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

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ordinal |
| 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?

A) For one coin, we have 2 outcomes i.e., Head and Tail

So here we have 3 coins and the total possibilities are, 2^3=2\*2\*2=8

Probability of 2 heads and 1 tail is 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 2 and 3

A) For one dice, we 6 outcomes

Here we have 2 dice and the total possibilities of the outcomes are 6^2=6\*6=36

So Total possibilities=36

1. The sum cannot be equal to 1 it is always more than 1

hence it is 0/36=0

1. Sum<= 4

there are 6 chances where the sum of two dice can be less than or equal to 4

i.e., 6/36 = 1/6

c. (1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6) are the outcomes which are divisible by 2 and 3 both.

i.e., 5/36

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?

A) Total number of balls = (2 + 3 + 2) = 7.

From 7 balls, we should select 2 balls.

No.of ways of drawing 2 balls of of 7 is 7c2

7\*6/2\*1=42/2=21

 Number of ways of drawing 2 balls out of (2 + 3) balls

5c2

5\*4/2\*1=10

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

A. 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

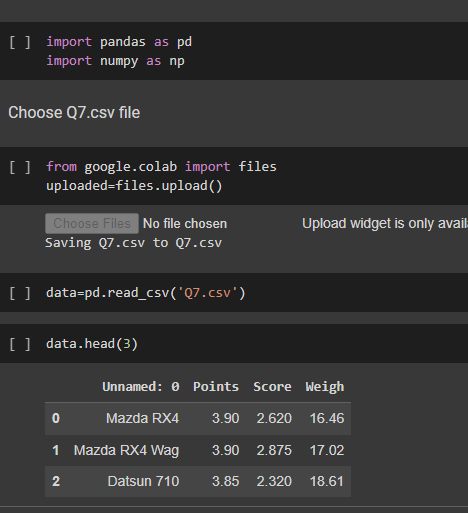
= 0.015 + 0.8  + 1.95 + 0.025 + 0.06 + 0.24

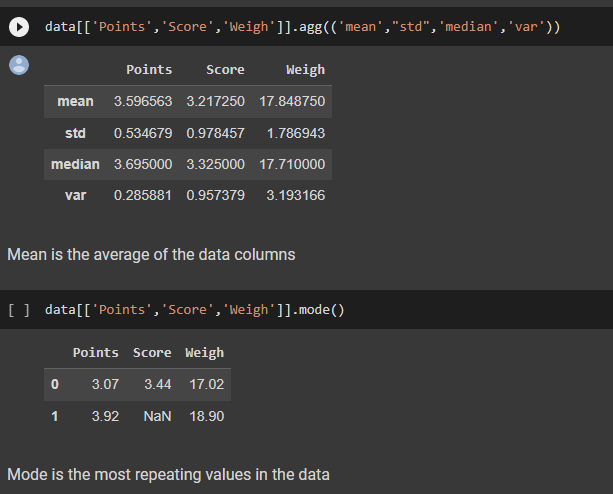
=       3.090

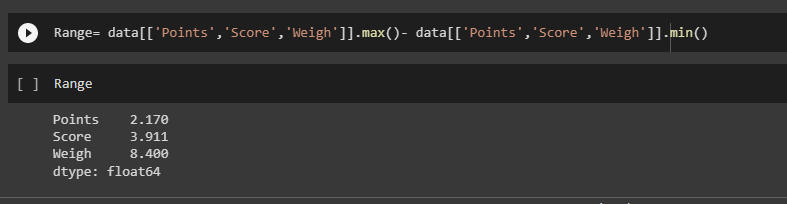
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

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







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?

1. there are 9 patients

Probability of selecting each patient = 1/9

Ex  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)

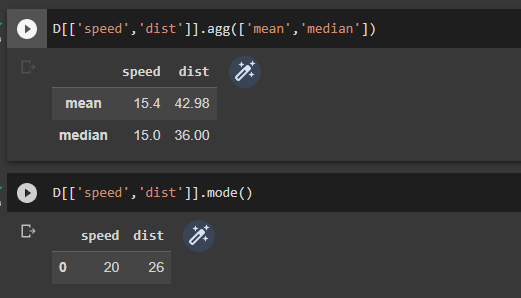
= (1/9)  (  1308)

= 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**



Calculated the Mean, Median and Mode

Mean!=Median!=Mode –>Asymmetric distribution/skewed distribution

#the speed is in negative value hence it is left tailed or negatively skewed

#dist is in positive value it is right tailed or posilively skewed

#Speed = -0.11 which is in between 0 and -0.5 i.e,., the distribution is approximately symetric

#dist = 0.78 which is in between 0.5 and 1 i.e., the distribution is Moderately skewed

#For SPEED Kurtosis<0 i.e., platykurtic kurtosis (Negative kurtosis) , outliers are less

