|  |  |
| --- | --- |
| 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 | Categorical |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Categorical |

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 |
| 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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| 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:**

When three coins are tossed, the total number of possible combinations are 23 = 8.

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

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

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

Therefore, the [Probability](https://www.cuemath.com/data/probability/) of getting two heads and one tails in the toss of three coins simultaneously is defined as=

P (2 heads and 1 Tail) / Number of desired outcomes

= 3 / 8= 0.375

=37.5%

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 2 and 3

**Ans:**

Two dice are thrown here N

N(s) =36

1. The sum is equal to 1is zero because they starts with (1,1)….so on. Other than in the dice we are not having zero.
2. The probability that sum is less than or equal to 4

According to above condition the outcomes are (1,1),(1,2),(1,3),(2,1)(2,2,)and(3,1)

So now the probability that sum is less than or equal to 4=6/36

=16.66%

1. The probability that sum is divisible by 2 and 3

For the above condition the outcome are:

(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)

So now the probability that sum is divisible that sum is divisible by 2 and 3=6/36

=16.66%

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:**

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

## N(S) =

## N(S) = (7 x 6) / (2 x 1) = 21

## Let E=Event of 2 balls, none of which is blue

## Therefore, N (E) = Number of ways of drawing 2 balls out of(2+3) balls

## N (E) =

## N (E) = (5 x 4) / (2 x 1) = 10

## Therefore, 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

= (1 x 0.015) + (4 x 0.20) + (3 x 0.65) + (5 x 0.005) + (6 x 0.01) + (2 x 0.12)

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

= 3.090

= 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:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

cars**=**pd**.**read\_csv("Database/Q7.csv")

cars

|  | **Unnamed: 0** | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- | --- |
| **0** | Mazda RX4 | 3.90 | 2.620 | 16.46 |
| **1** | Mazda RX4 Wag | 3.90 | 2.875 | 17.02 |
| **2** | Datsun 710 | 3.85 | 2.320 | 18.61 |
| **3** | Hornet 4 Drive | 3.08 | 3.215 | 19.44 |
| **4** | Hornet Sportabout | 3.15 | 3.440 | 17.02 |
| **5** | Valiant | 2.76 | 3.460 | 20.22 |
| **6** | Duster 360 | 3.21 | 3.570 | 15.84 |
| **7** | Merc 240D | 3.69 | 3.190 | 20.00 |
| **8** | Merc 230 | 3.92 | 3.150 | 22.90 |
| **9** | Merc 280 | 3.92 | 3.440 | 18.30 |
| **10** | Merc 280C | 3.92 | 3.440 | 18.90 |
| **11** | Merc 450SE | 3.07 | 4.070 | 17.40 |
| **12** | Merc 450SL | 3.07 | 3.730 | 17.60 |
| **13** | Merc 450SLC | 3.07 | 3.780 | 18.00 |
| **14** | Cadillac Fleetwood | 2.93 | 5.250 | 17.98 |
| **15** | Lincoln Continental | 3.00 | 5.424 | 17.82 |
| **16** | Chrysler Imperial | 3.23 | 5.345 | 17.42 |
| **17** | Fiat 128 | 4.08 | 2.200 | 19.47 |
| **18** | Honda Civic | 4.93 | 1.615 | 18.52 |
| **19** | Toyota Corolla | 4.22 | 1.835 | 19.90 |
| **20** | Toyota Corona | 3.70 | 2.465 | 20.01 |
| **21** | Dodge Challenger | 2.76 | 3.520 | 16.87 |
| **22** | AMC Javelin | 3.15 | 3.435 | 17.30 |
| **23** | Camaro Z28 | 3.73 | 3.840 | 15.41 |
| **24** | Pontiac Firebird | 3.08 | 3.845 | 17.05 |
| **25** | Fiat X1-9 | 4.08 | 1.935 | 18.90 |
| **26** | Porsche 914-2 | 4.43 | 2.140 | 16.70 |
| **27** | Lotus Europa | 3.77 | 1.513 | 16.90 |
| **28** | Ford Pantera L | 4.22 | 3.170 | 14.50 |
| **29** | Ferrari Dino | 3.62 | 2.770 | 15.50 |
| **30** | Maserati Bora | 3.54 | 3.570 | 14.60 |
| **31** | Volvo 142E | 4.11 | 2.780 | 18.60 |

*# mean*

cars**.**mean()

Points 3.596563

Score 3.217250

Weigh 17.848750

dtype: float64

*# Median*

cars**.**median()

