Analytics using Python

Learning outcomes

1. You will learn Python , a useful language

2. Use programming for problem solving

Great Lakes Institute of Management

A guide to learn python for analytics

P. V. Subramanian

**A workbook on Analytics using Python**

**P. V. SUBRAMANIAN**

**Visiting faculty – Great Lakes Institute of Management**

*Copyright @ 2019 Great Lakes Institute of Management : All contents in this book, including (but not limited to) its contents are proprietary to Great Lakes Institute of Management, Chennai. Any effort at re-using, modifying and circulating both through Digital or Physical means, without the express written permission of the author, will amount to infringement of copyrights, and will be liable for suitable action.*

**Chapter 5. Statistics using Python**

**Table of contents**

Table of Contents

[**1.** **Probability** 5](#_Toc8535503)

[**1.1** **Simple Probability** 5](#_Toc8535504)

[**1.2** **joint probability** 6](#_Toc8535505)

[**1.3** **Marginal probability** 8](#_Toc8535506)

[**1.4** **Conditional Probability** 10](#_Toc8535507)

[**1.5** **Bayes theorem¶** 12](#_Toc8535508)

[**2.** **Binominal** 15](#_Toc8535509)

[**2.1** Probability mass function 15](#_Toc8535511)

[**2.2** Cumulative Density function 15](#_Toc8535512)

[**2.3** Plot the binomial Density function 15](#_Toc8535513)

[**3.** **Poisson** 20](#_Toc8535514)

[**3.1** Properties: 20](#_Toc8535516)

[**3.2** Applications 20](#_Toc8535517)

[**3.3** important python functions for Poisson distribution problems 21](#_Toc8535518)

[**3.3.1** **Probability Mass Function** 21](#_Toc8535519)

[**3.3.2** **Cumulative Density Function** 21](#_Toc8535520)

[**4.** **Normal** 26](#_Toc8535521)

[**4.1** **Properties** 27](#_Toc8535523)

[**4.2** important python functions for normal distribution problems 28](#_Toc8535524)

[**4.2.1** **Cumulative Density Function (cdf)** 28](#_Toc8535525)

[**5.** **Sampling** 32](#_Toc8535526)

[**5.1** **Sampling methods** 32](#_Toc8535528)

[**5.1.1** **Simple Random Sampling** 32](#_Toc8535529)

[**5.1.2** **Stratified random sampling** 33](#_Toc8535530)

[**5.1.3** **Systematic sampling** 34](#_Toc8535531)

[**6.** **Central Limit Theorem** 35](#_Toc8535532)

[**6.1** **Implications of CLT** 36](#_Toc8535534)

[**7.** **Setting up of confidence intervals and hypothesis** 37](#_Toc8535535)

[**7.1** **Confidence Intervals** 37](#_Toc8535537)

[**7.1.1** **Confidence interval for population mean** 37](#_Toc8535538)

[**7.1.2** **Confidence interval for population proportion** 43](#_Toc8535539)

[**7.1.3** **Confidence interval for population mean when standard deviation is unknown** 45](#_Toc8535540)

[**7.1.4** **Confidence interval for population variance** 48](#_Toc8535541)

[**7.2** **Setting up hypothesis** 49](#_Toc8535542)

[7.2.1 Examples of some claims 49](#_Toc8535543)

[7.2.2 Type I Error, Type II error and power of the hypothesis test 49](#_Toc8535544)

[**7.2.3** Steps involved in solving the hypothesis testing 50](#_Toc8535545)

[**8.** **Sampling distributions such as t, Z** 55](#_Toc8535546)

[**8.1** **One sample and two sample tests** 55](#_Toc8535548)

[**8.1.1** **One sample test** 55](#_Toc8535549)

[**10.** **ANOVA** 65](#_Toc8535550)

[**10.1.** **Assumptions of ANOVA** 65](#_Toc8535563)

[**10.2.** **One-way ANOVA** 65](#_Toc8535564)

[**10.2.1** **Tukey’s Honest Significant Difference test** 69](#_Toc8535567)

[**11.** **Chi square** 74](#_Toc8535568)

[**11.1** **Properties of Chi Square distribution** 74](#_Toc8535570)

[**11.2** **Chi-square Goodness of fit tests** 75](#_Toc8535571)

[**11.3** **Chi-square tests of independence** 77](#_Toc8535572)

[**12.** **Exercises** 80](#_Toc8535584)

# **Probability**

**Simple, Joint, Marginal, Conditional probability and Bayes Theorem**

## **Simple Probability**

**Simple Probability refers to the probability of occurrence of a simple event.**

**Probability of an event X, P(X) is given by P(X)**

**= Number of observations in favorable event X / Total Number of observations**

**Example 1**

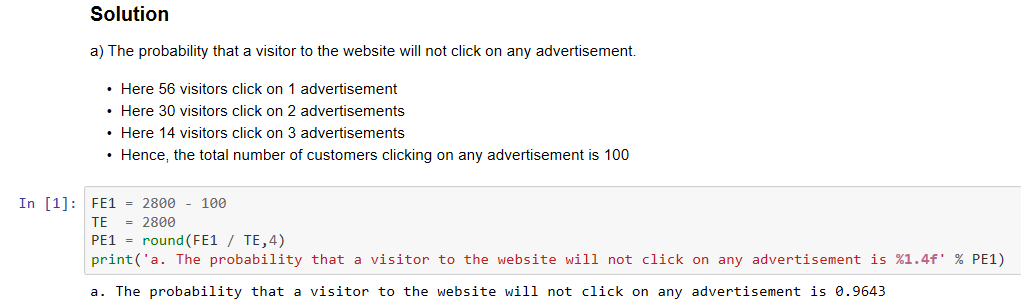
The data collected by an Advertisement agency has revealed that out of 2800 visitors, 56 visitors clicked on 1 Advertisement, 30 clicked on 2 advertisements and 14 clicked on 3 advertisements and the remaining did not click on any advertisement.

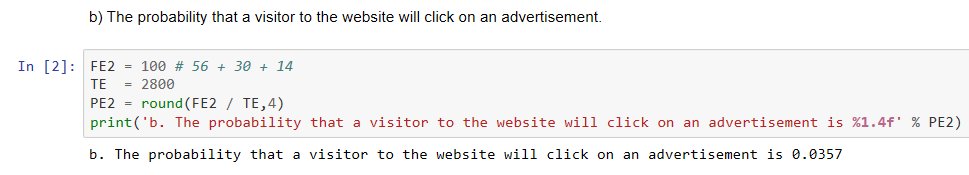
**Calculate**

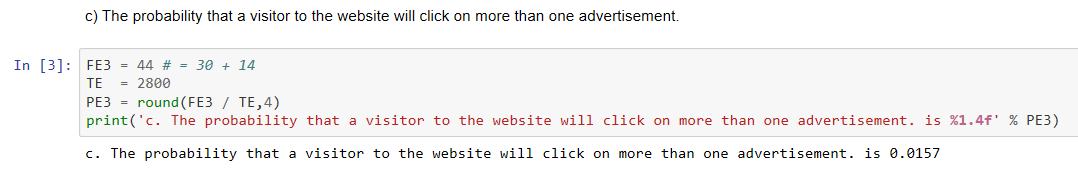
a) The probability that a visitor to the website will not click on any advertisement.

b) The probability that a visitor to the website will click on an advertisement.

c) The probability that a visitor to the website will click on more than one advertisement.

****

****

****

## **joint probability**

**Joint Probability refers to the probability of occurrence involving two or more events.**

**Example 2**

**Let A and B be the two events in a sample space. Then the joint probability if the two events denoted by P(A ∩B).**

**P(A ∩B) = Number of observations in A∩B / Total Number of observations**

*At a popular company service center, a total of 100 complaints were received. 80 customers complained about late delivery of the items and 60 complained about poor product quality.*

1. *Calculate the probability that a customer complaint will be about both product quality and late delivery.*
2. *What is the probability that a complaint will be only about late delivery?*

**Solution:**

1. *Calculate the probability that a customer complaint will be about both product quality and late delivery*

Let

L = Late delivery

Q = Poor quality

n(L) = Number of cases in favor of L = 80

n(Q) = Number of cases in favor of Q = 60

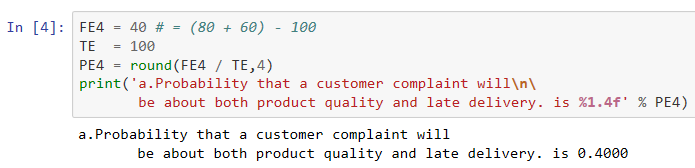
N = Total Number of complaints = 100

n(L∩Q) = (80 + 60) - 100 = 40

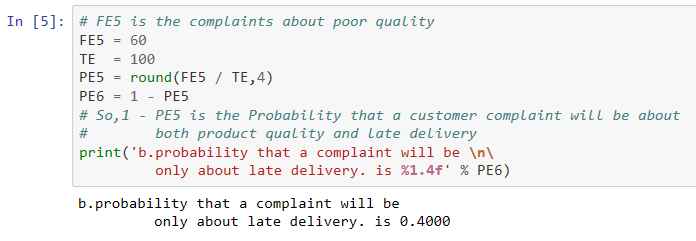
Probability that a customer complaint will be about both product quality and late delivery

= P(L∩Q)

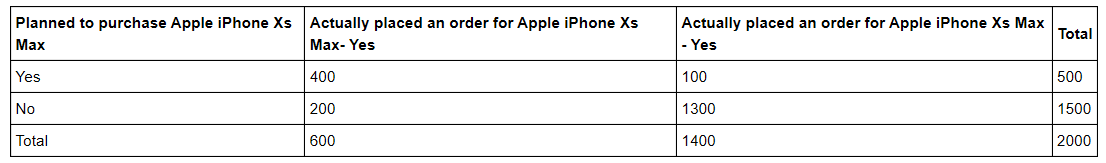
P(L∩Q) = n(L∩Q) / Total Number of observations



1. *What is the probability that a complaint will be only about late delivery?*

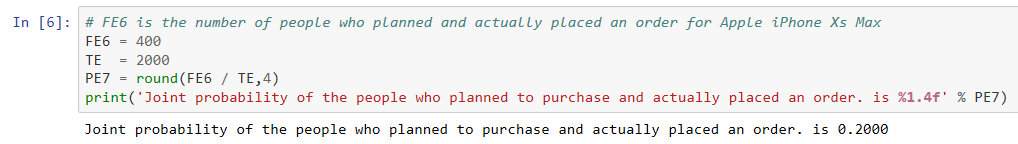


**Example 3**



*Calculate the joint probability of the people who planned to purchase and actually placed an order.*

*You observe from the above table, that 400 people planned to purchase and placed an order for Apple iPhone Xs Max is 400 out of 2000 people.*



## **Marginal probability**

**Marginal probability refers to the probability of an event without any condition.**

P(A) = P(A and B1) + P(A and B2) + P(A and B3) + ... + P(A and Bk)

where B1, B2, B3, ..., Bk are k mutually exclusive and collectively exhaustive events, defined as follows:

* Two events are mutually exclusive if both the events cannot occur simultaneously.
* A set of events are collectively exhaustive if one of the events must occur.

**Example 4**

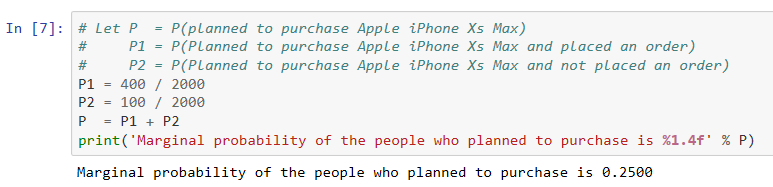
*Use the purchase of Apple iPhone Xs Max table.*

***What is the probability of planned to purchase Apple iPhone Xs Max?***

*P(planned to purchase Apple iPhone Xs Max) =*

*P(Planned to purchase Apple iPhone Xs Max and placed an order) +*

*P(Planned to purchase Apple iPhone Xs Max and not placed an order)*



Note that you get the same result by adding the number of outcomes that make up the simple event planned to purchase and calculate the probability of that simple event.

**General addition rule**

*To get the probability of the event A or B, you need to consider the occurrence of either event A or B or both A and B.*

*P(A or B) = P(A) + P(B) - P(A and B)*

***From set theory***

* *P(A∪B) is the event that either A or B or both occur.*
* *P(A∩B) is the event that both A and B occur at the same time.*
* *Events A and B are mutually exclusive if they cannot happen at the same time:*

*P(A∪B) = P(A) + P(B) - P(A ∩B)*

**Example 5**

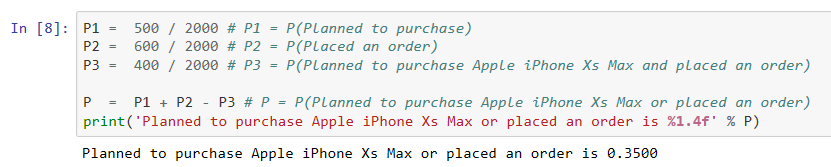
***Use the purchase of Apple iPhone Xs Max table.***

***What is the probability of planned to purchase Apple iPhone Xs Max or placed an order?***

***P(Planned to purchase Apple iPhone Xs Max or placed an order)***

***= P(Placed an order) + P(Planned to purchase Apple iPhone Xs Max) -***

***P(Planned to purchase Apple iPhone Xs Max and placed an order)***



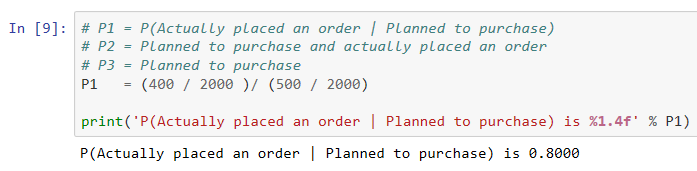
## **Conditional Probability**

***Conditional Probability refers to the probability of event A, given information about the occurrence of another event B*** .

