Hypothesis Testing in Machine Learning

- Hypothesis
- Null and Alternative Hypothesis
- One-Tailed and Two-Tailed Tests
- Z-Test
- T-Test
- Chi-Square Test
- ANOVA Test
- SciPy Library Functions for Tests

Smitesh Tamboli

NEXT ➡

Hypothesis Testing

A hypothesis is a claim or belief. Hypothesis testing **is a statistical process of either rejecting or retaining a claim or belief**, or association related to a business context, product, service etc. It plays an important role in providing evidence of an association relationship between an outcome variable and predictor variables.

Example:

- The new version of the eComm website has a better conversion rate
- The cash-On-Delivery payment method increases sales
- The average annual salary of machine learning experts differs for males and females.

Null and Alternative Hypothesis

The Null hypothesis is the default or baseline assumption that there is no effect, no difference, or no relationship between variables. It is denoted as ***H0.*** The Null hypothesis is the claim that is assumed to be true initially.

The alternative hypothesis is the complement of the Null hypothesis. The alternative hypothesis is the statement that indicates the presence of an effect, a difference, or a relationship between variables. It is denoted as ***HA***.

Example:

- The new version of the eComm website has a better conversion rate

Null Hypothesis (H0): The conversion rate of the new website version is equal to the conversion rate of the old website version.

Alternative Hypothesis (HA): The conversion rate of the new website version is better or higher than the conversion rate of the old website version.

- The cash-On-Delivery payment method increases sales

Null Hypothesis (H0): The sales amount with the cash-on-delivery payment method is equal to the sales amount with other payment methods.

Alternative Hypothesis (HA): The sales amount with the cash-on-delivery payment method is higher than the sales amount with other payment methods.

- The average annual salary of machine learning experts differs for males and females.

Null Hypothesis (H0): The average annual salary of male machine learning experts is equal to the average annual salary of female machine learning experts.

Alternative Hypothesis (HA): The average annual salary of male machine learning experts is different from the average annual salary of female machine learning experts.

Test Statistic, P-Value and Significance Level

Test Statistic

A test statistic is a standardized value that measures the distance between the observed sample statistic and the parameter specified in the null hypothesis. It is used to determine how far the sample data is from what we would expect if the null hypothesis were true. The test statistic is the standardized value used for calculating the p-value (probability value) in support of the null hypothesis.

P-Value

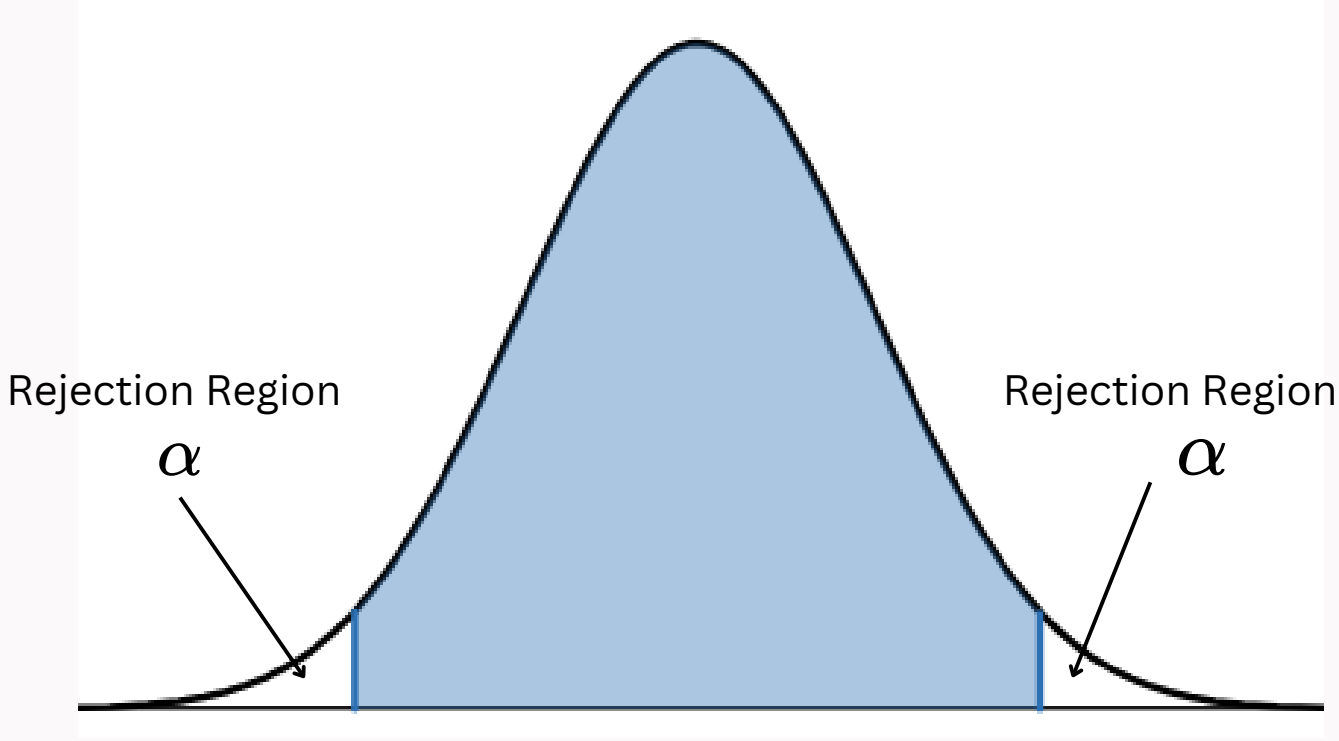
P-value is **a conditional probability of observing the statistic value given that the null hypothesis is true.** **The P-value is the evidence in support of the null hypothesis.**

