare the means of two independent samples. The null hypothesis states that there is no significant difference between the means of the two samples, while the alternative hypothesis states that there is a significant difference.
- **Paired t-test (dependent two-sample t-test)**: The paired t-test is used to compare the means of two samples that are dependent or paired, such as pre-test and post-test scores for the same group of subjects or measurements taken on the same subjects under two different conditions. The null hypothesis states that

there is no significant difference between the means of the paired differences, while the alternative hypothesis

Formula

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

Single(One- Sample T-test)



A one-sample t-test checks whether a sample mean differs from the population mean.

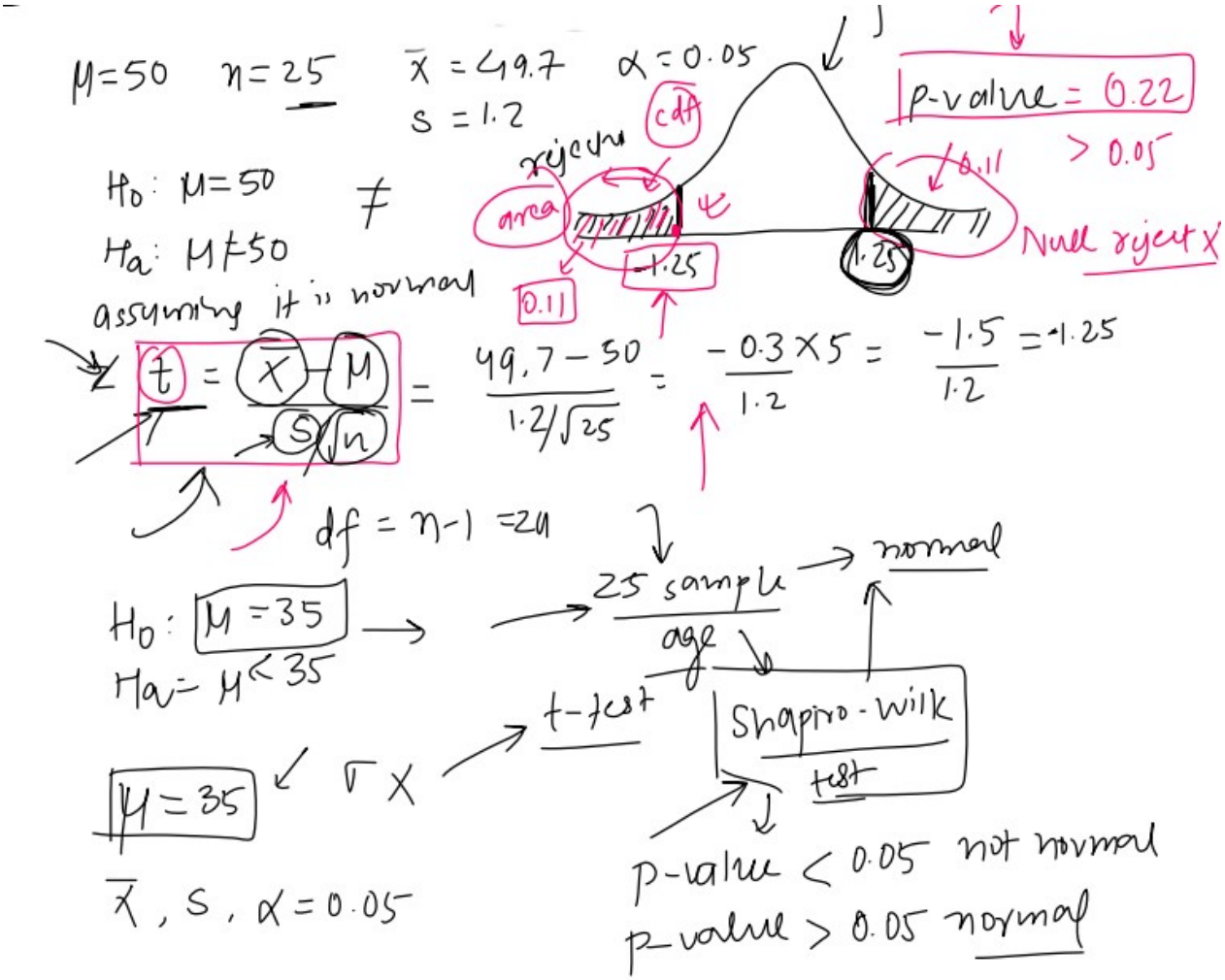
Assumptions for a single sample t-test

1. **Normality** - Population from which the sample is drawn is normally distributed
2. **Independence** - The observations in the sample must be independent, which means that the value of one observation should not influence the value of another observation.
3. **Random Sampling** - The sample must be a random and representative subset of the population.
4. **Unknown population std** - The population std is not known.