**Example 6**

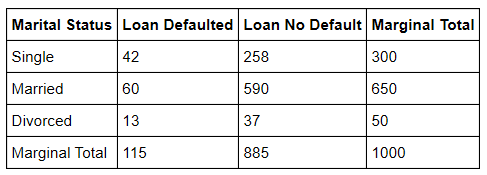
***Use the purchase of Apple iPhone Xs Max table.***

***Find the joint probability of the people who planned to purchase and actually placed an order, given that people planned to purchase.***

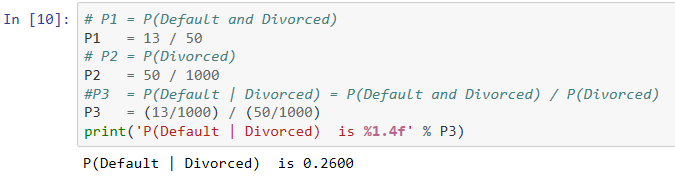


**Example 7**

***The following table describes loan default status at a bank and their marital status.***

******

***Based on the above table, calculate the probability of default given divorced.***

****

**Independent Events**

**Two events, A and B are independent if and only if P(A | B) = P(A),**

**where**

* P(A|B) is the conditional probability of A given B
* P(A) is the marginal probability of A

Example: A student getting A grade in both Final Stats exam and in final Marketing exam

**Example 8**

***What is the probability of getting a "6" in two consecutive trials when rolling a dice?***

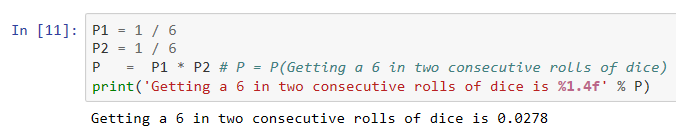
***For each roll of a dice:***

* Favorable events = {"6"}
* Total number of outcomes = {"1","2","3","4","5","6"}

Let P1 be the probability of getting a "6" in the first roll of dice.

Let P2 be the probability of getting a "6" in the second roll of dice.

Since first roll of dice does not influence the second roll of dice, these events are independent.



## **Bayes theorem**[**¶**](file:///D:\GL\Dsc_FullTime\Course_Workbook\Stats\day1_Statistics.html#Bayes-theorem)

***Bayes' Theorem is used to revise previously calculated probabilities based on new information***

*P*(*Bi*∣*A*)P(Bi∣A) =

*P(A∣Bi)P(Bi) / P(A∣B1)P(B1)+P(A∣B2)P(B2)+P(A∣B3)P(B3)+...+P(A∣Bk)P(Bk)*

***where* Bi *is the ith event of k mutually exclusive and collectively exhaustive events A is the new event that might impact P(*B*i )***

**Example 9**

A certain Electronic equipment is manufactured by three companies, X, Y and Z.

1. 75% are manufactured by X
2. 15% are manufactured by Y
3. 10% are manufactured by Z

*The defect rates of electronic equipment manufactured by companies X, Y and Z are 4%, 6% and 8%.*

*If an electronic equipment is randomly found to be defective, what is the probability that it is manufactured by X?*

**Solution:**

Let P(X),P(Y) and P(Z) be probabilities of the electronic equipment manufactured by companies X, Y and Z respectively.

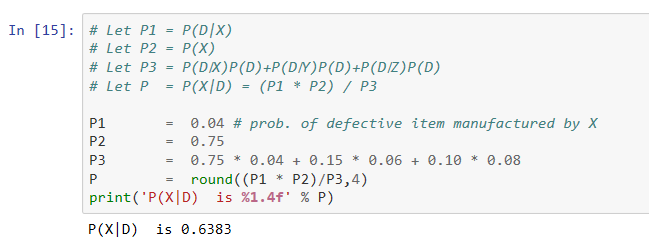
Let P(D) be the probability of defective electronic equipment.

We are interested in calculating the probability P(X|D).

P(X | D) = P(D | X)P(X) / P(D)

Bayes' rule in our case is given below:

P(X ∣ D) = P(D ∣ X)P(X) / {P(D ∣ X) P(D)+P(D ∣ Y) P(D)+P(D ∣ Z) P(D)}



**Example 10**

Given the following statistics, what is the probability that a woman has cancer if she has a positive mammogram result? We are given the following details:

a) 1% of over 50 have breast cancer

b) 90% of women who have breast cancer test positive on mammograms

c) 8% of women will have false positive

**Solution**

Let

* event A denote woman has breast cancer
* event ~A denote woman has no breast cancer
* event T denote mammogram test is positive
* event ~T denote mammogram test is negative

Let P(A) denote the probability of women over 50 years of age having breast cancer. P(A) = 0.01 So, P(~A) = 1 - 0.01 = 0.99.

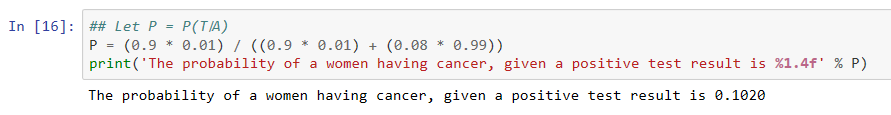
Let P(T|A) denote the conditional probability of women given positive result on mammograms and having breast cancer .

P(T|A) = 0.9

Let P(T|~A) denote the conditional probability of women given the positive result on mammograms and not having breast cancer .

P(T | ~A) = 0.08

P(T ∣ A) = P(A ∣ T)P(A) / {P(A ∣ T)P(A) + P(A ∣ T)P(T)



# **Binominal**

* It is widely used probability distribution of a discrete random variable.
* Plays major role in quality control and quality assurance function.

where P(X = x) is the probability of getting x successes in n trials and *π*π is the probability of an event of interest.

**Some important functions in Python for Binomial distribution:**



## Probability mass function

scipy.stats.binom.pmf gives the probability mass function for the binomial distribution

binomial = scipy.stats.binom.pmf (k,n,p),

where k is an array and takes values in {0, 1, ..., n}

n and p are shape parameters for the binomial distribution

The output, binomial, gives probability of binomial distribution function in terms of array.

## Cumulative Density function

cumbinomial = scipy.stats.binom.cdf(k,n,p) gives cumulative binomial distribution.

The output, cumbinomial, gives cumulative probability of binomial distribution function in terms of array.

## Plot the binomial Density function

The function, matplotlib.pyplot.plot(k, binomial, ‘o-’) gives us plot of the binomial distribution function.

**Example 11**

A LED bulb manufacturing company regularly conducts quality checks at specified periods on the products it manufactures. Historically, the failure rate for LED light bulbs that the company manufactures is 5%. Suppose a random sample of 10 LED light bulbs is selected. What is the probability that

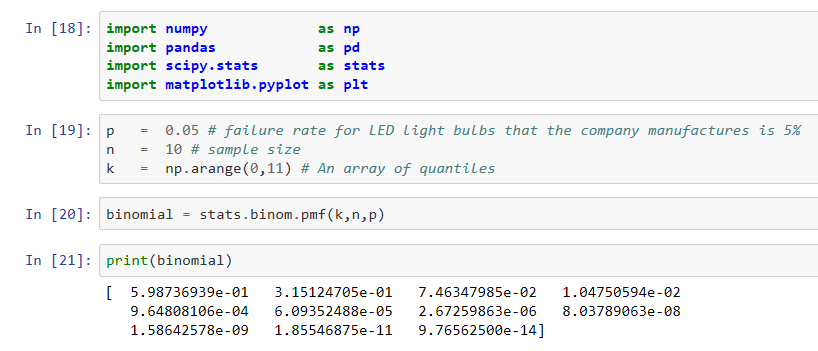
a) None of the LED bulbs are defective?

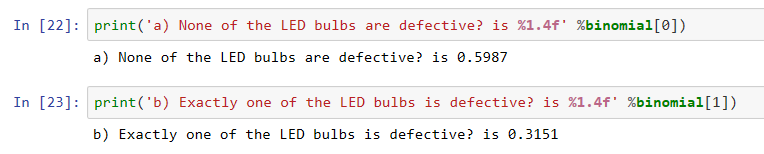
b) Exactly one of the LED bulbs is defective?

c) Two or fewer of the LED bulbs are defective?

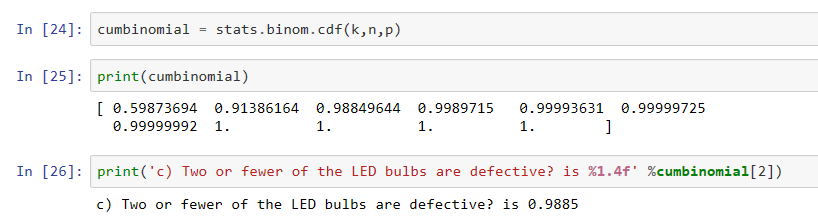
d) Three or more of the LED bulbs are defective

**Solution:**

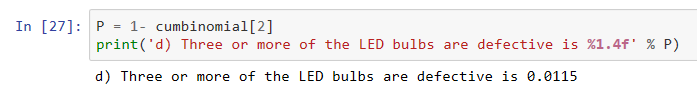


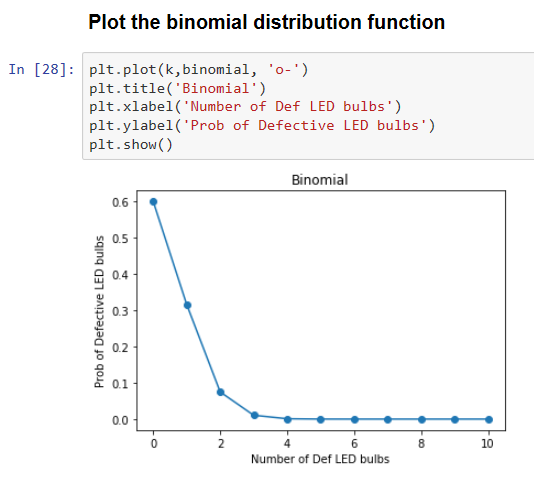


To answer the question c) Two or fewer of the LED bulbs are defective? we need to calculate cumulative probability of up to Two LED bulbs being defective.



To answer the question d) Three or more of the LED bulbs are defective, we need to subtract cumulative Probability upto 2 defective LED bulbs from 1.





**Example 12**

The percentage of orders filled correctly at Wendy's was approximately 86.8%. Suppose that you go to drive-through window at Wendy's and place an order. Two friends of yours independently place orders at the drive-through window at the same Wendy's.

What are the probabilities that

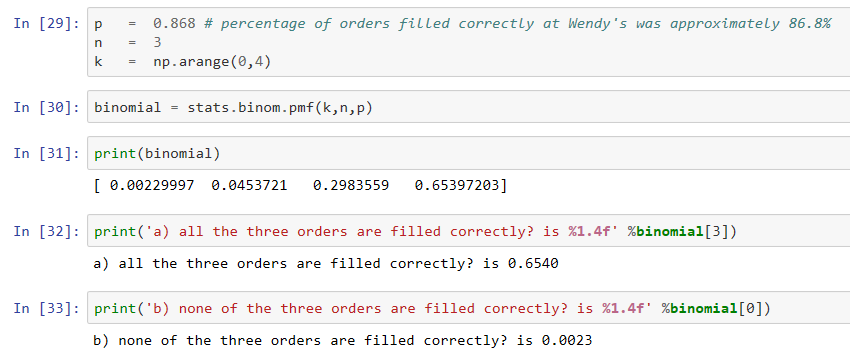
a) all three

b) none of the three

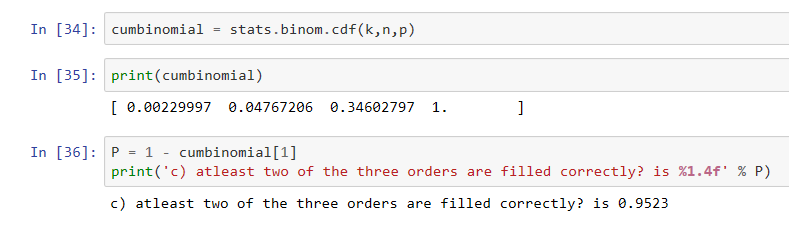
c) at least two of the three orders will be filled correctly.

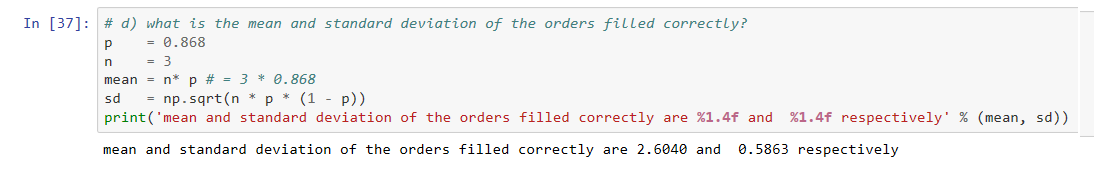
d) what is the mean and standard deviation of the orders filled correctly?

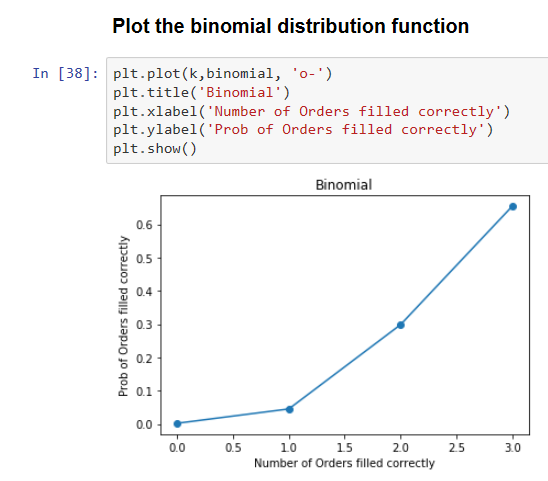
**Solution:**



To answer this question c) at least two of the three orders are filled correctly, we need to find out 1 - Probability of up to 1 failure.







# **Poisson**

* This discrete distribution which also plays a major role in quality control.
* The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time or space. In such areas of opportunity, there can be more than one occurrence. In such situations, Poisson distribution can be used to compute probabilities.
* Examples include number of defects per item, number of defects per transformer produced.
* Notes: Poisson Distribution helps to predict the arrival rate in a waiting line situation where a queue is formed, and people wait to be served and the service rate is generally higher than the arrival rate.



## Properties:

* Mean μ = λ
* Standard deviation σ = √ μ
* The Poisson distribution is the limit of binomial distribution as n approaches ∞and p approaches 0

where

* P(x) = Probability of x successes given an idea of λ
* λ = Average number of successes
* e = 2.71828 (based on natural logarithm)
* x = successes per unit which can take values 0,1,2,3, ... ∞

## Applications

* Car Accidents
* Number of deaths by horse kicking in Prussian Army (first application)
* Birth defects and genetic mutation

**Note:**

* If there is a fixed number of observations, n, each of which is classified as an event of interest or not an event of interest, use the binomial distribution.
* If there is an area of opportunity, use the Poisson distribution.

