Q1) Identify the Data type for the Following:

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

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
Socioeconomic 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	Nominal
Time Of Day	Ordinal
Time on a Clock with Hands	Interval
Number of Children	Ordinal
Religious Preference	Nominal
Barometer Pressure	Interval
SAT Scores	Interval
Years of Education	Ratio

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

Ans: Probability is 3/8

Explanation: When three coins are tossed the total number of possible combinations are $2^3 = 8$.

These combinations are HHH, HHT, HTH, THH, TTH, THT, HTT, TTT.

The number of combinations which have two heads and one tail are:

HHT, HTH, TTH which makes them 3 in number.

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

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

Ans: 4a) Probability is 0

Explanation: If two dices were rolled, then total possible cases= 36

Total Favorable cases (Having sum =1) = 0

As minimum sum is 2 for outcome (1,1).

Hence, probability is 0.

Ans: 4b) Probability is 1/6

Explanation: When we roll two dice, the possibility of getting <= 4 is (1, 1), (1, 2), (1,3), (2, 1), (2, 2), (3, 1) so,

The number of favorable outcomes = 6

Total number of possibilities = 36

Probability = The number of favorable outcomes / Total number of possibilities = 6 / 36 = 1/6.

Ans: 4c) Probability is 5/36

Explanation**:** The probability of getting the sum which is divisible by 2 & 3 is 5/36.

Favorable outcomes = (1, 5), (3, 3), (4, 2), (5, 1), (6, 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: The probability of getting no blue balls is 10/21.
Explanation:
Total number of balls = (2+3+2) = 7
Let S be the sample space.
Then, n(S) = Number of ways of drawing 2 balls out of 7
$= {}^{7}C_{2} = (7\times6) / (2\times1) = 21$

Let E = Event of drawing 2 balls, none of which is blue.
∴ n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.
$= {}^{5}C_{2} = (5\times4) / (2\times1) = 10$

∴ P(E) = n(E) / n(S) = 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:
Expected number of candies for a randomly selected child = 3.09
Explanation:
Expected number of candies for a randomly selected child
= 1 * 0.015 + 4*0.20 + 3 *0.65 + 5*0.005 + 6 *0.01 + 2 * 0.12
= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24
= 3.09

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

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

Use Q7.csv file

Ans:

Name	Point	Score	Weight
Mean	3.60	3.22	17.85
Median	3.70	3.32	17.71
Mode	3.07	3.44	17.02
Variance	0.29	0.96	3.19
Std Deviation	0.53	0.98	1.79
Range	2.76 - 4.93	1.51 - 5.42	14.50 - 22.90

Using python script find the above value. Those code written as below:
For Mean, Median, Std, Range: data.describe().round(2)
For Mode: data.mode().round(2)
For Variance: data.var().round(2)

Q8) Calculate Expected Value for the problem below

a) 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:

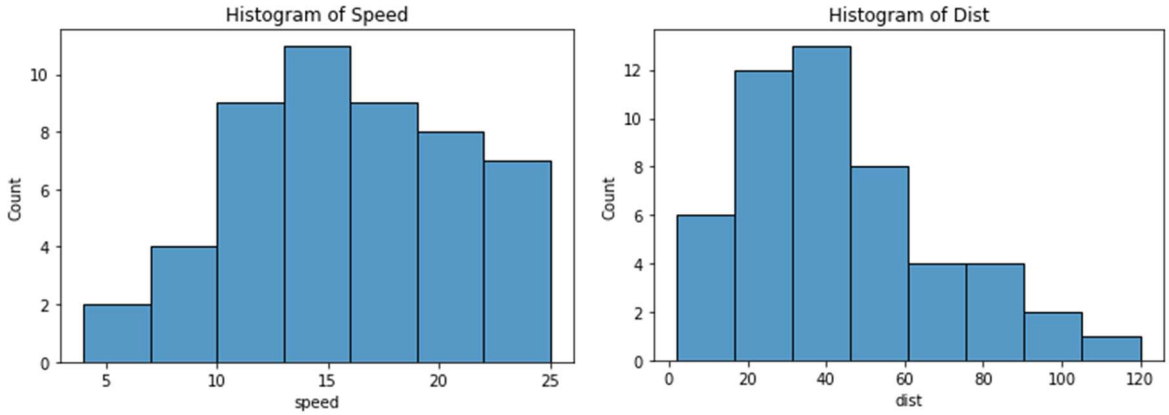
Expected value of the weight for a randomly selected patient = 145.33
Explanation:
Expected Value = ∑ (probability * Value)
Probability of selecting each patient = 1/9
Expected Value = (1/9) (108) + (1/9) (110) + (1/9) (123) + (1/9) (134) + (1/9) (135) + (1/9) (145) + (1/9 (167) + (1/9) (187) + (1/9) (199)
= (1/9) (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)
= (1/9) (1308) = 145.33