Example

Suppose a manufacturer claims that the average weight of their new chocolate bars is **50** grams, we highly doubt that and want to check this so we drew out a sample of **25** chocolate bars and measured their weight, the sample mean came out to be **49.7** grams and the sample std deviation was **1.2** grams. Consider the significance level to be **0.05**

Solution



In [1]:
```
# code

from scipy.stats import t

# Set the t-value and degrees of freedom ( left side value)
t_value =  -1.25
df = 21  # Replace this with your specific degrees of freedom ( n-1)

# Calculate the CDF value

cdf_value = t.cdf(t_value, df)

print(cdf_value)
```

```
0.11252538445659269
```

```
so , 0.11 + 0.11 =0.22 , its greater than 0.05
we cant reject Null Hypothesis
```

Case Study : titanic-single-sample-t-test

In [2]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
train_df = pd.read_csv("D:\\datascience\\Nitish sir\\Inferential Statistics\\titanic_train.csv
test_df = pd.read_csv("D:\\datascience\\Nitish sir\\Inferential Statistics\\titanic_test.csv"
```

In [4]:
```
# Concat

df = pd.concat([train_df.drop(columns=['Survived']), test_df]).sample(1309)
```

In [5]:
```
df.head()
```

Out[5]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
279	1171	2	Oxenham, Mr. Percy Thomas	male	22.0	0	0	W./C. 14260	10.5000	NaN	S
146	1038	1	Hilliard, Mr. Herbert Henry	male	NaN	0	0	17463	51.8625	E46	S
627	628	1	Longley, Miss. Gretchen Fiske	female	21.0	0	0	13502	77.9583	D9	S
655	656	2	Hickman, Mr. Leonard Mark	male	24.0	2	0	S.O.C. 14879	73.5000	NaN	S
624	625	3	Bowen, Mr. David John "Dai"	male	21.0	0	0	54636	16.1000	NaN	S

In [6]:
```
# Population

pop = df['Age'].dropna()
```

In [7]:
```
pop.mean()
```

Out[7]:
```
29.881137667304014
```

In [8]:
```
# Samples

sample_age = pop.sample(25).values
```

In [9]:
```
sample_age
```

Out[9]:
```
array([30., 54., 16., 21., 23., 44., 29., 26., 22., 32., 67., 43., 28.,
       23., 28., 40., 46., 47., 45., 50., 47., 46., 18., 57.,  2.])
```

```
 Hypothesis Testing :
----------------------
 H0 -> The mean age is 35
 H1 -> The mean is less than 35
```

In [10]:
```
# check for normality using Shapiro Wilk test

from scipy.stats import shapiro

shapiro_age = shapiro(sample_age) # Normal distribution)

print(shapiro_age)
```

```
ShapiroResult(statistic=0.9718778729438782, pvalue=0.6929346919059753)
```

In [11]:
```
pop_mean = 35 # Assuming
```

In [12]:
```
import scipy.stats as stats

t_statistic, p_value = stats.ttest_1samp(sample_age, pop_mean) # T-Test_1sample

print("t-statistic:", t_statistic) # sides of normal distribution

print("p-value:", p_value/2) #  divide p value with 2 because its one tailed test
```

```
t-statistic: 0.11856225782628638
p-value: 0.45330447345089925
```

In [13]:
```
alpha = 0.05

if p_value < alpha:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
```

```
Fail to reject the null hypothesis.
```

Independent 2 sample t-test

An independent two-sample t-test, also known as an unpaired t-test, is a statistical method used to compare the means of two independent groups to determine if there is a significance difference between them.

Assumptions for the test:

1. **Independence of observations**: The two samples must be independent, meaning there is no relationship between the observations in one group and the observations in the other group. The subjects in the two groups should be selected randomly and independently.

2. **Normality**: The data in each of the two groups should be approximately normally distributed. The t-test is considered robust to mild violations of normality, especially when the sample sizes are large (typically n ≥ 30)

and the sample sizes of the two groups are similar. If the data is highly skewed or has substantial outliers, consider using a non- parametric test, such as the Mann-Whitney U test.

3. **Equal variances (Homoscedasticity)**: The variances of the two populations should be approximately equal. This assumption can be checked using F-test for equality of variances. If this assumption is not met, you can use Welch's t-test, which does not require equal variances.

4. **Random sampling**: The data should be collected using a random sampling method from the respective populations. This ensures that the sample is representative of the population and reduces the risk of selection bias.

Fomula :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Example

Suppose a website owner claims that there is no difference in the average time spent on their website between desktop and mobile users. To test this claim, we collect data from **30** desktop users and **30** mobile users regarding the time spent on the website in minutes. The sample statistics are as follows:

In [14]:
```
desktop_users = [[12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25,
                  17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14]]
```

In [15]:
```
mobile_users = [[10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17,
11, 18, 15, 10, 16, 15, 13, 16, 11]]
```