## important python functions for Poisson distribution problems

### **Probability Mass Function**

* poisson = scipy.stats.poisson.pmf(n, rate)

where n is where n is an array like quantiles and rate is the mean It gives poisson distribution result in the form of an array.

### **Cumulative Density Function**

* poisson = scipy.stats.poisson.cdf(n,rate)

where n is where n is an array like quantiles and rate is the mean It gives cumulative density function result in the form of an array.

**Example 13**

The number of work-related injuries per month in a manufacturing plant is known to follow a Poisson distribution, with a mean of 2.5 work-related injuries a month.

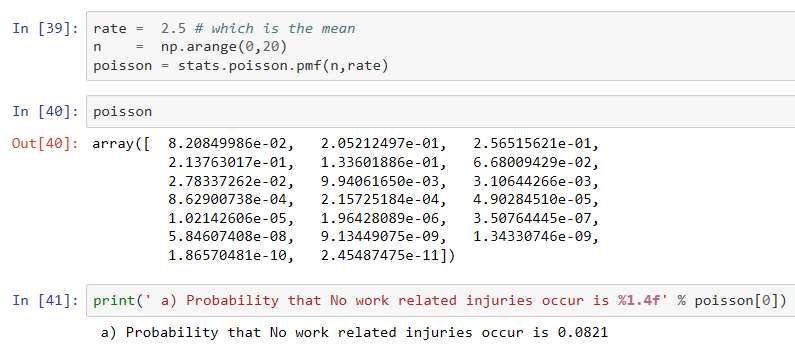
***What is the probability that in a given month,***

***a) No work related injuries occur?***

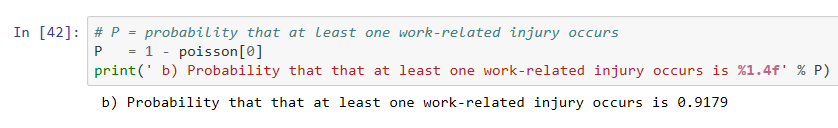
***b) At least one work-related injury occurs?***

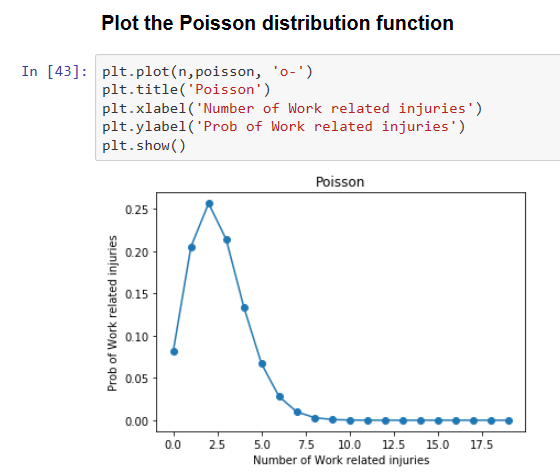
**Solution:**

**Here, λ = 2.5 injuries**



***To calculate the probability that at least one work-related injury occurs, we need to subtract probability of no work related injury from 1.***





**Example 14**

A Life Insurance agent sells on the average 3 life insurance policies per week.

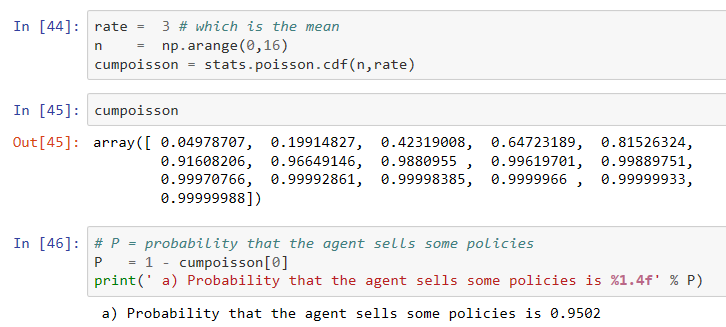
Use the Poisson law to calculate the probability that in a given week, he will sell

1. Some policies
2. 2 or more but less than 5 policies?

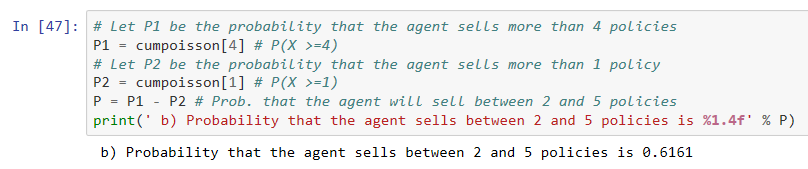
**Solution:**

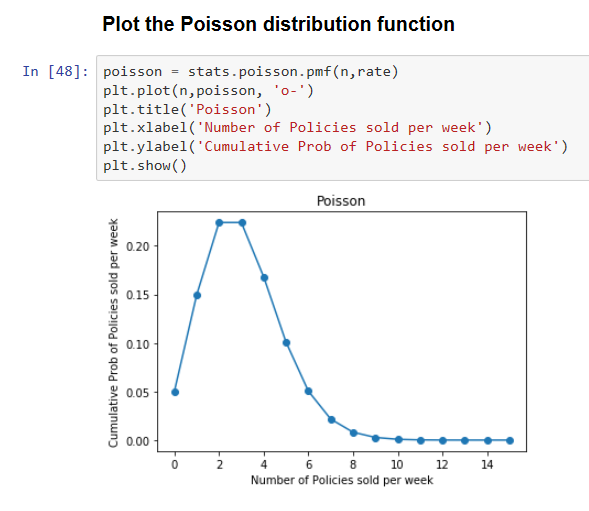
Here, λ = 3

* 1. Agent sells some policies (at least one)



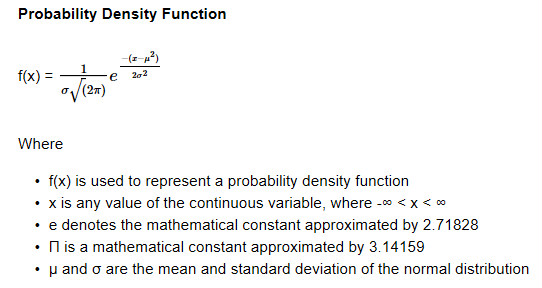
* 1. Agent sells 2 or more but less than 5 policies



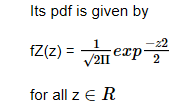


# **Normal**

* One of the most popular continuous distribution in Analytics field.
* Normal distribution is observed across many naturally occurring measures such as birth weight, height and intelligence etc.



* For a continuous function, the probability density function is the probability that the value has the value x.
* Since for continuous distributions the probability at a single point is zero, this is expressed in terms of an integral of its probability density function P(X<= x) = F(x) = ∫x−∞f(t)dt
* Standardize normal variable Compute Z by subtracting the mean, mu from a normally distributed variable, divide by the standard deviation, sigma.
* Z = (X - μ) / σ Z is in standard units. Z ~ N(0,1) and the variable, Z always has mean 0 and standard deviation 1

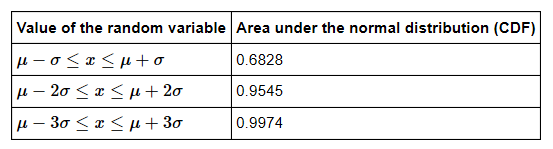


* The (1 / √2Π ) is there to make sure that the area under the PDF is 1.
* For a normal distribution, mu is the location parameter, which locates (center) the distribution on the horizontal axis.
* Sigma is the scale parameter, which defines the spread of the normal distribution.
* Normal distribution has no shape parameter since all the normal distribution curves have bell shape and are symmetrical.



## **Properties**

* Theoretical normal density functions are defined between -∞ and ∞
* There are two parameters, location (μ which is the mean) and scale (σ which is standard deviation).
* It has a symmetrical (bell shape) around the mean. mean = median = mode
* Areas between specific values are measured in terms of μ and σ
* Any linear transformation if a normal random variable is also normal random variable.
* If X1 is an independent normal random variable with mean μ1 and variance σ12 and X2 is another independent normal random variable with mean μ2 and σ22, then X1 + X2 is also a normal distribution with mean μ1 + μ2 and variance σ12+ σ22 .



## important python functions for normal distribution problems

### **Cumulative Density Function (cdf)**

* scipy.stats.norm.cdf(z) # Here z is an attribute. Cumulative distribution function at z used to calculate the area under the curve up to the Z score
* stats.norm.cdf(z2) – stats.norm.cdf(z1) # Here z is an attribute. This is used to calculate the area under the curve between two scores or the probability that a score would turn out to be between two scores.
* stats.norm.isf(0.99) # Inverse Survival function gives the value given a probability. This is used to calculate the z score that corresponds to this probability.

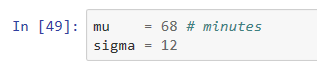
**Example 15**

A survey on use of smart phones in India was conducted and it is observed the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes.

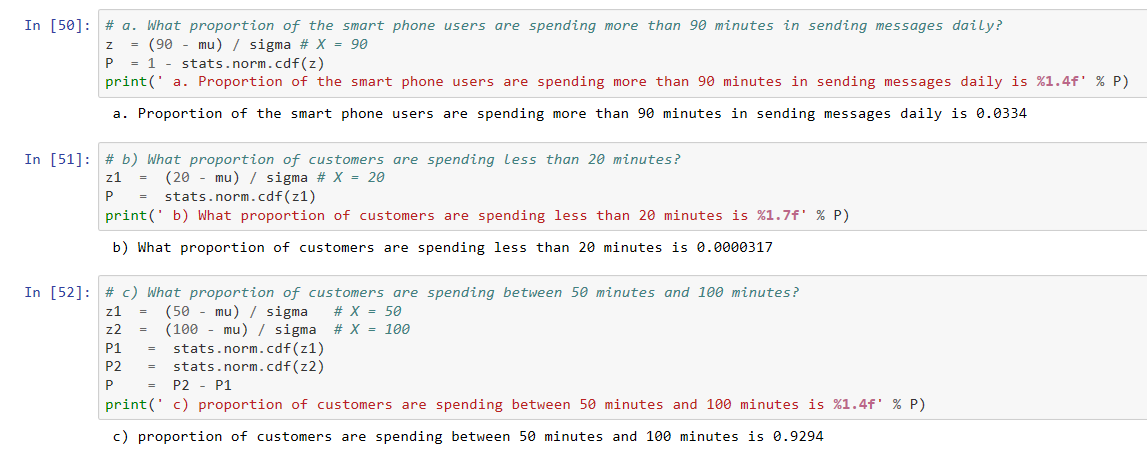
Assume that the time spent in sending messages follows a normal distribution.

1. What proportion of the smart phone users are spending more than 90 minutes in sending messages daily?
2. What proportion of customers are spending less than 20 minutes?
3. What proportion of customers are spending between 50 minutes and 100 minutes?

**Solution:**

****

*Compute Z by subtracting the mean, mu from a normally distributed variable, divide by the standard deviation, sigma*

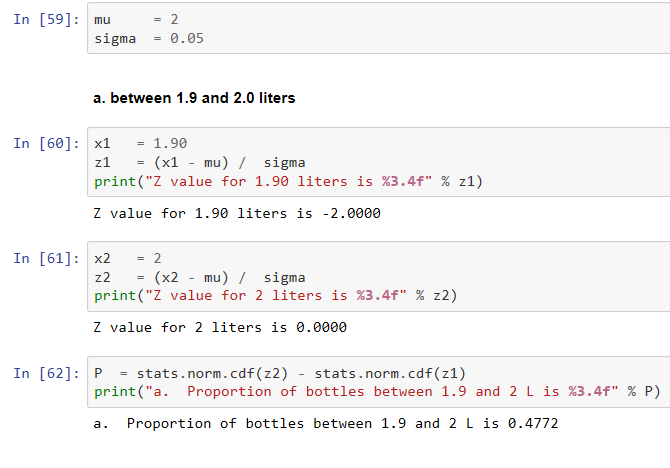


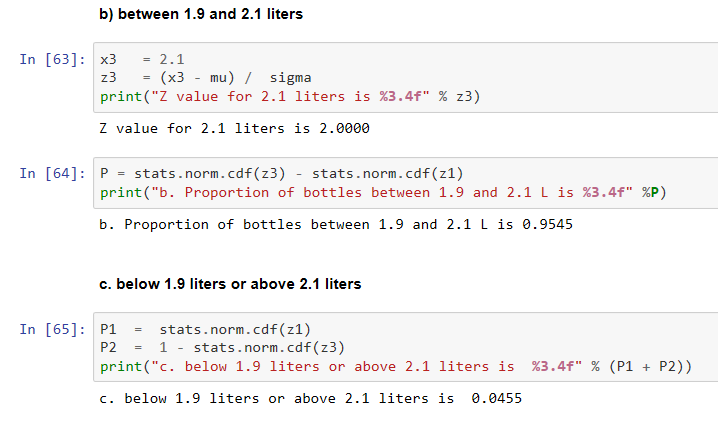
**Example 16**

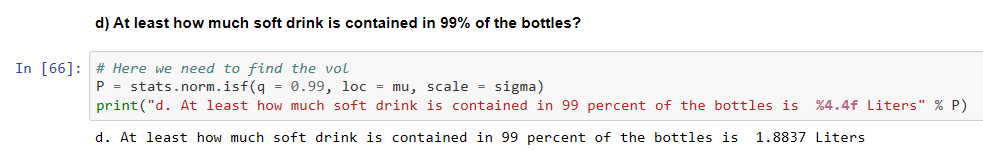
The fill amount in 2-liter soft drink bottles is normally distributed, with a mean of 2.0 liters and a standard deviation of 0.05 liter. If the bottles contain less than 95% of the listed net content (1.90 liters, in our case), the manufacturer may be subject to penalty by the state office of consumer affairs. Bottles that have a net content above 2.1 liters may cause excess spillage upon opening. What is the proportion of bottles that will contain

1. between 1.9 and 2.0 liters
2. between 1.9 and 2.1 liters
3. below 1.9 liters or above 2.1 liters
4. At least how much soft drink is contained in 99% of the bottles?

