|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | discrete |
| Results of rolling a dice | discrete |
| Weight of a person | continuous |
| Weight of Gold | continuous |
| Distance between two places | continuous |
| Length of a leaf | continuous |
| Dog's weight | continuous |
| Blue Color | discrete |
| Number of kids | discrete |
| Number of tickets in Indian railways | discrete |
| Number of times married | discrete |
| Gender (Male or Female) | discrete |

**Q1)** Identify the Data type for the Following:

**Q2)** Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socio economic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Ordinal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Ordinal |
| Barometer Pressure | Interval |
| SAT Scores | Ratio |
| Years of Education | Ratio |

**Q3)** Three Coins are tossed, find the probability that two heads and one tail are obtained?

**ANS -**Possible outcomes for the three coins tossed : (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT )

Probability that two heads and one tail is : (HHT, HTH, THH)

**Hence P = 3/8**

**Q4)** Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2and 3

**ANS -**S = { (1,1),(1,2),(1,3),(1,4),(1,5),(1,6)  
 (2,1),(2,2),(2,3),(2,4),(2,5),(2,6)  
 (3,1),(3,2),(3,3),(3,4),(3,5),(3,6)  
 (4,1),(4,2),(4,3),(4,4),(4,5),(4,6)  
 (5,1),(5,2),(5,3),(5,4),(5,5),(5,6)  
 (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) }

Possible total number of outcomes = 6 x 6 = 36

1. **Equal to 1** : When the sum is equal to '1' the Probability will be Zero.

P(E) = n(E) / n(S) = 0 / 36 **= 0**

(b) **Less than or equal to 4**

P(AorB)=P(A)+P(B)

Three possible outcomes give a sum less than 4: E = {(1,1),(1,2),(2,1)}, hence. P(A) =3/36

Three possible outcomes give a sum equal to 4: E = {(1,3),(2,2),(3,1)}, hence. P(B)=3/36

P(AorB)=P(A)+P(B) = 3/36+3/36=6/36**=1/6**

(c) **Sum is divisible by 2 and 3**

Total number of events =36

Interested number of events = {(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)} =6

P(Sum is divisible by 2 and 3) =6/36 **=1/6.**

**Q5)**A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

**ANS – DATA ={R1,R2,G1,G2,G3,B1,B2}**

Total number of events = 5C2=(5\*4)/(2\*1)=10

SET= {(R1,R2),(R1,G1),(R1,G2),(R1,G3),(R2,G1),(R2,G2),(R2,G3),(G1,G2),(G1,G3),(G2,G3)}=10

Interested number of events =10

TOTAL SET COUNT=7C2=(7\*6)/(2\*1)=21

**Probability =10/21**

**Q6)** Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children(ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

**ANS –**

The formula for the Expected Value for a random variable is:

= **Xi**- is the number of trials and **P(x)**- is the probability of success.

Expected number of candies for a randomly selected child is given by

∑Xi\*P(x)=1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

=0.015+0.8+1.95+0.025+0.06+0.24

**∑Xi\*P(x)=3.09**

**Q7)** Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**ANS-**

**Median= for points-3.69,for score-3.325,for weigh-17.71**

**Variance= for points0.285,for score-0.957,for weigh-3.193**

**Mode= for points-3.07,for score-3.44,for weigh-17.02**

**Range=Min:Max**

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **count** | 32.000000 | 32.000000 | 32.000000 |
| **mean** | 3.596563 | 3.217250 | 17.848750 |
| **std** | 0.534679 | 0.978457 | 1.786943 |
| **min** | 2.760000 | 1.513000 | 14.500000 |
| **25%** | 3.080000 | 2.581250 | 16.892500 |
| **50%** | 3.695000 | 3.325000 | 17.710000 |
| **75%** | 3.920000 | 3.610000 | 18.900000 |
| **max** | 4.930000 | 5.424000 | 22.900000 |

If the mean is greater than the median, the distribution is positively skewed.

