|  |  |
| --- | --- |
| 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 | Interval |
| Celsius Temperature | Interval |
| Weight | Interval |
| 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?

* **Answer:**

Total number of events = (HHH, HHT, HTH, HTT, TTT, TTH, THT, THH) = 8

Total number of Interested Events = (HHT, HTH, THH) = 3

Probability that two heads and one tail are obtained:

= (No. Interested Events) / (Total no. of events)

=3/8 = **0.375**

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

* **Answer:**

Total number of outcomes when two dices are rolled:

(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)}

Total number of events = 36

a) Sum is Equal to 1:

No. of events where sum = 1 are: No such event = 0

Therefore, probability that the sum is 1 when two dice are rolled is 0/36 = **0**.

b) Sum is less than or equal to 4:

No. of events where sum =<4 are:

{(1,1),(1,2),(1,3),(2,1),(2,2)(3,1)} = 6

Probability that sum<=4 is: 6/36 = 1/6 = **0.1666**

c) Sum is divisible by 2 and 3:

No. of events when sum is divisible by 2 and 3:

{(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)} = 6

Probability that sum is divisible by 2 and 3: 6/36 =1/6 = **0.1666**

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?

* **Answer:**

Total number of events = =

= =21

Interested events = = =10

Probability that none of the balls is blue =

(Interested Events) / (Total number of events) = 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

* **Answer:**

Expected Value = E(x)

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

Expected Value = **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**

* **Answer:**

****

**For Points:**

* **Mean: df.Points.mean()= 3.5965**
* **Median: df.Points.median()= 3.6950**
* **Mode: df.Points.mode() = 0 3.07**

**1 3.92**

* **Variance: df.Points.var()= 0.2858**
* **Standard Deviation : df.Points.std() = 0.5346**
* **Range: Range= df.Points.max()-df.Points.min() = 2.17**

**For Score:**

* **Mean: df.Score.mean()= 3.2172**
* **Median: df.Score.mode()= 3.325**
* **Mode: df.Score.median()= 3.44**
* **Variance: df.Score.var()= 0.9573**
* **Standard Deviation : df.Score.std()=0.9784**
* **Range: Range = df.Score.max() - df.Score.min()=3.9110**

**For Weights:**

* **Mean: df.Weight.mean()= 17.8487**
* **Median: df.Weight.median()= 17.71**
* **Mode: df.Weight.mode()= 0 17.02**

**1 18.90**

* **Variance: df.Weight.var()= 3.1931**
* **Standard Deviation : df.Weight.std() = 1.7869**
* **Range: Range= df.Weight.max()-df.Weight.min() = 8.3999**

**Inferences:**

* For the Column “Points”, the data is Bi-modal, since it has 2 modes, 3.07 & 3.92. Here, we can see that the mean value is slightly greater than the median value.
* For the Column “Score”, the data has single mode. The median value is slightly greater than the mean value.
* For the column “Weight”, the data is Bi-modal, since it has 2 modes, 17.02 & 18.90. Here, the Mean value is slightly greater than 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?