**Solution:**







# **Sampling**

***Sampling theory*** *is the field of statistics that is involved with the collection, analysis and interpretation of data gathered from random samples of a population under study.*



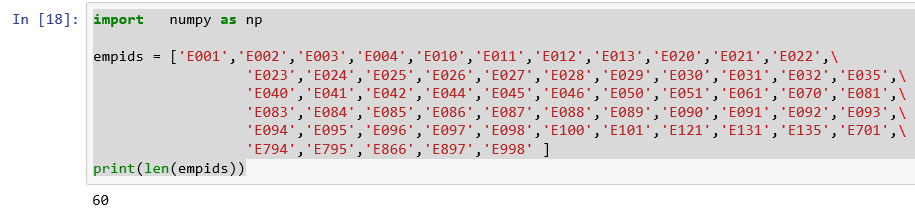
## **Sampling methods**

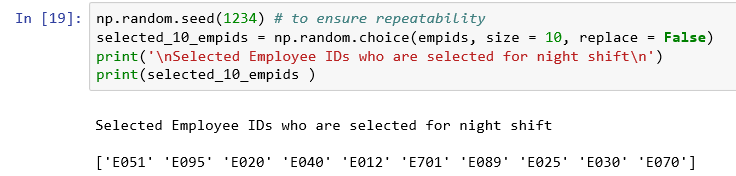
### **Simple Random Sampling**

* This is the simplest type of sampling.
* *Consider a population of size, n and you want to sample k units from this population. Then a simple random sample of size, k has the property that every possible sample of size k being chosen.*

**Example 17**

From the list of 60 employees, select 10 of them randomly for performing night duties next month.

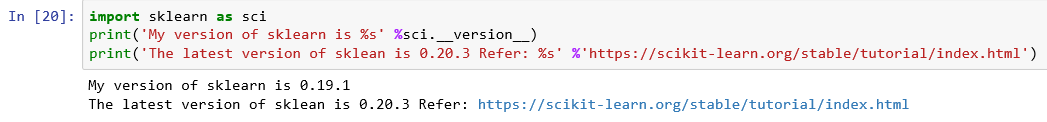




*This is a simple random sample of size 10 where each possible sample of size 10 has the same chance of being chosen.*

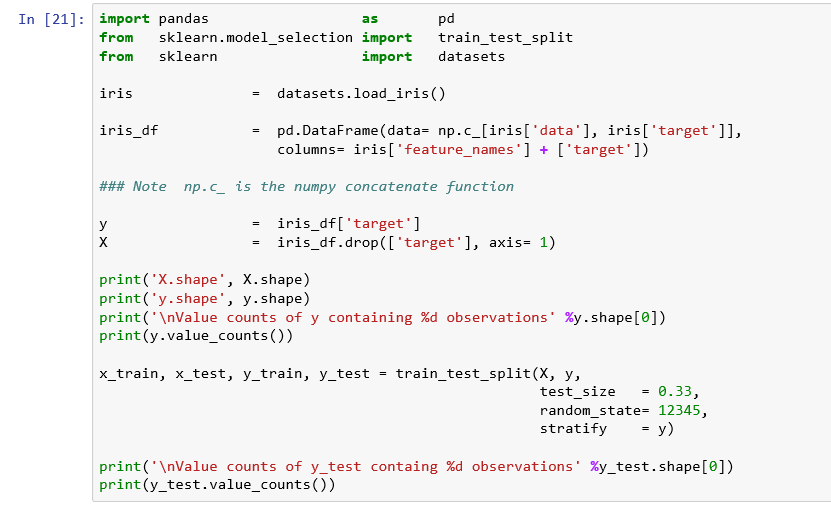
### **Stratified random sampling**

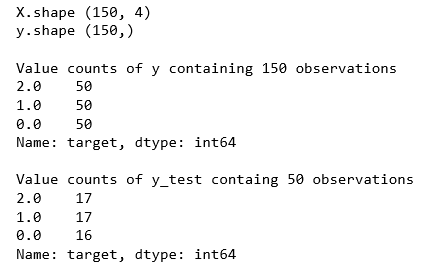
* Stratified random sampling is a method of sampling that involves the division of a population into relatively homogeneous subsets called strata and then random samples are taken from each stratum. Here, we get estimates within each stratum unlike simple random sampling.
* Ref: Book Business Analytics - Data Analysis and Decision Making - by Albright and Winston.
* Scikit-learn is an open source Python library that implements a range of machine learning, preprocessing, cross-validation and visualization algorithms using a unified interface.
* The parameter, stratify takes value such as array-like or None (default is None). If not None, data is split in a stratified fashion, using this as the class labels.



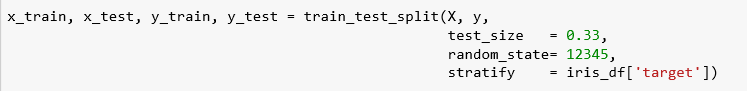
**Example 18**

*Use the iris data set preloaded in sklearn datasets. Split the data into 67%:33% so that we have a proper stratified random sampling and show the value counts of y and y\_test.*





***You can also mention the value of the parameter, stratify as follows to get the same result.***



### **Systematic sampling**

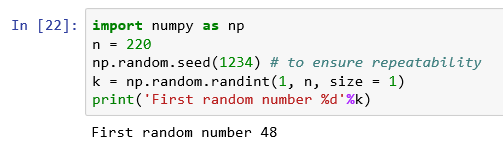
* Systematic sampling is a probability sampling method where the elements are chosen from a target population by selecting a random starting point, k and selecting other members after k, a fixed sampling interval.
* Here k = N/n where N is the population size and n is the desired sample size.

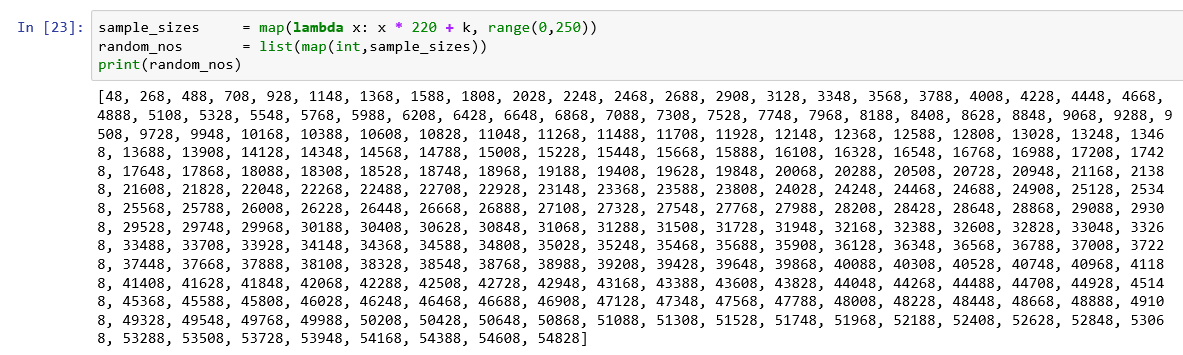
**Example 19**

*Assume there are 55000 customers and their list is arranged in the decreasing order of order volumes. You are asked to select 250 customers. So you divide the population size by sample size = 55000 / 250 = 220.*

**SOLUTION:**

*We have got the population of customers subdivided into 250 customer samples of size 220. So, the sample size, n= 220. So, we need to first choose a random number between 1 and 220 (both inclusive) and add the sample size to get other numbers.*





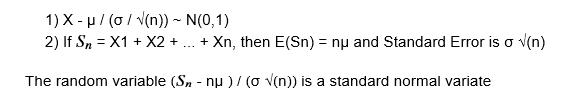
*Now you choose a number randomly between 1 and 220. Assume that number is 48. So, you would choose 48, 268, 488, and so on.*

# **Central Limit Theorem**

* Central Limit Theorem (CLT) is one of the most important theorems in Statistics due to its applications in testing of hypothesis.
* CLT states that for a large sample drawn from a population with mean μ and standard deviation σ, the sampling distribution of mean, follows an approximate normal distribution with mean, μ and standard error σ / √(n) irrespective of the distribution of the population for large sample size.
* Let S1, S2, ..., Sk be samples of size n, drawn from an independent and identically distributed population with mean, μ and standard deviation, σ.
* 
* 
* As a rule, statisticians have found that for many population distributions, when the sample size is at least 30, the sampling distribution of the mean is approximately **normal.**



## **Implications of CLT**



**Example 20**

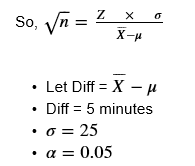
**A hospital is interested in estimating the average time it takes to discharge a patient after the doctor signs the discharge summary sheet.**

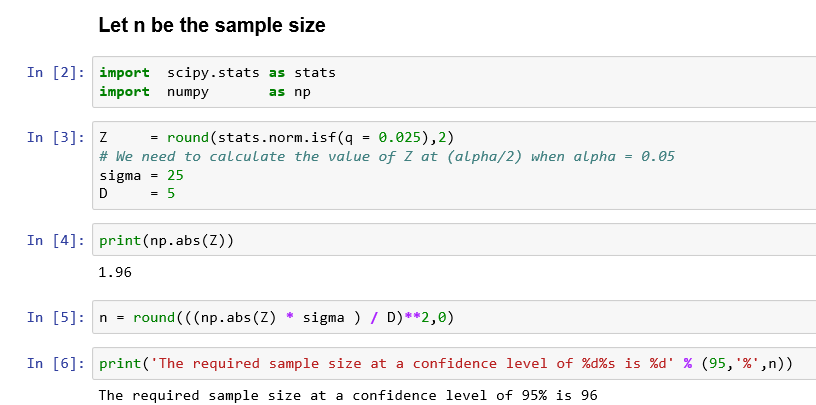
**Calculate the required sample size at a confidence level of 95%. Assume that the population standard deviation is 25 minutes.**

**Solution**

From the CLT (Central Limit theorem), we know that the sampling distribution of the mean follows a normal distribution with mean μ and standard deviation σ/ √n.







# **Setting up of confidence intervals and hypothesis**



## **Confidence Intervals**

* When there is an uncertainty around measuring the value of an important population parameter, it is better to find the range in which the range in which the value of the parameter is likely to lie rather than predicting a point estimate (single value).
* Confidence interval is the range in which the value of a population parameter is likely to lie with certain probability.
* Confidence interval provides additional information about the population parameter that will be useful in decision making.

### **Confidence interval for population mean**

* Let X1, X2, X3, ..., Xn be the sample means of samples, S1, S2, S3, ..., Sn that are drawn from an independent and identically distributed population with mean, μ and standard deviation, σ.
* From the Central Limit Theorem, we know that the sample means, Xi follows a normal distribution with mean, μ and standard deviation σ√n.



* We can distribute αα (probability of not observing true population parameter mean in the interval) equally (α/2α/2) on either side of the distribution.
* For α = 0.05 or α/2 = 0.025, that is 95% confidence interval, we can calculate lower and upper values of the confidence interval from the standard normal distribution.
* scipy.stats.norm.isf(q = 0.025) gives the value of Z for which the area under the normal distribution is less than 0.025.
* The corresponding value is approximately 1.96 as shown in the previous example.
* Using the transformation relationship between standard normal random variable Z and normal random variable X, we can write the 95% confidence interval for population mean when population standard deviation (σ) is known as:



* In general, (1 - α) 100% the confidence interval for the population mean when population standard deviation is known can be written as:

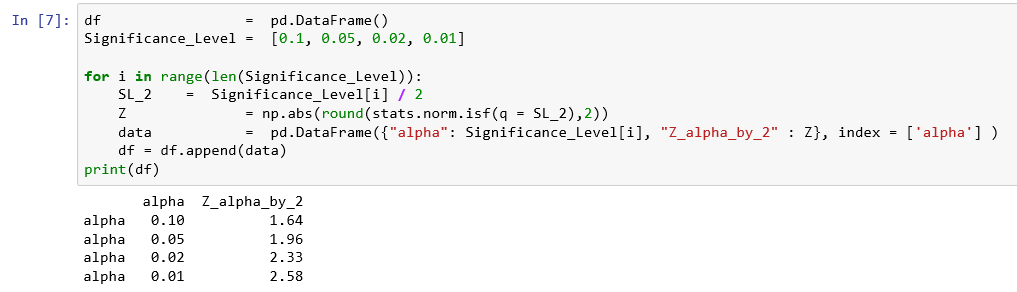
****

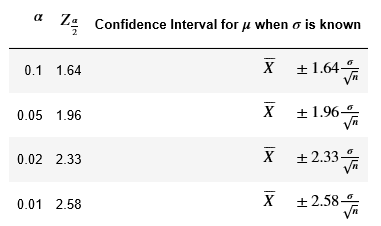
* This equation is valid for large sample sizes, irrespective of the distribution of the population.
* This is equivalent to the following:



**Example 21**

**Compute the absolute values of Zα/2 and show it in a table.**





**Example 22**

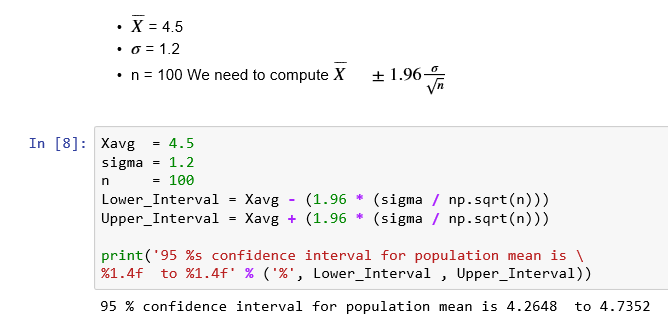
**A sample of 100 diabetic patients was chosen to estimate the length of stay at a local hospital. The sample was 4.5 days and the population standard deviation was known to be 1.2 days.**

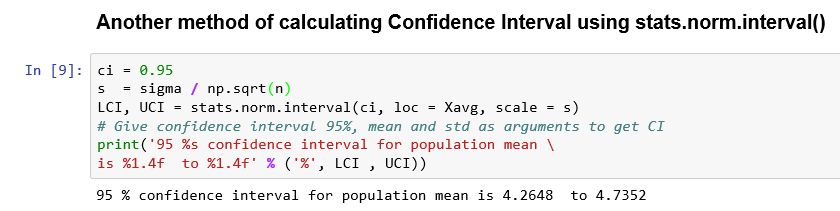
1. *Calculate the 95% confidence interval for the population mean.*
2. *What is the probability that the population mean is greater than 4.73 days?*