P-value = P(Observing test statistics value | Null hypothesis is true)

Significance Value α

The primary task in hypothesis testing is to make a decision to either reject or fail to reject the Null hypothesis. **Significance level provides criteria used for making a decision regarding the null hypothesis reject or fail to reject (retain) based on calculated P-value.**

The significance value is the maximum threshold for the P-value. Usually, the value of significance level α = 0.05. The reason for choosing a very low value of 0.05 is that we start the process of hypothesis testing with an assumption that the Null hypothesis is true. Unless there is strong evidence against this assumption, we will not reject the Null hypothesis.



Criteria	Decision
P-value < α	Reject the Null hypothesis
P-value >= α	Retain or fail to reject the Null hypothesis

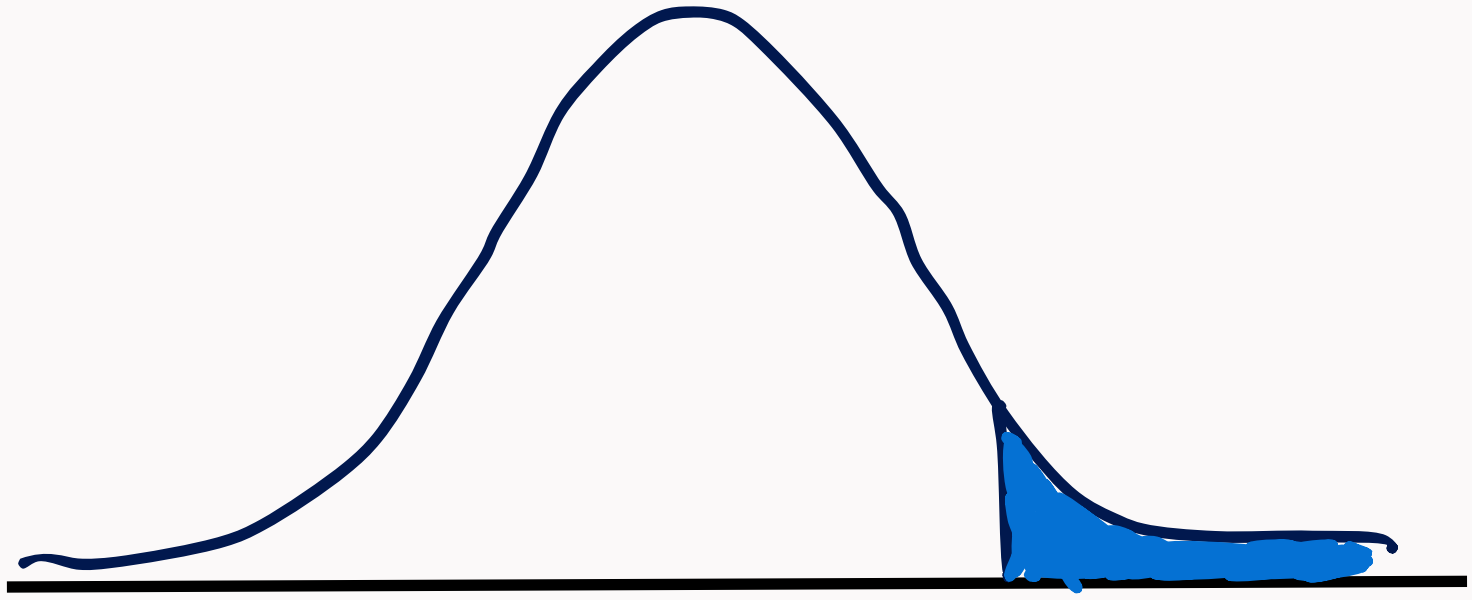
One-Tailed and Two-Tailed Test

One-Tailed Test

A one-tailed test is a statistical test where **the critical area of the distribution is on one side (either left or right) of the distribution** so that it is either greater than or less than a certain value, but not both.