#For Distance kurtosis>0 i.e., leptokurtic kurtosis (positive kurtosis) , outliers are more

#The outlier frequency is less for Platykurtic kurtosis and more for leptokurtic kurtosis

**SP and Weight(WT)**

**Use Q9\_b.csv**

Skewness for SP is 1.5814536794423764

Skewness for WT is -0.6033099322115126

--------------------------------------------------------------------------------------------SP is positively or right skewed as skewness value is 1.5814536794423764

WT weight is negatively/left skewed as skewness value is -0.6033099322115126

--------------------------------------------------------------------------------------------

SP is highly skewed as skewness is greater than 1 or -1: 1.5814536794423764

WT is Moderately skewed as skewness value is in between -1.0 to -0.5 and 1 to 0.5 -0.6033099322115126

--------------------------------------------------------------------------------------------

Kutosis for SP is 2.7235214865269244

Kutosis for WT is 0.8194658792266849

SP is leptokurtic kurtosis as it is greater than 0, it is Narrow peaked or positivly peaked and the outliers are more

WT is leptokurtic kurtosis as it is greater than 0, it is Narrow peaked or positivly peaked and the outliers are more

--------------------------------------------------------------------------------------------

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



* Chick weight data is right skewed
* Around 200 chicks have the weight between 50-100
* Around 0-5 chicks have the maximum weight i.e., 400
* Most of the chicks weigh between 50 to 150
* Around 80 chicks have the least weight i.e., 0-50



* There are 7 outliers at Upper Extreme
* The boxplot is Right or Positively skewed
* Maximum of the data is present on left side
* Median is lower than mean

**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?

1. Estimate average weight of an adult male in Mexico,

Mu=?

n=2000 (random sample)

x̄ = 200 (sample average)

s = 30 (sample standard deviation)

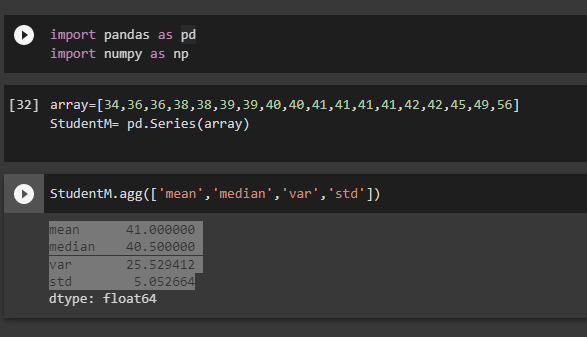
Sigma is unknown, i.e., the standard deviation for the population is unknown. Hence we use T test

stats.norm.interval(0.95, x̄,s)

**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?



* Average marks the student scored in the test is 41

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

1. Zero skewness

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

1. Positively skewed

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

1. Negatively skewed

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

1. it is leptokurtic kurtosis and it is fat tailed i.e., the outliers are more.

Peak is more than mesokurtic kurtosis

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

1. Platyukurtic kurtosis, it has wide peak and has thin outliers, i.e., less outliers are present

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

* Above boxplot is not normally distributed,
* It is Left skewed/tailed or Negatively skewed as median is greater than mean.
* IQR=UQ-LQ
* IQR = 18-10=8

Q19) Comment on the below Boxplot visualizations?



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

* Both the boxplots are Normal distribution.
* The median is same for two boxplot’s which is around 262.5
* No outliers in both
* Data is less is Boxplot1 and more in Boxplot2

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

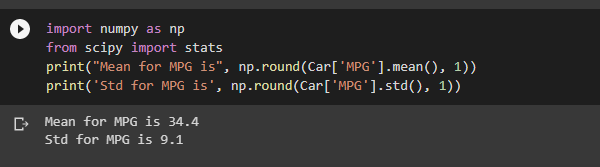
Data \_set: Cars.csv

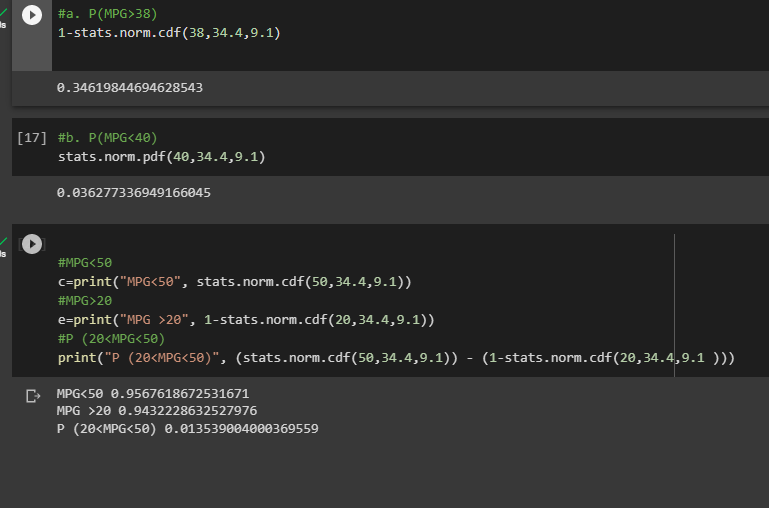
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