**Solution**

1. *Calculate the 95% confidence interval for the population mean.*

**It is known that**





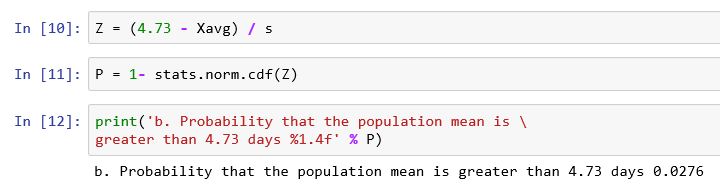
1. *What is the probability that the population mean is greater than 4.73 days?*

*We need to do the following:*

1. *Calculate Z value corresponding to 4.73,*

*by subtracting Xavg and divide by s.*

1. *Find out the probability corresponding to the Z value using scipy.stats.norm.cdf*
2. *Subtract from 1 since cdf gives cumulative probability upto the Z value as we are interested in finding the probability that the population mean is greater than Z.*



**Example 23**

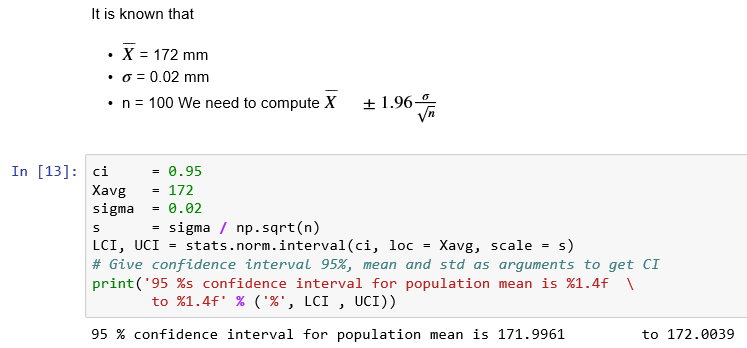
**Hindustan Pencils Pvt. Ltd. is an Indian manufacturer of pencils, writing materials and other stationery items, established in 1958 in Mumbai. Nataraj brand of pencils manufactured by the company is expected to have a mean length of 172 mm and the standard deviation of the length is 0.02 mm.**

**To ensure quality, a sample is selected at periodic intervals to determine whether the length is still 172 mm and other dimensions of the pencil meet the quality standards set by the company.**

**You select a random sample of 100 pencils and the mean is 170 mm.**

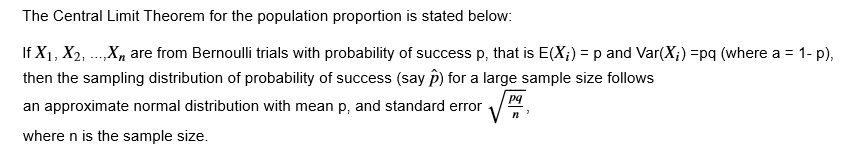
**Construct a 95% confidence interval for the pencil length.**

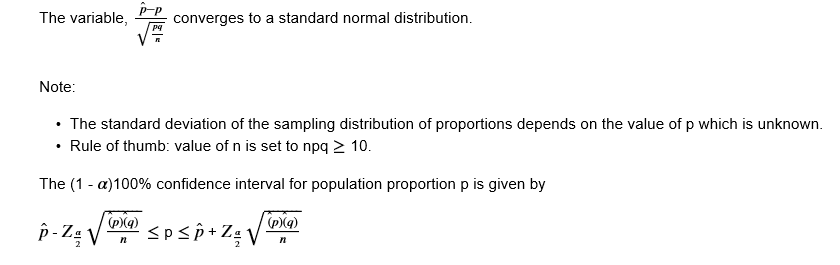
**Solution:**

****

### **Confidence interval for population proportion**

**The Central Limit Theorem for the population proportion is stated below:**

****

****

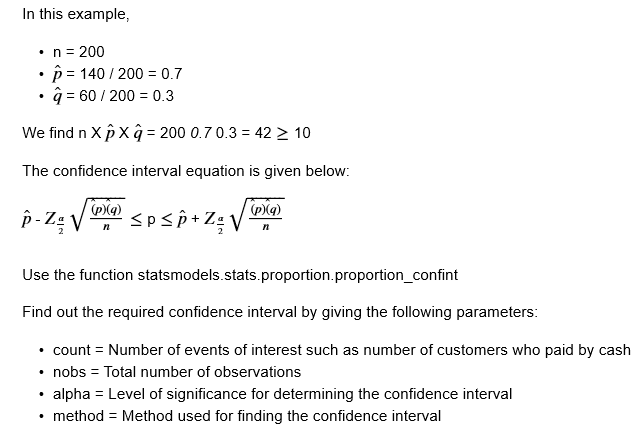
**Example 24**

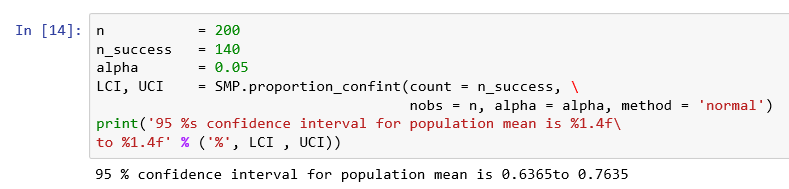
**A medical pharmacy was interested in finding the proportion of customers who pay cash for their medicines as against digital cash or plastic money.**

**From a sample of 200 customers, it was found that 140 customers paid by cash.**

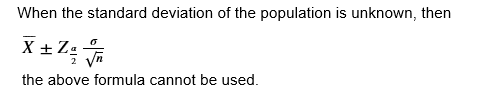
**Calculate the 95% confidence interval for proportions who pay by cash.**

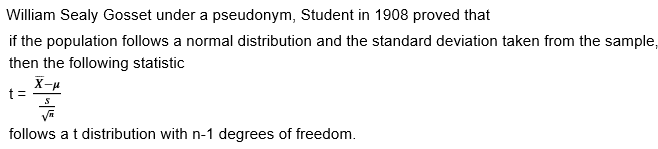
**Solution:**

****

****

### **Confidence interval for population mean when standard deviation is unknown**





Here, S is the standard deviation (aka standard error) estimated from the sample.

*μ* is the population mean and n is the sample size

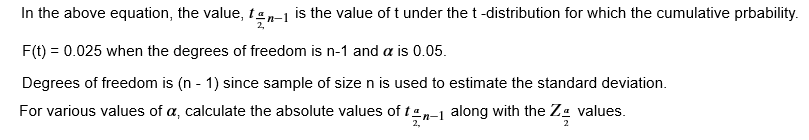
**Note:** n - 1 is the degrees of freedom. Degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated to arrive at the estimate in question. In other words, the number of values in the final calculation of a statistic that are free to vary.

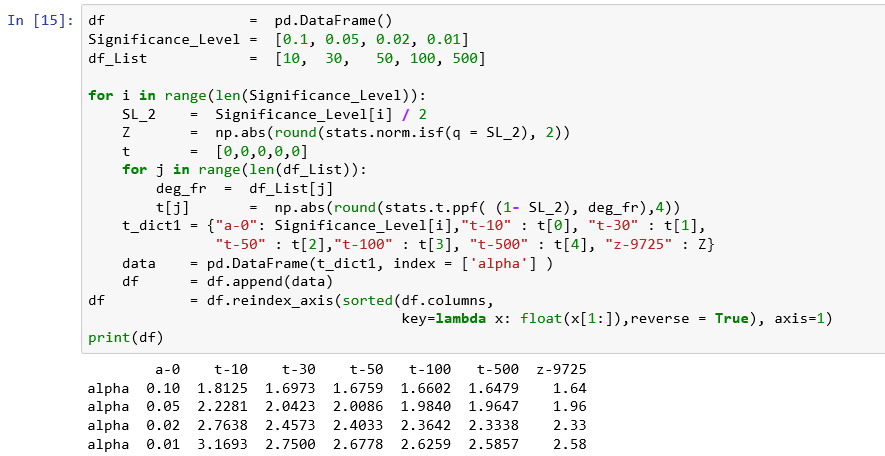
The t distribution is very similar to the standard normal distribution.

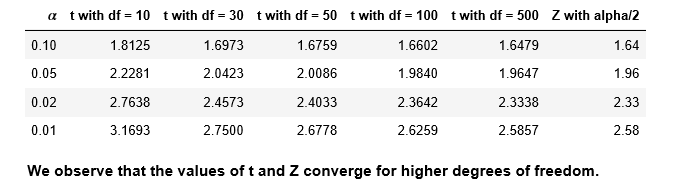
As the degrees of freedom increases, the t distribution converges to standard normal distribution.

The (1 - α) 100% confidence interval for population mean when the population standard deviation is unknown is given as:









**Example 25**

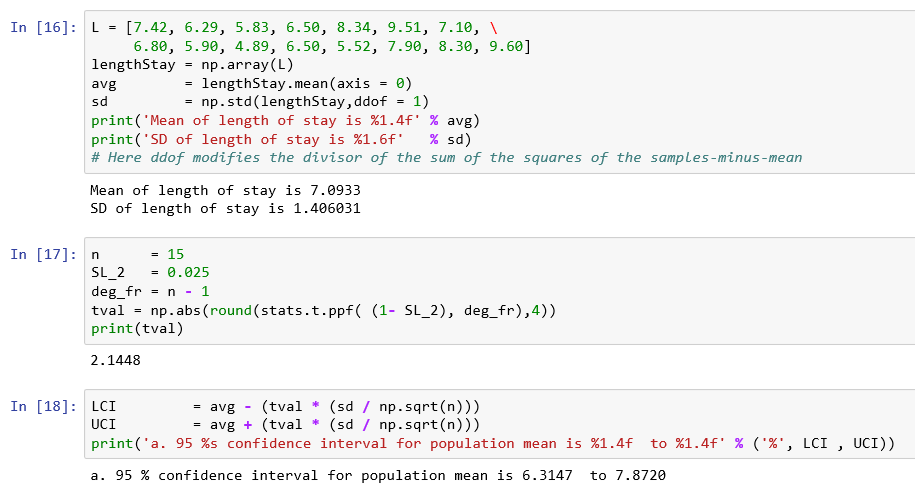
**The following table contains the length of stay in minutes of each customer at a Fast Food restaurant.**

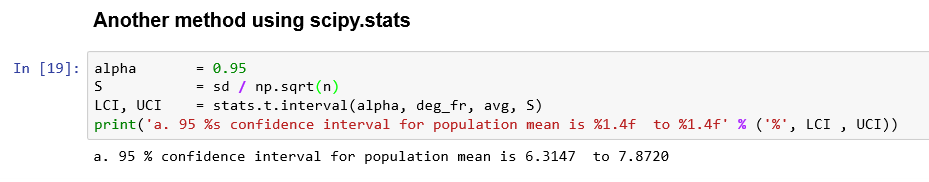
**7.42 6.29 5.83 6.50 8.34**

**9.51 7.10 6.80 5.90 4.89**

**6.50 5.52 7.90 8.30 9.60**

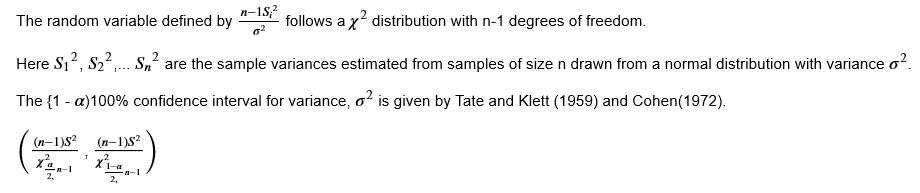
1. Construct 95% confidence interval estimate for the population mean length of stay at Fast Food restaurant, assuming a normal distribution.
2. Interpret the interval constructed at a.





1. ***You can be 95% confident that the mean length of stay at a Fast Food restaurant lies between 6.31 minutes to 7.87 minutes.***

### **Confidence interval for population variance**

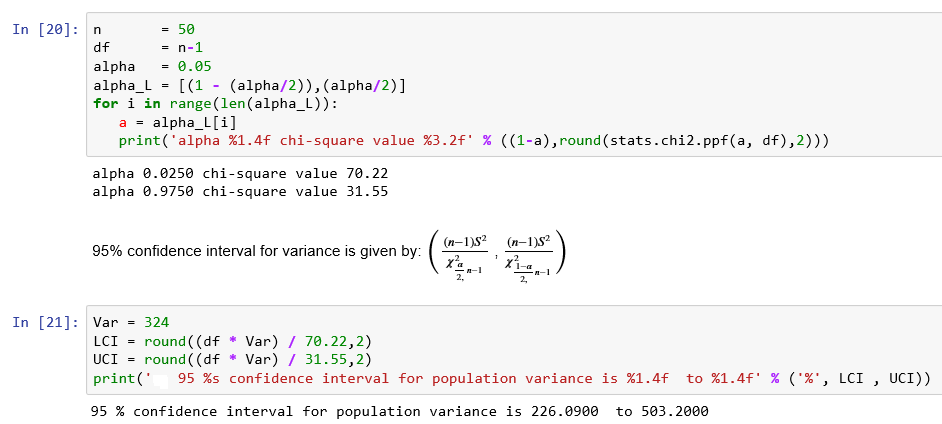


**Example 26**

**The variance of volume of 20 liter water can is estimated to be 324 ml based on a sample of 50 water cans.**

**Calculate a 95% confidence interval for the variance in water cans.**

**Solution:**

****

## **Setting up hypothesis**

* *Beware of the problem of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confessions obtained under duress may not be admissible in the court of scientific opinion - Stephen M Stigler*
* *Hypothesis is a claim made by a person / organization.*
* *The claim is usually about the population parameters such as mean or proportion and we seek evidence from a sample for the support of the claim (Example: average salary of Data Scientist with 1 year experience is Rs 5 Lakhs per annum).*
* *Hypothesis testing is a process used for either rejecting or retaining null hypothesis.*

### Examples of some claims

* *If you drink Horlicks, you can grow taller, stronger and sharper.*
* *Two - minute for cooking noodles. (or eating !!)*
* *Married people are happier than singles (Anon - 2015).*
* *Smokers are better sales persons.*
* *Hypothesis testing is used for checking the validity of the claim using evidence found in sample data*

### Type I Error, Type II error and power of the hypothesis test

**Type I error**

* *It is the conditional probability of rejecting a null hypothesis when it is true, is called* ***Type I error or False positive.***
* *α, the level of significance is the value of Type I error.*
* *P(Reject null hypothesis | H0 is true) = α*

**Type II error**

* *It is the conditional probability of retaining a null hypothesis when it is true, is called Type II error or False Negative.*
* *β, is the value of Type II error.*
* *P(Retain null hypothesis | H0 is false) = β*

**Power of the test**

* *(1 -* β*) is known as the* ***power of the test****.*
* *It is P(Reject null hypothesis |* H*0 is false) = 1-* β

### Steps involved in solving the hypothesis testing

**Step 1: Define null and alternative hypotheses**

* *Null hypothesis means no relationship or status quo*
* *Alternative hypothesis is what the researcher wants to prove*

**Example**:

Write the null and alternative hypothesis from the following hypothesis description:

a. Average annual salary of Data Scientists is different for those having Ph.D in Statistics and those who do not.