The alternative hypothesis specifies that the parameter is either greater than or less than a certain value, but not both. The critical region for rejecting the null hypothesis is located entirely in one tail of the distribution either the left tail or the right rail.

Right-Tailed Test

In a right-tailed test, the critical region for rejecting the Null hypothesis is located in the right tail of the probability distribution. **This test is used when the alternative hypothesis HA specifies that the parameter of interest is greater than a certain value.**



Examples:
A pharmaceutical company claims that their new drug increases patient recovery rate more than the standard treatment.

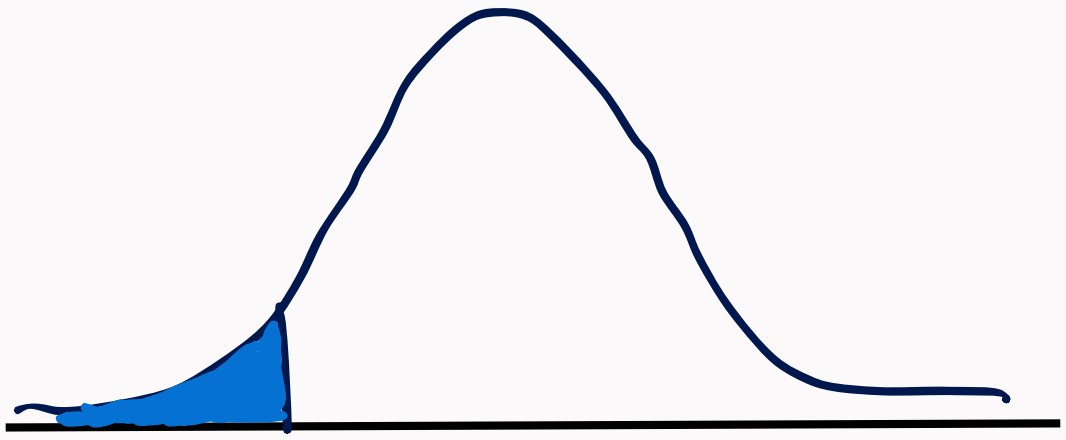
- **Null Hypothesis H0**: The new drug does not increase the recovery rate more than the standard treatment. H0: $\mu_{new} \leqslant \mu_{standard}$
- **Alternate Hypothesis HA**: The new drug increases the recovery rate more than the standard treatment. $H_A : \mu_{new} > \mu_{standard}$

Examples:
A new marketing campaign increases the average number of customers visiting their website daily compared to the old campaign.