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

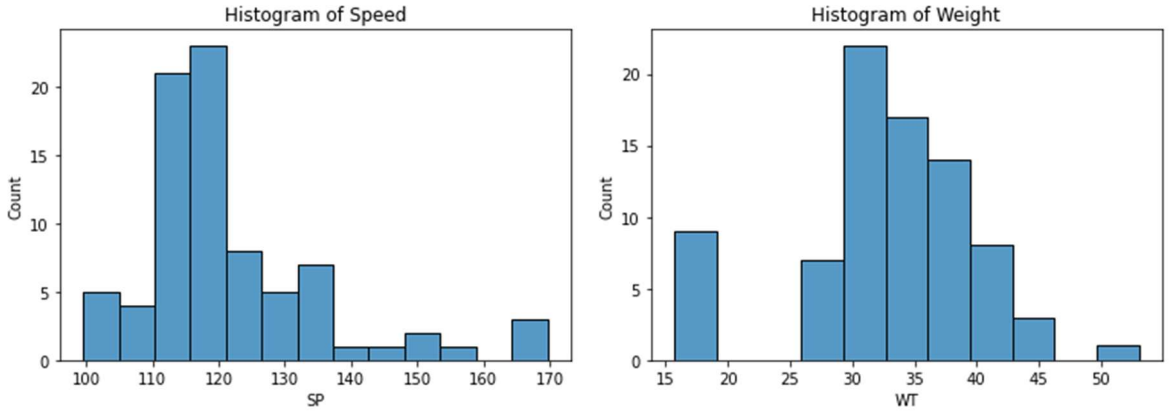
- Car's speed and distance
  Use Q9_a.csv
- SP and Weight (WT)
  Use Q9_b.csv

Ans:

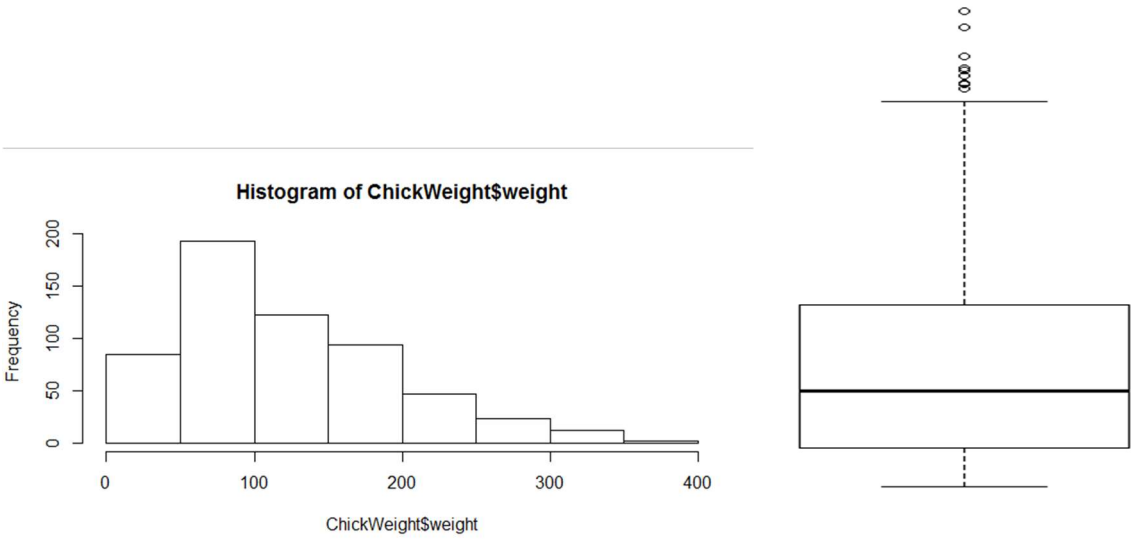
Skewness and Kurtosis of Car speed and distance is as follows:
Skewness = -0.117 (car speed), 0.806 (distance)
Kurtosis = -0.508 (car speed), 0.405 (distance)
Explanation: Inferences of Q9_a file



Skewness and Kurtosis of SP and Weight (WT) data are as follows:
Skewness = 1.611 (SP), -0.614 (weight)
Kurtosis = 2.977 (SP), 0.950 (weight)
Explanation: Inferences of Q9_b file



Q10) Draw inferences about the following boxplot & histogram



Ans: The histogram and boxplot figure are positively skewed on right side. i.e., Mean and median of the data is greater than mode.