Let μPhD be the average annual salary of a Data scientist with Ph.D in Statistics.

Let μNoPhD be the average annual salary of a Data scientist without Ph.D in Statistics.

Null hypothesis: H0: μPhD = μNoPhD

Alternative hypothesis: HA: μPhD ≠ μNoPhD

Since the rejection region is on either side of the distribution, it will be a **two-tailed test**.

b. Average annual salary of Data Scientists is more for those having Ph.D in Statistics and those who do not.

Null hypothesis: H0: μPhD ≤ μNoPhD

Alternative hypothesis: HA: μPhD > μNoPhD

Since the rejection region is on the right side of the distribution, it will be a **one-tailed test**.

**Step 2: Decide the significance level**

* *You control the Type I error by determining the risk level, α, the level of significance that you are willing to reject the null hypothesis when it is true.*
* *Traditionally, you select a level of 0.01, 0.05 or 0.10.*
* *The choice of selection for making Type I error depends on the cost of making a Type I error.*
* *One way to reduce the probability of making a Type II error is by increasing the sample size.*
* *For a given level of α, increasing the sample size decreases β resulting in increasing the power of the statistical test to detect that null hypothesis is false.*

**Step 3: Identify the test statistic**

The test statistic will depend on the probability distribution of the sampling distribution

**Step 4: Calculate the p-value or critical values**

*P-value is the conditional probability of observing the test statistic value or extreme than the sample result when the null hypothesis is true.*

**Critical value approach**

*Critical values for the appropriate test statistic are selected so that the rejection region contains a total area of α when H0 is true and the non-rejection region contains a total area of 1 - α when H0 is true.*

**Step 5: Decide to reject or accept null hypothesis**

*Reject null hypothesis when test statisic lies in the rejection region; retain null hypothesis otherwise*

*or*

*Reject null hypothesis when p-value < α; retain null hypothesis otherwise.*

**Example 27**

**A beverages company produces mineral water and available in 250 ml, 500 ml, 1 liter and 2 liter bottles, 5 liter, 15 liter and 20 liter jars. Let us consider 2 liter bottles. Company specification require a mean volume of 2 liter per bottle. You must adjust the water filling process when the mean volume in the population of bottles differs from 2 liters. Adjusting the process requires shutting down the water filling production line completely, so you do not want to make any adjustments without any reason unnecessarily.**

**Assume a sample of 50 water bottles indicate a sample mean,**

** of 2.001 liters and the population standard deviation, σ is 15 ml.**

**Solution:**

**Hypothesis testing using the critical value approach**

**Step 1: Define null and alternative hypotheses**

*In testing whether the mean volume is 2 liters, the null hypothesis states that mean volume, μ equals 2 liters. The alternative hypothesis states that the mean volume, μ is not equal to 2 liters.*

*H*0: *μ* = 2

*HA*: *μ* ≠ 2

**Step 2: Decide the significance level**

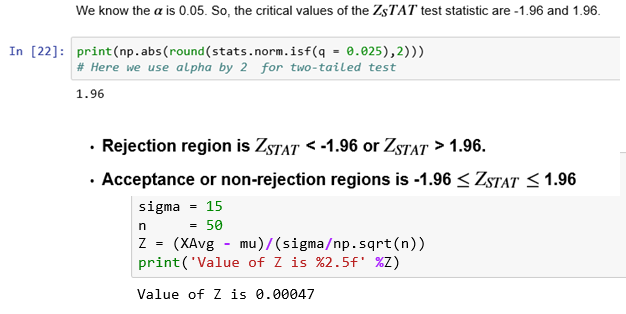
* *Choose the* α*, the level of significance according to the relative importance of the risks of committing Type I and Type II errors in the problem.*
* *In this example, making a Type I error means that you conclude that the population mean is not 2 liters when it is 2 liters. This implies that you will take corrective action on the filling process even though the process is working well (false alarm).*
* *On the other hand, when the population mean is 1.98 liters and you conclude that the population mean is 2 liters, you commit a Type II error. Here, you allow the process to continue without adjustment, even though an adjustment is needed (missed opportunity).*
* *Here, we select* α *= 0.05 and n, sample size = 50.*

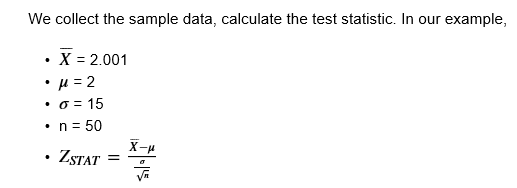
**Step 3: Identify the test statistic**

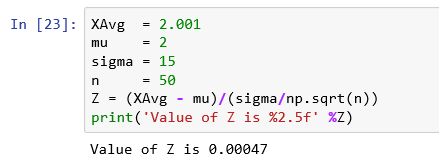
*We know the population standard deviation and the sample is a large sample, n>30. So, you use the normal distribution and the* ZSTAT *test statistic.*

**Step 4: Calculate the critical value**

*We know the* α *is 0.05. So, the critical values of the* ZSTAT *test statistic are -1.96 and 1.96.*



****

****

**Step 5 Decide to reject or accept null hypothesis**

Here, Z = 0.00047 lies in the acceptance region since -1.96 < Z = 0.00047 < 1.96.

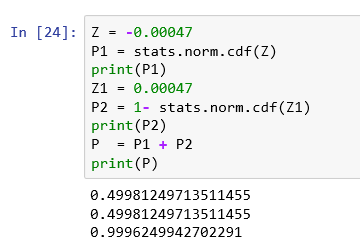
**Our statistical decision is not to reject the null hypothesis. Our conclusion is So there is no enough evidence to prove that the mean fill is different from 2 liters.**

**Hypothesis testing using p-value approach**

*Use the example 15. To use the p-value approach, find the probability that the test statistic ZSTAT is equal to or more extreme than 0.00047 standard error units from the center of the standard normal distribution.*

**Step 4: Calculate the p-value**

* *We need to compute P(Z < -0.00047) and P(Z > 0.00047).*
* *P value for the two-tail test is P(Z < -0.00047) + P(Z > 0.00047).*



**Step 5: Decide to reject or accept null hypothesis**

*Since the P value for this two tail test is 0.9996 and it is greater than 0.05, our level of significance, we do not reject the null hypothesis and conclude that there is no sufficient evidence to prove that the mean fill is different from 2 litres.*

# **Sampling distributions such as t, Z**



## **One sample and two sample tests**

### **One sample test**

*In one sample test, we compare the population parameter such as mean of a single sample of data collected from a single population.*

#### **Z test**

*A one sample Z test is one of the most basic types of hypothesis test.*

**Example 28**

**A principal of a prestigious city college claims that the average intelligence of the students of the college is above average.**

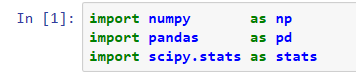
**A random sample of 100 students IQ scores have a mean score of 115. The mean population mean IQ is 100 with a standard deviation of 15.**

**Is there enough evidence to support the principal's claim?**

**Solution:**

**Let us work through the several required steps.**

**Load all the required modules.**

****

**Step 1: Define null and alternative hypotheses**

In testing whether the mean IQ of the students is more than 100, the null hypothesis states that mean IQ, μ equals 100. The alternative hypothesis states that the mean IQ, μ is greater than 100.

H0: μ = 100

HA: μ > 100

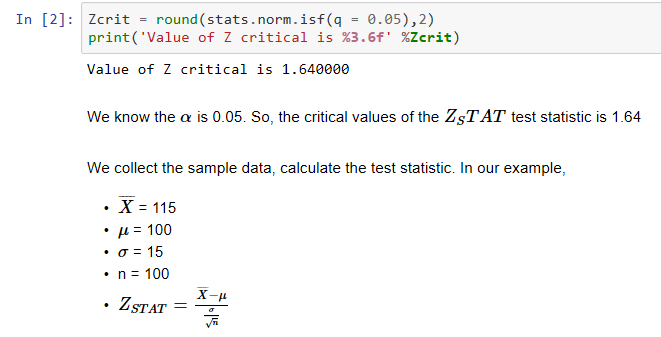
**Step 2: Decide the significance level**

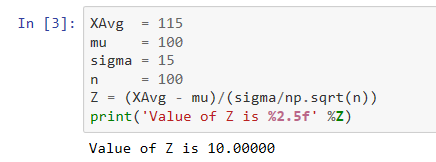
Here we select α = 0.05 and it is given that n, sample size = 100.

**Step 3: Identify the test statistic**

We know the population standard deviation and the sample is a large sample, n>30. So you use the normal distribution and the ZSTAT test statistic.

**Step 4: Calculate the critical value and test statistic**





**Step 5 Decide to reject or accept null hypothesis**

*In this example, Z = 10 lies in the rejection region because, Z = 0.00047 > 1.64.*

*So, the statistical decision is to reject the null hypothesis.*

***We conclude that the mean average intelligence of the students of the college is above average based on enough evidence.***

#### **t test**

***Very rarely we know the variance of the population.***

***A common strategy to assess hypothesis is to conduct a t test. A t test can tell whether two groups have the same mean.***

***A t test can be estimated for:***

1. ***One sample t test***
2. ***Two sample t test (including paired t test)***

***We assume that the samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances.***

***One sample t test***

**Example 29**

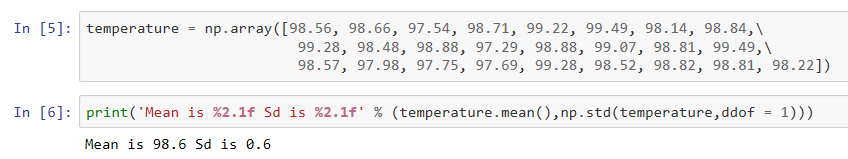
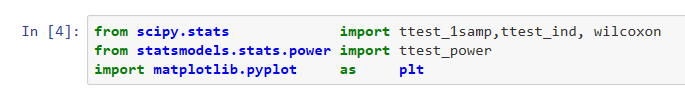
**Suppose that a doctor claims that 17 year old patients have an average body temperature that is higher than the commonly accepted average human temperature of 98.6 degree F.**

**A simple random statistical sample of 25 people, each of age 17 is selected.**

| **ID** | **Temperature** |
| --- | --- |
| 1 | 98.56 |
| 2 | 98.66 |
| 3 | 97.54 |
| 4 | 98.71 |
| 5 | 99.22 |
| 6 | 99.49 |
| 7 | 98.14 |
| 8 | 98.84 |
| 9 | 99.28 |
| 10 | 98.48 |
| 11 | 98.88 |
| 12 | 97.29 |
| 13 | 98.88 |
| 14 | 99.07 |
| 15 | 98.81 |
| 16 | 99.49 |
| 17 | 98.57 |
| 18 | 97.98 |
| 19 | 97.75 |
| 20 | 97.69 |
| 21 | 99.28 |
| 22 | 98.52 |
| 23 | 98.82 |
| 24 | 98.81 |
| 25 | 98.22 |

**Solution**

1. **Load the required modules.**

****

**Step 1: Define null and alternative hypotheses**

**In testing whether 17 year old patients have an average body temperature that is higher than 98.6 deg F,**

**The null hypothesis states that mean body temperature, μ equals 98.6.**

**The alternative hypothesis states that the mean body temperature, μ is greater than 98.6.**

*H*0: *μ* = 98.6

*HA*: *μ* > 98.6

**Step 2: Decide the significance level**

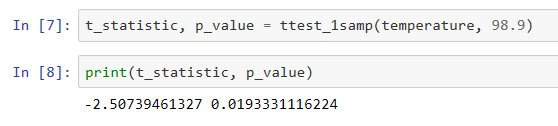
Here we select α = 0.05 and it is given that n, sample size = 25.

**Step 3: Identify the test statistic**

We do not know the population standard deviation and the sample is not a large sample, n < 30. So, you use the t distribution and the tSTAT test statistic.

**Step 4: Calculate the p - value and test statistic**

scipy.stats.ttest\_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and two-tailed p value.



**Step 5 Decide to reject or accept null hypothesis**

For a two-tailed t test, p value is 0.01933. For one-tailed t test, it is 0.009665

In this example, p value is 0.009665 and it is less than 5% level of significance. So, the statistical decision is to reject the null hypothesis at 5% level of significance.

**So, there is sufficient evidence to prove that 17 year olds have an average body temperature that is higher than the commonly accepted average human temperature of 98.6 degree F.**

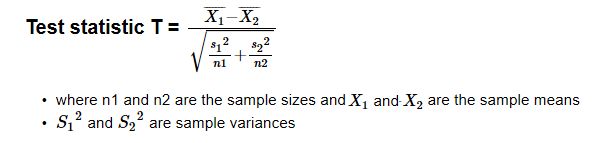
***Two sample test***

* Two sample t test (Snedecor and Cochran 1989) is used to determine if two population means are equal.
* A common application is to test if a new treatment or approach or process is yielding better results than the current treatment or approach or process.
* Data is paired - For example, a group of students are given coaching classes and effect of coaching on the marks scored is determined.
* Data is not paired - For example, find out whether the miles per gallon of cars of Japanese make is superior to cars of Indian make.