- **Null Hypothesis H0**: The new marketing campaign does not increase the average number of daily website visitors compared to the old campaign. $H_0 : \mu_{new} \leq \mu_{old}$
- **Alternate Hypothesis HA**: The new marketing campaign increases the average number of daily website visitors compared to the old campaign. $H_A : \mu_{new} > \mu_{old}$

Left-Tailed Test

In a left-tailed test, the critical region for rejecting the Null hypothesis is located in the left tail of the probability distribution. **This test is used when the alternative hypothesis HA specifies that the parameter of interest is less than a certain value.**



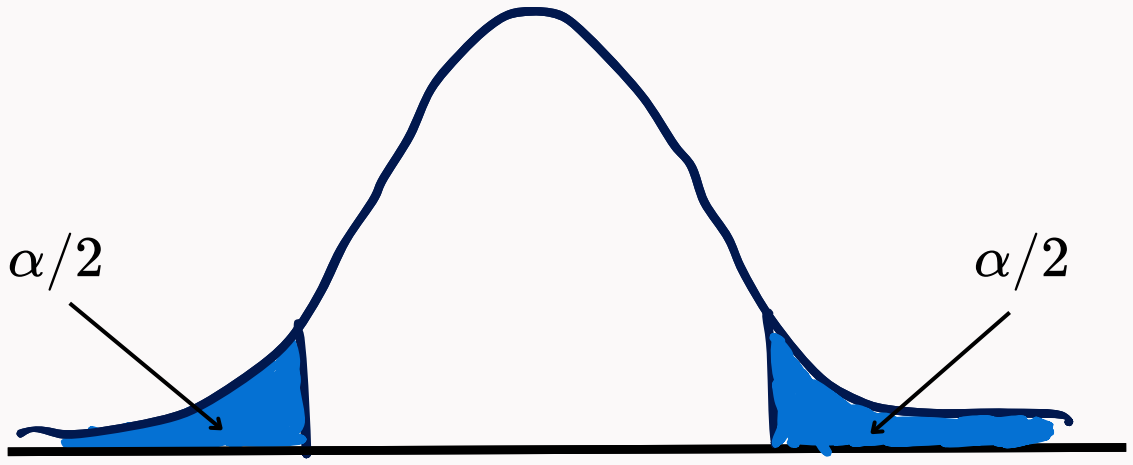
Examples:
An educational researcher wants to determine if a new teaching method decreases the average number of student failures compared to the traditional teaching method.

- **Null Hypothesis H0**: The new teaching method does not decrease the average number of student failures compared to the traditional method. $H_0 : \mu_{new} \geq \mu_{traditional}$
- **Alternate Hypothesis HA**: The new teaching method decreases the average number of student failures compared to the traditional method. $H_A : \mu_{new} < \mu_{traditional}$

Two-Tailed Test

A two-tailed test is a type of statistical test where **the critical area of the distribution falls in both tails (left and right) of the distribution**. This test is used when the alternative hypothesis does not specify the direction of the effect but states that there is a difference.

- If the test statistic falls into either of these regions, the Null hypothesis is rejected
- The significance level α is split equally the two tails, so each tails has an area of $\alpha/2$



Examples:

A researcher wants to test whether a new diet program has a different effect on weight loss compared to the standard diet program, without specifying if it is more or less effective.

- **Null Hypothesis H0**: The new diet program has the same effect on weight loss as the standard diet program. $H_0 : \mu_{new} = \mu_{standard}$
- **Alternate Hypothesis HA**: The new diet program has a different effect on weight loss compared to the standard diet program. $H_A : \mu_{new} \neq \mu_{standard}$

Hypothetic Condition	Tailed-Test
$H_0 : \mu_{new} \leq \mu_{old}$ $H_A : \mu_{new} > \mu_{old}$	Right-Tailed Test
$H_0 : \mu_{new} \geq \mu_{old}$ $H_A : \mu_{new} < \mu_{old}$	Left-Tailed Test
$H_0 : \mu_{new} = \mu_{old}$ $H_0 : \mu_{new} \neq \mu_{old}$	Two-Tailed Test

Steps To Perform Hypothesis Test

1

Describe the hypothesis in words

e.g. The cash-On-Delivery payment method increases sales

2

Define Null and Alternate Hypothesis

- **Null Hypothesis (H0)**: The sales amount with the cash-on-delivery payment method is equal to the sales amount with other payment methods.
- **Alternative Hypothesis (HA)**: The sales amount with the cash-on-delivery payment method is higher than the sales amount with other payment methods.

