|  |  |
| --- | --- |
| 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 | 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: Total number of events= {hhh, hht, htt, ttt, tth, thh, hth, tht} =8

Interested events=3

Probability=3/8.

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

1. Equal to 1  
   Ans: 0/36 = Zero
2. Less than or equal to 4  
   Ans: 6/36 = 0.1666 = 16.66%
3. Sum is divisible by 2 and 3  
   Ans: 24/36 = 0.6666 = 60.66%

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:  (5/7)\*(4/6)=20/42=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 can be calculated using the formula

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

=  3.09

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.



Ans:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | Range | Variance | Standard Deviation |
| Points | 115.09/32 = 3.596 | = {(n+1)2} = (32+1)/2 = 33/2 = 16.5th value = (3.69+3.7)/2= 7.39/2 = 3.695 | 3.92 | Max-Min= (4.93-2.76) = 2.17 | 0.277 | 0.526 |  |
|  |  |  |  |  |  |  |  |
| Score | 102.952/32 = 3.217 | {(n+1)2} = (32+1)/2 = 33/2 = 16.5th value = (3.215+3.435)/2= 6.65/2 = 3.325 | 3.44 | Max-Min= (5.424-1.513) = 3.911 | 0.927 | 0.963 |  |
|  |  |  |  |  |  |  |  |
| Weigh | 571.16/32 = 17.849 | {(n+1)2} = (32+1)/2 = 33/2 = 16.5th value = (17.6+17.82)/2 = 35.42/2 = 17.71 | 17.02 | Max-Min= (22.9-14.5) = 8.4 | 3.093 | 1.759 |  |

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: Expected Value=(108+110+123+134+135+145+167+187+199)/9 = 145.333

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

**Cars speed and distance**



Ans: NOTE: SOLUTION IS CALCULATED BY CODING IN PYTHON I HAVE ATTACHED THE PYTHON CODE ALONG WITH THIS DOCUMENT

Skewness(speed) = -0.11750986144663393

Ie., Speed is having Negative Skewness

Skewness(dist) = 0.8068949601674215

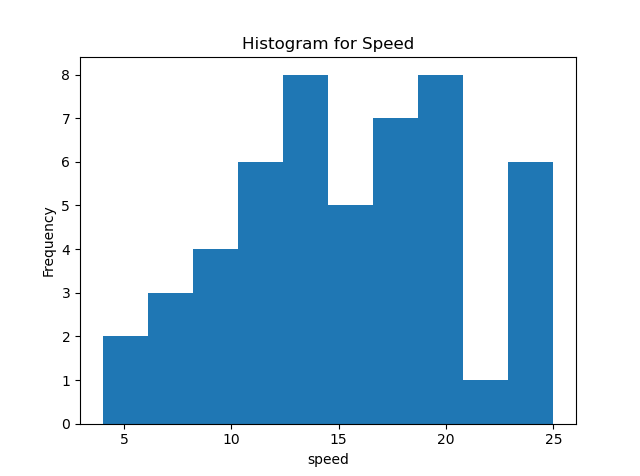
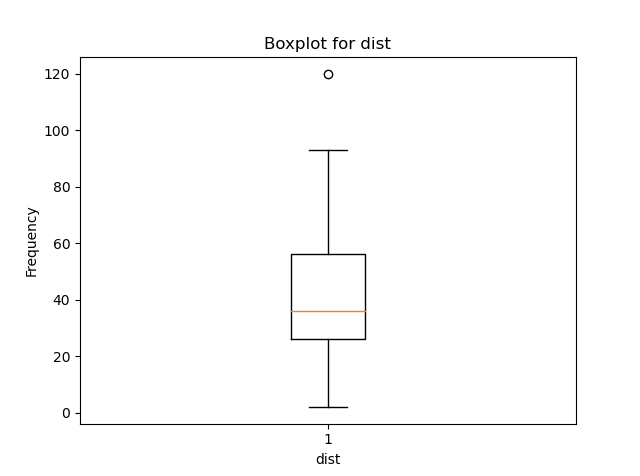
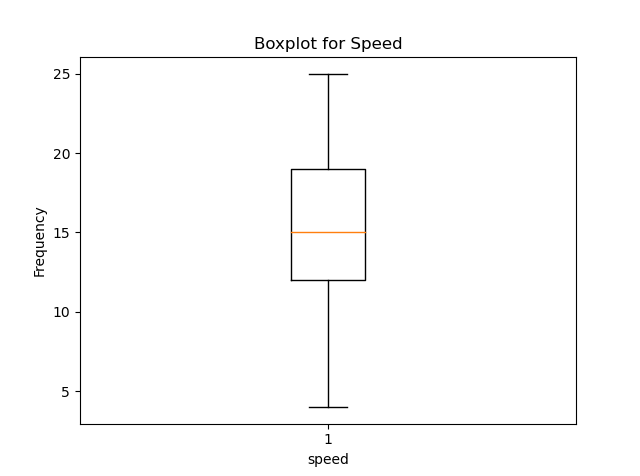
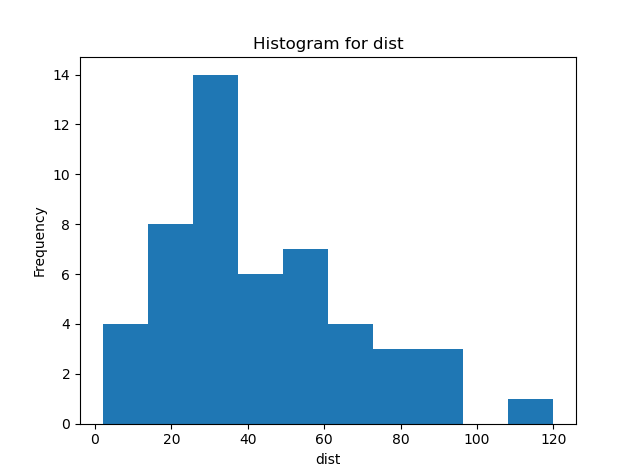
Ie., Dist is having Positive Skewness