Two sample t test for unpaired data is defined as

H0: μ1 = μ2

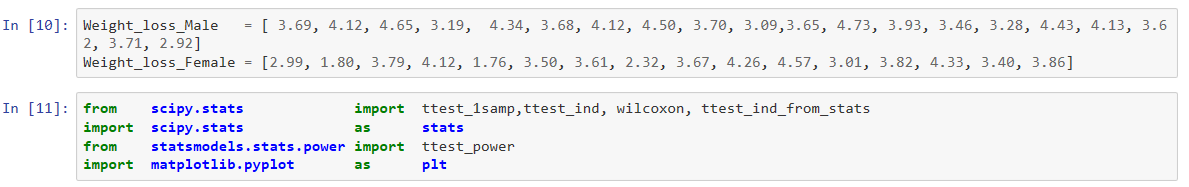
Ha: μ1≠ μ2



**Example 30**

**Compare two unrelated samples. Data was collected on the weight loss of 16 women and 20 men enrolled in a weight reduction program. At αα = 0.05, test whether the weight loss of these two samples is different.**

**Solution:**



**Step 1: Define null and alternative hypotheses**

In testing whether weight reduction of female and male are same,

* The null hypothesis states that mean weight reduction, μM equals μF
* The alternative hypothesis states that the weight reduction is different for Male and Female, μM ≠ μF

H0: μM - μF = 0

HA: μM - μF ≠ 0

**Step 2: Decide the significance level**

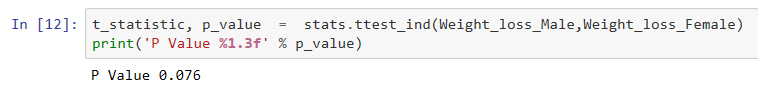
Here we select αα = 0.05 and sample size < 30 and population standard deviation is not known.

**Step 3: Identify the test statistic**

* We have two samples and we do not know the population standard deviation.
* Sample sizes for both samples are not same.
* The sample is not a large sample, n < 30. So, you use the t distribution and the tSTAT test statistic for two sample unpaired test.

**Step 4: Calculate the p - value and test statistic**

* We use the scipy.stats.ttest\_ind to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.
* This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.



**Step 5: Decide to reject or accept null hypothesis**

* In this example, p value is 0.076 and it is more than 5% level of significance
* So, the statistical decision is to accept the null hypothesis at 5% level of significance.

**So, there is no enough evidence to reject the null hypothesis that the weight loss of these men and women is same.**

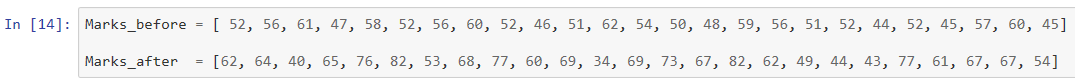
***Two sample t test for paired data***

**Example 31**

**Compare two related samples. Data was collected on the marks scored by 25 students in their final practice exam and the marks scored by the students after attending special coaching classes conducted by their college.**

**At 5% level of significance, is there any evidence that the coaching classes has any effect on the marks scored?**

**Solution:**

****

**Step 1: Define null and alternative hypotheses**

* We are testing whether coaching has any effect on marks scored.
* The null hypothesis states that difference in marks, μAfter equals μBefore
* The alternative hypthesis states that difference in marks is more than 0, μAfter ≠ μBefore

H0: μAfter - μBefore = 0

HA: μAfter - μBefore ≠ 0

**Step 2: Decide the significance level**

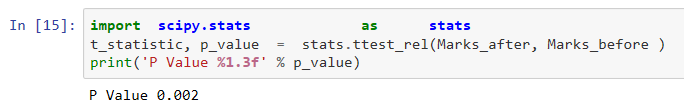
Here we select α = 0.05 and sample size < 30 and population standard deviation is not known.

**Step 3: Identify the test statistic**

* Sample sizes for both samples are same.
* We have two paired samples and we do not know the population standard deviation.
* The sample is not a large sample, n < 30. So, you use the t distribution and the tSTAT test statistic for two sample paired test.

**Step 4: Calculate the p - value and test statistic**

* We use the scipy.stats.ttest\_rel to calculate the T-test on TWO RELATED samples of scores.
* This is a two-sided test for the null hypothesis that 2 related or repeated samples have identical average (expected) values.
* Here we give the two sample observations as input.
* This function returns t statistic and two-tailed p value.



**Step 5: Decide to reject or accept null hypothesis**

* In this example, p value is 0.002 and it is less than 5% level of significance
* So, the statistical decision is to reject the null hypothesis at 5% level of significance.
* So, there is sufficient evidence to reject the null hypothesis that there is an effect of coaching classes on marks scored by students.

# **ANOVA**

* ANOVA is a hypothesis testing technique tests the equality of two or more population means by examining the variances of samples that are taken.
* ANOVA tests the general rather than specific differences among means.



## **Assumptions of ANOVA**

1. **All populations involved follow a normal distribution**
2. **All populations have the same variance**
3. **The samples are randomly selected and independent of one another**

## **One-way ANOVA**

**Example 32**

**Consider the monthly income of members from three different gyms - fitness centers given below:**

Gym 1 (n = 22): [60, 66, 65, 55, 62, 70, 51, 72, 58, 61, 71, 41, 70, 57, 55, 63, 64, 76, 74, 54, 58, 73]

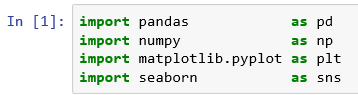
Gym 2 (n = 18): [56, 65, 65, 63, 57, 47, 72, 56, 52, 75, 66, 62, 68, 75, 60, 73, 63, 64]

Gym 3 (n = 23): [67, 56, 65, 61, 63, 59, 42, 53, 63, 65, 60, 57, 62, 70, 73, 63, 55, 52, 58, 68, 70, 72, 45]

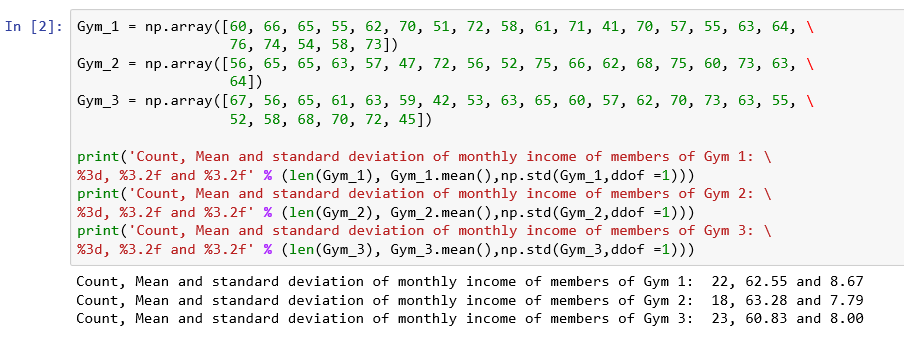
**Using ANOVA, test whether the mean monthly income is equal for each Gym.**

**Solution:**

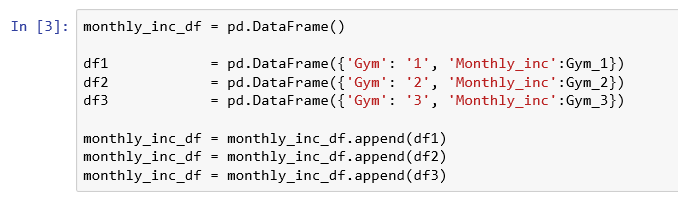
**Load the required packages.**

****

**Find the mean and standard deviation for each group.**

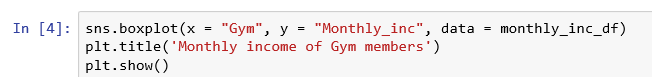


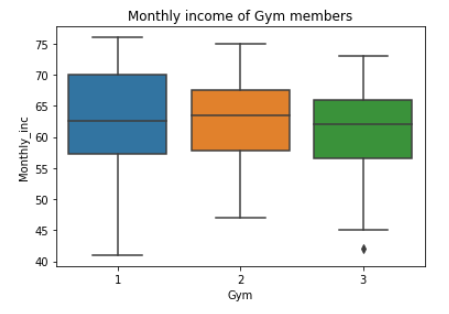
**Create a data frame with these groups to make the analysis easier.**



**Let us explore the data graphically.**

**A side by side boxplot is one of the best ways to compare group locations, spreads and shapes.**

****

****

**The boxplots show almost similar shapes, location and spreads and group 3 has a low outlier.**

**Step 1: State the null and alternative hypothesis:**

*H0: μ1 = μ2 = μ3*

*HA: At least one μ differs*

**Step 2: Decide the significance level**

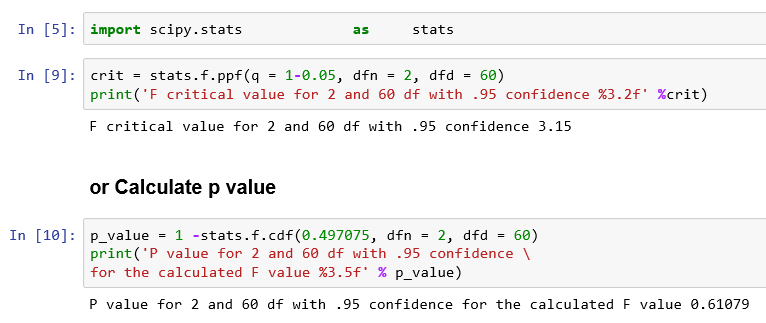
*Here we select α = 0.05*

**Step 3: Identify the test statistic**

*Here we have three groups. Analysis of variance can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.*

**Step 4: Calculate F, a test statistic**

* *scipy.stats.f.ppf gives the critical value at a given level of confidence with a pair of degrees of freedom.*
* *scipy.stats.f.cdf gives the cumulative distribution function for the given random variable - given the calculated F value at a given level of confidence with a pair of degrees of freedom.*

****

**Or formulate an ANOVA table using statsmodels**

* *statsmodels.formula.api.ols creates a model from a formula and dataframe*
* *statsmodels.api.sm.stats.anova\_lm gives an Anova table for one or more fitted linear models*

**In the formula. we know that**

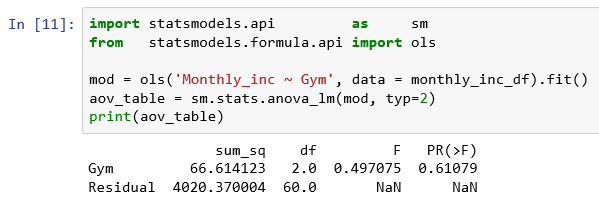
1) ~ separates the left hand side of the model from the right hand side

2) + adds new columns to the design matrix

3) : adds a new column to the design matrix with the product of the other two columns

4) \* also adds the individual columns multiplied together along with their product

5) c() operator denotes that the variable enclosed in c() will be treated explicitly as categorical variable.

****

**Step 5: Decide to reject or accept null hypothesis**

* *In this example, calculated value of F ( = 0.497075) is more than Critical value of F( = 3.15)*
* *So, the statistical decision is to fail to reject the null hypothesis at 5% level of significance.*
* *So, there is no sufficient evidence to reject the null hypothesis that at least one mean monthly income of a gym is different from others.*

### **Tukey’s Honest Significant Difference test**

* *An ANOVA test can tell you if your group mean differences are significant overall, but it won’t tell you exactly where those differences lie.*
* *After you have run an ANOVA and found significant results, then you can run Tukey’s HSD to find out which specific groups’s means (compared with each other) are different.*
* *The test compares all possible pairs of group means.*

[**https://www.statisticshowto.datasciencecentral.com/tukey-test-honest-significant-difference/**](https://www.statisticshowto.datasciencecentral.com/tukey-test-honest-significant-difference/)

**Example 33**

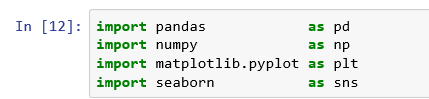
* National Transportation Safety Board (NTSB) examines the safety of compact cars, midsize cars and full-size cars.
* Using the fake data given below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car at 5% level of significance.

meanpressure compact\_car = [643, 655,702]

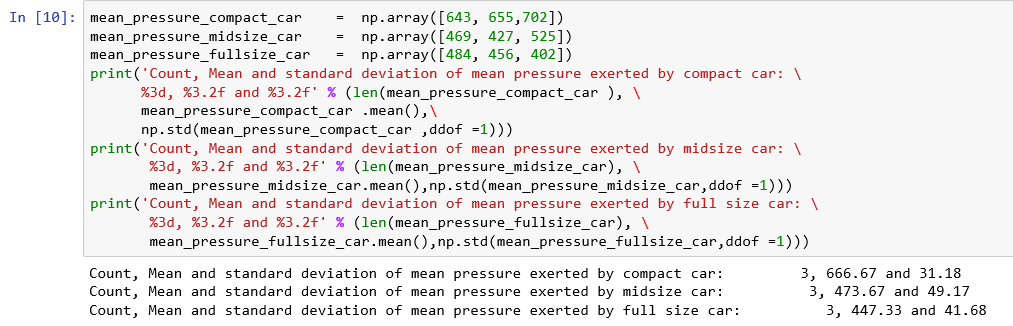
meanpressure midsize\_car = [469, 427, 525]

meanpressure fullsize\_car = [484, 456, 402]

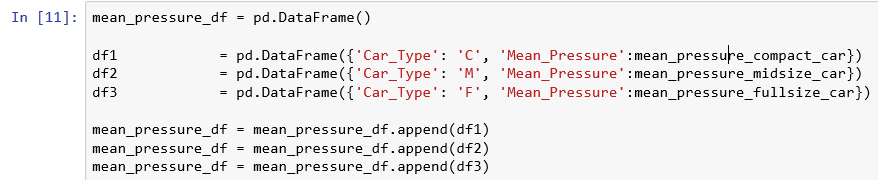
**Load the required packages**

****

**Find the mean and standard deviation for each group**

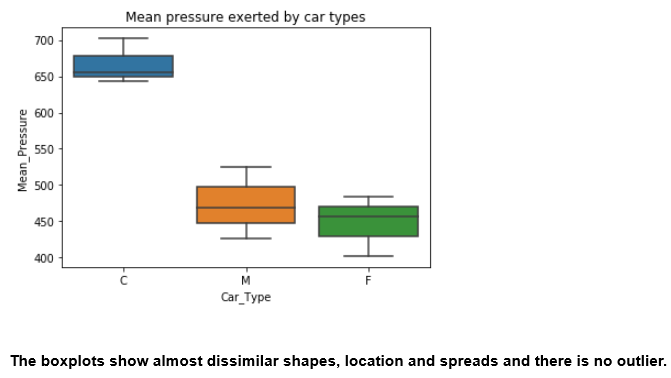
****

**Create a data frame with these groups to make the analysis easier**

****

**Let us explore the data graphically**

**A side by side boxplot is one of the best way to compare group locations, spreads and shapes.**

****

**Step 1: State the null and alternative hypothesis:**

H0: μ1 = μ2 = μ3

HA: At least one μ differs

Here μ1, μ2 and μ3 are the mean pressure applied to the driver's head during crash test by Compact car, Midsize car and Full size car respectively.