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: x = 200, s = 30, n = 2000

1. The 94% confidence interval is (198.738, 201.261)
2. The 96% confidence interval is (198.439, 201.560)
3. The 98% confidence interval is (198.622, 201.377)

```
In [30]: import numpy as np
         import pandas as pd
         from scipy import stats
         from scipy.stats import norm

In [34]: # Avg. weight of Adult in Mexico with 94% CI
         stats.norm.interval(0.94,200,30/(2000**0.5))

Out[34]: (198.738325292158, 201.261674707842)

In [35]: # Avg. weight of Adult in Mexico with 98% CI
         stats.norm.interval(0.98,200,30/(2000**0.5))

Out[35]: (198.43943840429978, 201.56056159570022)

In [36]: # Avg. weight of Adult in Mexico with 96% CI
         stats.norm.interval(0.96,200,30/(2000**0.5))

Out[36]: (198.62230334813333, 201.37769665186667)
```

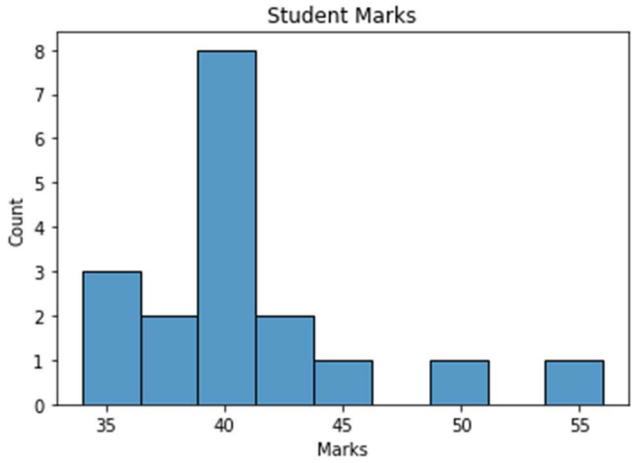
Q12) Below are the scores obtained by a student in tests

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.
2) What can we say about the student marks?

Ans:

1. Mean = 41
   Median = 40.5
   Variance = 25.529
   Standard Deviation = 5.052



2. From above plot we can say that mean of marks of student is 41 which is slightly greater than median.
   Most of the students got marks in between 41-43.

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

Ans: If the mean is equal to the median as well as the mode, hence the skewness is zero. If the distribution is symmetric, the mean equals to median, and the skewness of the distribution is zero.

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

Ans: If the mean is greater than the median, then 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.

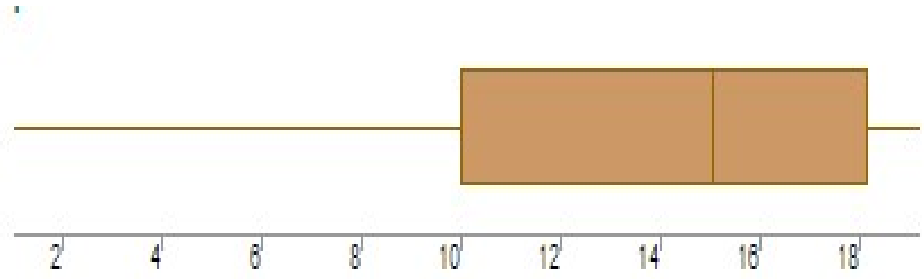
Q16) What does positive kurtosis value indicates for a data?

Ans: Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.

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

Ans: If a distribution has negative kurtosis, it is said to be platykurtic**,** which means that it has a flatter peak and thinner tails compared to a normal distribution. This simply means that more data values are located near the mean and less data values are located on the tails. Negative kurtosis is the uniform distribution, which has no peak at all and is a completely flat distribution.

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



1)What can we say about the distribution of the data?

Ans: The distribution in which more values are concentrated on the right side (tail) of the graph is called Negatively Skewed Distribution, while the left tail of the distribution graph is longer.

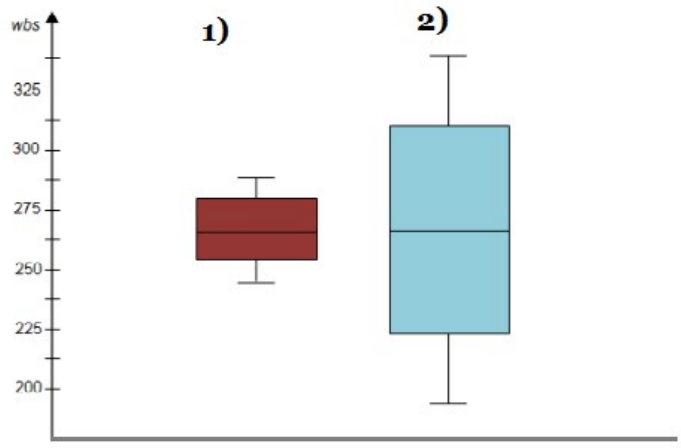
2)What is nature of skewness of the data?