3

Identify the test statistic for validity of Null Hypotheis

z-test **t-test** **chi-square test** **ANOVA test**

4

Decide the criteria for rejection and retention of Null hypothesis

Define Significance level α

Generally 0.05 or 5%, but can vary based on criticality of problem

5

Calculate p-value

p-value is the evidence in support of the Null hypothesis.

6

Decide reject or retain the Null hypothesis

Based on p-value and significance level α decide either reject or fail to reject (retain) Null hypothesis.

p-value < α -> Reject Null hypothesis

p-value >= α -> Retain Null hypothesis

Smitesh Tamboli

NEXT

Type-I and Type-II Errors

In hypothesis testing we end up with two decisions

- Reject Null hypothesis
- Fail to reject (or retain) Null hypothesis

Type-I Error (False Positive)

The conditional probability of rejecting a Null hypothesis when it is true. A Type-I error occurs when the Null hypothesis H0 is True, but we incorrectly reject it.

$$\text{Type-I Error} = \alpha = P(\text{Rejecting Null hypothesis} \mid H0 \text{ is True})$$

Significance Level α is the probability of committing a Type-I error. The common choices for α is 0.05 meaning there is a 5% risk of rejecting the Null hypothesis when it is actually True.

Type-II Error (False Negative)

The conditional probability of failing to reject (or retain) a Null hypothesis when it is false (or Alternate Hypothesis is true. A Type-II error occurs when the Null hypothesis H0 is False, but we fail to reject it. It is the error of not detecting a significant effect or difference when there is one.

$$\text{Type-II Error} = \beta = P(\text{Retain Null hypothesis} \mid H0 \text{ is False})$$

	Decision on Null hypothesis based on hypothesis test	
Actual Value of H0	**Reject H0**	**Retain H0**
H0 is True	Type-I Error P(Reject H0 | H0 is True) α	Correct Decision P(Retain H0 | H0 is True)
H0 is False	Correct Decision P(Reject H0 | H0 is False)	Type-II Error P(Retain H0 | H0 is False) β

Z-Test

A z-test is a statistical test used to determine whether there is a significant difference between the means of two groups, or to test if a sample mean significantly differs from a known population mean. It is called a z-test because it follows a normal distribution (z-distribution) under the Null hypothesis.

When to use z-test?

- We need to test the value of the population mean, given that population variance is known.
- The population is a normal distribution and the population variance is known.
- The sample size is large and the population variance is known. i.e. sample size n > 30.

One-sample z-test: used when the sample mean is significantly different from a known population mean

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

$\bar{X}$ = Sample Mean
μ = Population Mean
σ = Population Standard Deviation
n = Sample size

Two-sample z-test: used when the mean of two independent samples are significantly different

$$Z = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

$\bar{x}_1, \bar{x}_2$ = Mean of sample 1 and sample 2
σ_1, σ_2 = s.d. of populations 1 and 2
n1, n2 = sample sizes of sample 1 and sample 2

Example

Suppose an e-commerce platform receives an average of 100 visitors per day (known population mean). We want to test if the average number of visitors for a recent sample of 40 days is significantly different from this known average. Verify the claim at significance level alpha = 0.05.

- Null Hypothesis (H0): The average number of visitors for the sample period is not significantly different from the population mean. I.e. Sample Mean = 100
- Alternative Hypothesis (HA): The average number of visitors for the sample period is significantly different from the population average. i.e. Sample Mean $\neq$ 100