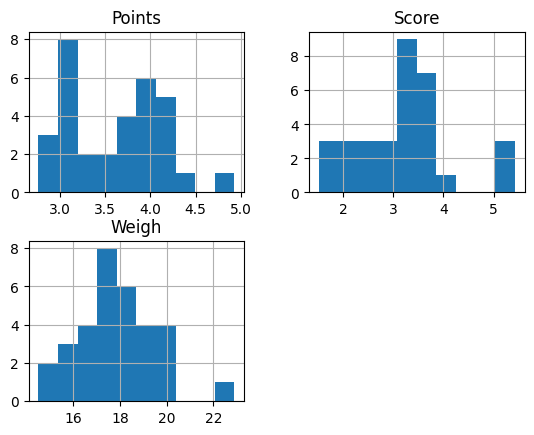
If the mean is less than the median, the distribution is negatively skewed.

**FOR POINTS🡪 Negative Skewed (Left), As Mean<Median.**

**FOR SCORE🡪 Negative Skewed (left), As Mean<Median.**

**FOR WEIGH🡪 Positive Skewed (Right), As Mean>Median.**

**From histo🡪**



**FOR POINTS🡪 Positive Skewed (Right).**

**FOR SCORE🡪 Negative Skewed (left).**

**FOR WEIGH🡪 Negative Skewed (left).**

**Also, there are outlier present in all the dataset(points, score, weigh).**

**Code🡪**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

data = pd.read\_csv("/content/Q7.csv",index\_col=0)

data

print(np.median(data[['Points']]))

print(np.median(data[['Score']]))

print(np.median(data[['Weigh']]))

data.var()

data.mode()

data.describe()

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**ANS-** Random patients chosen =mean(x) = **145.3333**

**Code🡪**

from statistics import mean

 d1 = [108,110,123,134,135,145,145,167,187,199]

 mean(d1)

**Q9)** Calculate Skewness, Kurtosis & draw inferences on the following data

Cars speed and distance

Use Q9\_a.csv

**ANS-**

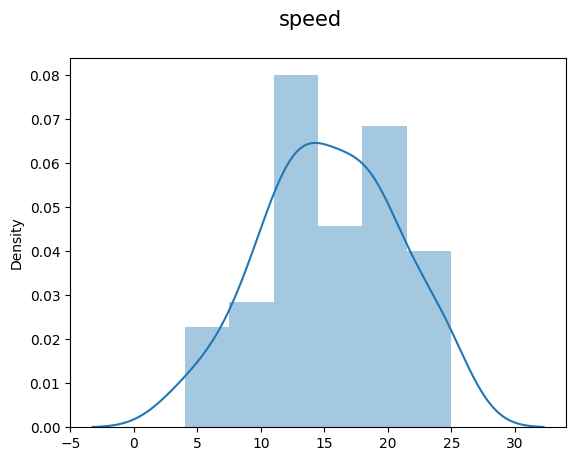
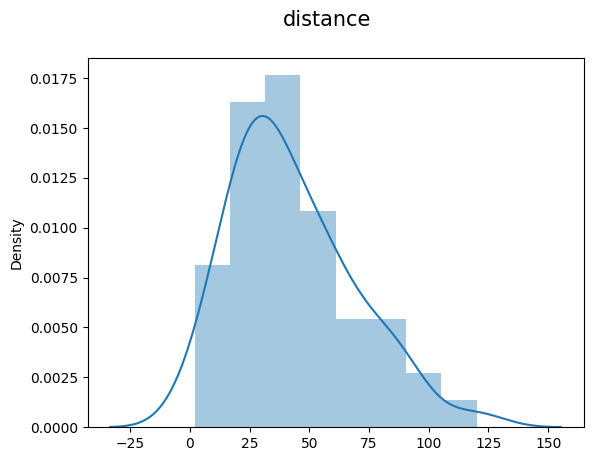
**Speed**-

**Skewness 🡪For Speed= -0.1139548, For Distance=0.7824835**

**Kurtosis 🡪for Speed= -0.57714742, for Distance= 0.24801866**

**Negative kurtosis indicates that the distribution is flatter than the normal distribution.**

**Positive kurtosis indicates that the distribution is more peaked than the normal distribution.**

For Speed🡪The speed distribution is negative skewed or left skewed.

For Distance🡪 The distance distribution is positive skewed or right skewed.