Kurtosis(speed) = -0.5089944204057617

Ie., speed is having Negative Kurtosis

Kurtosis(dist) = 0.4050525816795765

Ie., Dist is having Positive Kurtosis

****

**SP and Weight(WT)**



Skewness(SP) = 1.6114501961773555

Ie., SP is having Positive Skewness

Skewness(WT) = -0.6147533255357768

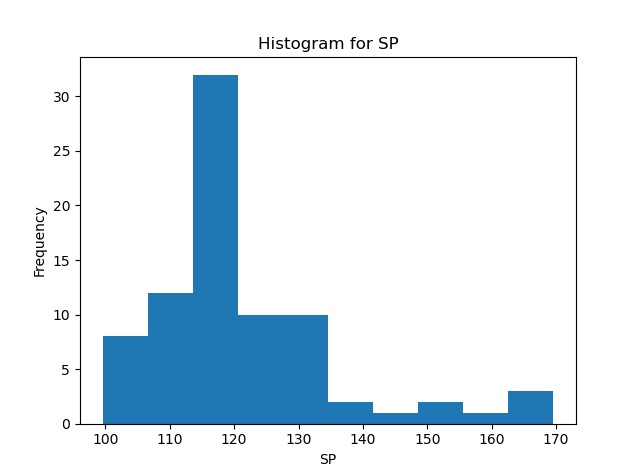
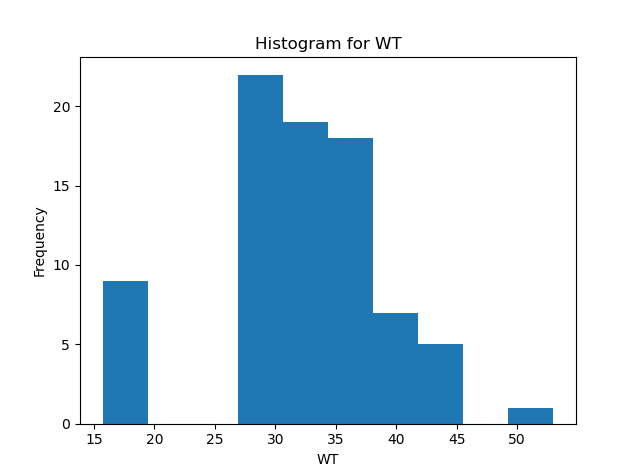
Ie., WT is having Negative Skewness