One-Sample Z-Test

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

```python
# Function to perform one-sample z-test
# z = (sample_mean - pop_mean) / (pop_sd / sqrt(sample_size))
def z_test(data, pop_mean, pop_sd):
    sample_mean = data.mean()
    sample_size = len(data)
    z_score = (sample_mean - pop_mean) / (pop_sd / np.sqrt(sample_size))
    p_value = 2 * (1 - stats.norm.cdf(np.abs(z_score)))

    return z_score, p_value
```

```python
import numpy as np
import pandas as pd
from scipy import stats

# Known population parameters
pop_mean = 100  # average visitors (population mean)
pop_sd = 15     # population standard deviation

# Sample data (recent 40 days of visitor counts)
visitors = [117, 119, 132, 106, 106, 133, 121, 102, 118,
            103, 103, 113, 81, 84, 101, 94, 114, 96, 88,
            131, 106, 111, 88, 101, 111, 92, 115, 100,
            105, 100, 137, 109, 94, 122, 91, 113, 80, 90, 112]

# Create DataFrame
df = pd.DataFrame({"visitors": visitors})

# Perform z-test
z_score, p_value = z_test(df['visitors'], pop_mean, pop_sd)

print(f"Z-Score: {z_score}")
print(f"P-Value: {p_value}")

# Conclusion
alpha = 0.05  # significance level
if p_value < alpha:
    print("Reject the null hypothesis:
           The average number of visitors for the sample period
           is significantly different from the known average.")
else:
    print("Fail to reject the null hypothesis:
           The average number of visitors for the sample period
           is not significantly different from the known average.")
```

```
Output
==========================================
Z-Score: 2.55137525062772
P-Value: 0.010729872895687054
Reject the null hypothesis: The average number of visitors for the sample
period is significantly different from the known average.
```

Two-Sample Z-Test

Example

Suppose an e-commerce platform runs two different campaigns to drive traffic to their website. We want to test if there is a significant difference in the average number of daily visitors between the two campaigns.

- Null Hypothesis (H0): There is no significant difference in the average number of daily visitors between the two campaigns. I.e. $\mu_1 = \mu_2$
- Alternative Hypothesis (HA): There is a significant difference in the average number of daily visitors between the two campaigns. i.e. $\mu_1 \neq \mu_2$

$$Z = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

```
import numpy as np
import pandas as pd
from scipy import stats
```

```python
# Sample data for the two campaigns
campaign1_data = [157,147,159,172,146,146,173,161,142,158,
143,143,153,121,124,141,134,154,136,128,
171,146,151,128,141,151,132,155,140,145,
140,177,149,134,162,131,153,120,130,152]

campaign2_data = [178,167,181,195,166,166,196,183,162,179,
162,162,174,137,140,160,152,175,154,145,
194,166,171,145,160,171,150,176,159,165,
159,201,169,152,183,149,173,136,147,173]

# Convert data to DataFrames
df_camp1 = pd.DataFrame({"visitors":campaign1_data})
df_camp2 = pd.DataFrame({"visitors":campaign2_data})

# Campaign-1 parameters (mean, std, sample size)
camp1_mean = df_camp1['visitors'].mean()
camp1_std = df_camp1['visitors'].std()
n1 = len(df_camp1)

# Campaign-2 parameters (mean, std, sample size)
camp2_mean = df_camp2['visitors'].mean()
camp2_std = df_camp2['visitors'].std()
n2 = len(df_camp2)

# Z-test for two sample formaula
# z-score = (sample1_mean - sample2_mean) / sqrt( (sample1_std**2 / n1) + (sample2_std**2 / n2) )
z_score = (camp1_mean - camp2_mean) / np.sqrt(((camp1_std ** 2) / n1) + ((camp2_std ** 2) / n2))

# p-value
p_value = 2 * (1 - stats.norm.cdf(np.abs(z_score)))

print(f"Campaign 1 Mean: {camp1_mean}")
print(f"Campaign 2 Mean: {camp2_mean}")
print(f"Z-Score: {z_score}")
print(f"P-Value: {p_value}")

# Conclusion based on alpha = 0.05
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis:
          There is a significant difference in the
          average number of daily visitors between the two campaigns.")
else:
    print("Fail to reject the null hypothesis:
          There is no significant difference in the
          average number of daily visitors between the two campaigns.")
```

```
Campaign 1 Mean: 146.15
Campaign 2 Mean: 165.825
Z-Score: -5.771857099686556
P-Value: 7.840259330649246e-09
Reject the null hypothesis: There is a significant difference
in the average number of daily visitors between the two
campaigns.
```

T-Test

A T-test is a statistical test used to compare the means of two groups and determine if they are significantly different from each other.