* **Answer:**

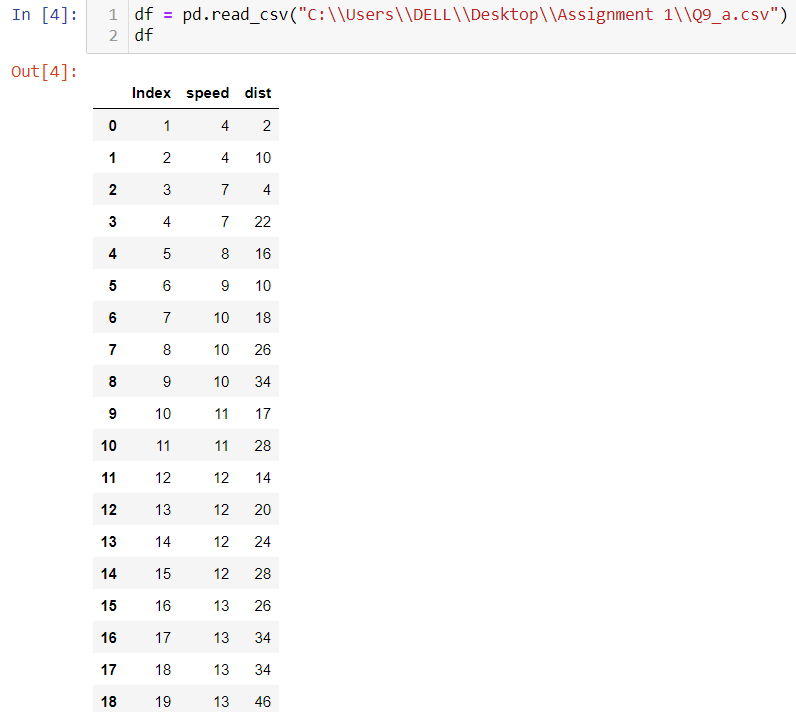
Expected Value =

Expected Value =

Expected Value= **145.33**

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

* **Answer:**
* **Cars’ Speed and Distance. (Use Q9\_a.csv)**



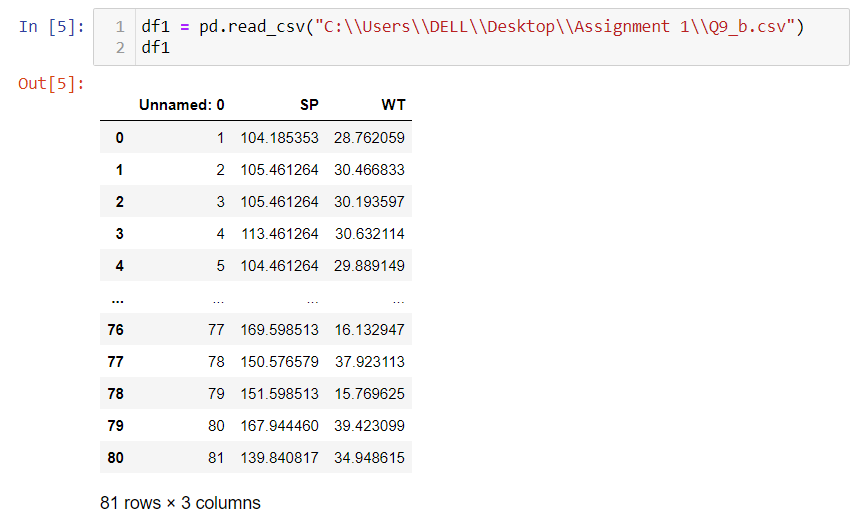
1. **For “Speed” Column:**

* **Skewness: df.speed.skew() = -0.1139**
* **Kurtosis: df.speed.kurt() = -0.5771**
* **Inference:** Here, skewness is -0.11 which indicates that the data is negatively skewed as tail dragging towards the left and kurtosis is -0.57 which is slightly

flatter than normal distribution.

1. **For “Distance” Column:**

* **Skewness: df.dist.skew()= 0.7824**
* **Kurtosis: df.dist.kurt() = 0.2480**
* **Inference:** Here, the skewness is 0.78 which indicates that the data is positively skewed as tail dragging towards the right and kurtosis is 0.24 which is slightly pointy than normal distribution.
* **SP and Weight (WT). (Use Q9\_b.csv)**

****

1. **For “SP” Column:**

* **Skewness: df.SP.skew() = 1.5814**
* **Kurtosis: df.SP.kurt() = 2.7235**
* **Inference:** Here, the skewness is 1.58 which indicates that the data is positively skewed as tail dragging towards the right and kurtosis is 2.72 which is very pointy than normal distribution.

1. **For “WT” Column:**

* **Skewness: df.WT.skew() = -0.6033**
* **Kurtosis: df.WT.kurt() = 0.8194**
* **Inference:** Here, the skewness is -0.60 which indicates that the data is negatively skewed as tail dragging towards the left and kurtosis is 0.81 which is slightly pointy than normal distribution.