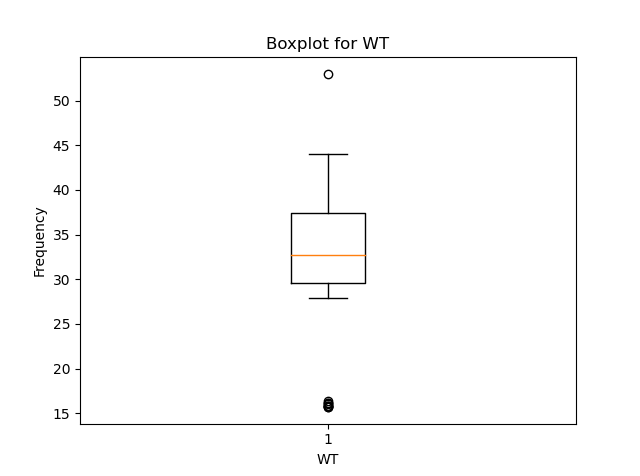
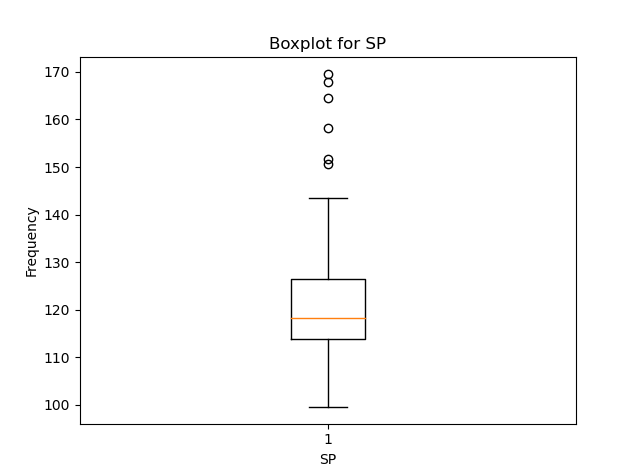
Kurtosis(SP) = 2.9773289437871764

Ie., SP is having Positive Kurtosis

Kurtosis(WT) = 0.9502914910300326

Ie., WT is having Positive Kurtosis

****

****

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



Ans:

* The histogram of ChickWeight$weight is Right Skewed and Positive Kurtosis.
* The Box Plot is used to identify outliers in the data set. The outliers are above upper extreme in case of the figure above and the lower whisker length is smaller than the upper whisker length.

**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:

sample statistic = 200 pounds

The standard error = 30/(square root of 2000) = 0.671

* For confidence interval = 94%

Compute alpha (α): α = 1 - (confidence level / 100) = 1-(94/100)=0.06

* critical probability = p\* = 1 - α/2 = 1 - 0.06/2 = 0.97
* Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 2000 - 1 = 1999
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.97. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), critical value = 1.882

Compute margin of error (ME): ME = critical value \* standard error = 1.882 \* 0.671 = 1.26

* The range of the confidence interval is defined by the *sample statistic* + *margin of error*. And the uncertainty is denoted by the confidence level. Therefore, this 95% confidence interval = 200 + 1.26
* For confidence interval = 98%

Compute alpha (α): α = 1 - (confidence level / 100) = 1-(98/100)=0.02

* critical probability = p\* = 1 - α/2 = 1 - 0.02/2 = 0.99
* Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 2000 - 1 = 1999
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.99. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), critical value = 2.328

Compute margin of error (ME): ME = critical value \* standard error = 2.328 \* 0.671 = 1.562

* The range of the confidence interval is defined by the *sample statistic* + *margin of error*. And the uncertainty is denoted by the confidence level. Therefore, this 95% confidence interval = 200 + 1.562
* For confidence interval = 96%