**Code🡪**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import scipy

from scipy.stats import skew

from scipy.stats import kurtosis

data = pd.read\_csv('/content/Q9\_a.csv',index\_col=0)

data

skew(data, axis=0, bias=True)

kurtosis(data, axis=0, bias=True)

import seaborn as sn

plt.suptitle("speed",fontsize=15)

print(sn.distplot(data[['speed']]))

plt.suptitle("distance",fontsize=15)

print(sn.distplot(data[['dist']]))

**SP and Weight(WT) Use Q9\_b.csv**

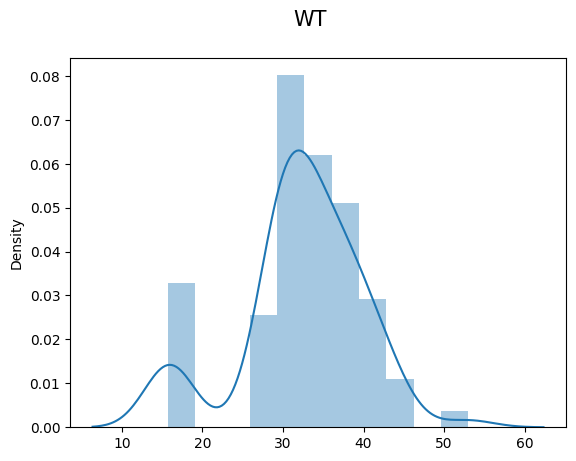
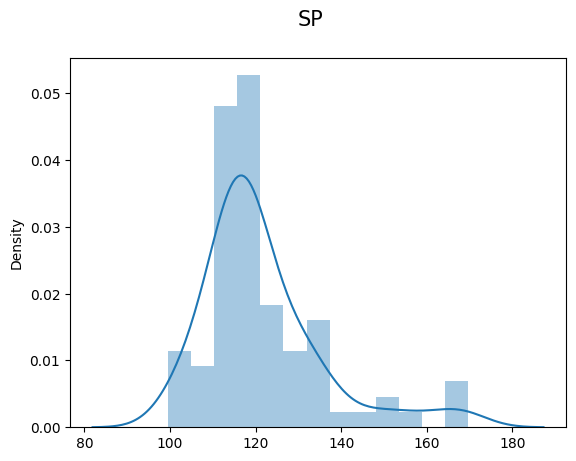
**ANS-**

**Skewness 🡪For SP= 1.5814568, For WT=-0.60330993**

**Kurtosis 🡪for SP= 2.72352149, for WT=0.81946588**

**Negative kurtosis indicates that the distribution is flatter than the normal distribution.**

**Positive kurtosis indicates that the distribution is more peaked than the normal distribution.**



**The SP distribution is positive skewed or right skewed, mean & median of data is on right side.**

**The weight distribution is negative skewed or left skewed, mean & median of data is on left side.**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import scipy

from scipy.stats import skew

from scipy.stats import kurtosis

data1 = pd.read\_csv('/content/Q9\_b.csv',index\_col=0)

data1

skew(data1, axis=0, bias=True)

kurtosis(data1, axis=0, bias=True)

import seaborn as sn

plt.suptitle("SP",fontsize=15)

print(sn.distplot(data1[['SP']]))

plt.suptitle("WT",fontsize=15)

print(sn.distplot(data1[['WT']]))

**Q10)** Draw inferences about the following boxplot & histogram



**ANS -**

Histogram :- Chick weight data is **right skewed or positively skewed**, More than 50% Chick Weight is between 50 to 200, Most of the chick weight is between 50 to 100, mean of data is in right side, because the data is positive skewed.



**ANS -**

Boxplot :- The data is right skewed, there are outliers at upper side.

So that, mean-mode-median is present on the right side of the distribution.

**Q11)**Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

**ANS –**

given data :

n = 2000

sample mean = 200

sample SD = 30

Alpha= (0.94,0. 98,0.96)

1.For Avg. weight of Adult in Mexico with 94%CI

Formula in python 🡪 st.norm.interval(alpha=0.94, mean, scale )

**= (198.738325292158, 201.261674707842)**

**2.** For Avg. weight of Adult in Mexico with 98%CI

Formula in python 🡪 st.norm.interval(alpha=0.98, mean, scale )

=**= (198.43943840429978, 201.56056159570022)**

3. For Avg. weight of Adult in Mexico with 96%CI

Formula in python 🡪 st.norm.interval(alpha=0.96, mean, scale )