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

c. P (20<MPG<50)





Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

* 1. import pandas as pd

import seaborn as sn

import matplotlib.pyplot as plt

from google.colab import files

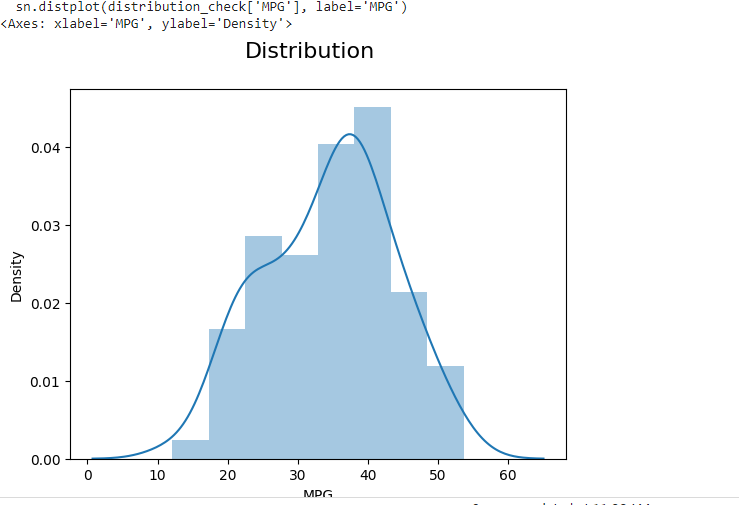
uploaded=files.upload()

distribution\_check.head(2)

#Density distribution plot

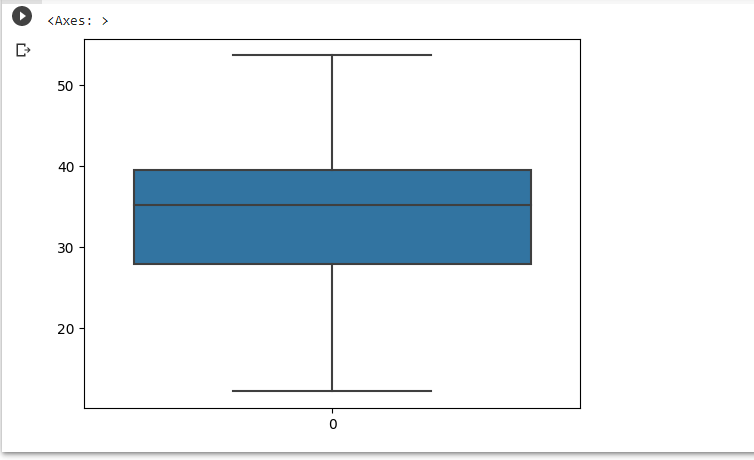
plt.suptitle('Distribution', fontsize=16)

sn.distplot(distribution\_check['MPG'], label='MPG')



#boxplot

sn.boxplot(distribution\_check['MPG'])



#Check skewness

import scipy

from scipy import stats

import numpy as np

sw=scipy.stats.skew(distribution\_check['MPG'])

print("Skewness is", np.round(sw, 1))

o/p: Skewness is -0.2

The skewness is between +0.5 and -0.5, hence the distribution is **approximately symmetric.**

i.e., By looking at the density plot,box plot and skewness we can tell it is approximately symmetric or approximately Normal distribution.

**b.** import pandas as pd

import scipy

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sn

from google.colab import files

uploaded=files.upload()

WCAT=pd.read\_csv('wc-at.csv')

WCAT.head(3)

#lets check skewness

skew\_waist= scipy.stats.skew(WCAT[['Waist']])

print('Skewness for Waist is:', np.round(skew\_waist, 1))

skew\_at=scipy.stats.skew(WCAT[['AT']])

print('Skewness for AT is:', np.round(skew\_at, 1))

Output: Skewness for Waist is: [0.1]

Skewness for AT is: [0.6]

scipy.stats.kurtosis(WCAT['AT'])