When to use T-test?

- The T-test is used when the population follows a normal distribution and the population standard deviation σ is unknown and is estimated from the sample.
- The sample size is small (n < 30).

One-Sample T-Test

Used when comparing the mean of a single sample to a known population mean and variance unknown.

Example: An eComm platform believes that the average number of daily visitors is 150. Test if the average number of visitors for a sample of 20 days is significantly different from this value.

- **H0**: The sample mean is not significantly different from the population mean
- **HA**: The sample mean is significantly different from the population mean

```
# Number of visitors for 20 days
visitors = [157, 147, 159, 172, 146, 146, 173, 161, 142, 158,
            143, 143, 153, 121, 124, 141, 134, 154, 136, 128]

# Population mean
pop_mean = 150

# Perform One-Sample T-Test
t_stat, p_value = stats.ttest_1samp(visitors,popmean=pop_mean)

print(f"T-Statistic: {t_stat}")
print(f"P-Value: {p_value}")

# Conclusion
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")

OUTPUT
====================
T-Statistic: -0.9679294783122818
P-Value: 0.34524012885098876
Fail to reject the null hypothesis.
```

Two-Sample T-Test

Two-sample T-test is used to determine if there is a significant difference between the means of two independent groups.

Example: A dataset contains Male and Female students marks. We need to test the means marks of Male students are not equal to Female students

- **H0**: The means of Marks for Male and Female students are equal
- **HA**: The means of Marks for Male and Female students are not equal

```
import pandas as pd
from scipy import stats

data = {
  'Gender': ['Female', 'Male', 'Female', 'Female', 'Male', 'Male', 'Male',
    'Male', 'Female', 'Female', 'Male', 'Male', 'Female', 'Male', 'Female',
    'Female','Female','Male', 'Male','Female', 'Female', 'Male', 'Male',
    'Male', 'Male','Female', 'Female','Male', 'Male', 'Male', 'Female','Female',
    'Female', 'Female', 'Female', 'Male', 'Male', 'Female', 'Female','Male'],
 'Marks': [77, 89, 89, 91, 76, 85, 79, 78, 93,
    88, 91, 77, 81, 88, 86, 80, 82, 95, 87,  83,
    79, 94, 84, 73, 85, 85, 78, 88, 81, 82,  84,
    86, 75, 83, 87, 80, 90, 90, 92, 92]
 }

# Create DataFrame
df = pd.DataFrame(data)

# Separate the data into two groups
male_marks = df[df['Gender'] == 'Male']['Marks']
female_marks = df[df['Gender'] == 'Female']['Marks']

# Perform independent two-sample t-test
t_statistic, p_value = stats.ttest_ind(male_marks, female_marks)

print(f"T-Statistic: {t_statistic}")
print(f"P-Value: {p_value}")

# Conclusion
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis:
               There is a significant difference
               in the means of Marks for males and females.")
else:
    print("Fail to reject the null hypothesis:
                There is no significant difference in the
                means of Marks for males and females.")

OUTPUT
======================
T-Statistic: 0.13733874423305975
P-Value: 0.8914881726038391
Fail to reject the null hypothesis: There is no significant difference in the
means of Marks for males and females.
```

Chi-Square Tests

The chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables.

Example: Consider a dataset that contains preferred payment methods (Credit Card, Cash, PayPal) and the satisfaction level (Satisfied, Not Satisfied) of customers on an eComm Platform. We need to test whether there is any association between payment methods and satisfaction levels.

- **H0**: There is no association between the mode of payment and the satisfaction level of customers.
- **HA**: There is an association between the mode of payment and the satisfaction level of customers.

```python
import pandas as pd
from scipy.stats import chi2_contingency

# Create the DataFrame
data = {
    "PaymentMode": ["Credit Card", "PayPal", "Cash", "Credit Card", "Credit Card",
                    "PayPal", "Cash", "PayPal", "Cash", "Credit Card", "Cash",
                    "Credit Card", "PayPal", "PayPal", "Cash"],
    "SatisfactionLevel": ["Satisfied", "Satisfied", "Not Satisfied", "Satisfied",
                          "Satisfied", "Not Satisfied", "Not Satisfied", "Satisfied",
                          "Satisfied", "Not Satisfied", "Not Satisfied", "Not Satisfied",
                          "Not Satisfied", "Satisfied", "Satisfied"]
}

df = pd.DataFrame(data)

# Create the contingency table
contingency_table = pd.crosstab(df["PaymentMode"], df["SatisfactionLevel"])

# Perform the chi-square test of independence
chi2, p, dof, expected = chi2_contingency(contingency_table)

print("\nChi-Square Statistic:", chi2)
print("P-Value:", p)
print("Degrees of Freedom:", dof)

# Conclusion
alpha = 0.05
if p < alpha:
    print("\nReject the null hypothesis: There is an association between the mode of
payment and the satisfaction level of customers.")
else:
    print("\nFail to reject the null hypothesis: There is no association between the mode
of payment and the satisfaction level of customers.")

OUTPUT
==========================
Chi-Square Statistic: 0.5357142857142857
P-Value: 0.7650170614485746
Degrees of Freedom: 2
Fail to reject the null hypothesis: There is no association between the mode of payment and
the satisfaction level of customers.
```

ANOVA Tests

ANOVA or Analysis of Variance is a statistical test used to compare the means of three or more groups to determine if there are statistically significant differences between them. It assesses whether the means of several groups are equal or not by examining the variation between and within groups.

Example: An eComm Platform wants to analyze the effect of different shipping options (Standard, Express, Same-Day) on customers' purchase amounts.

- **H0**: There is no significant difference between the group means.
- **HA**: There are significant differences between the group means.

```
import pandas as pd
from scipy.stats import f_oneway

# Example dataset
data = {
    "Shipping Option": ["Standard", "Express", "Same-Day", "Standard", "Express",
"Same-Day", "Standard", "Express", "Same-Day"],
    "Purchase Amounts": [50, 70, 90, 55, 75, 85, 60, 80, 95]
}

df = pd.DataFrame(data)

standard_shipping = df[df["Shipping Option"] == "Standard"]["Purchase Amounts"]
express_shipping = df[df["Shipping Option"] == "Express"]["Purchase Amounts"]
same_day_shipping = df[df["Shipping Option"] == "Same-Day"]["Purchase Amounts"]

# Perform one-way ANOVA
f_statistic, p_value = f_oneway(standard_shipping,
                                express_shipping,
                                same_day_shipping)

print("F-Statistic:", f_statistic)
print("P-Value:", p_value)

# Conclusion
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There are significant differences between the
group means.")
else:
    print("Fail to reject the null hypothesis: There are no significant differences
between the group means.")

OUTPUT
===============
F-Statistic: 37.00000000000006
P-Value: 0.0004218749999999983
Reject the null hypothesis: There are significant differences between the group
means.
```

Summary

Z-Test

- We need to **test the value of the population mean, given that population variance is known**.
- The population is a normal distribution and the population variance is known.
- The sample size is large and the population variance is known. i.e. sample size n > 30.

One-sample z-test:

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

Two-sample z-test:

$$Z = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

T-Test

- The T-test is used when the population follows a normal distribution and the **population standard deviation σ is unknown** and is estimated from the sample.
- The sample size is small (n < 30).

One-sample t-test: Used when comparing the mean of a single sample to a known population mean and variance unknown.

```
t_stat, p_value = stats.ttest_1samp(visitors,popmean=pop_mean)
```

Two-sample t-test: Two-sample T-test is used to determine if there is a significant difference between the means of two independent groups.

```
t_statistic, p_value = stats.ttest_ind(male_marks, female_marks)
```

Chi-Square Test

The chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables.

```
chi2, p, dof, expected = chi2_contingency(contingency_table)
```

ANOVA Test

compare the means of three or more groups to determine if there are statistically significant differences between them.

```
f_statistic, p_value = f_oneway(standard_shipping, express_shipping,same_day_shipping)
```

Thank You

Smitesh Tamboli