= **(198.62230334813333, 201.37769665186667)**

**CODE🡪**

import numpy as np

import pandas as pd

from scipy import stats

stats.norm.interval(0.94,200,30/2000\*\*0.5)

stats.norm.interval(0.98,200,30/2000\*\*0.5)

stats.norm.interval(0.96,200,30/2000\*\*0.5)

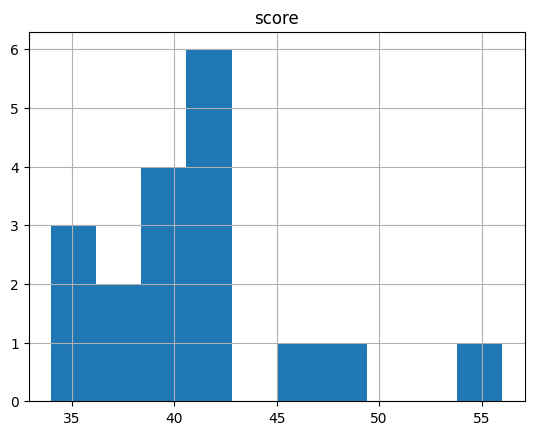
**12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean,median,variance,standard deviation.

**ANS-**

|  | **score** |
| --- | --- |
| **count** | 18.000000 |
| **mean** | 41.000000 |
| **std** | 5.052664 |
| **min** | 34.000000 |
| **25%** | 38.250000 |
| **50%** | 40.500000 |
| **75%** | 41.750000 |
| **max** | 56.000000 |



Median =**40.5**

Variance =**25.52941**

1. What can we say about the student marks?

**ANS –** The data is positive skewed, there are 2 outliers.

Code🡪

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

data=pd.DataFrame ({'score':[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]})

data

data.describe()

data.var()

np.median(data)

data.hist()

**Q13)**What is the nature of skewness when mean, median of data are equal?

**ANS-** If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median = mode.

**Q14)**What is the nature of skewness when mean >median ?

**ANS -** The mean, mode and median can be used to figure out if you have a positively or negatively skewed distribution. If the mean is greater than the median, the distribution is positively skewed.

**Q15)** What is the nature of skewness when median > mean?

**ANS**- If the mean is less than the median, the distribution is negatively skewed.

**Q16)**What does positive kurtosis value indicates for a data ?

**ANS**-Positive values of kurtosis indicate that a distribution is peaked and possess thick tails.

**Q17)** What does negative kurtosis value indicates for a data?

**ANS**- Negative values of kurtosis indicate that a distribution is flat and has thin tails. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution, with fewer values in its shorter (i.e. lighter and thinner) tails.

**Q18)** Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

**ANS –** The data is negative skewed or left skewed& not normal distribution(**Asymmetrical**)

What is nature of skewness of the data?

**ANS –** **left skewed or negative skewed**.

What will be the IQR of the data (approximately)?

**ANS –** Third quartile- First quartile=18-10= **IQR=8**  
  
**Q19)** Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

**ANS -**By observing both the plots whisker’s level is high in boxplot 2, mean and median are equal hence distribution is **symmetrical**.

**Q 20)** Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

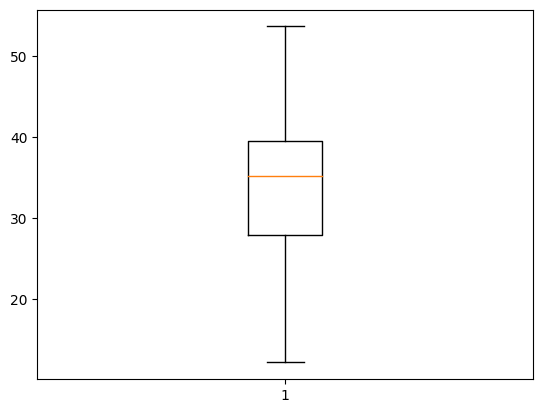
**ANS –**

**Q 21)** Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANS –**



|  | **MPG** |
| --- | --- |
| **count** | 81.000000 |
| **mean** | 34.422076 |
| **std** | 9.131445 |
| **min** | 12.101263 |
| **25%** | 27.856252 |
| **50%** | 35.152727 |
| **75%** | 39.531633 |
| **max** | 53.700681 |

**Mean= 34.42**

**Median = 35.15**

MPG of Cars do follow **normal distribution** approximately (as mean and median are approx. same)