Points 3.695

Score 3.325

Weigh 17.710

dtype: float64

*# Mode*

cars**.**Points**.**mode()

0 3.07

1 3.92

dtype: float64

cars**.**Score**.**mode()

0 3.44

dtype: float64

cars**.**Weigh**.**mode()

0 17.02

1 18.90

dtype: float64

*# Variance*

cars**.**var()

Points 0.285881

Score 0.957379

Weigh 3.193166

dtype: float64

*# Satndard Deviation*

cars**.**std()

Points 0.534679

Score 0.978457

Weigh 1.786943

dtype: float64

# Range

cars**.**describe()

|  | **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 |

Points\_Range**=**cars**.**Points**.**max()**-**cars**.**Points**.**min()

Points\_Range

2.17

Score\_Range**=**cars**.**Score**.**max()**-**cars**.**Score**.**min()

Score\_Range

3.9109999999999996

Weigh\_Range**=**cars**.**Weigh**.**max()**-**cars**.**Weigh**.**min()

Weigh\_Range

8.399999999999999

f,ax**=**plt**.**subplots(figsize**=**(15,5))

plt**.**subplot(1,3,1)

plt**.**boxplot(cars**.**Points)

plt**.**title('Points')

plt**.**subplot(1,3,2)

plt**.**boxplot(cars**.**Score)

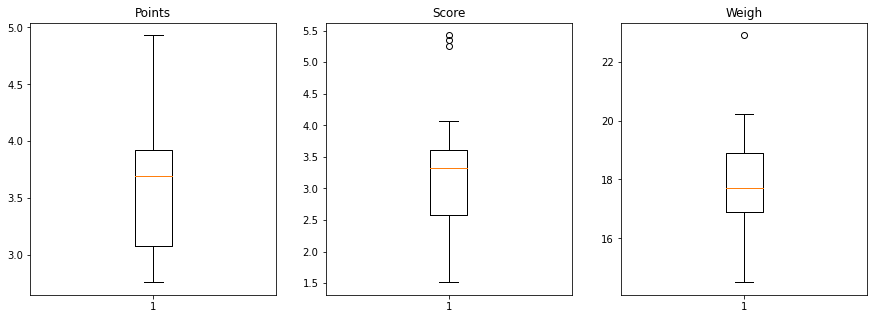
plt**.**title('Score')

plt**.**subplot(1,3,3)

plt**.**boxplot(cars**.**Weigh)

plt**.**title('Weigh')

plt**.**show()



**Inferences:**

Here in this case of different models of cars data, most type of cars have average points of 3.596563, score of 3.217150 and weight of 17.848750. Also here in this scenario the standard deviation is very low in points and score so chances of presence of outliers I both the case is very low and comparing to weigh there is little bit higher standard deviation so may be some outliers are present.

Somehow data points in every case have less spread so most of the data points lie near to the median.

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:**

The weights (X) of patients at a clinic (in pounds), are 108, 110, 123, 134, 135, 145, 167, 187, 199

one of the patients is chosen at random.

Expected Value  =  ∑ ( probability  x Value )

 ∑ P(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

E(x) =  108, 110, 123, 134, 135, 145, 167, 187, 199

P(x) = 1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  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

Expected Value of the Weight of that patient = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

data=pd.read\_csv("Database/Q9\_a.csv")

data

*# Skewness*

data**.**skew()

Index 0.000000

speed -0.117510

dist 0.806895

dtype: float64

*# Kurtosis*

data**.**kurt()

Index -1.200000

speed -0.508994

dist 0.405053

dtype: float64

**Inferences:**

* For car speed, skewness is negative and the kurtosis is negative, which suggests that the distribution is more towards left. It means the distribution is left skewed or negative skewed. Here in negative skewed mean is less than median. As considering kurtosis, it shows that the distribution has broad peak and thin tail.
* For the distance, travel by the car skewness is positive and the kurtosis is positive, which suggests that the distribution is more towards right. It means the distribution is right skewed or positive skewed. Here in positive skewed mean is greater than median. As considering kurtosis, it shows that the distribution has pointed peak and wide tail.

|  |
| --- |

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans :**

data**=**pd**.**read\_csv('Database/Q9\_b.csv')

data2**=**data**.**iloc[:,1:]

data2

*# Skewness*

data2**.**skew()

SP 1.611450

WT -0.614753

dtype: float64

*# Kurtosis*

data2**.**kurt()

SP 2.977329

WT 0.950291

dtype: float64

**Inferences:**

* For SP skewness is positive and the kurtosis is positive, which suggests that

the distribution is more towards right. It means the distribution is right

skewed or positive skewed. Here in positive skewed mean is greater than

median. As considering kurtosis, it shows that the distribution has pointed

peak and wide tail.