```
Desktop_users :
-------------------------
Sample size (n1): 30
Sample mean (mean1): 18.5 minutes
Sample standard deviation (std_dev1): 3.5 minutes
```

```
  mobile_users :
-------------------------
 Sample size (n2): 30
Sample mean (mean1): 14.3 minutes
Sample standard deviation (std_dev1): 2.7 minutes
```

```
  Hypothesis Testing :
 ----------------------
H0 -> Avg time spend on desktop = Avg time spend on mobile usage is same
H1 -> Avg time spend on desktop  ≠ Avg time spend on mobile usage is not same
```

Normality Test

In [16]:
```
# The Shapiro-Wilk test returns two values: the test statistic (W) and the p-value.
# If the p-value is greater than your chosen significance level (a = 0.05),
# you can assume the data comes from a normally distributed population.
# If the p-value is less than or equal to the significance level, the normality assumption is
# and you should consider applying a data transformation or using a non-parametric test like t


from scipy.stats import shapiro

# Input the data as lists
desktop_users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 
mobile_users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 1

# Perform the Shapiro-Wilk test for both desktop and mobile users
shapiro_desktop = shapiro(desktop_users)
shapiro_mobile = shapiro(mobile_users)

print("Shapiro-Wilk test for desktop users:", shapiro_desktop)
print("Shapiro-Wilk test for mobile users:", shapiro_mobile)
```

```
Shapiro-Wilk test for desktop users: ShapiroResult(statistic=0.9783118963241577, pvalue=0.77
91106104850769)
Shapiro-Wilk test for mobile users: ShapiroResult(statistic=0.9714352488517761, pvalue=0.579
1513919830322)
```

```
        BOTH has P Value > 0.005 = Normal Distribution
```

In [17]:
```
# If the p-value from Levene's test is greater than your chosen significance level (a = 0.05),
#  If the p-value is less than or equal to the significance level, the assumption of equal var
# and you should consider using Welch's t-test instead of the regular independent two-sample t

from scipy.stats import levene

# Input the data as lists
desktop_users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 
mobile_users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 1

# Perform Levene's test
levene_test = levene(desktop_users, mobile_users)
print(levene_test)
```

```
LeveneResult(statistic=2.94395488191752, pvalue=0.09153720526741761)
```

* HERE P value is greater than 0.05 , Its Variances of Pop.Variance of A = Pop.Variance of B

* Equal Variances (Homoscedasticity)

In [18]:
```
from scipy.stats import t

# Set the t-value and degrees of freedom

t_value =  -5.25 # from formula

df = 58  # Replace this with your specific degrees of freedom ( n1 + n2 -2 )

# Calculate the CDF value
cdf_value = t.cdf(t_value, df)
print(cdf_value*2) # 2 sides
```

```
2.256369746933224e-06
```

WE Reject Null Hypothesis

Example : titanic-2-sample-t-test

In [19]:
```
train_df = pd.read_csv("D:\\datascience\\Nitish sir\\Inferential Statistics\\titanic_train.csv
test_df = pd.read_csv("D:\\datascience\\Nitish sir\\Inferential Statistics\\titanic_test.csv"
```

In [20]:
```
# Concat

df = pd.concat([train_df.drop(columns=['Survived']), test_df]).sample(1309)
```

In [21]:
```
pop_male = df[df['Sex'] == 'male']['Age'].dropna()
pop_female = df[df['Sex'] == 'female']['Age'].dropna()
```

In [22]:
```
pop_female
```

Out[22]:
```
518    36.0
322    30.0
119    29.0
133    29.0
53     29.0
       ...
416    34.0
275    63.0
248    29.0
56     21.0
865    42.0
Name: Age, Length: 388, dtype: float64
```

In [23]:
```
sample_male = pop_male.sample(25)
sample_female = pop_female.sample(25)

alpha = 0.05
```

```
   Hypothesis Testing
-----------------------------------------
H0 - Mean age of male and female are similar
H1 - Mean age of male is higher than female
```

In [24]:
```
pop_male.mean()
```

Out[24]: 30.58522796352584

In [25]:
```
pop_female.mean()
```

Out[25]: 28.68708762886598

In [26]:
```
## Normality


from scipy.stats import shapiro

# Perform the Shapiro-Wilk test for both desktop and mobile users

shapiro_male = shapiro(sample_male)
shapiro_female = shapiro(sample_female)

print("Shapiro-Wilk test for desktop users:", shapiro_male)
print("Shapiro-Wilk test for mobile users:", shapiro_female)
```