Ans: The Mean of negatively skewed data will be less than the Median.

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

Ans: The IQR describes the middle 50% of values when ordered from lowest to highest. The Interquartile Range (IQR) = Q (3) – Q (1). In above example of data, the IQR = (18-10).

Q19) Comment on the below Boxplot visualizations?



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

Ans:

Both the plots infer that their data is normally distributed.

We can say that box plot 1 is for sample distribution and box plot 2 is for population or a sample with larger size.

In there are no outliers.

Q1 is 25%, Q3=75%, IQR is 50% for both the box plots. So, we can say both the distribution s follow normal distribution (mean=median=mode).

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

Data _set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

a) P(MPG>38)
b) P(MPG<40)
c) P (20<MPG<50)

Ans. a) 0.34 b) 0.72 c) 0.89

Explanation:

a) 34% of the cars are greater than 38 MPG.
b) 72% of the cars are less than 40 MPG.
c) 89% of the cars are in between 20 to 50 MPG.

```
In [105]: cars.describe()
```

Out[105]:

	HP	MPG	VOL	SP	WT
count	81.000000	81.000000	81.000000	81.000000	81.000000
mean	117.469136	34.422076	98.765432	121.540272	32.412577
std	57.113502	9.131445	22.301497	14.181432	7.492813
min	49.000000	12.101263	50.000000	99.564907	15.712859
25%	84.000000	27.856252	89.000000	113.829145	29.591768
50%	100.000000	35.152727	101.000000	118.208698	32.734518
75%	140.000000	39.531633	113.000000	126.404312	37.392524
max	322.000000	53.700681	160.000000	169.598513	52.997752

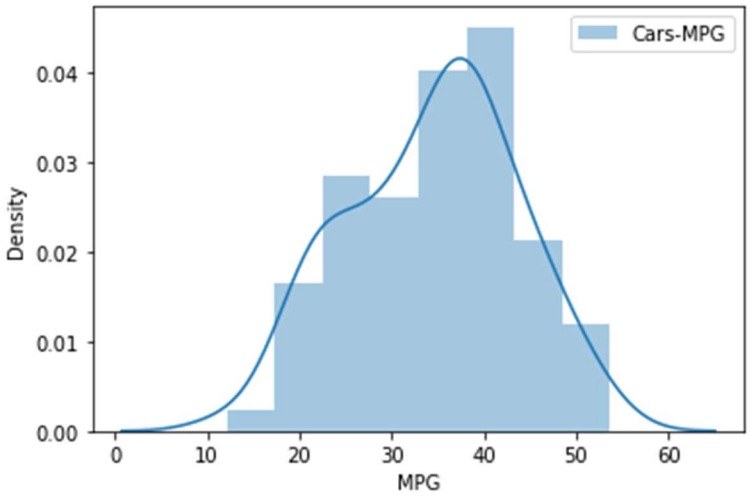
```
In [106]: # P(MPG>38)
          1- stats.norm.cdf(x = 38,loc = 34.42, scale = 9.13)
Out[106]: 0.34748702501304063

In [107]: # P(MPG<40)
          stats.norm.cdf(x = 40,loc = 34.42, scale = 9.13)
Out[107]: 0.7294571279557076

In [109]: # P (20<MPG<50)
          stats.norm.cdf(x = 50,loc = 34.42, scale = 9.13) - stats.norm.cdf(x = 20,loc = 34.42, scale = 9.13)
Out[109]: 0.8989177824549222
```

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

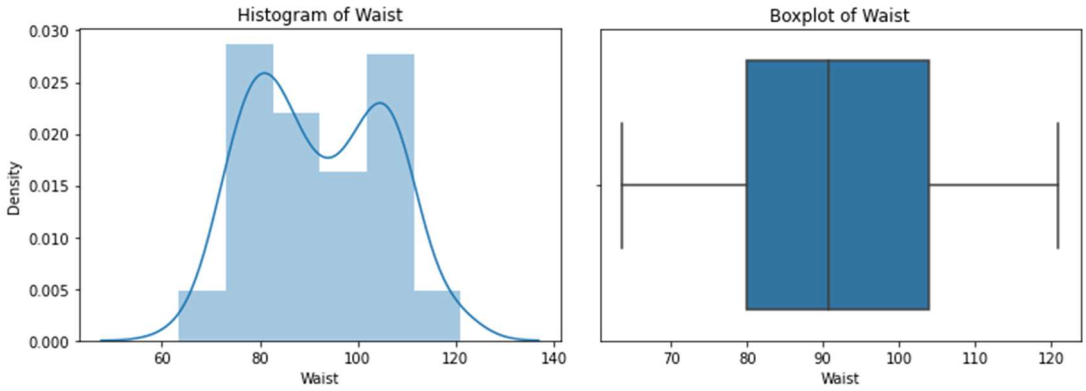
Ans:

The MPG of cars not following Normal Distribution.
Skewness = -0.177

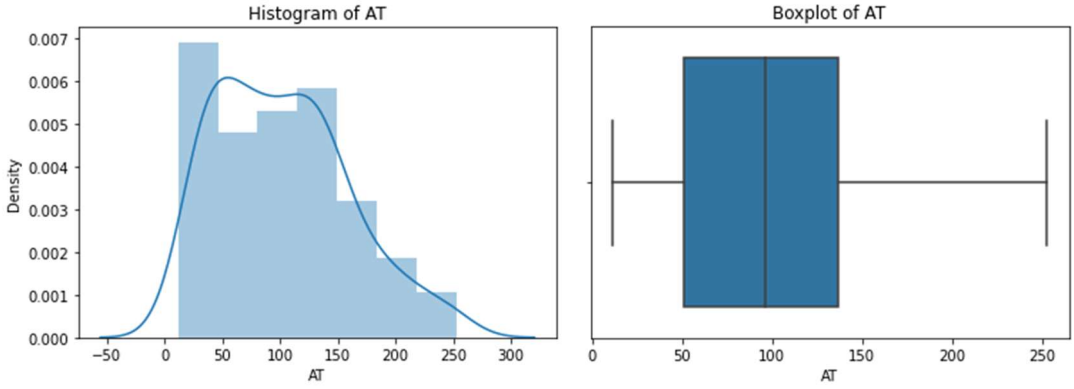


b) Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution.
Dataset: wc-at.csv

Ans: mean greater than median, both the whisker is of same length, median is slightly shifted towards left. Data is fairly symmetric.



Mean greater than median, right whisker is larger than left whisker, data is positively skewed.



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

Ans:

Z score of 60% Confidence Interval = 0.841
Z score of 90% Confidence Interval = 1.644
Z score of 94% Confidence Interval = 1.880

```
In [133]: # Z-score of 60% confidence interval
          stats.norm.ppf(0.8)

Out[133]: 0.8416212335729143

In [134]: # Z-score of 90% confidence interval
          stats.norm.ppf(0.95)

Out[134]: 1.6448536269514722

In [135]: # Z-score of 94% confidence interval
          stats.norm.ppf(0.97)

Out[135]: 1.8807936081512509
```

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

Ans:

Confidence Interval	T Score
95%	2.06
96%	2.17
99%	2.79

```
In [62]: # t scores of 95% confidence interval for sample size of 25
         stats.t.ppf(0.975,24)  # df = n-1 = 24
Out[62]: 2.0638985616280205

In [63]: # t scores of 96% confidence interval for sample size of 25
         stats.t.ppf(0.98,24)
Out[63]: 2.1715446760080677

In [64]: # t scores of 99% confidence interval for sample size of 25
         stats.t.ppf(0.995,24)
Out[64]: 2.796939504772804
```

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 2

Hint:

rcode → pt (tscore,df)

df → degrees of freedom

Ans:

Population mean = 270 days
Sample mean = 260 days
Sample SD = 90 days
Sample n = 18 bulbs
df = n-1 = 17

t = {(260-270) / (90/√18)}
t = (-1 * √2) / 3
t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

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

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

In [66]:
```
# find t-scores at x=260; t=(s_mean-P_mean)/(s_SD/sqrt(n))
t=(260-270)/(90/18**0.5)
t
```

Out[66]: -0.4714045207910317

Find P(X>=260) for null hypothesis

In [68]:
```
# p_value=1-stats.t.cdf(abs(t_scores),df=n-1)... Using cdf function
p_value=1-stats.t.cdf(abs(-0.4714),df=17)
p_value
```

Out[68]: 0.32167411684460556

In [70]:
```
#  OR p_value=stats.t.sf(abs(t_score),df=n-1)... Using sf function
p_value=stats.t.sf(abs(-0.4714),df=17)
p_value
```

Out[70]: 0.32167411684460556

THE END!!