**Step 2: Decide the significance level**

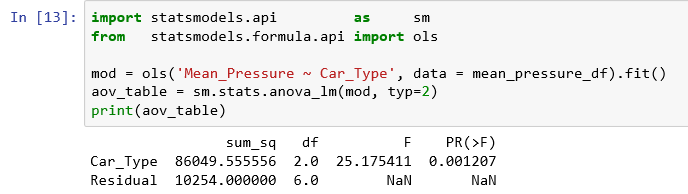
Here we select α = 0.05

**Step 3: Identify the test statistic**

Here we have three groups. Analysis of variance can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.

**Step 4: Calculate p value using ANOVA table**

* *statsmodels.formula.api.ols creates a model from a formula and dataframe*
* *statsmodels.api.sm.stats.anova\_lm gives an Anova table for one or more fitted linear models*

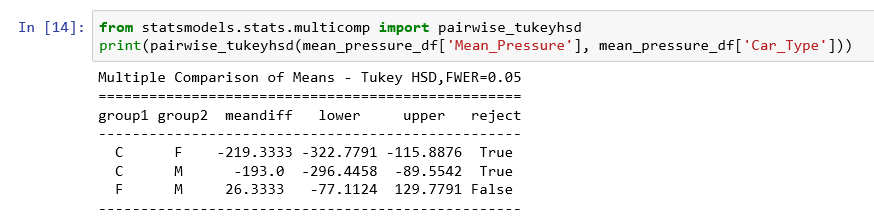
****

**Step 5: Decide to reject or accept null hypothesis**

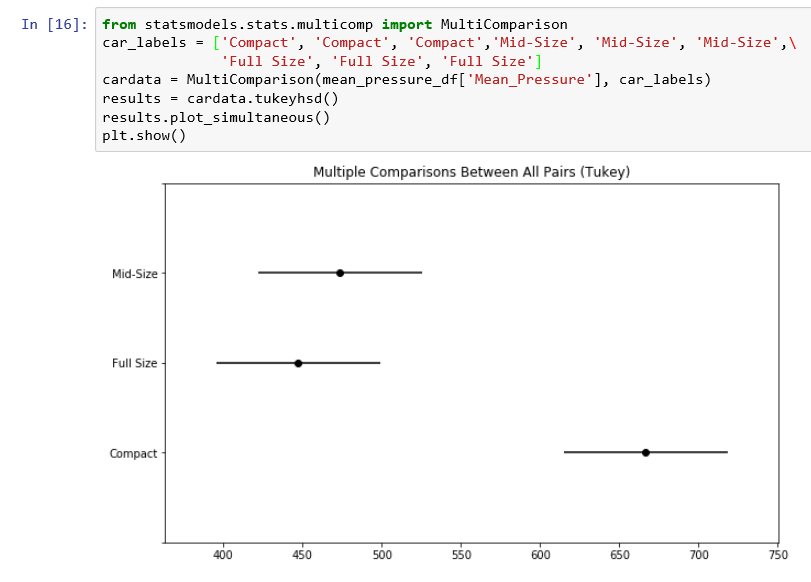
* *In this example, p value is 0.001207 and it is less than our chosen level of significance at 5%*
* *So, the statistical decision is to reject the null hypothesis at 5% level of significance.*
* *So, there is sufficient evidence to reject the null hypothesis that at least one mean pressure of car is different from others .*

**Determine which mean(s) is / are different**

* *An ANOVA test will test that at least one mean is different.*
* *You have rejected the null hypothesis but do not know which mean(s) is / are different.*
* *We use Tukey-krammer HSD test to detect which mean(s) is / are different.*
* *pairwise\_tukeyhsd() method in statsmodels.stats.multicomp calculates all pairwise comparisons with TuLeyHSD confidence intervals.*
* *It can also compute and plot additional post-hoc evaluations using the results class.*



**Plot additional post-hoc evaluations using the results**

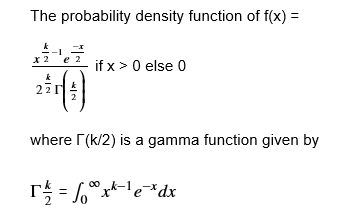
****

**Inference**

* *Compact Car Vs Full-size cars: Since the result - reject is true, mean head pressure is statistically different between Compact and Full-size cars.*
* *Compact Car Vs Mid-size cars: Since the result - reject is true, mean head pressure is statistically different between Compact and Mid-size cars.*
* *Full-size Car Vs Mid-size cars: Since the result - reject is false, mean head pressure is statistically equal to Full-size and Mid-size cars.*

# **Chi square**

* *A chi-square distribution with k degrees of freedom is given by sum of squares of standard normal random variables Z1, Z2, ... Zk obtained by transforming normal standard variables*
* *X1, X2, ... Xk with mean values μ1, μ2, ... μk and corresponding standard deviation σ1, σ2, ... σk*
* *χk2 = Z12 + Z22 + … + Zk2*





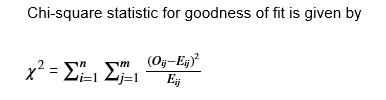
## **Properties of Chi Square distribution**

* *The mean and standard deviation of a chi-square distribution are k and √2k respectively, where k is the degrees of freedom.*
* *As the degrees of freedom increases, the probability density function of a chi-square distribution approaches normal distribution.*
* *Chi-square goodness of fit is one of the popular tests for checking whether a data follows a specific probability distribution.*
* *Chi square test is a right tailed test.*

## **Chi-square Goodness of fit tests**

* *Goodness of fit tests are hypothesis tests that are used for comparing the observed distribution of data with expected distribution of the data to decide whether there is any statistically significant difference between the observed distribution and a theoretical distribution (for example, normal, exponential, etc.).*
* *This is based on the comparison of observed frequencies in the data and the expected frequencies if the data follows a specified theoretical distribution.*

|  |  |
| --- | --- |
| **Hypothesis** | **Description** |
| Null hypothesis | There is no statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution. |
| Alternative hypothesis | There is statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution. |

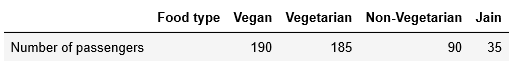


* *This test is invalid when the observed or expected frequencies in each category are too small. A typical rule is that all of the observed and expected frequencies should be at least 5.*

**Example 34**

**A1 airlines operated daily flights to several Indian cities. The operations manager believes that 30% of their passengers prefer vegan food, 45% prefer vegetarian food , 20% prefer non-veg food 5% request for Jain food.**

**A sample of 500 passengers was chosen to analyze the food preferences and the data is shown in the following table:**

****

**At 5% level of significance, can you confirm that the meal preference is as per the belief of the operations manager?**

**Solution:**

**Step 1: State the null and alternative hypothesis:**

* *Null hypothesis: H0: Meal preference is as per the perceived ratios of the operations manager*
* *Alternative hypothesis: HA: Meal preference is different from the perceived ratios of the operations manager*

**Step 2: Decide the significance level**

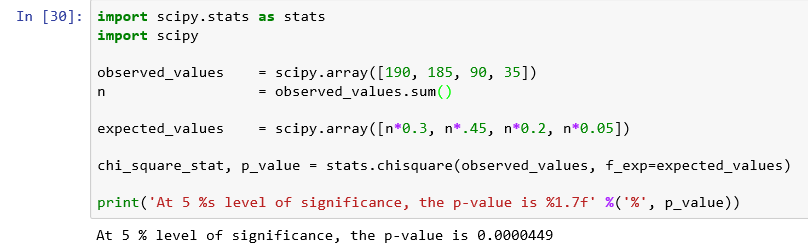
*Here we select α = 0.05*

**Step 3: Identify the test statistic**

*Since we have observed frequencies of meal preference and we can calculate the expected frequencies, we can use chi-square goodness of fit for this problem.*

**Step 4: Calculate p value or chi-square statistic value**

* *Use the scipy.stats.chisquare function to compute Chi square goodness of fit by giving the observed values and expected values as input.*
* *The first value in the returned tuple is the χ2 value itself, while the second value is the p-value computed using ν = k−1 where k is the number of values in each array.*
* *We can calculate the expected frequency as follows:*
* *Compute the total number of passengers. It will be 500.*
* *We expect 30% of them prefer Vegan food, so the expected frequency for Vegan Food is = 0.3 \* 500 = 150*
* *Similarly, we can calculate the expected frequencies of the rest of them.*

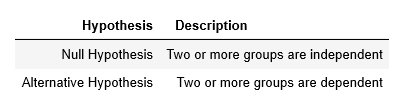
****

**Step 5: Decide to reject or accept null hypothesis**

* *In this example, p value is 0.0000449 and < 0.05 so we reject the null hypothesis.*
* *So, we conclude that Meal preference is not defined in the null hypothesis.*

## **Chi-square tests of independence**

*Chi-square test of independence is a hypothesis test in which we test whether two or more groups are statistically independent or not.*

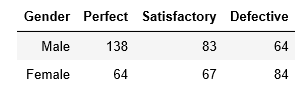




* *The corresponding degrees of freedom is (r - 1) \* ( c - 1) , where r is the number of rows and c is the number of columns in the contingency table.*
* *scipy.stats.chi2\_contingency is the Chi-square test of independence of variables in a contingency table.*
* *This function computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table observed. The expected frequencies are computed based on the marginal sums under the assumption of independence.*

**Example 35**

**The table below contains the number of perfect, satisfactory and defective products are manufactured by both male and female:**

****

**Do these data provide sufficient evidence at the 5% significance level to infer that there are differences in quality among genders (Male and Female)?**

**Solution:**

**Step 1: State the null and alternative hypothesis:**

* Null hypothesis: H0: There is no difference in quality of the products manufactured by male and female
* Alternative hypothesis: HA: There is a significant difference in quality of the products manufactured by male and female

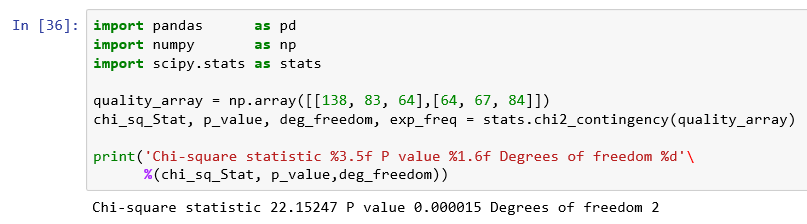
**Step 2: Decide the significance level**

Here we select α = 0.05

**Step 3: Identify the test statistic**

We use the chi-square test of independence to find out the difference of categorical variables

**Step 4: Calculate p value or chi-square statistic value**



**Step 5: Decide to reject or accept null hypothesis**

* In this example, p value is 0.000015 and < 0.05 so we reject the null hypothesis.
* So, we conclude that there is a significant difference in quality of the products manufactured by male and female.



# **Exercises**

1. The mean weight of a morning breakfast cereal pack is 0.295 kg with a standard deviation of 0.025 kg. The random variable weight of the pack follows a normal distribution.
   1. What is the probability that the pack weighs less than 0.280 kg?
   2. What is the probability that the pack weighs more than 0.350 kg?
   3. What is the probability that the pack weighs between 0.260 kg to 0.340 kg?
2. If X is a random variable that follows normal distribution, then 5X + 40 is a
   1. Chi-square distribution
   2. Poisson distribution
   3. Binomial distribution
   4. Normal distribution
3. The number of warranty claims received at an Electronic equipment manufacturer follows a Poisson distribution with a rate of 10 per day.

Calculate the following:

* 1. The probability that the warranty claims is below 5 on a given day.
  2. The probability that the warranty claims will exceed 10 on a given day.

1. Hours spent on studies by 20 students on a course is given in the following table:

**4.7 9.2 9.3 11.2 8**

**7.6 7.4 4.9 9.2 5.3**

**1.7 2.8 7.2 12.3 8.6**

**10.6 9 5.7 6.9 3.8**

Assume that the population of hours spent follows a normal distribution and the standard deviation is 3.1 hours, calculate the 90% confidence interval for the mean hours spent by the students.

1. Assume the food label on a Lays chips bag states that there is at most 182 mg of sodium in a single chip. Assume the actual mean amount of sodium per chip is 182.09 mg and the population standard deviation is 0.20 mg.

At 5% significance level, what is the probability of type II error for a sample of 40 chips?

1. The delivery time of Pizza from an online food delivery service firm and the home delivery from a local restaurant are given below: At 5% level of significance, is the mean delivery time for online delivery food service firm is less than the mean delivery time for the home delivery from a local restaurant.

Pizza\_delivery\_online = [16.8, 11.7, 15.6, 16.7, 17.5, 18.1, 14.1, 21.8, 13.9, 20.8]

Pizza\_delivery\_local = [22.0, 15.2, 18.7, 15.6, 20.8, 19.5, 17.0, 19.5, 16.5, 24.0]

Hint: Use paired t test

1. A company makes three types of electronic device. Lifetime in hours for each type of electronic device is given below:

life\_type\_A = [ 407, 411, 409 ]

life\_type\_B = [ 404, 406, 408, 405, 402 ]

ife\_type\_C = [ 410, 408, 406, 408]

Is there any significant difference in the average lifetime of the electronic devices?

1. A Cable service provider company is interested in checking whether the customer churn depends on customer segment. Use 5% as level of significance and check.

| **Customer Segment** | **Churned** | **Retained** |
| --- | --- | --- |
| S1 | 15 | 142 |
| S2 | 24 | 400 |
| S3 | 30 | 389 |