```
Shapiro-Wilk test for desktop users: ShapiroResult(statistic=0.9555267095565796, pvalue=0.33
25960636138916)
Shapiro-Wilk test for mobile users: ShapiroResult(statistic=0.955476701259613, pvalue=0.3317
6112174987793)
```

In [27]:
```
from scipy.stats import levene

# Perform Levene's test
levene_test = levene(sample_male, sample_female)

print(levene_test)
```

```
LeveneResult(statistic=2.2973682580197363, pvalue=0.13615173958979532)
```

In [28]:
```
import scipy.stats as stats

t_statistic, p_value = stats.ttest_ind(sample_male, sample_female)


# Calculate t-statistic and p-value using independent t-test

print("t-statistic:", t_statistic)
print("p-value:", p_value/2)
```

```
t-statistic: -0.9263760980404362
p-value: 0.17944254910462965
```

In [29]:
```
alpha = 0.05

if p_value < alpha:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
```

```
Fail to reject the null hypothesis.
```

Paired 2 sample t-test

```
A paired two-sample t-test, also known as a dependent or paired-samples t-test, is a
statistical test used to compare
the means of two related or dependent groups.
```

Common scenarios where a paired two-sample t-test is used include:

1. **Before-and-after studies**: Comparing the performance of a group before and after an intervention or treatment.
2. **Matched or correlated groups**: Comparing the performance of two groups that are matched or correlated in some way, such as siblings or pairs of individuals with similar characteristics.

Assumptions

- **Paired observations**: The two sets of observations must be related or paired in some way, such as before-and-after measurements on the same subjects or observations from matched or correlated groups.
- **Normality**The differences between the paired observations should be approximately normally distributed. This assumption can be checked using graphical methods (e.g., histograms, Q-Q plots) or statistical tests for normality (e.g., Shapiro-Wilk test). Note that the t-test is generally robust to moderate violations of this assumption when the sample size is large.

- **Independence of pairs**: Each pair of observations should be independent of other pairs. In other words, the outcome of one pair should not affect the outcome of another pair. This assumption is generally satisfied by appropriate study design and random sampling.

Example

Let's assume that a fitness center is evaluating the effectiveness of a new **8**-week weight loss program. They enroll **15** participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants' weight.

```
Before the program:

[80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91]

After the program:

[78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]
```

Significance level (α) = 0.05

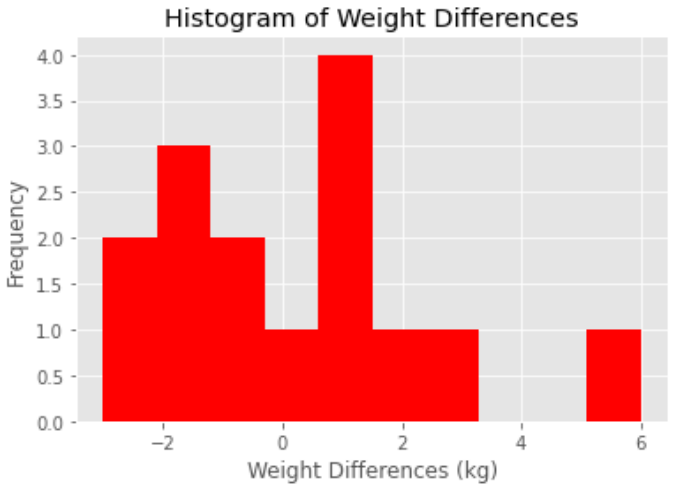
In [30]:
```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
plt.style.use('ggplot')

before = np.array([80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91])
after = np.array([78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88])
```

In [31]:
```
differences = after - before
```

In [32]:
```
plt.hist(differences , color ='red')
plt.title("Histogram of Weight Differences")
plt.xlabel("Weight Differences (kg)")
plt.ylabel("Frequency")
plt.show()

shapiro_test = stats.shapiro(differences)
print("Shapiro-Wilk test:", shapiro_test)
```




```
Shapiro-Wilk test: ShapiroResult(statistic=0.9220570921897888, pvalue=0.20704729855060577)
```

In [33]:
```
mean_diff = np.mean(differences)

std_diff = np.std(differences, ddof=1)
```

In [34]:
```
mean_diff
```

Out[34]:
```
0.06666666666666667
```

In [35]:
```
std_diff
```

Out[35]:
```
2.4630604269214893
```

In [ ]: