Advanced Statistics

Agenda

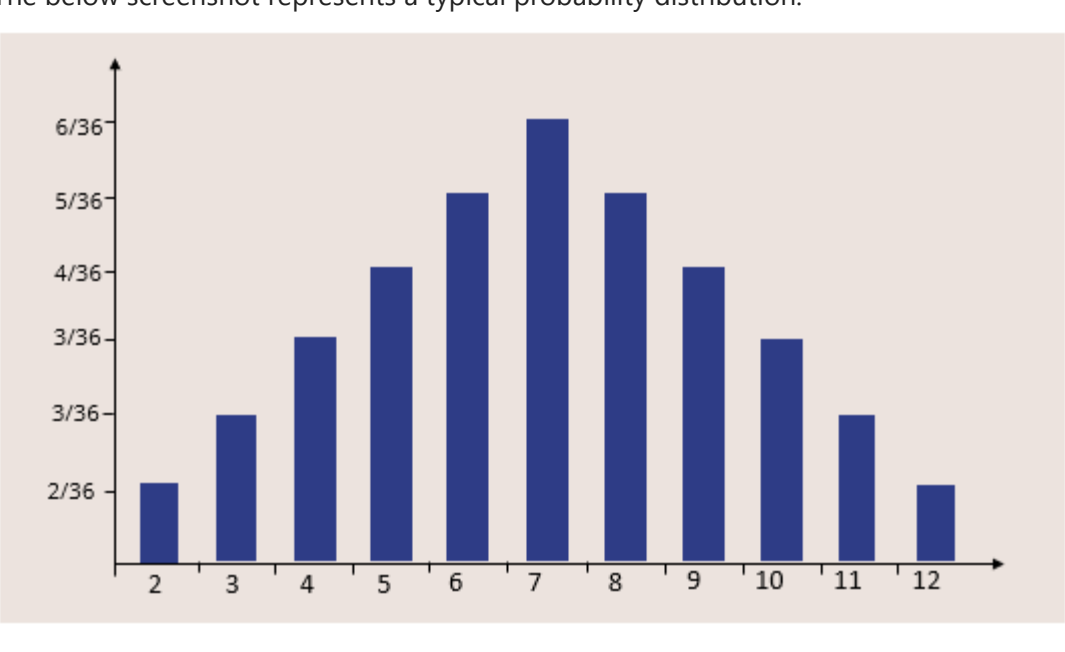
In this lesson, we will cover the following concepts with the help of a business use case:

- Distribution with its types
- Skewness and Kurtosis
- Student's T-Distribution
- Hypothesis Testing and Mechanism
- Null and Alternative Hypotheses
- Confidence Interval
- Margin of Error
- Confidence Levels
- Comparing and Contrasting T-Test and Z-Test
- Bayes' Theorem
- Chi-square Distribution
- Analysis of Variance or ANOVA
- Types of ANOVA
- Partition of Variance
- F-Distribution

What Is Distribution?

The probability distribution function is a statistical distribution function that describes all possible and likely values that a random variable can assume in a given range. The possible values of this distribution function are determined by factors like mean, standard deviation, skewness, and kurtosis.

The below screenshot represents a typical probability distribution:



Types of Probability Distribution

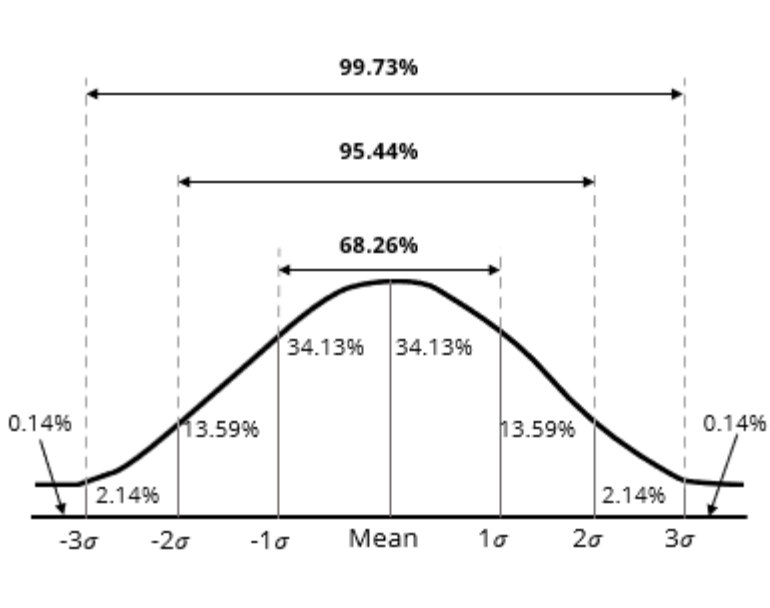
The probability distribution can be classified into different types:

- Normal Distribution
- Binomial Distribution
- Chi-square Distribution
- Poisson Distribution

Note: Normal Distribution is the most commonly used distribution. It is used in almost every field like finance, science, and engineering.

Normal Distribution

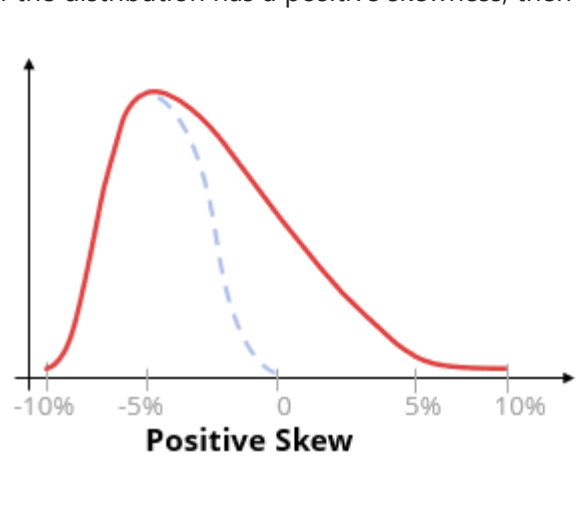
Normal Distribution, also known as Gaussian Distribution, is a probability distribution that is characterized by symmetry about the mean. That is, the frequency of the possibility of data occurring in and around the mean is higher than ends of the distribution. A normal distribution typically resembles a bell and is often referred to as a Bell Curve.



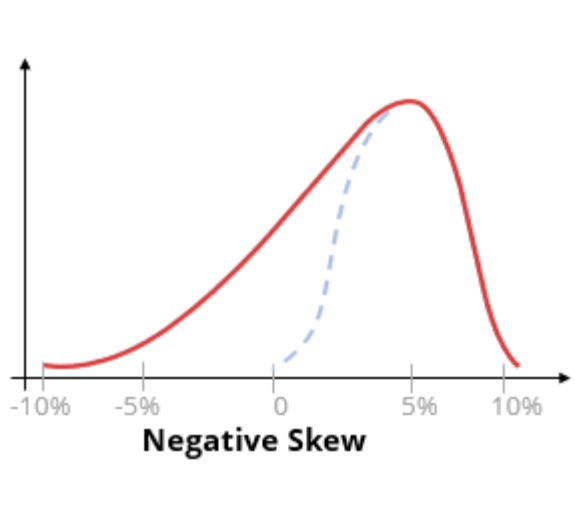
Skewness and Kurtosis

In real-life data is never distributed normally, even though such a distribution is called a normal distribution. Skewness and kurtosis are the two main parameters that affect the bell shape of the curve. If the bell curve shifts towards its left or towards its right, it is called 'skewed'. Skewness is represented by a numerical value that is either negative or positive.

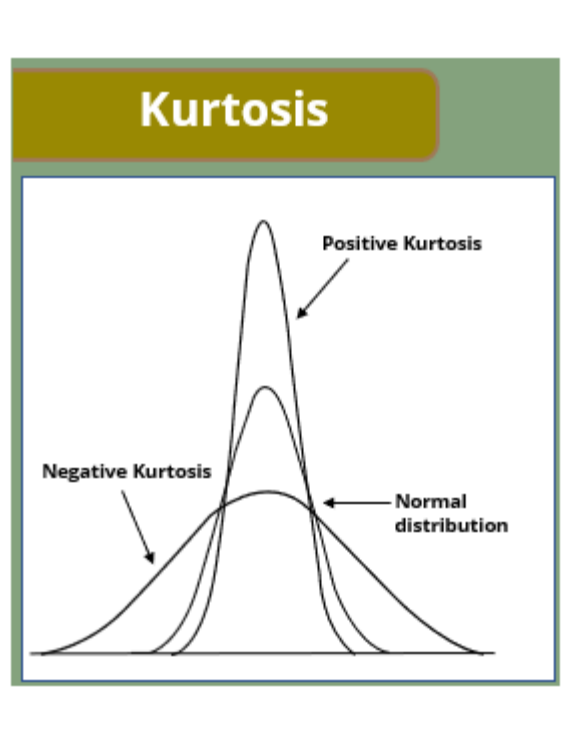
If the distribution has a positive skewness, then the curve will appear shifted towards the left of its bell peak, as shown below:



On the other hand, if the distribution has a negative skewness, then the curve will appear shifted towards the right of the bell peak, as shown below:

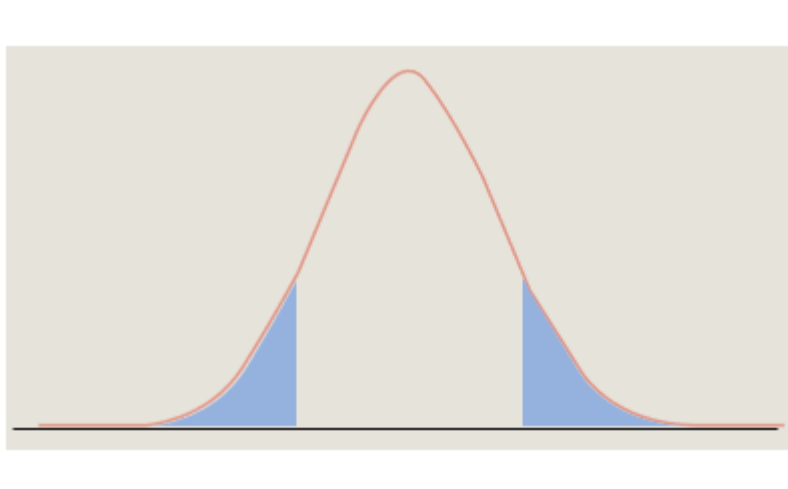


A normal distribution is superimposed with a positive Kurtosis and a distribution with a negative Kurtosis. A distribution with a positive Kurtosis will be narrower and taller around the mean with a flat tail, while a distribution with a negative Kurtosis will be flatter and shorter around the mean, but with a thicker distribution around the tails.



What Is Student's T-Distribution?

A Student's T-Distribution, also known as T-Distribution, is similar to a normal distribution like a bell curve but differs in that it has a thicker tail. Technically, Student's T-Distributions have a greater chance for extreme values than normal distributions, which is indicated by their flatter tails, as shown below:

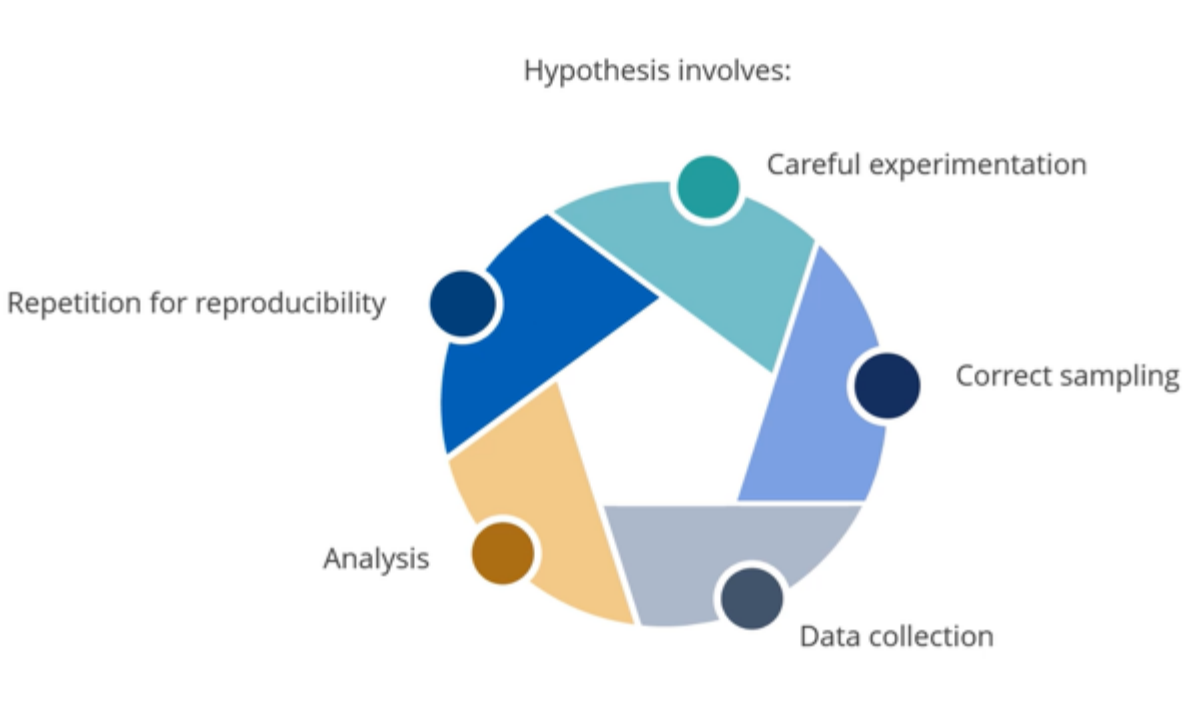


The tail in a normal distribution qualifies its classification as a Student's T-Distribution. Technically, a Student's T-Distribution is characterized by a parameter called degrees of freedom (or DOF), a mean of 0, and a standard deviation of 1.

Hypothesis Testing and Mechanism

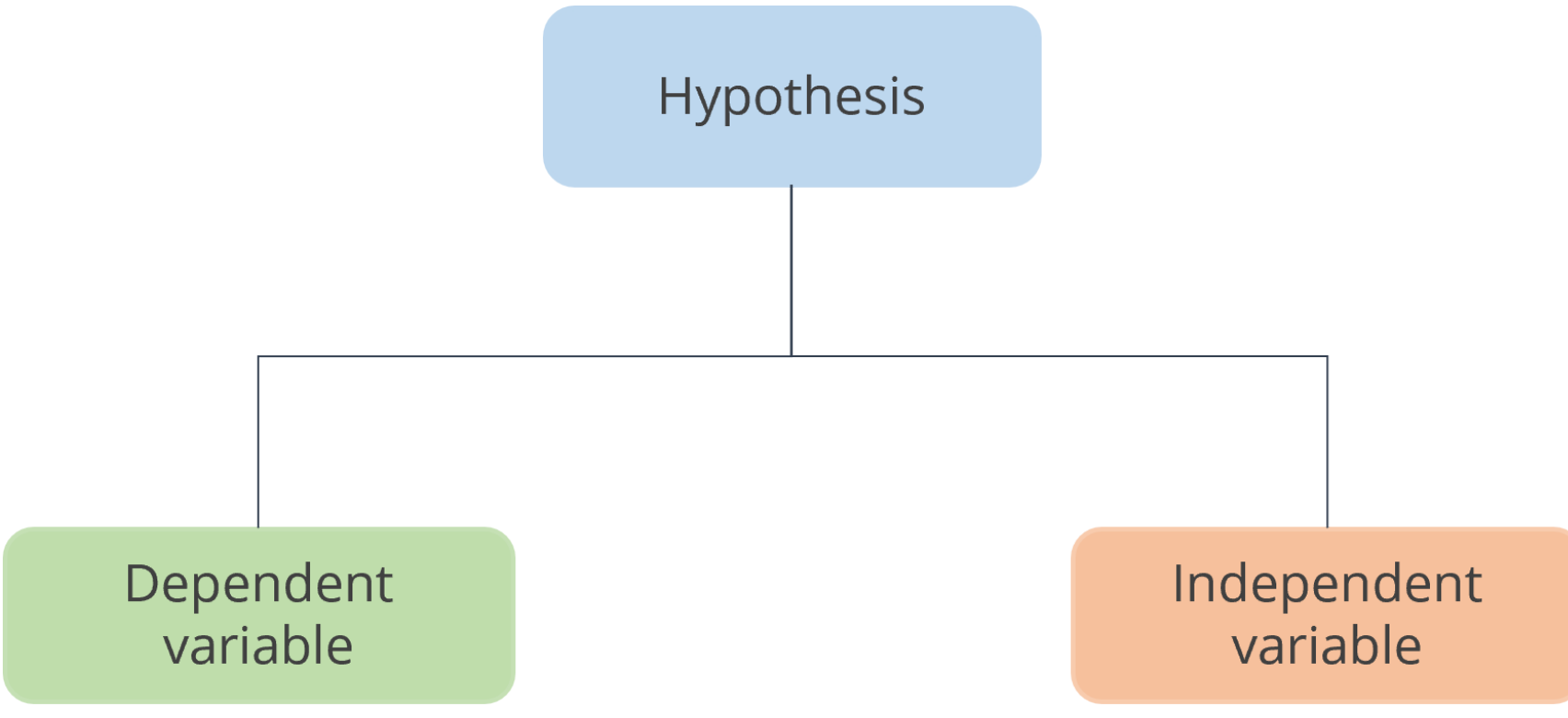
Introduction

A hypothesis is a statement that introduces both a question and proposes an answer. It involves careful experimentation, correct sampling, data collection, analysis, and repetition for reproducibility under various conditions.



Hypothesis Components

A hypothesis has two components:



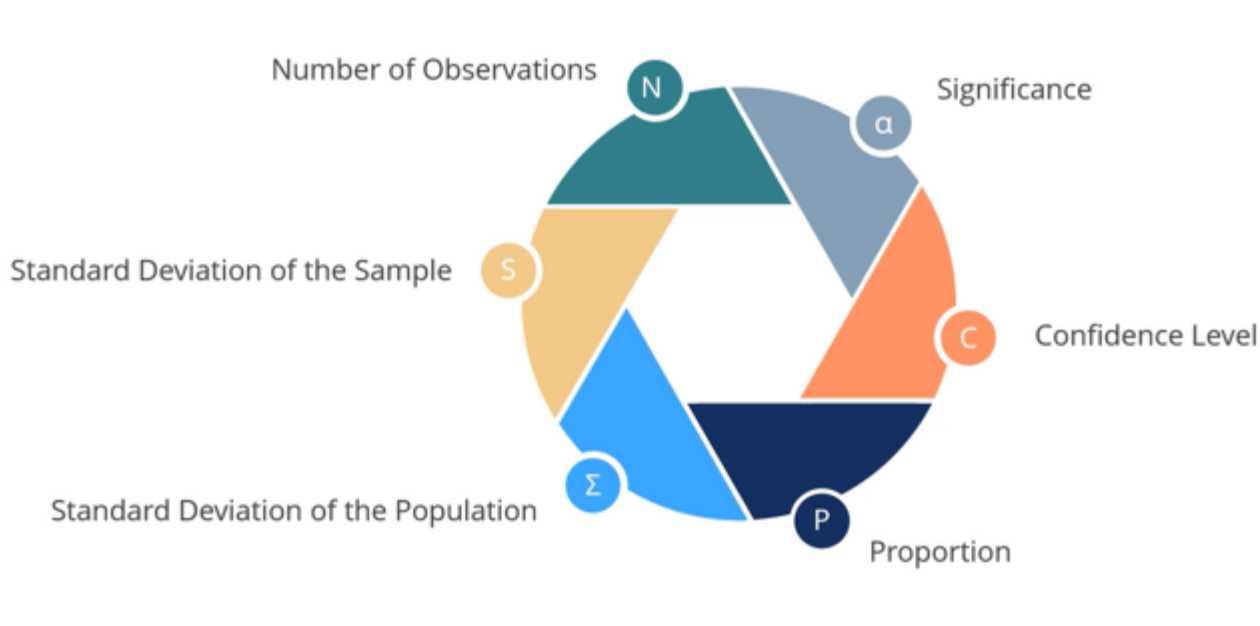
The independent variable affects or causes an action on the dependent variable.

Example:

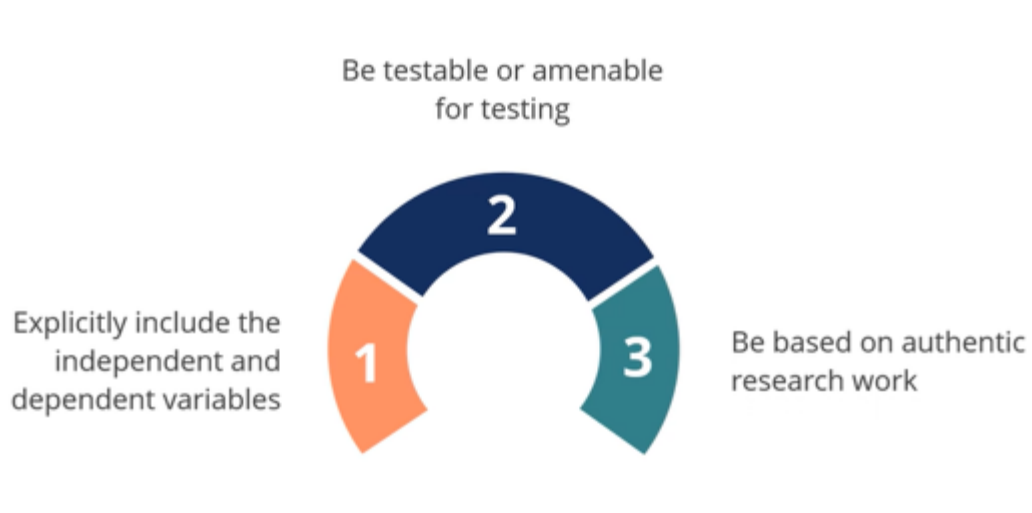
'If you don't clean tank water once every third day, it is possible that the fish won't survive for more than 3 months.'

The phrase 'don't clean tank water once every third day' is the independent variable and 'fish won't survive for more than 3 months' is the dependent variable.

Some of the components included in the proposition and expression of hypotheses are:



A good hypothesis must:



Hypothesis Testing

Hypothesis testing is a verification of the plausibility of a hypothesis using 'sample data'. A sample data may come from a larger population of data, or even from data-generating experimentation.

A statistical analyst tests a hypothesis by measuring a random set of sample data from the population being analyzed.

A hypothesis test always involves four steps. These are:

1. State the hypotheses:A statement of the correctness of the two hypotheses (null or alternative)

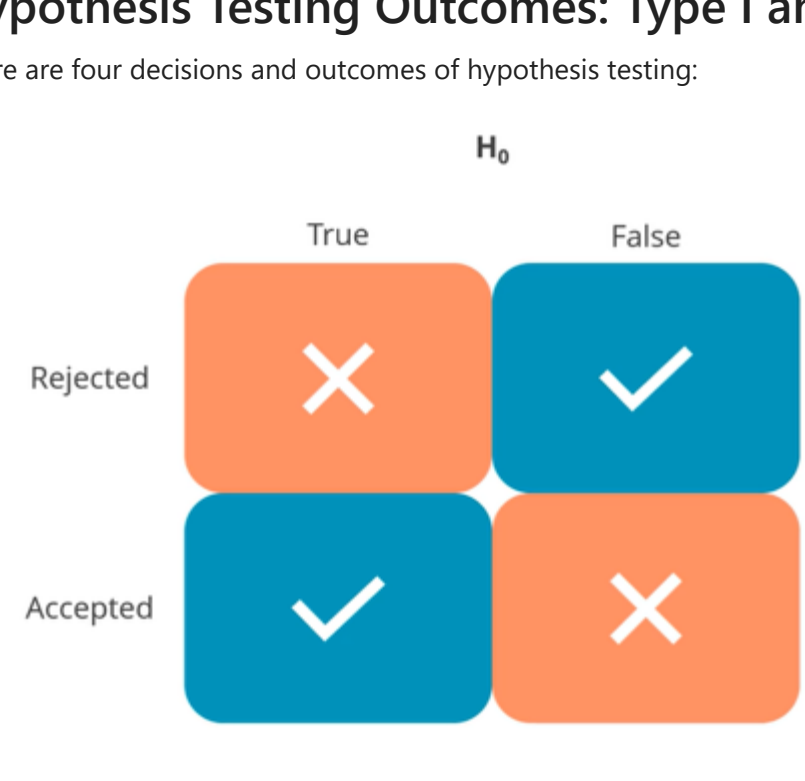
2. Set the criteria for a decision: A plan that outlines how to evaluate data

3. Compute sample statistics: Execution of the plan to physically carry out the analysis

4. Make a decision: Analysis of the result that rejects the null hypothesis or states that the null hypothesis is plausible

Hypothesis Testing Outcomes: Type I and Type II Errors

There are four decisions and outcomes of hypothesis testing:



1. H0 (Null Hypothesis) is TRUE and it is rejected
2. H0 (Null Hypothesis) is TRUE and it is accepted
3. H0 (Null Hypothesis) is FALSE and it is rejected
4. H0 (Null Hypothesis) is FALSE and it is accepted

The first and fourth outcomes are incorrect inferences and are referred to as type I errors.

The second and third are correct inferences and are referred to as type II errors.

The probability of the occurrence of type I errors is denoted by Alpha and the probability of type II errors is denoted by Beta.

The two values of Alpha and Beta cannot be zero simultaneously when inferences are based on samples. However, it can be achieved with complete enumeration. If one of them is set to zero, the other becomes one. Having said that, this does not imply that alpha and beta are equal to one. Hence, Alpha and Beta are assigned low values. In most situations, the value of Alpha is set at 0.05 or 0.01, and large sample sizes are used so that Beta also has a low value.

The selected value of Alpha is known as the level of significance.

For example, when Alpha is equal to 0.05, the level of significance is five percent.

That was a simple and effective way to estimate the outcomes of hypothesis testing.

Null and Alternative Hypotheses

There are two types of hypotheses:

1. Null hypothesis (H0)
2. Alternative hypothesis (H1)

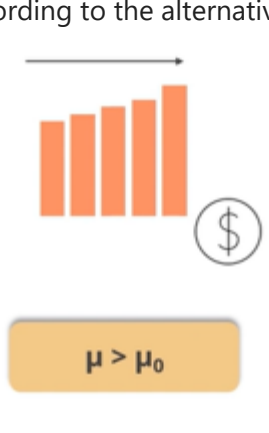
Example 1:

Sales Promotion Drive

According to the hypothesis, the sales promotion drive will increase the average monthly sales (μ) by 500 units.

According to the null hypothesis (H0), the sales promotion drive has an insignificant impact on the average monthly sales. The historical average (μ0) holds μ = μ0

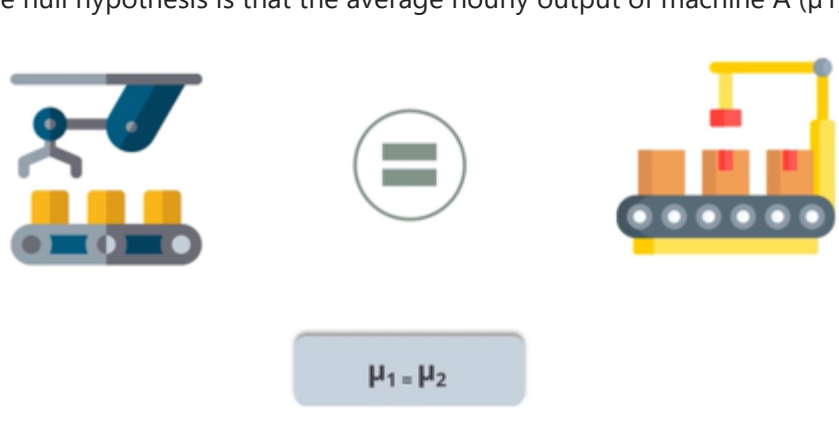
According to the alternative hypothesis (H1), the sales promotion drive has a significant impact on the average monthly sales. Thus, μ > μ0



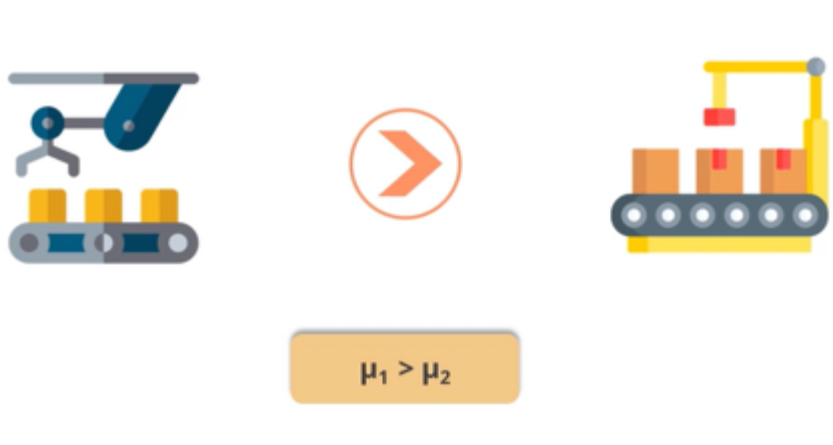
Example 2:

Hourly Output of Two Machines

The null hypothesis is that the average hourly output of machine A (μ1) differs insignificantly from machine B (μ2). So, μ1 = μ2



The alternative hypothesis is that the average hourly output of machine A (μ1) is significantly larger than that of machine B (μ2). So, μ1 > μ2



Generally, a null hypothesis is the negation of the assertion, while an alternative hypothesis is an assertion itself.

Steps Involved in Hypothesis Testing

1. Create a null hypothesis and the alternate hypothesis
2. Decide a level of significance i.e. alpha = 5% or 1%
3. Choose the type of test you want to perform as per the sample data (z test, t test, chi squared etc.)
4. Calculate the test statistics (z-score, t-score etc.) using the respective formula of test chosen
5. Obtain the critical value in the sampling distribution to construct the rejection region of size alpha using z-table, t-table, chi table etc.
6. Compare the test statistics with the critical value and locate the position of the calculated test statistics i.e. is it in rejection region or non-rejection region.
7. *a.* If the critical value lies in the rejection region, we will reject the hypothesis i.e. sample data provides sufficient evidence against the null hypothesis and there is significant difference between hypothesized value and observed value of the parameter.

   b. If the critical value lies in the non- rejection region, we will not reject the hypothesis i.e. sample data does not provide sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter is due to fluctuation of the sample.

Confidence Interval

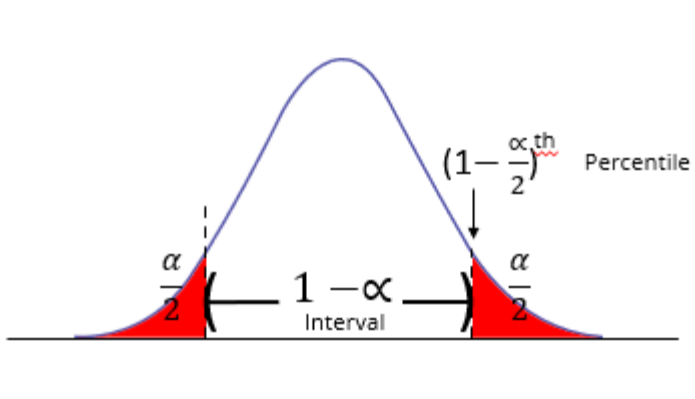
A Confidence Interval (CI) generally indicates the amount of uncertainty in any distribution. It is usually expressed as a number or a set or pair of numbers, and it can even be computed for a given distribution statistic.

CI is the probability that a particular population parameter will fall between a set of values for a certain period.

CIs can take any number of probability limits, the most common being 95%, and in some cases even 99%. When the behavior of a population is not known, then it is required to deduce the confidence intervals based on the sample data, using statistical methods like a T-Test.

A confidence interval is essentially a range of values that bind the statistic's mean value, which could in turn contain an unknown population parameter.

A typical confidence interval in a statistical distribution is as shown below:



As we can see, CIs are essentially numbers, consisting of an upper limit and a lower limit. On either side of the distribution, these are marked.

For example:

Survey a group of car owners to see how many gallons of gas they fill in the car in a year. Next, test the statistic at a 95% Confidence Level and get a Confidence Interval of (200, 450). This indicates that car owners buy anywhere between 200 and 450 gallons of gas a year.

Margin of Error

The Margin of Error (MoE) indicates by how many percentage points the results will differ from the real population value.

Consider the following statement:

A 95% confidence level with a 3% margin of error implies that the statistic distribution data is within 3% points of the real population value with 95% of the time.

The MoE is thus an important part of the confidence interval, without which one can't accept the inference from statistical analysis. The lower the margin of error, the better the acceptability of the population statistic. MoE is popularly used in pool and election surveys. A pool survey MoE must be scrutinized before accepting the Confidence Interval.

For Example:

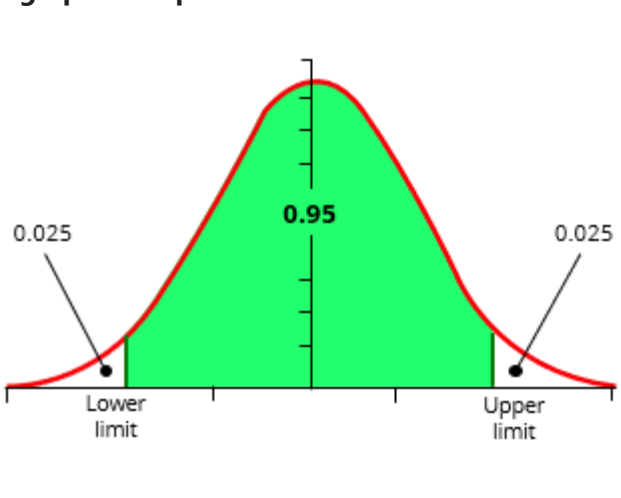
Consider the Gallup poll survey conducted in the 2012 US Presidential elections. The survey indicated 49% voting in favor of Mitt Romney, and 47% in favor of Barack Obama, with 95% CI and +/- 2% MoE. However, Barack Obama polled 51%, while Mitt Romney got 47% in the actual election. The results were even outside the range of the Gallup poll's MoE of +/-2%. This illustrates the need for statistics while taking CI, CL, and MoE into consideration.

Confidence Levels

A confidence level is the percentage of probability, or certainty, that the confidence interval would contain the true population parameter, when a random sample is drawn repeatedly.

In statistics, confidence levels are expressed as a percentage (for example, as 99%, 95% or 80% confidence level). However, for the purpose of supporting or disposing of the null hypothesis, scientists, and engineers usually work with a level of 95% or more. On the other hand, most governmental organizations and departments use 90% as the limit for Confidence Level.

A graphical representation of a Confidence Level of 95% is shown below:



A confidence level of 95% means when the experiment is repeated (or the poll survey conducted repeatedly), the survey results will match the results from a population 95% of the time. Confidence levels for different fields are different and are usually adopted in consultation with domain experts. This is done to ensure that the prediction from the statistic is reliable.

Comparing and Contrasting T-Test and Z-Test

T-Test

T-Test is a way of testing a hypothesis. They are chosen when:

- The population variance is unknown
- The sample size is comparatively small (n < 30)

T-Test is either a one-sample or two-sample test. In a one-sample T-Test, standard deviation of the sample is used instead of population standard deviation. Similarly, we perform a two-sample test for comparing the means of two samples.

For a one-sample T-Test, the formula is:

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

Where x̄ is the sample mean, s is the standard deviation of the sample, μ is the mean of the population and n is the sample size.

For a two-sample test, the formula is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

P-Value

P-Value is the smallest level of significance at which a null hypothesis can be rejected. Today, many tests gives a P-Value since it provides more information than the critical value and is therefore recommended.

For right-tailed test:

P-Value = P[Test statistics >= observed value of the test statistic]

For left-tailed test:

P-Value = P[Test statistics <= observed value of the test statistic]

For two-tailed test:

P-Value = 2 * P[Test statistics >= |observed value of the test statistic|]

Decision Making Using P-Value

The p-value is compared to the significance level(alpha) for decision making on null hypothesi

1. If P-Value is greater than alpha, we do not reject the null hypothesis.
1. If P-Value is smaller than alpha, we reject the null hypothesis.

One-Sample T-Test: Example

For a particular organization, the average age of the employees was claimed 30 years. The authorities collected a random sample of 10 employees' age data to check the claim made by the organization. Construct a hypothesis test to validate the hypothesis at a significance level of 0.05.

Ho = mean = 30

Ha = mean ≠ 30

In [9]:

```python
# Import the Ages.csv dataset
from scipy.stats import ttest_1samp
import pandas as pd
import numpy as np
ages = pd.read_csv("Ages.csv")
ages_mean = np.mean(ages)
print(ages_mean)
tset, pval = ttest_1samp(ages, 30)
print("p-values",pval)
if pval < 0.05:    # alpha value is 0.05 or 5%
   print(" we are rejecting null hypothesis")
else:
  print("we are accepting null hypothesis")
```

```
ages    43.75
dtype: float64
p-values [0.01012962]
 we are rejecting null hypothesis
```

Paired Sample T-Test: Example

For a particular hospital, it is advertised that a particular chemotherapy session does not affect the patient's health based on blood pressure.

It is to be checked if the blood pressure before the treatment is equivalent to the BP after the treatment.

Perform a statistical significance at alha=0.05 to help validate the claim.

H_0 = mean difference between two sample is 0

H_1 = mean difference between two sample is not 0

In [10]:

```python
import pandas as pd
from scipy import stats
df = pd.read_csv("blood_pressure.csv")
df.head()
```

Out[10]:

	patient	sex	agegrp	bp_before	bp_after
0	1	Male	30-45	143	153
1	2	Male	30-45	163	170
2	3	Male	30-45	153	168
3	4	Male	30-45	153	142
4	5	Male	30-45	146	141

In [11]:

```python
df[['bp_before','bp_after']].describe()
ttest,pval = stats.ttest_rel(df['bp_before'], df['bp_after'])
print(pval)
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

```
0.0011297914644840823
reject null hypothesis
```

Z-Test

Z-Test is also a way of testing a hypothesis. They are chosen when:

- The population variance is known
- The population variance is unknown, but the sample size is comparatively large (n ≥ 30)

Z-Test is either one-sample or two-sample test. One-sample Z–Test is chosen to compare a population mean with the sample mean. On the other hand, Two-sample Z-Test is chosen to compare the mean of two different samples.

The one-sample Z-score is computed using the formula:

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

Where x̄ is the sample mean, σ is the standard deviation of the population, μ is the mean of the population and n is the sample size.

Similarly, the two-sample Z-Test is computed using the following formula:

$$z\ score = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example:

A school principal claims that the students in his school are more intelligent than those of other schools. A random sample of 50 students' IQ scores has a mean score of 110. The mean population IQ is 100 with an Standard deviation of 15. State whether the claim of the principal is right or not at a 5% significance level.

Ho = the average population IQ score is 100

Ha = the average population IQ score is above 100

In [12]:

```python
# import the libraries
import math
import numpy as np
from numpy.random import randn
from statsmodels.stats.weightstats import ztest

# Generate a random array of 50 numbers having mean 110 and standard deviation of 15
# similar to the IQ scores data
mean_iq = 110
sd_iq = 15/math.sqrt(50)
alpha =0.05
null_mean =100
data = sd_iq*randn(50)+mean_iq
# Print mean and SD
print('mean=%.2f stdv=%.2f' % (np.mean(data), np.std(data)))

# Now we perform the test, and In this function, we passed data in the value parameter
# We passed mean value in the null hypothesis and will check if the mean is larger in the
# alternative hypothesis

ztest_Score, p_value= ztest(data,value = null_mean, alternative='larger')
# The function outputs a p_value and z-score corresponding to that value, we compare the
# p-value with alpha, and if it is greater than alpha, then we do accept the null hypothesis else we reject it.

if(p_value < alpha):
    print("Reject Null Hypothesis")
else:
    print("Fail to Reject NUll Hypothesis")
```

```
mean=110.71 stdv=2.29
Reject Null Hypothesis
```

Choosing between T-Test and Z-Test

Both the T-Test and Z-Test work with one-sample and two-sample scenarios. Also, both use the mean and standard deviation, however, tests are selected based on the sample or population into consideration.

If the sample size is large enough, then both T-Test and Z-Test will correlate with the same results. However, for a large sample size, sample variance provides a better estimate of population variance. Therefore, even if population variance is unknown, we can choose Z-Test using sample variance.

For a large sample size, we have higher degrees of freedom. Also, since T-distribution approaches the normal distribution, the difference between the T-score and Z-score is small.

T-Test and P-Value Using Python

T - Test

When is it used:

1. When we want to determine if means of 2 groups are different or not
2. When we want to compare means of two samples
3. Sample size <=30
4. Data follows normal distribution

Independent Sample T-Test:

1. Check the average of 2 independent unrelated groups.
2. Samples should be from 2 different populations.

- Null Hypothesis : $\mu_a = \mu_b$
- Alternate Hypothesis : $\mu_a \neq \mu_b$

Paired Sample T-Test:

1. Average of 2 samples taken from the same population but in different points in time.

- Null Hypothesis : μ_d = 0
- Alternate Hypothesis : $\mu_d \neq$ 0

One Sample T-Test:

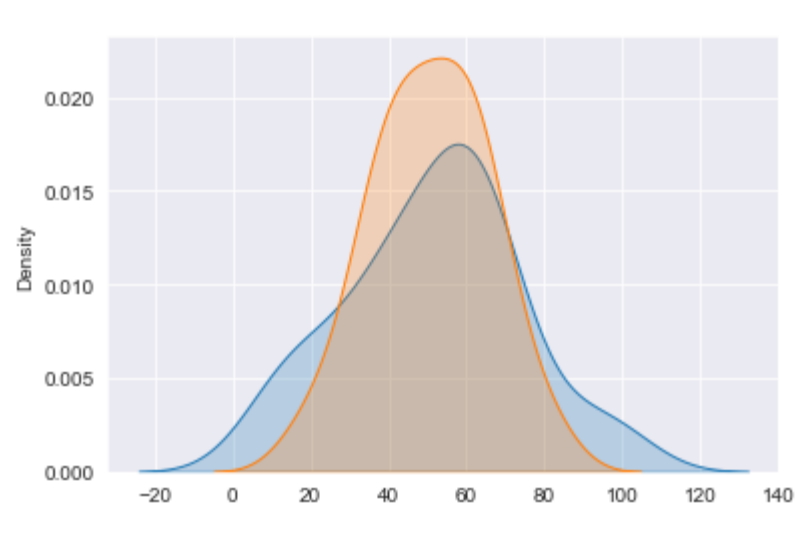
1. Average of a single group is different from known average

- Null Hypothesis : μ_a = X
- Alternate Hypothesis : $\mu_a \neq$ X

In [8]:
```
# Create sample data
import random
random.seed(100)
a = [ random.gauss(50,20) for x in range(30)]
b = [ random.gauss(55,15) for x in range(30)]
```

In [9]:
```
import seaborn as sns
sns.set_style('darkgrid')
sns.kdeplot(a,shade = True)
sns.kdeplot(b,shade = True)
```

Out[9]:
```
<AxesSubplot:ylabel='Density'>
```



Independent Sample T-Test

In [10]:
```
import scipy.stats as stats
```

In [11]:
```
t_stat,p_value = stats.ttest_ind(a,b,equal_var=False)
```

In [12]:
```
t_stat,p_value
```

Out[12]:
```
(0.10279425917534266, 0.9185308294235142)
```

p >= 0.05 and hence we accept the null hypothesis

In [13]:
```
import numpy as np
```

In [14]:
```
# Confirming the results of the test
np.mean(a), np.mean(b)
```

Out[14]:
```
(51.18263934085782, 50.67895224399505)
```

In [15]:
```
a = [ random.gauss(50,20) for x in range(30)]
b = [ random.gauss(60,25) for x in range(30)]
```

Paired Sample T-Test

In [16]:
```
t_stat,p_value = stats.ttest_rel(a,b)
```

In [17]:
```
t_stat,p_value
```

Out[17]:
```
(-2.7477434424821494, 0.010210498285614093)
```

p <=0.05 and hence we reject the null hypothesis

In [18]:
```
# Confirming the results of the test
np.mean(b) - np.mean(a)
```

Out[18]:
```
14.301311923126583
```

One Sample T-Test

In [19]:
```
# Checking the mean to give the population mean value
np.mean(a)
```

Out[19]:
```
46.074961678768354
```

In [20]:
```
t_stat,p_value = stats.ttest_1samp(a,100,axis = 0)
```

In [21]:
```
t_stat,p_value
```

Out[21]:
```
(-15.726161201628097, 9.860930052233548e-16)
```

p <=0.05 and hence we reject the null hypothesis

Z-Test and P-Value Using Python

In [22]:
```
# Import ztest library
import random
random.seed(20)
from statsmodels.stats.weightstats import ztest as ztest
```

Known mean and standard deviation

$\mu = 100$
$\sigma = 15$

One Sample Z-test

- Null Hypothesis : $\mu_s = value$
- Alternate Hypothesis : $\mu_s \neq value$

In [23]:
```
# Create sample with same mean
a = [random.gauss(100,15) for x in range(40)]
```

In [24]:
```
ztest(a,value = 100)
```

Out[24]:
```
(1.3430675401429018, 0.17925010504664385)
```

p value <=0.05 and hence we reject the null hypothesis

In [25]:
```
# Sampe with same mean
a = [random.gauss(100,15) for x in range(40)]
b = [random.gauss(100,15) for x in range(40)]
```

In [26]:
```
ztest(a,b,value = 0)
```

Out[26]:
```
(-0.9878523542697073, 0.32322496274152)
```

Use Case: Determine the Diameter of Two Units

Problem Statement:

An administrator wants to determine whether there is any significant difference in the diameter of the cutlet between two units. A randomly selected sample of cutlets was collected from both units and measured? Analyze the data and draw inferences at 5% significance level. Please state the assumptions and tests that you've carried out to check validity of the assumptions.

H_o = mean < 0.05

H_a = mean >= 0.05

Dataset:

Unit defines the diameter of cutlet.

The dataset consists of two columns, i.e. Unit A and Unit B, with 35 values.

Solution:

In [28]:
```
# Import the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy as sp
```

In [29]:
```
df = pd.read_csv('Cutlets.csv')
df.head()
```

Out[29]:

	Unit A	Unit B
0	6.8090	6.7703
1	6.4376	7.5093
2	6.9157	6.7300
3	7.3012	6.7878
4	7.4488	7.1522

In [30]:
```
df.describe()
```

Out[30]:

	Unit A	Unit B
count	35.000000	35.000000
mean	7.019091	6.964297
std	0.288408	0.343401
min	6.437600	6.038000
25%	6.831500	6.753600
50%	6.943800	6.939900
75%	7.280550	7.195000
max	7.516900	7.545900

In [31]:
```
df_unit_A = df['Unit A'].mean()
df_unit_A
```

Out[31]:
```
7.01909142857143
```

In [32]:
```
df_unit_B = df['Unit B'].mean()
df_unit_B
```

Out[32]:
```
6.964297142857142
```

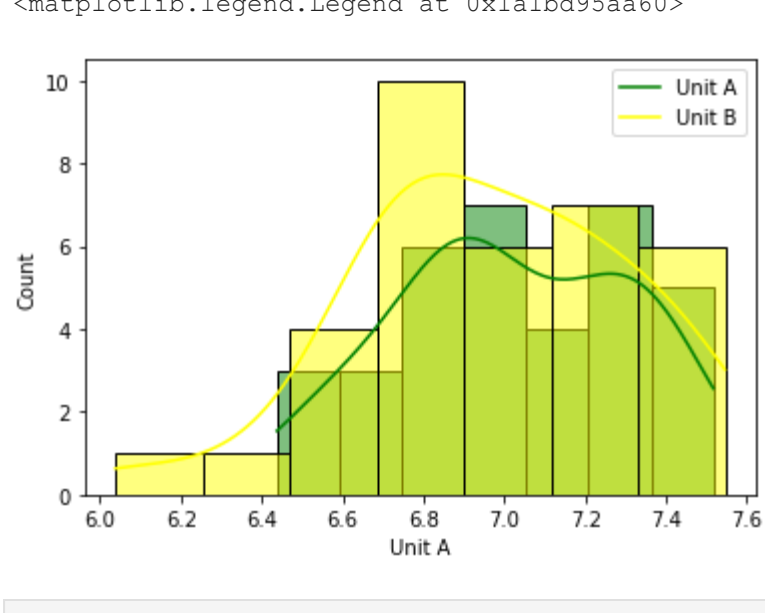
In [33]:
```
df_unit_A > df_unit_B
```

Out[33]:
```
True
```

Plotting the data

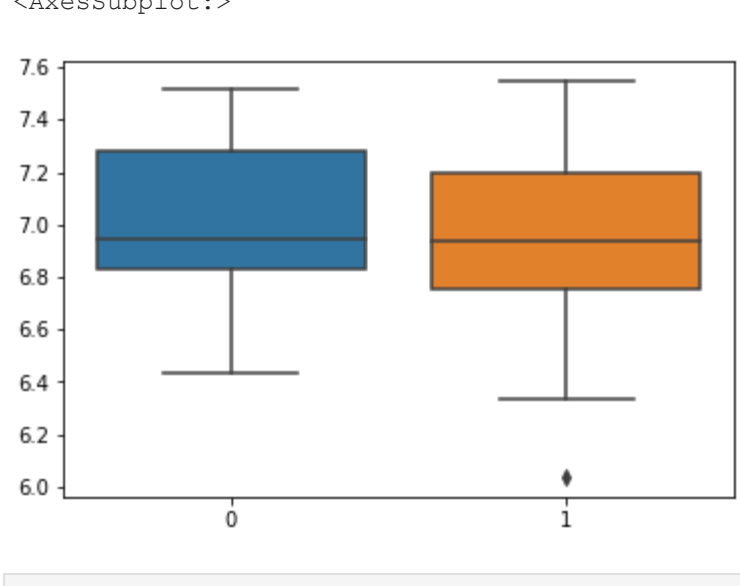
In [34]:
```
sns.histplot(data=df['Unit A'],color = 'green', kde=True)
sns.histplot(data=df['Unit B'],color = 'yellow', kde=True)
plt.legend(['Unit A','Unit B'])
```

Out[34]:
```
<matplotlib.legend.Legend at 0x1a1bd95aa60>
```



In [35]:
```
sns.boxplot(data=[df['Unit A'],df['Unit B']])
```

Out[35]:
```
<AxesSubplot:>
```



In [36]:
```
# There is one outlier present in Unit B of cutlet data
```

In [37]:
```
# Calculating t-value and p-value using scipy
tStat,pValue =sp.stats.ttest_ind(df['Unit A'],df['Unit B'])
```

In [38]:
```
tStat
```

Out[38]:
```
0.7228688704678063
```

In [39]:
```
pValue
```

Out[39]:
```
0.4722394724599501
```

In [40]:
```
if pValue < 0.05:
    print('We reject Null Hypothesis')
else:
    print('We accept Null Hypothesis')
```

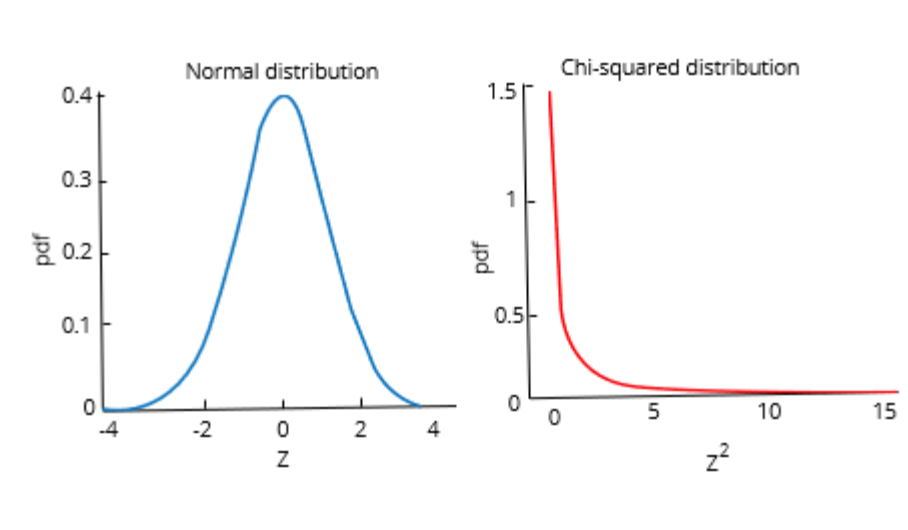
```
We accept Null Hypothesis
```

The inference is that there is no significant difference in the diameters of Unit A and Unit B.

Chi-square Distribution

A Chi-square distribution, pronounced 'khai squared', is a continuous probability distribution widely used in statistical inference. The Greek letter χ is often used, and χ2 is termed as chi-square.

The χ2 distribution and the standard normal distribution are related. If a random variable Z has a standard normal distribution, then Z2 has the χ2 distribution with one degree of freedom.



Multiple standard random variables are possible. Mathematically, for k variables,

Z12 + Z22 has 2 degrees of freedom.

Z12 + Z22 + Z32 + Z42 + Z52 + Z62 has 6 degrees of freedom, etc.

For k degrees of freedom, we have:

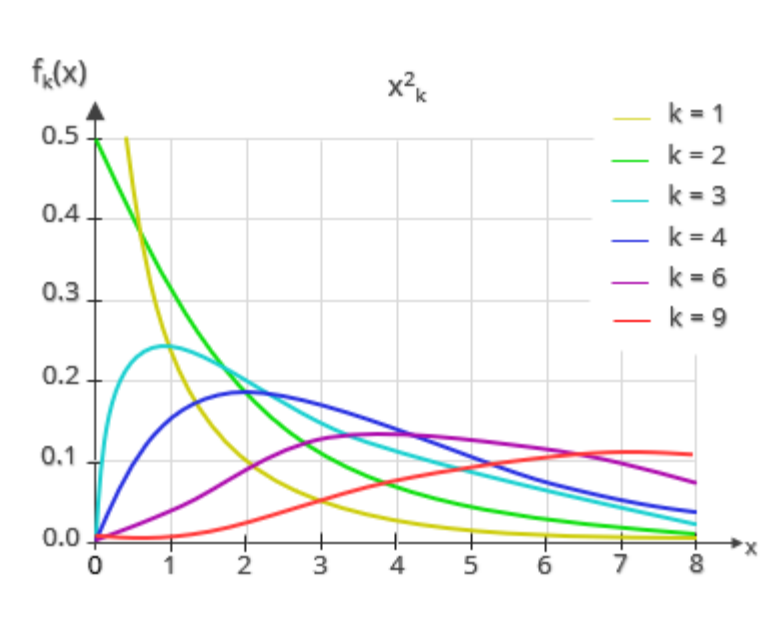
Z12 + Z22 + Z32 + Z42 + Z52 + Z62 + ... + Zk2, and this has a χ2 distribution of k degrees of freedom.

The equation for the probability density function (PDF) of the χ2 distribution with k degrees of freedom is:

$$f(x;k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, \\ 0, \end{cases}$$

The function is valid for all positive values of x, and k is the number of degrees of freedom. As the PDF has the gamma function Γ, the χ2 distribution of k degrees of freedom is also a gamma function.

The χ2 distribution of k degrees of freedom is:



With increasing degrees of freedom, the shape of the χ2 distribution varies. For k = 1, the PDF is infinity, when χ2 = 0. For k = 2, PDF is 0.5 for χ2 = 0. For higher values of k (3 or more), the χ2 distribution changes to a positively skewed standard normal distribution, and with higher degrees of freedom, the skewness and the kurtosis of the χ2 distribution changes, with the distribution becoming increasingly symmetric.

Note: In any χ2 distribution, the mean (μ) is k, the number of degrees of freedom, and the variance is 2k.

For example, for k = 3 in the diagram, μ = 3, while the variance is 2 x k, or 6.

The mode of the distribution will occur at k-2 for the distributions with k = 3 and above. So, when k = 4, the mode is at k – 2, which is 2.

Chi-Squared Distribution Using Python

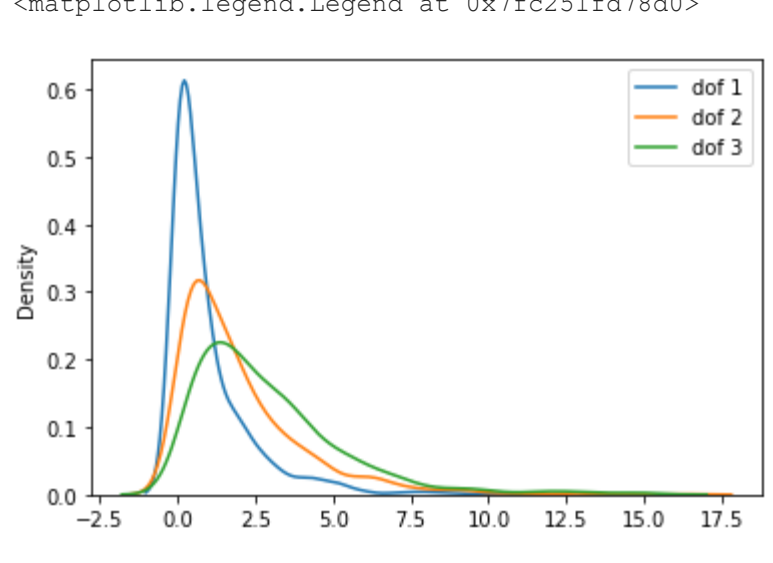
In [3]:
```
# Create Chi-Square Distribution of varying degrees of freedom
data1 = np.random.chisquare(df = 1,size = 1000)
data2 = np.random.chisquare(df = 2,size = 1000)
data3 = np.random.chisquare(df = 3,size = 1000)
print(data1[:10])
```

```
[5.25527182e-02 4.84313828e-01 4.23308042e+00 1.57520781e+00
 5.86342405e-08 9.41443423e-01 3.14066795e-03 2.59859313e+00
 6.05183396e-01 2.52410170e+00]
```

In [4]:
```
import matplotlib.pyplot as plt
import seaborn as sns
# Plot the distributions
sns.distplot(data1, hist = False, label = 'dof 1')
sns.distplot(data2, hist = False, label = 'dof 2')
sns.distplot(data3, hist = False, label = 'dof 3')
plt.legend()
```

```
/usr/local/lib/python3.7/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `kdeplot` (an axes-level function for kernel density plots).
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `kdeplot` (an axes-level function for kernel density plots).
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `kdeplot` (an axes-level function for kernel density plots).
  warnings.warn(msg, FutureWarning)
```

Out[4]:
```
<matplotlib.legend.Legend at 0x7fc251fd78d0>
```



In [27]:
```
from scipy.stats import chi2_contingency
```

In [28]:
```
data = [[10,20,30],[6,9,17]]
```

In [29]:
```
stat, p_value, dof, chi_array = chi2_contingency(data)
```

In [30]:
```
p_value
```

Out[30]:
```
0.873028283380073
```

p-value > 0.05, reject null hypothesis as there is relationship between the two groups.

In [31]:
```
data = [[10,20,30],[9,1,8]]
```

In [32]:
```
stat, p_value, dof, chi_array = chi2_contingency(data)
```

In [33]:
```
p_value
```

Out[33]:
```
0.005595003173072842
```

Chi-square Test and Goodness-of-Fit

Goodness-of-Fit

Goodness-of-Fit is a statistical test that examines how closely the sample data fits a population with a normal distribution.

There are multiple ways to determine goodness-of-fit, like the Shapiro-Wilk test, Kolmogorov-Smirnov test, Chi-square test, etc.

Chi-square Test The Chi-square test is the most common and popular goodness-of-fit test. It determines whether categorical data are related. It's a non-parametric test and is also called Pearson's Chi-square test.

In order to ascertain the goodness-of-fit test, it's important to establish an alpha value, such as the p-value, for the Chi-square test. The p-value refers to the probability of getting results close to the extremes of the observed distribution. This assumes that the null hypothesis is correct.

Steps for Chi-Square Test

1. Define the Null and Alternate hypotheses based on the data. H0 implies that the data met the expected distribution, while H1 implies that it did not
2. State the alpha value. As mentioned earlier, we usually work with a value of 0.05
3. Calculate the degrees of freedom, k. It depends on the number of categories or groups, and is usually K – 1, where K is the number of frequencies
4. State the decision rule. Calculate the 'decision value' based on the alpha value and degrees of freedom. Based on this value, either reject H0 or reject H1
5. Calculate the test statistic for χ2 using the formula:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

Here O is the observed value and E is the expected value from the sample. K is the degree of freedom.

1. Finally, compare the decision value computed in step 4 with that of step 5 to either accept or reject the null hypothesis
2. Based on step 6, conclude the domain observation

Example:

In a study about the election survey, voters might be classified by gender (male or female) and voting preference (democrat, republican, or independent).

Using alpha = 0.05, perform a chi-square test for independence to determine whether gender is related to voting preference.

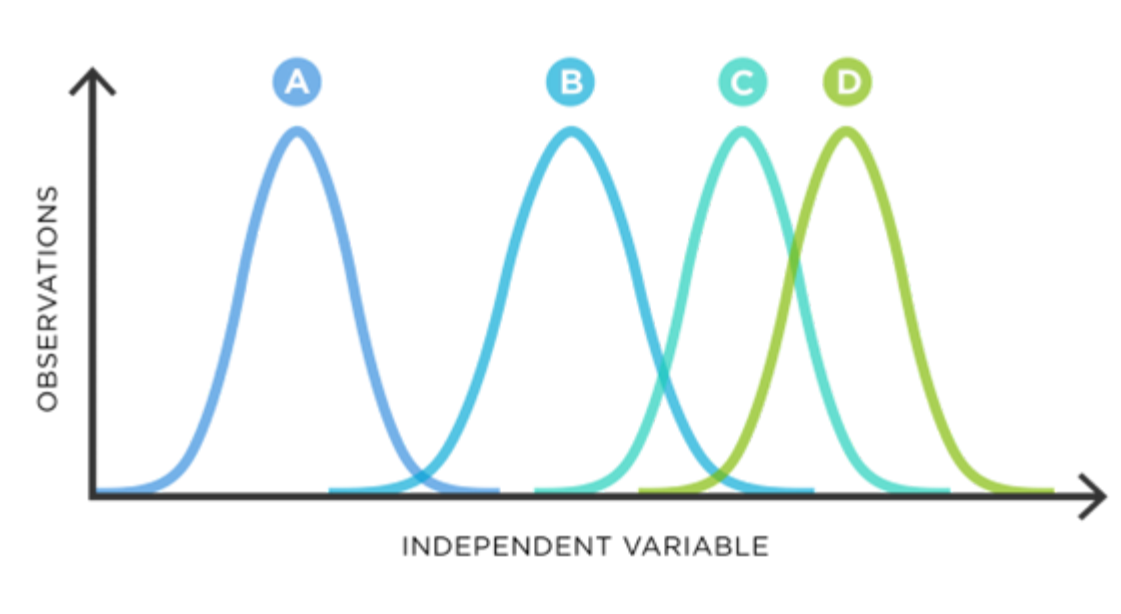
In [13]:
```
from scipy.stats import chi2_contingency
import pandas as pd
# Load the chi-test.csv file
df_chi = pd.read_csv('chi-test.csv')
contingency_table=pd.crosstab(df_chi["Gender"],df_chi["Shopping"])
print('contingency_table :-\n',contingency_table)
# Observed Values
Observed_Values = contingency_table.values
print("Observed Values :-\n",Observed_Values)
b=stats.chi2_contingency(contingency_table)
Expected_Values = b[3]
print("Expected Values :-\n",Expected_Values)
no_of_rows=len(contingency_table.iloc[0:2,0])
no_of_columns=len(contingency_table.iloc[0,0:2])
ddof=(no_of_rows-1)*(no_of_columns-1)
print("Degree of Freedom:-",ddof)
alpha = 0.05
from scipy.stats import chi2
chi_square=sum([(o-e)**2./e for o,e in zip(Observed_Values,Expected_Values)])
chi_square_statistic=chi_square[0]+chi_square[1]
print("chi-square statistic:-",chi_square_statistic)
critical_value=chi2.ppf(q=1-alpha,df=ddof)
print('critical_value:',critical_value)
# P-value
p_value=1-chi2.cdf(x=chi_square_statistic,df=ddof)
print('p-value:',p_value)
print('Significance level: ',alpha)
print('Degree of Freedom: ',ddof)
print('chi-square statistic:',chi_square_statistic)
print('critical_value:',critical_value)
print('p-value:',p_value)
if chi_square_statistic>=critical_value:
    print("Reject H0,There is a relationship between 2 categorical variables")
else:
    print("Retain H0,There is no relationship between 2 categorical variables")

if p_value<=alpha:
    print("Reject H0,There is a relationship between 2 categorical variables")
else:
    print("Retain H0,There is no relationship between 2 categorical variables")
```

```
contingency_table :-
 Shopping  No  Yes
Gender
Female     2    3
Male       2    2
Observed Values :-
 [[2 3]
 [2 2]]
Expected Values :-
 [[2.22222222 2.77777778]
 [1.77777778 2.22222222]]
Degree of Freedom:- 1
chi-square statistic:- 0.09000000000000008
critical_value: 3.841458820694124
p-value: 0.7641771556220945
Significance level:  0.05
Degree of Freedom:  1
chi-square statistic: 0.09000000000000008
critical_value: 3.841458820694124
p-value: 0.7641771556220945
Retain H0,There is no relationship between 2 categorical variables
Retain H0,There is no relationship between 2 categorical variables
```

Analysis of Variance or ANOVA

ANOVA, or Analysis of Variance, is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. It was developed by the statistician, Ronald Fisher in 1918. Please note that the term 'variance' is used to indicate the 'variation' or the 'dispersion' and must not be confused with the common statistical term 'variance'.

The ANOVA test extends the T and the Z test, as they are constrained to allowing the nominal level variable to have only two categories. ANOVA is also called the Fisher analysis of variance and is suitable for performing simultaneous tests on sets of data drawn from different populations.



As you can see on the screen, there are four groups: A, B, C, and D, and their sample distributions are compared.

How Does ANOVA Works?

Statistical tests such as T-Test, Z-Test, etc. Most of them are good for univariate situations. On the other hand, the ANOVA test compares the means of different groups and reveals any statistical differences between them.

ANOVA is used in the analysis of complex multivariate situations, and most of the bio-chemical, pharmaceutical, and science researchers depend heavily on ANOVA for studying the effect on the dependent variable by multiple independent variables.

To understand how ANOVA works, let's consider a medical experiment example:

To study the effectiveness of different hypertension medicines, scientists usually plan and conduct experiments to understand the relationship between the type of medicine and the resulting blood pressure. The sample population is, obviously, a set of people or patients. The scientists generally divide the sample population into multiple groups, and each group is administered a particular medicine for a trial period.

In the end, hypertension levels are measured for each of the individual patients. For each group, the mean hypertension level is calculated. ANOVA then helps in comparing these group 'means' to find out if they are statistically varying or similar.

ANOVA Terminologies

ANOVA uses many concepts and compares and produces the outcomes.

- **Sample and Grand Means:** A sample mean is the average value for a specific group. The grand mean, on the other hand, is the average of the sample means from various samples.
- **Dependent Variable:** It is the item or subject under investigation that is supposed to be influenced by many other independent variables.
- **Independent Variables:** There are many of them. Each of them could have an impact on or influence the Dependent Variable.
- **Factor:** Independent variables are frequently referred to as factors.
- **Levels:** It denotes the different values of factors that are used in any typical experiment. For example, the different medicines cited in the previous section.
- **F-Statistic:** It is also known as F-Ratio. This is the outcome of ANOVA.
- **Fixed-factor Model:** In this model, experiments use only a discrete set of levels of factors.
- **Random-factor Model:** In this model, a random value of level is drawn from multiple possible values of the factor.

ANOVA Outcome

ANOVA's outcome is known as the 'F statistic'. This is a ratio that shows the difference between the variation in the inter-group and the intra-group. With the help of this ratio, one can conclude whether the Null Hypothesis is true and either accept it or reject it.

Assumptions and Types of ANOVA

Purpose and Procedure of ANOVA ANOVA is an omnibus test statistic. The null hypothesis for ANOVA is that there is no significant difference in the 'mean' values among the groups. The alternate hypothesis, on the other hand, concludes that there are at least some significant differences.

ANOVA makes the following assumptions about the probability distribution of the responses:

1. Independence of observations
2. Normalcy
3. Equality of variances

The first one is obvious and needs no further description. The normalcy indicates that the distribution of the 'residuals' is normal. The equality of variances essentially means that the variance of data in the groups should be the same.

Note: Assumptions also depend on the types of ANOVA used

Types of ANOVA

There are two types of ANOVA:

1. One-way ANOVA

1. Two-way ANOVA

These are also referred to as one-factor ANOVA and two-factor ANOVA respectively.

One-Way ANOVA

This is also known as simple ANOVA or just ANOVA. This test is suitable for experiments with only one factor (independent variable) with two or more levels. A one-way ANOVA assumes the following:

- Independence
- Normalcy
- Variance
- Continuous data

The one-way ANOVA needs to operate on continuous data.

Example:

Three different categories of plants can be differentiated on the basis of their weights. A dataset with various plants and their weights are given. Construct a hypothetical test to determine the category of a plant at a significance level of 0.05.

In [35]:
```
import pandas as pd
from scipy import stats
from statsmodels.stats import weightstats as stests
df = pd.read_csv('plant.csv')
```

In [38]:
```
df = df[['weight','group']]
groups = pd.unique(df.group.values)
data = {grp:df['weight'][df.group == grp] for grp in groups}

F, p = stats.f_oneway(data['ctrl'], data['trt1'], data['trt2'])
print("p-value for significance is: ", p)
if p<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

```
p-value for significance is:  0.0159099583256229
reject null hypothesis
```

Two-Way ANOVA

The two-way ANOVA is used when there are two or more independent variables. It is also known as full factorial ANOVA or two-factor ANOVA. This test assumes importance when every possible permutation of factors and their levels are used. It is based on the following assumptions:

- Independence
- Normalcy
- Variance
- Continuous data
- Categories

For two-way ANOVA, the independent variables should be in separate groups or categories.

N-Way ANOVA

With increasing independent variables, ANOVA moves into a more complex category of problems which are tackled by N-Way ANOVA or MANOVA. The term MANOVA stands for multivariate analysis of variance.

For example:

Analysis of voter preferences based on gender, age, ethnicity, etc., can be studied using MANOVA

Example:

The yield of crops depends on factors like fertilizers and water usage. A researcher wants to find conditions for a higher crop yield by each factor as well as with the factors grouped together.

In [39]:
```
# Import crop_yield.csv dataset
import statsmodels.api as sm
from statsmodels.formula.api import ols
df_anova2 = pd.read_csv("crop_yield.csv")
model = ols('Yield ~ C(Fert)*C(Water)', df_anova2).fit()
print(f"Overall model F({model.df_model: .0f},{model.df_resid: .0f}) = {model.fvalue: .3f},
      p = {model.f_pvalue: .4f}")
res = sm.stats.anova_lm(model, typ= 2)
res
```

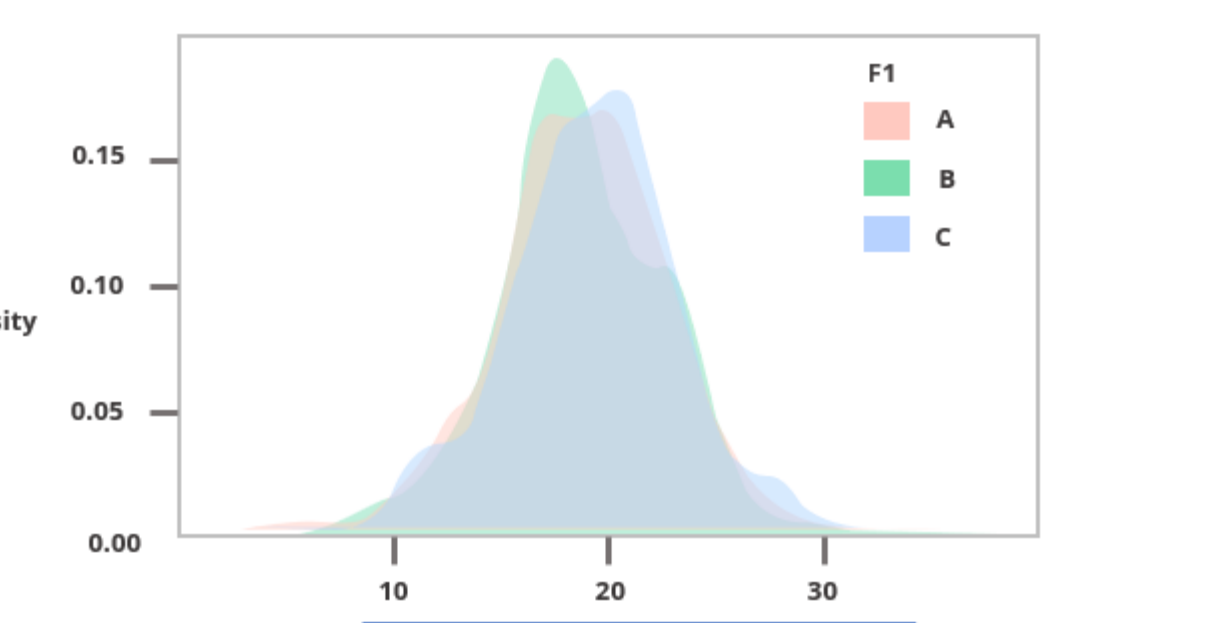
```
Overall model F( 3, 16) =  4.112, p =  0.0243
```

Out[39]:

	sum_sq	df	F	PR(>F)
C(Fert)	69.192	1.0	5.766000	0.028847
C(Water)	63.368	1.0	5.280667	0.035386
C(Fert):C(Water)	15.488	1.0	1.290667	0.272656
Residual	192.000	16.0	NaN	NaN

Partition of Variance

Partition of variance is a statistical analysis of the distributions of two or more samples in a population.



The figure shows the distribution of three different groups, A, B, and C from a typical population. The PDFs are marked in different colors for clarity. They all have mean = 20 and a standard deviation = 3.

The spread in A, B, and C is between 10 and 30, and is denoted by var1.

Now, compare this with the other figure. Group A, B, and C have means of 15, 20 & 25, respectively. However, the standard deviation for these three distributions is 3

There are similar distribution groups marked A, B, and C, but the variance range, var2, is between 5 and 35.

For var1, the mean is almost the same for A, B, and C, around 20. In var2, the means for A, B, and C are approximately 15, 20, and 25, respectively. This means that as the difference between the means of the groups varies the dispersion range also varies.

This concept is cemented by the variance values of the groups in each case. Variance for var1 is around 8.69, and for var2 it is 20.68.

Two recognizable components of variances are inter-group (within a group) and intra-group (between two or more groups). The former is also called 'error' or 'residual variance'.

Partitioning

The sum of squares is used to indirectly partition variance.

The equation for Variance is shown below:

$$\sigma^2 = \Sigma(x - \bar{x})^2/(n-1)$$

Here the group mean, called GM, is considered.

The equation for Total Sum of Squares or SS_{total} is shown below:

$$SS_{total} = \Sigma(x - \bar{x})^2$$

Similarly, the Error Sum of Squares:

$$For A, SS_{error} = \Sigma(x_A - \bar{x}_A)^2$$

$$For B, SS_{error} = \Sigma(x_B - \bar{x}_B)^2$$

$$For C, SS_{error} = \Sigma(x_C - \bar{x}_C)^2$$

The SS_{error} = $\Sigma(x_A - \bar{x}_A)^2$ + $\Sigma(x_B - \bar{x}_B)^2$ + $\Sigma(x_C - \bar{x}_C)^2$

The SS_{total} or the sum of squares between groups and the error is the difference $SS_{total} - SS_{error}$

F-Distribution

The F-distribution is similar to and related to χ2 distribution. the F-distribution or F-ratio, also known as Snedecor's F distribution or the Fisher–Snedecor distribution.

F-distribution is essentially a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA) and other F-Test.

Unlike the chi-square distribution, F-distribution deals with multiple random variables.

Consider a situation of two independent random variables, R1 and R2 that have an χ2 distribution with Degrees of Freedom (DOF) d1 and d2 respectively. The F-distribution or the F-ratio for this situation is expressed as below:

F = (R1/d1)/(R2/d2)

The Probability Density Function of F-Distribution

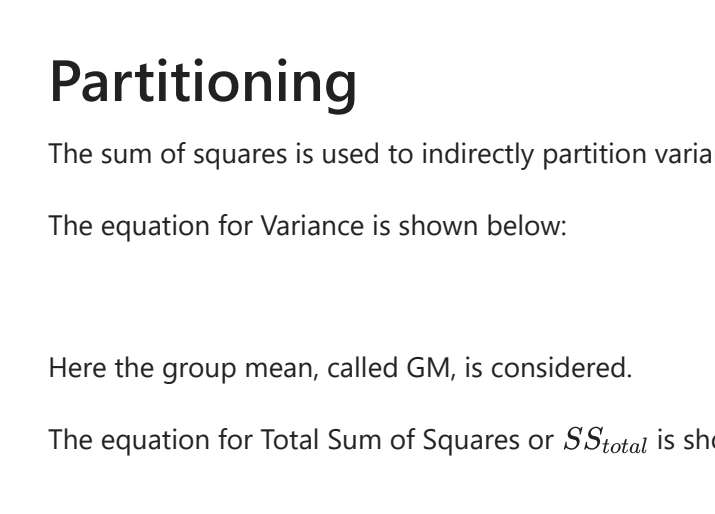
The probability density function for F-distribution can be computed using the below formula:

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x \, B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

The formula is valid for all positive values of x and the Degree of Freedom, d1 and d2. The 'B' in the formula represents the 'beta function'.

Graph of F-distribution's PDF

A typical PDF of F-distribution is shown below:



The figure suggests that the shape of the F-distribution curve depends on the two Degrees of Freedom, d1 and d2.

Description of F-distribution:

F-distribution is always described by stating the number of DOFs associated with the standard deviation in the numerator of the f statistic first.

Thus, f(6, 8) refers to an F-distribution with $d_1 = 6$ and $d_2 = 8$ Degrees of Freedom. Likewise, f(8, 6) also refers to an F-distribution with $d_1 = 8$ and $d_2 = 6$ degrees of freedom.

Note: The curves represented by f(6, 8) and f(8, 6) are different from each other.

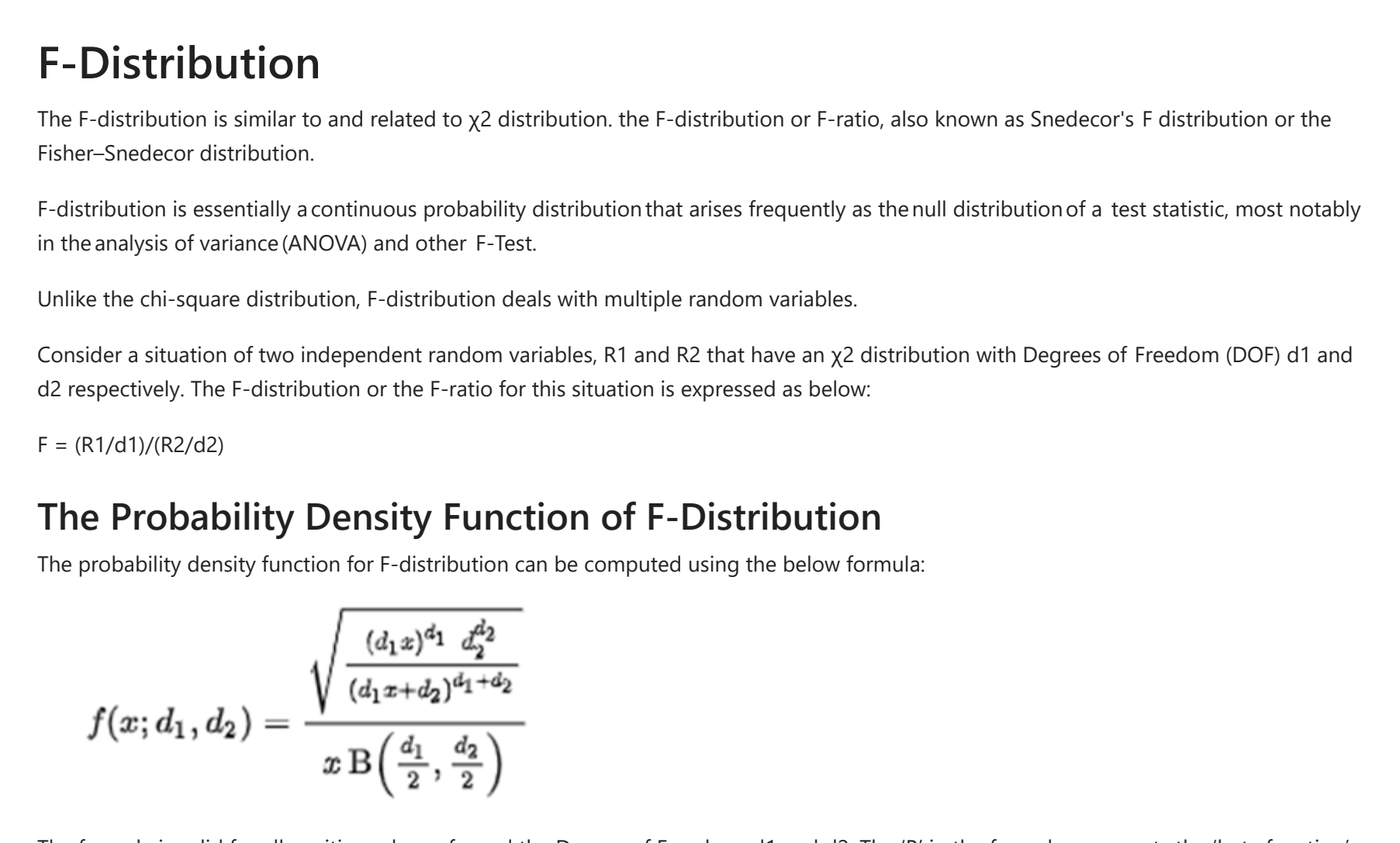
F-Distribution Using Python

A random variable

$F \sim \frac{\chi_1^2/v1}{\chi_2^2/v2}$

In [40]:
```
# Code for F-distribution
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import f
x = np.linspace(0, 4.5, 1000)
data_1 = f(1, 1, 0)
data_2 = f(5,8, 0)
data_3 = f(4, 4, 0)
data_4 = f(100, 200, 0)
plt.figure(figsize=(12, 6),dpi = 150)
plt.plot(x,data_1.pdf(x), label = '1,1')
plt.plot(x, data_2.pdf(x), label = '5,8')
plt.plot(x, data_3.pdf(x), label = '4,4')
plt.plot(x, data_4.pdf(x), label = '100,200')
plt.legend()
```

Out[40]:
```
<matplotlib.legend.Legend at 0x1c87969a610>
```



$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$

In [41]:
```
import random
random.seed(20)
```

In [42]:
```
# Generate data with same std
x = np.array([random.gauss(100,15) for x in range(20)])
y = np.array([random.gauss(100,15) for x in range(20)])
```

In [43]:
```
# Calculate f_test_statistic
f_test_stat = np.var(x,ddof=1) / np.var(y,ddof = 1)
```

In [44]:
```
dfn = x.size - 1
dfd = y.size - 1
```

In [45]:
```
# Calculate p value
p_value = 1 - f.cdf(f_test_stat,dfn,dfd)
```

In [46]:
```
p_value
```

Out[46]:
```
0.027388728443720423
```

p_value < 0.05 | Reject the Null Hypothesis | Variance is same

In [47]:
```
x = np.array([random.gauss(100,15) for x in range(20)])
y = np.array([random.gauss(200,15) for x in range(20)])
```

In [48]:
```
f_test_stat = np.var(x,ddof=1) / np.var(y,ddof = 1)
dfn = x.size - 1
dfd = y.size - 1
p_value = 1 - f.cdf(f_test_stat,dfn,dfd)
```

In [49]:
```
p_value
```

Out[49]:
```
0.6890214865041783
```

p_value > 0.05 | Accept Null Hypothesis | Variance is not same

F-Test

F-Test is any statistical test where the test statistic has an F-distribution under the null hypothesis. It is mostly used when comparing statistical models fitted to a data set and identifying the model that best fits the population from which the data were sampled.

F-Test always implies the comparison of two variances. It resembles ANOVA.

Comparing Two Variances

The F Test compares two variances, s_1 and s_2, by calculating their ratio. The result is always positive, as variances are always positive.

The equation for the F-Test is shown below:

$$F = {s_1}^2 / {s_2}^2$$

When the variances are equal, their ratio is 1.

Example:

If two data sets are available with sample 1 (variance of 8) and sample 2 (variance of 8), the ratio would be 10/10 = 1.

As a thumb rule, It is required to test whether the population variances are equal when running an F-Test. In other words, if the samples are from the same population, the variances are 1. Then the null hypothesis will always be that the variances are equal.

Assumptions for F-Test

1. The population distribution must be approximately fitting to the normal distribution (it should resemble a bell curve).
2. The samples must be independent of each other.
3. The larger variance is always in the numerator, to force the test into a right-tailed test. Right-tailed tests are easier to compute.
4. For two-tailed tests, alpha is to be taken at half-of-its value before finding the right critical value.
5. If the standard deviations are available, square them to get the value of variances.
6. If the degrees of freedom aren't available in the F Table, use the larger critical value to avoid Type I errors.

Steps to Perform F-Test

There are four simple steps for performing an F-Test:

Step 1. State the Null Hypothesis and Alternate Hypothesis

Step 2. Compute the F-value using the residual sum of squares, number of restrictions(m) and number of independent variables(k)

Formula: F = (SSE_1 – SSE_2 / m) / (SSE_2 / n-k)

Step 3. Find the F-statistic, the critical value

F-statistic = (variance of the group means) / (mean of the within group variances)

Step 4. Based on the results, support or reject the null hypothesis

Note: F-critical is calculated using the F-table, degree of freedoms and Significance level.If observed value of F is greater than the F-critical value then we reject the null hypothesis.

Use Case: Determine the Percentage of Defective User Data

Problem Statement:

TeleCall uses four centers around the globe to process customer order forms. They audit a certain percentage of the customer order forms. Any error in order form renders it defective and has to be reworked before processing. The manager wants to check whether the defective percentage varies by centre. Analyze the data at 5% significance level and help the manager draw appropriate inferences

Dataset:

Dimensions:

- 4 variables
- 300 observations

Attribute Information:

The dataset contains data from four different locations of accurate and erroneous user data.

In [63]:
```
customer = pd.read_csv('Customer+OrderForm.csv')
customer.head()
```

Out[63]:

	Phillippines	Indonesia	Malta	India
0	Error Free	Error Free	Defective	Error Free
1	Error Free	Error Free	Error Free	Defective
2	Error Free	Defective	Defective	Error Free
3	Error Free	Error Free	Error Free	Error Free
4	Error Free	Error Free	Defective	Error Free

In [64]:
```
customer.describe()
```

Out[64]:

	Phillippines	Indonesia	Malta	India
count	300	300	300	300
unique	2	2	2	2
top	Error Free	Error Free	Error Free	Error Free
freq	271	267	269	280

In [67]:
```
Phillippines_value=customer['Phillippines'].value_counts()
Indonesia_value=customer['Indonesia'].value_counts()
Malta_value=customer['Malta'].value_counts()
India_value=customer['India'].value_counts()
print(Phillippines_value)
print(Indonesia_value)
print(Malta_value)
print(India_value)
```

```
Error Free    271
Defective      29
Name: Phillippines, dtype: int64
Error Free    267
Defective      33
Name: Indonesia, dtype: int64
Error Free    269
Defective      31
Name: Malta, dtype: int64
Error Free    280
Defective      20
Name: India, dtype: int64
```

Based on the Error Free value and Defective value we will use chi-square Test

In [65]:
```
chiStats = sp.stats.chi2_contingency([[271,267,269,280],[29,33,31,20]])
print('T value is: ',chiStats[0])
print('P value is: ',chiStats[1])
```

```
T value is:  3.858960685820355
P value is:  0.2771020991233135
```

In [66]:
```
if chiStats[1] < 0.05:
    print('We reject Null Hypothesis')
else:
    print('We accept Null Hypothesis')
```

```
We accept Null Hypothesis
```

Since variables are not rejected hence, we cannot reject null hypothesis. Also, the proportion of defective percentage across the TeleCall center is same.

Powered by simplilearn

Note: In this lesson, we saw the use of the Linear regression,fundamentals statistics, probability, and advanced statistics in details, and in the next upcoming lessons we are going to use these techniques.