Assignment 1:

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 | Nominal |
| 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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| 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 :- Let S be the sample. 8 probabilities happen here

Then S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Probability of getting two heads

Then: two heads = {HHT, HTH, THH}

Two heads = 3

P (two heads) = 3 / 8

Probability of getting one tail

Then: one tail = {HHT, HTH, THH}

One tail = 3

P (one tail) = 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

Ans: Total number of outcomes when two dice are rolled=6\*6=36

(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)

*P = Number of favourable outcomes/Total number of possible outcomes*

1. the sum is equal to 1 is zero (0%) probability

because they start with (1,1) other than in the dice we are not having zero.

1. the sum is equal to 4 the possible outcomes are (1,1) (1,2) (1,3) (2,1) (2,2) (3,1)

therefore n(b) = 6/36 = 1/6

1. sum is divisible by 2 and 3 is = 6/36

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

Therefore n(c) =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:

Red: R1, R2

Green: G1, G2, G3

Blue: B1, B2

Total ball = 7

Two balls are drawn randomly

2 balls can be selected from 7 balls in 7C2 ways

7C2 = 7 \* 6 / 2 \*1

7C2 = 42 / 2

7C2 = 21 ways

Favourable cases that two nonblue balls are drawn = 5C2 ways

5C2 = 5\*4 / 1\*2

5C2 = 20 / 2 = 10 ways

*P = Number of favourable outcomes/Total number of possible outcomes*

P = 10/21

P = 0.47

So, the probability that none of the balls drawn are 10/21 = 0.47

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

Expected Value = x \* P(x)

= (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

= 3.09

Expected number of candies for a randomly selected child = 3.09

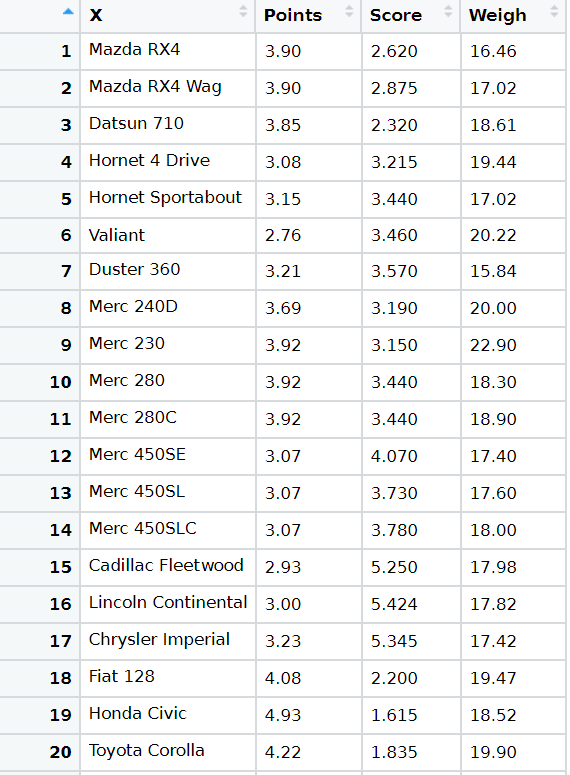
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weigh>

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

**Use Q7.csv file**

**Ans:**



Mean of the given dataset is

points = 115.09/32

= 3.59

scores = 102.952/32

= 3.21725

weigh = 27.16/32

= 17. 84875

Median of the given dataset is:

points = (3.69+3.7)/2

= 7.39/2

= 3.695

Scores = (3.215+3.435)/2

= 6.65/2

= 3.325

weigh = (17.6+17.82)/2

= 35.42/2

= 17.71

Mode of the given dataset is:

points = 3.92

scores = 3.44

weigh = 17.02

Variance of the given dataset is:

points = σ =8.862/31

= 0.2858814

score = σ = 29.678748/31

= 0.957379

weigh = σ = 98.98815/31

= 3.193166

Standard Deviation of the given dataset is:

Points = 0.5346787

Scores = 0.9784574

Weigh = 1.786943

Range of the given dataset is:

Points = 2.17

Scores = 3.911

Weigh = 8.4

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:

one of the patients is chosen at random then expected value of weight of patient is

first, we calculate the mean

Expected value = 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199 /9

= 1308/9

= 145.333

Expected value of weight of that patient is 145.33

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

Cars speed and distance

Use Q9\_a.csv

SP and Weight(WT)

Use Q9\_b.csv

Ans:

install.packages("moments")

library(moments)

a <- read.csv(file.choose())

View(a)

b <- read.csv(file.choose())

View(b)

# -------------skewness-----------------

skewness(a$SP)

# skewness of sp is 1.581454

skewness(a$WT)

# skewness of weight is -0.6033099

#------------------kurtosis----------------

kurtosis(b$SP)

# kurtosis of sp is = 5.723521

kurtosis(b$WT)

# kurtosis of weight is = 3.819466

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



Ans:

Histogram represents the frequency distribution of data, how many observations to take the value within central interval.

In histogram the data is Left side more and long right tail then it is positive skewed.

Positive skewness implies mass of the data concentrated on the left.

In X axis weight is given and Y axis Frequency is given.

Major weight of 50 to 100 on this histogram.



The above boxplot suggests that the distribution has lots of outliers towards upper extreme.

A box plot is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum.

In this box plot here, we rotate by 90degree clockwise direction.

We have seen a right skewed graph with a long tail and outliers.

The dispersion of the boxplot is not high because the IQR of the boxplot is not much stretched and in the median is not in the centre.

In the middle 3rd quartile is more and few outliers.

Which implies Q1-Q2 is not equal to Q3-Q2

**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:

n = 2000, std deviation = 30, N= 3,000,000, x = 200

find the std error:

sigma(x) = sigma / sqrt(n)

std error = 30 / sqrt (2000)

std error = 0.6708204

Interval Estimate = Point Estimate +- Margin of error

**Calculate 94% confidence interval**

alpha = 1 -(confidence interval / 100)

alpha = 1 -(94/100)

alpha = 0.06

find the probability:

p = 1 - alpha / 2

p = 1 - 0.05 / 2

p = 0.97

Calculate z value:

z value = qnorm (0.97) = 1.8807

Interval Estimate = Point Estimate + - Margin of error

= [200 - 0.67 \*1.88 200 + 0.67 \* 1.88]

= [198.7404 201.2596 ]

94% of confidence interval is [198.74 201.26]

**Calculate 96% confidence interval**

alpha = 1 -(confidence interval / 100)

alpha = 1 -(96/100)

alpha = 0.04

find the probability:

p = 1 - alpha / 2

p = 1 - 0.04 / 2

p = 0.98

Calculate z value:

z value = qnorm (0.98) = 2.053749

Interval Estimate = Point Estimate + - Margin of error

= [200 - 0.67 \*2.05 200 + 0.67 \* 2.05]

= [198.6265 201.3735 ]

96% of confidence interval is [198.62 201.37]

**Calculate 98% confidence interval**

alpha = 1 -(confidence interval / 100)

alpha = 1 -(98/100)

alpha = 0.02

find the probability:

p = 1 - alpha / 2

p = 1 - 0.02 / 2

p = 0.99

Calculate z value:

z value = qnorm (0.99) = 2.326348

Interval Estimate = Point Estimate + - Margin of error

= [200 - 0.67 \*2.32 200 + 0.67 \* 2.32]

= [198.4456 201.5544 ]

98% of confidence interval is [198.44 201.55]

|  |
| --- |
|  |
| |  | | --- | |  | |

**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. Find mean, median, variance, standard deviation.
2. Mean:

Sum = (34 + 36 + 36 + 38 + 38 + 39 + 39 + 40 + 40 + 41 + 41 + 41 + 41 + 42 + 42 + 45 + 49 + 56)

Sum = 738

Mean = 738/18

Mean = 41

1. Median:

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

Median = (40+41)/2

= 40.5

1. Variance:

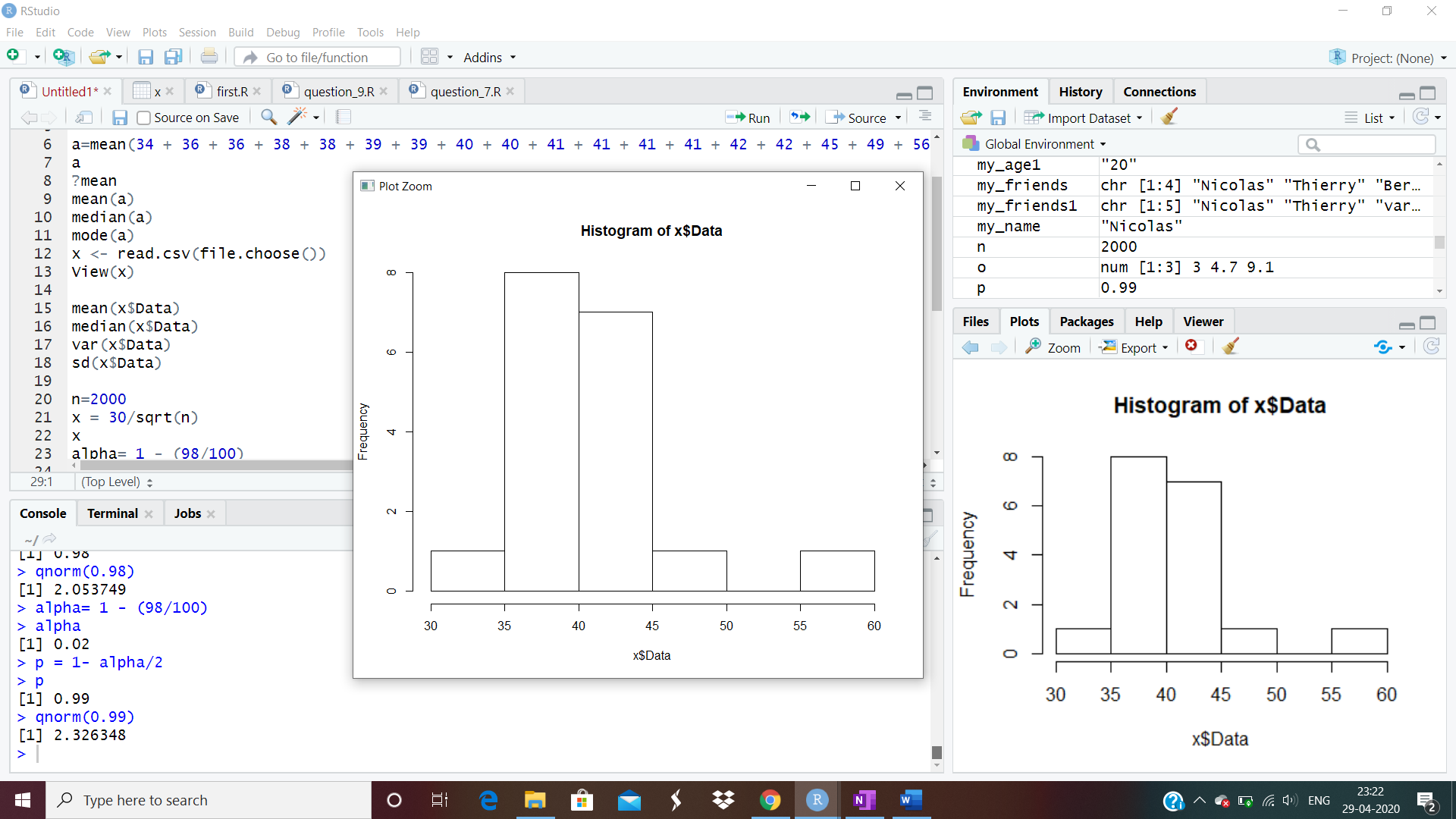
Var = 25.52941

1. Standard Deviation:

Standard deviation = 5.052664

1. What can we say about the student marks?

Ans: We can say about student marks are skewed data.



The student mark is uniformly distributed around the data average in a normal distribution curve.

In histogram the data is left side more and long right tail then it is positively skewed.

A positively skewed data set has its tail extended towards the right.

It is an indication that both the mean, median and the mode are all different.

Mean > Median, this implies that the distribution is slightly skewed towards right.

No outliers are present.

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

Ans:

If the mean, median, and mode are approximately equal to each other, the distribution can be assumed to be approximately symmetrical.

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

Ans:

If the mean > median then the distribution will be skewed to the right.

Then it is positively skewed.

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

Ans:

If the median > mean then the distribution will be skewed to the left.

Then it is negatively skewed.

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

Ans:

A distribution with a positive kurtosis value indicates that the distribution has

heavier tails and a sharper peak and less variation than the normal distribution.

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

Ans:

A distribution with a negative kurtosis value indicates that the distribution has lighter tails and a less peakness (Broad peak) and more variation than the normal distribution.

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



Q. What can we say about the distribution of the data?

Ans:

It is not a Normal Distribution.

The distribution is skewed to the left, most values are 'large', but there are a few exceptionally small ones. Those exceptional values will impact the mean and pull it to the left, so that the mean will be less than the median.

The box plot will look as if the box was shifted to the right so that the left tail will be longer, and the median will be closer to the right line of the box in the box plot.

Q. What is nature of skewness of the data?

Ans:

The mean is less than the median, then the distribution is negatively skewed.

skewness is negative, the data are negatively skewed or skewed left.

Meaning is that the left tail is longer. It is left skewed.

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

Inter Quartile Range = Upper Quartile- Lower Quartile

IQR = 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:

A box plot visualization allows to examine the distribution of data. One box plot appears for each attribute element.

Each box plot displays the minimum, first quartile, median, third quartile, and maximum values.

1) The median of the two boxplots are same approximately 260.

2) The boxplots are not skewed in positive or negative direction.

3) Outliers does not exist in both of boxplots.

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:

mean = 34.42208

std\_dev = 9.131445

P(MPG>38) = 1-pnorm(38,34.42208,9.131445)

= 0.3475941

P(MPG<40) = pnorm(40,34.42208,9.131445)

= 0.7293497

P (20<MPG<50) =

pnorm(50,34.42208,9.131445)-pnorm(20,34.42208,9.131445)

=0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: Follows Normal distribution as indicated by qq-plot.



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 follows Normal Distribution from the below QQ-plot

qqnorm(wc\_at$Waist)

qqline(wc\_at$Waist)



Adipose Tissue follows normal distribution

qqnorm(wc\_at$AT)

qqline(wc\_at$AT)



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

Ans:

Z score of 90% Confidence Interval =

Alpha = 1 + Confidence Level/2

= 1 + 0.90 /2

= 1.9

= 0.95

So, Z score of 0.95 is 1.64

Z score of 94% Confidence Interval =

Alpha = 1 + Confidence Level /2

= 1 + 0.94 /2

= 1.94 / 2

= 0.97

So, Z score of 0.97 is 1.88

Z score of 60% Confidence Interval =

Alpha = 1 + Confidence Level /2

= 1 + 0.60 /2

= 1.6 / 2

= 0.8

So, Z score of 0.97 is 0.84

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

Ans:

Sample size is 25

Degrees of Freedom = n-1

Degrees of Freedom= 25 -1 = 24

t score of 95% confidence interval =

conf int =1- alpha/100 = 0.05

1 - 0.05/2 = 0.975

t\_distribution(97.5,24) = 2.063899

t score of 96% confidence interval =

conf int =1- alpha/100 = 0.04

1 - 0.04/2 = 0.98

t\_distribution(98,24) = 2.171545

t score of 99% confidence interval =

conf int =1- alpha/100 = 0.01

1 - 0.051/2 = 0.995

t\_distribution(99.5,24) = 2.79694

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:

x = 260

population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

t = x – mean / (s/sqrt(n))

t = 260 – 270 / (90 /sqrt (18))

t = -10 / (90 / 4.24264)

t = -10 / 21.2132

t = - 0.4714045

t = - 0.471

pt(tscore,df)

pt (t, n-1)

pt(- 0.471, 17)

0.3216725

Probability = 32%

For probability calculations, the number of degrees of freedom is n - 1,

so here 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.