**Code🡪**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

data = pd.read\_csv('/content/Cars.csv',index\_col=0)

data

data[['MPG']].describe()

np.median(data[['MPG']])

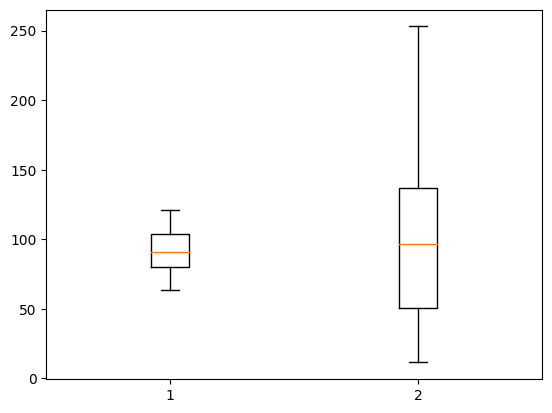
plt.boxplot(data[['MPG']])

z

1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

**ANS –**



**Waist AT**

**count** 109.000000 109.000000

**mean 91.901835 101.894037**

**std** 13.559116 57.294763

**min** 63.500000 11.440000

**25**% 80.000000 50.880000

**50**% 90.800000 96.540000

**75**% 104.000000 137.000000

**max** 121.000000 253.000000

**median 90.8 96.54**

Both the Adipose Tissue (AT) and Waist Circumference(Waist) data set do not follow the normal distribution approximately (as mean and median of both the data are approximately different)

Code🡪

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

data=pd.read\_csv('/content/wc-at.csv')

data

data.describe()

np.median(data[['AT']])

np.median(data[['Waist']])

plt.boxplot(data)

**Q 22)** Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval.

**ANS–**

Z-score of 90% confidence interval , proportion=0.95

stats.norm.ppf(proportion) = **1.6448**

Z-score of 94% confidence interval proportion=0.97

stats.norm.ppf(proportion) = **1.88079**

Z-score of 60% confidence interval proportion=0.80

stats.norm.ppf(proportion) = **0.8416**

**Code🡪**

import numpy as np

import pandas as pd

from scipy import stats

from scipy.stats import norm

stats.norm.ppf(0.95)

stats.norm.ppf(0.97)

stats.norm.ppf(0.80)

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

**ANS –**

t scores of 95% confidence interval for sample size of 25

stats.t.ppf(0.975,df, loc = mean, scale = std\_dev) .…loc &scale taken as default

df = n-1 = 24

stats.t.ppf(0.975,24) = **2.063898**

t scores of 96% confidence interval for sample size of 25

stats.t.ppf(0.98,df, loc = mean, scale = std\_dev) .…loc &scale taken as default

df = n-1 = 24

stats.t.ppf(0.98,24) = **2.171544**

t scores of 99% confidence interval for sample size of 25

stats.t.ppf(0.995,df, loc = mean, scale = std\_dev) .…loc &scale taken as default

df = n-1 = 24

stats.t.ppf(0.995,24) = **2.79693**

**Code**🡪

import numpy as np

import pandas as pd

from scipy import stats

from scipy.stats import norm

stats.t.ppf(0.975,24)

stats.t.ppf(0.98,24)

stats.t.ppf(0.995,24)

**Q 24)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪pt(tscore,df)

df 🡪 degrees of freedom

**ANS –**

given data

Population Mean = 270

Sample Mean = 260

Sample SD = 90

n = 18

Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

**Find t-scores at x=260;**

t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

t = (260-270)/(90/18\*\*0.5)

= **-0.4714045207910317**

**Find P\_value**

The df argument is degrees of freedom, which is usually the sample size minus 1 for a single population sampling problem.

P\_value = stats.t.cdf(-0.4714045207910317, df=17)

= **0.32167411684460556**

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32.17%Assuming significance value α = 0.05 (Standard Value) (If p\_value <α; Reject Ho and accept Ha or vice-versa) Thus, as p-value >α; Accept Ho i.e. The CEO claims are false and the avg life of bulb > 260 days

**Code🡪**

import numpy as np

import pandas as pd

from scipy import stats

from scipy.stats import norm

t=(260-270)/(90/18\*\*0.5)

t

p\_value=1-stats.t.cdf(abs(-0.4714),df=17) #using cdf fuction

p\_value