* For WT skewness is negative and the kurtosis is positive, which suggests

that the distribution is more towards left. It means the distribution is left

skewed or negative skewed. Here in negative skewed mean is less than

median.

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

**Ans:**



**Inference:**

From this above Histogram and Box plot, it shows that the distribution has outliers at the end(means in histogram tail side and in box plot at in upper extreme) .The distribution is positive skewed or right skewed.

**Q11)** Suppose we want to estimate the average weight of an adult male in 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:**

Here total no. of sample mean(n) = 2000

The average weight of person in sample() = 200

Standard deviation of sample(δ) =30

Confidence interval =

For 94% of CI value Z score = 1.89

Confidence interval for 94% = 200 ± (1.89 x (30/√2000))

= 198.73 to 201.27

For 98% of CI value Z score = 2.33

Confidence interval for 98% = 200 ± (2.33 x (30/√2000))

= 198.43 to 201.56

For 96% of CI value Z score = 1.06

Confidence interval for 96% = 200 ± (2.06 x (30/√2000))

= 198.62 to 201.38

**In python code:**

**import** numpy **as** np

**import** pandas **as** pd

**from** scipy **import** stats

**from** scipy.stats **import** norm

*# Avg. weight of Adult in Mexico with 94% CI*

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

(198.738325292158, 201.261674707842)

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

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

(198.43943840429978, 201.56056159570022)

*# Avg. weight of Adult in Mexico with 96% CI*

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

(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.

**Ans:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

x**=**pd**.**Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

*# Mean*

x**.**mean()

41.0

*# Median*

x**.**median()

40.5

*# Variance*

x**.**var()

25.529411764705884

*# Standard Deviation*

x**.**std()

5.05266382858645

1. What can we say about the student marks?

**Ans:**

The student score 41 mark most of the time. He scores average 41 mark.

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

**Ans:**

The nature of skewness is zero.

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

**Ans:**

When the mean > median, the nature of skewness is positive. It means right skewed.

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

**Ans:**

When the median > mean, the nature of skewness is negative. It means left skewed.

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

**Ans:**

Positive kurtosis indicates that the distribution is peaked and possess thick tails. It means most of the data located on the tail side. And it also indicates that the distribution has heavier tails than the normal distribution.

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

**Ans:**

Negative kurtosis value for a data indicates that the distribution has lighter tails than the normal distribution.

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



What can we say about the distribution of the data?

**Ans:**

Most of the data distributed between of 10 to 18.3.

What is nature of skewness of the data?

**Ans:**

Nature of skewness of the data is Negative Skewness. It means left skewed.

What will be the IQR of the data (approximately)?   
**Ans:**

IQR = 8.2.

50% of data lies in between IQR range.

Q19) Comment on the below Boxplot visualizations?



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

**Ans:**

From the both box plot the mean is around 262.5 for both cases. The distribution in both the cases looks like symmetric distribution. We can also say that both the box plot are normally distributed.

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

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

**Ans:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

from scipy.stats import norm

cars=pd.read\_csv('Database/Cars.csv')

cars

|  | HP | MPG | VOL | SP | WT |
| --- | --- | --- | --- | --- | --- |
| 0 | 49 | 53.700681 | 89 | 104.185353 | 28.762059 |
| 1 | 55 | 50.013401 | 92 | 105.461264 | 30.466833 |
| 2 | 55 | 50.013401 | 92 | 105.461264 | 30.193597 |
| 3 | 70 | 45.696322 | 92 | 113.461264 | 30.632114 |
| 4 | 53 | 50.504232 | 92 | 104.461264 | 29.889149 |
| ... | ... | ... | ... | ... | ... |

*# P(MPG>38)*

1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

0.3475939251582705

*# P(MPG<40)*

stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

0.7293498762151616

*# P (20<MPG<50)*

stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())