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



* **Answer:**

**Histogram:** As we can see, from the histogram, the maximum observations (nearly 200) lie between ChickWeight range of 50-100, and the data has a tail (asymmetry in the distribution) on the right side of the distribution(if drawn). The data seems to be positively skewed (Right skewed).

Also, Approximate minimum point = 25

Approximate maximum point = 375

Range = 375-25 = 350 approximately.

**BoxPlot:** From the boxplot, we can conclude that the given data has more number of outliers in it and the data is positively skewed (Right Skewed).

**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?

* **Answer:**

Population Size = 3,00,000.

Sample Size = n = 2,000.

Sample mean = = 200

Sample Standard Deviation = = 30

Standard Error = / = 30 / = 0.6708.

1. **Confidence Interval for 94% :**

Z-score for (1-0.94)/2= (0.06)/2 = 0.03 probability is

Confidence Interval Estimate at 94% = Z

= 200 1.88\*(30/)

=200 1.88\*(0.6708)

= 2001.26

Range of the Confidence Interval Estimate at **94%** is **198.74-201.26.**

1. **Confidence Interval for 98%:**

Z-score for (1-0.98)/2= (0.02)/2 = 0.01 probability is

Confidence Interval Estimate at 94% = Z

= 200 2.33\*(30/)

= 200 2.33\*(0.6708)

= 200 1.56

Range of the Confidence Interval Estimate at **98%** is **198.54-201.56.**

1. **Confidence Interval for 96%:**

Z-score for (1-0.96)/2= (0.04)/2 = 0.02 probability is

Confidence Interval Estimate at 96% = Z

= 200 2.05\*(30/)

= 200 2.05\*(0.6708)

= 200 1.38

Range of the Confidence Interval Estimate at **96%** is **198.62-201.38.**

**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?

* **Answer:**

1. Mean = df.mean() = **41.0**

Median = df.median() = **40.5**

Variance = df.var() = **25.5294**

Standard Deviation = df.std() = **5.0526**

1. **About Students Marks :**

Here, we can see that Mean > Median, this implies that given data is a positively skewed data meaning the tail is dragged more on the right side. Most of the values end up being left of the mean. Most extreme values in the distribution fall on the right side. Also, there are no outliers present in the data.

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

* **Answer:** The data is said to be normally distributed when Mean = Median = Mode. There is no skewness in the data.

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

* **Answer:** When the Mean > Median, the data is said to be positively skewed and there is a tail on the right side of the curve.

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

* **Answer:** When the Median > Mean, the data is said to be negatively skewed and there is a tail on the left side of the curve.

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

* **Answer:** Positive Kurtosis value for a data indicates that the curve is more peaked meaning it has a sharp peak and there is less variation in the data.

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

* **Answer:** Negative Kurtosis value for a data indicates that the curve is less peaked meaning it has a broad or flat peak and there is more variation in the data.

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



What can we say about the distribution of the data?

* **Answer:** It is not a normally distributed data.

What is nature of skewness of the data?

* **Answer:** The data is negatively skewed meaning left skewed.

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

* **Answer:**

IQR is given by Q3 – Q1.

Here, Q3 = 18 & Q1 = 10

Therefore, 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.

* **Answer:**

Looking at both the boxplots, we can conclude that:

1) The median value of the two boxplots are approximately similar & equal to 260.

2) Both the boxplots do not show skewness.

3) The range of data in Boxplot 1 is lesser than the range of the data in Boxplot 2 meaning the spread of the values in Boxplot 2 is much wider than the spread of the values in Boxplot1.

4) There are no outliers in any of the given Boxplots.

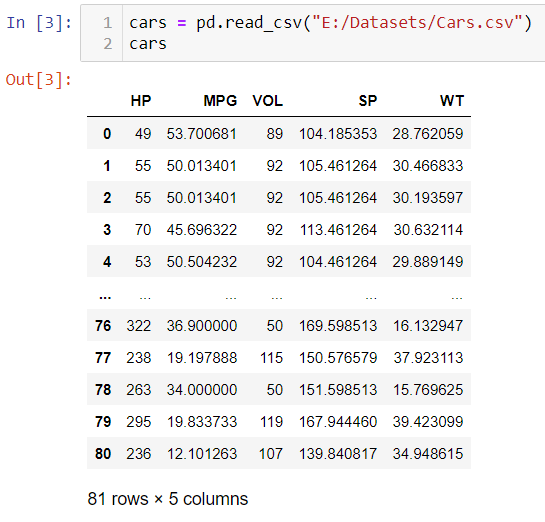
Q 20) Calculate probability from the given dataset for the below cases

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)
* **Answer:**

****

* **Mean** = = cars['MPG'].mean() = **34.4220**
* **Standard Deviation** = = cars['MPG'].std() = **9.1314**

1. **P(MPG>38) :**

Python code :

**1-stats.norm.cdf(38, loc = cars.MPG.mean(), scale = cars.MPG.std())**

**P(MPG>38) = 0.3475**

1. **P(MPG<40):**

Python code:

**stats.norm.cdf(40, loc = cars.MPG.mean(), scale = cars.MPG.std())**

**P(MPG<40) = 0.7293**

1. **P(20<MPG<50):**

Python code:

**stats.norm.cdf(50, loc = cars.MPG.mean(), scale = cars.MPG.std()) - stats.norm.cdf(20, loc = cars.MPG.mean(), scale = cars.MPG.std())**

**P(20<MPG<50) = 0.8988**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

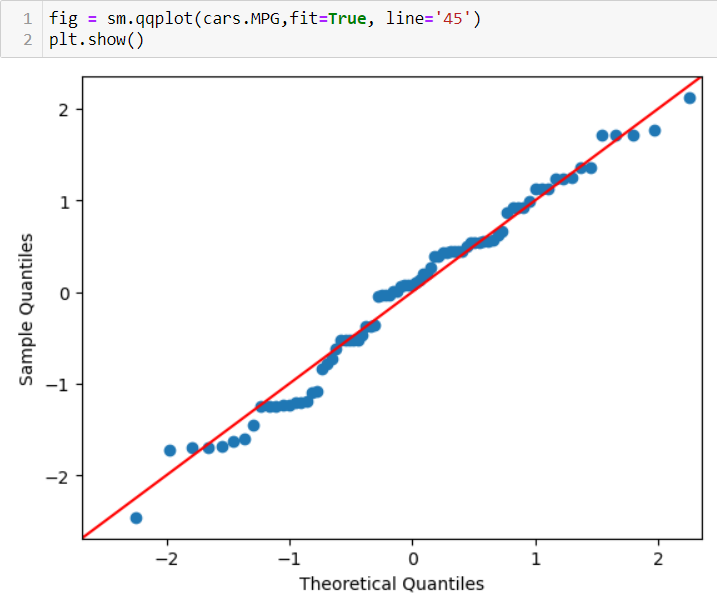
(Dataset: Cars.csv)

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

* **Answer:**

1. **Dataset:** Cars.csv

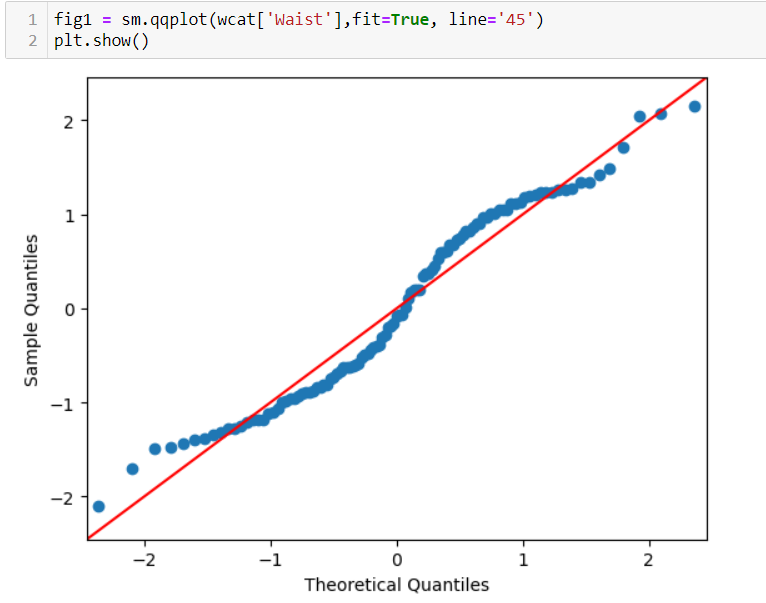
* Q-Q Plot for MPG Column:

****

As we can see from the above Q-Q plot that maximum data points in the MPG column lie on a straight line. This means that the data is normally distributed.

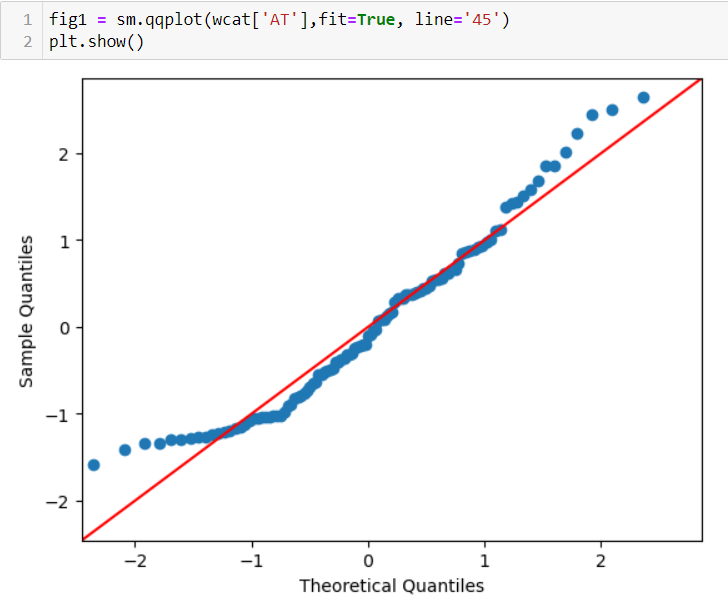
1. **Dataset:** wc\_at.csv

* Q-Q plot for Waist Circumference (Waist):



Here in the wc\_at.csv file, we can see from the Q-Q Plot for the Waist column that maximun data points lie on a straight line. Therefore, the data in the Waist column is normally distributed.

* Q-Q plot for the Adipose Tissue (AT):



Also, for the AT column in the wc-at.csv file, the data points when plotted using a Q-Q Plot, seem to lie on a single line. Therefore, the data in the AT column is normally distributed.

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

* **Answer:**

**Z-Score** for**:**

1. **90**% Confidence Interval:

Z-Score at (1+0.**90**)/2 = 1.90/2 = 0.95 is **1.64**

1. **94**% Confidence Interval:

Z-Score at (1+0.**94**)/2 = 1.94/2 = 0.97 is **1.88**

1. **60**% Confidence Interval:

Z-Score at (1+0.**60**)/2 = 1.60/2 = 0.80 is **0.84**

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

* **Answer:**

Here, Sample Size = **n** = **25**.

**df** = **n-1** = 25-1 = **24**

**T-Score for:**

1. **95**% Confidence Interval:

T-score (95% , df = 24) is **2.0638**

1. **96%** Confidence Interval:

T-score (96%, df = 24) is **2.1715**

1. **99%** Confidence Interval:

T-score (99%, df = 24) is **2.7969**

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

* **Answer :**

**Given:**

Sample Mean = = 260

Population mean = **µ** = 270

Standard Deviation = = 90

Sample size = **n** = 18

Degrees of Freedom = **df** = n-1 = 17

**t** = = = **- 0.47**

Therefore, using python code, we shall find the probability value for the t-score of -0.47 such that bulbs would have an average life of no more than 260 days.

Python code:

stats.t.cdf(t,df=17)

P = **0.3221**

P = **32.21 %**

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is **32.21 %.**