-0.3271884938021854

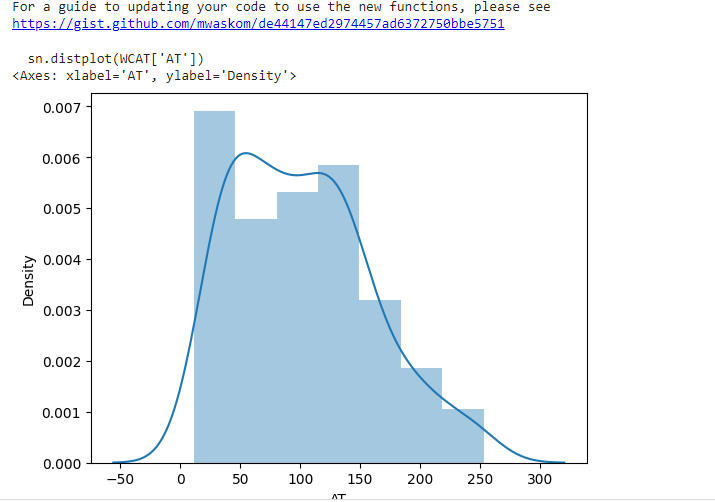
scipy.stats.kurtosis(WCAT['Waist'])

-1.1072764806858817

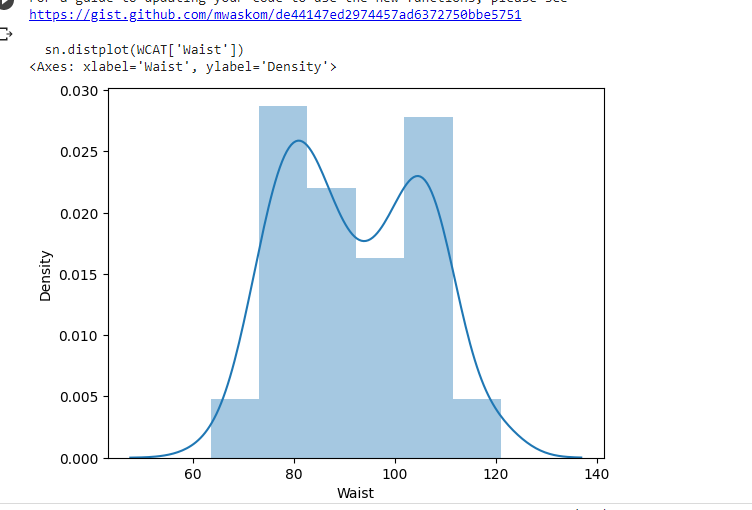
For waist skewness is between 0 and 0.5 i.e., it is approximately symmetrical

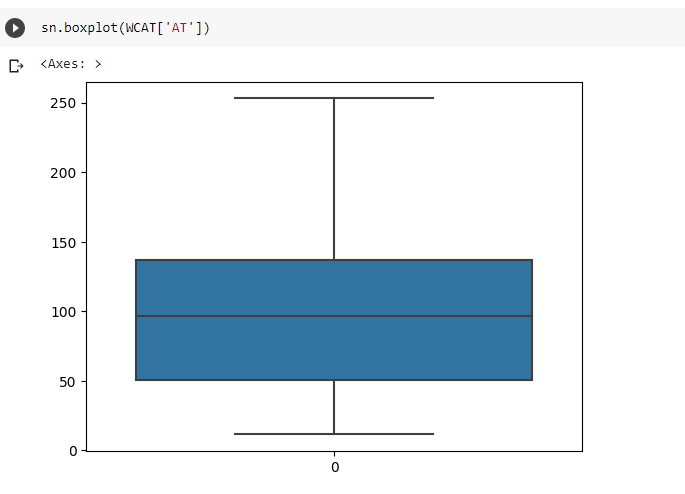
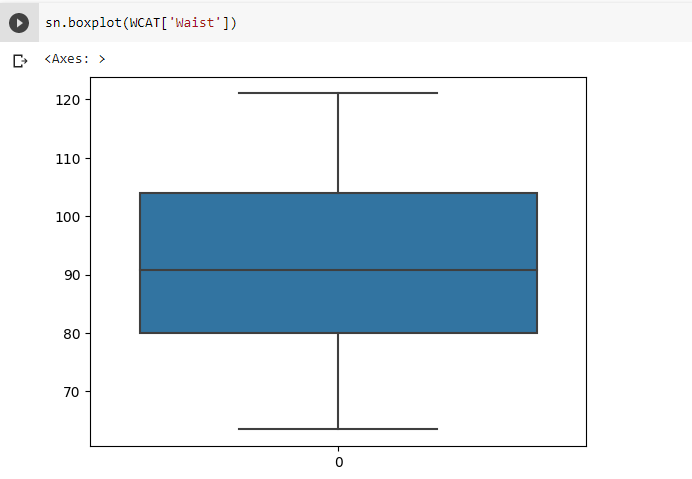
For AT it is moderately right skewed and both has wide peak i.e., platykurtic kurtosis

sn.distplot(WCAT['AT'])



sn.distplot(WCAT['Waist'])



For ATFor Waist

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

A. 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

A.

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

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

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

stats**.**t**.**ppf(0.975,24) *# df = n-1 = 24*

2.0638985616280205

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

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

2.1715446760080677

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

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

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

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

1. 0.321 percent probability that 18 randomly selected bulbs would have an average life of no more than 260 days