1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

cars**=**pd**.**read\_csv('Database/Cars.csv')

cars

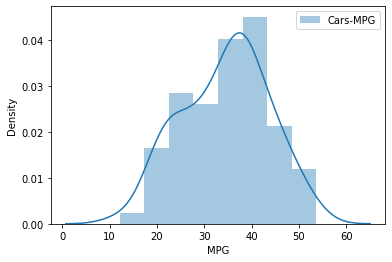
|  |  | **HP** | **MPG** | **VOL** | **SP** | **WT** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** |  | 49 | 53.700681 | 89 | 104.185353 | 28.762059 |
| **1** |  | 55 | 50.013401 | 92 | 105.461264 | 30.466833 |
| **2** |  | 55 | 50.013401 | 92 | 105.461264 | 30.193597 |
| **3** |  | 70 | 45.696322 | 92 | 113.461264 | 30.632114 |
| **4** |  | 53 | 50.504232 | 92 | 104.461264 | 29.889149 |
| **...** |  | ... | ... | ... | ... | ... |
|  |  |  |  |  |  |  |

sns**.**distplot(cars**.**MPG, label**=**'Cars-MPG')

plt**.**xlabel('MPG')

plt**.**ylabel('Density')

plt**.**legend();



cars**.**MPG**.**mean()

34.422075728024666

cars**.**MPG**.**median()

35.15272697

From this above distribution plot we can say the MPG of cars follows normal distribution.

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

Dataset: wc-at.csv

**Ans:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

wcat**=**pd**.**read\_csv('Database/wc-at.csv')

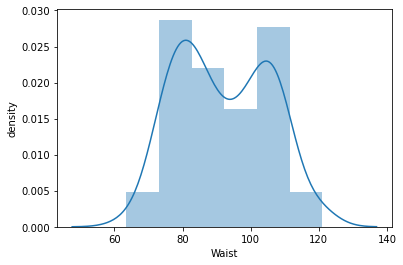
wcat

|  | **Waist** | **AT** |
| --- | --- | --- |
| **0** | 74.75 | 25.72 |
| **1** | 72.60 | 25.89 |
| **2** | 81.80 | 42.60 |
| **3** | 83.95 | 42.80 |
| **4** | 74.65 | 29.84 |
| **...** | ... | .. |

*# plotting distribution for Waist Circumference (Waist)*

sns**.**distplot(wcat**.**Waist)

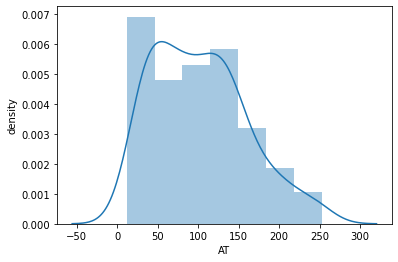
plt**.**ylabel('density');



*# plotting distribution for Adipose Tissue (AT)*

sns**.**distplot(wcat**.**AT)

plt**.**ylabel('density');



*# WC*

wcat**.**Waist**.**mean() , wcat**.**Waist**.**median()

(91.90183486238533, 90.8)

*# AT*

wcat**.**AT**.**mean() , wcat**.**AT**.**median()

(101.89403669724771, 96.54)

From the above distribution plot for both AT and Waist of wc-at data set , it shows that both AT and Waist follows normal distribution.

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

**Ans:**

from scipy import stats

from scipy.stats import norm

*# Z-score of 90% confidence interval*

stats.norm.ppf(0.95)

1.6448536269514722

*# Z-score of 94% confidence interval*

stats.norm.ppf(0.97)

1.8807936081512509

*# Z-score of 60% confidence interval*

stats.norm.ppf(0.8)

0.8416212335729143

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

**Ans:**

from scipy import stats

from scipy.stats import norm

*# t-score of 95% confidence interval*

stats.t.ppf(0.975,24)

t score of 95% confidence interval for sample size of 25 = 2.064

*# t-score of 96% confidence interval*

stats.t.ppf(0.98,24)

t score of 96% confidence interval for sample size of 25 = 2.172

*# Z-score of 99% confidence interval*

stats.norm.ppf(0.995,24)

t score of 99% confidence interval for sample size of 25 = 2.797

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:**

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.

**To find:**

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

**Solution:**

t - statistics for the data is given as follows:

https://tex.z-dn.net/?f=t%3D%5Cdfrac%7Bx-%5Cmu%7D%7B%5Cfrac%7Bs%7D%7B%5Csqrt%20n%7D%7D

x = mean of the sample of bulbs =  260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

https://tex.z-dn.net/?f=t%3D%5Cdfrac%7B260-270%7D%7B%5Cfrac%7B90%7D%7B%5Csqrt%2018%7D%7D

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-10%7D%7B%5Cfrac%7B90%7D%7B3%5Csqrt%202%7D%7D

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-10%7D%7B%5Cfrac%7B30%7D%7B%5Csqrt%202%7D%7D

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-1%20%5Ctimes%20%5Csqrt%202%7D%7B3%7D

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.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.