Compute alpha (α): α = 1 - (confidence level / 100) = 1-(96/100)=0.04

* critical probability = p\* = 1 - α/2 = 1 - 0.04/2 = 098
* Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 2000 - 1 = 1999
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.98. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), critical value = 2.055

Compute margin of error (ME): ME = critical value \* standard error = 2.055 \* 0.671 = 1.379

* The range of the confidence interval is defined by the *sample statistic* + *margin of error*. And the uncertainty is denoted by the confidence level. Therefore, this 95% confidence interval = 200 + 1.379

**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: Mean = 738/18 = 41

Median = 81/2 = 40.5

Variance = 434/18 = 24.11

Standard Deviation = sqrt(24.11) = 4.91

1. What can we say about the student marks?

Ans:

* Least Score obtained in the test is 34.
* Highest score obtained in the test is 56.
* The average score that students obtained in the test is 41.
* Students Score range between 34 and 56.

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

Ans: Symmetric (or) Zero Skewness

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

Ans: Positively Skewed

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

And: Negatively Skewed

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

Ans: The positive kurtosis value for a data indicates that the distribution has heavier tails and a sharper peak than the normal distribution.

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

Ans: The negative kurtosis value for a data indicates that the distribution has lighter tails and a flatter peak than the normal distribution.

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)?   
  
  
Ans:

* The distribution is left skewed(Negatively Skewed) and data lies between 1 and 19. 50% of the data lies between 10 and 18.2, Median = 15.2
* Left skewed (Negatively Skewed)
* IQR = 18.2 - 10 = 8.2

Q19) Comment on the below Boxplot visualizations?



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

Ans: Boxplot 1

* The data set lies between 240 and 288
* Minimum = 240, Maximum = 288, Median = 263, Q1= 253, Q3= 278
* Data is normally distributed

Boxplot 2

* The data set lies between 187 and 338
* Minimum= 187, Maximum=338, Median= 263, Q1=220, Q3=310
* Data is 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)

Ans: 31/81 = 0.383

* 1. P(MPG<40)

Ans: 61/81 = 0.753

* 1. P (20<MPG<50)

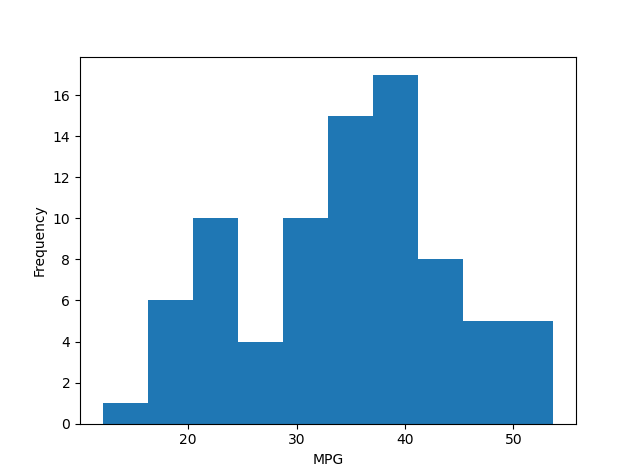
Ans: 69/81 = 0.852

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:

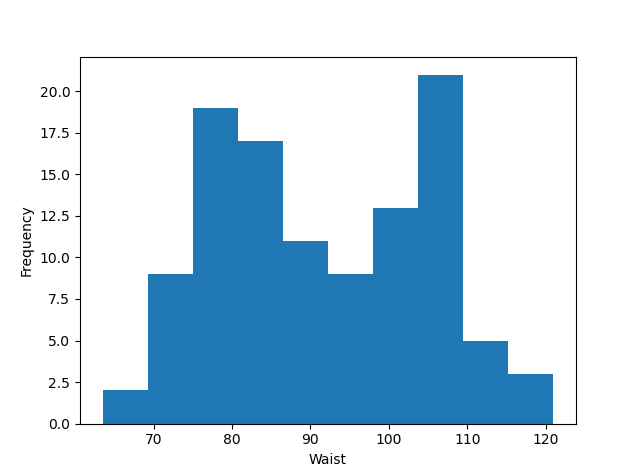


The data is not normally distributed for MPG data

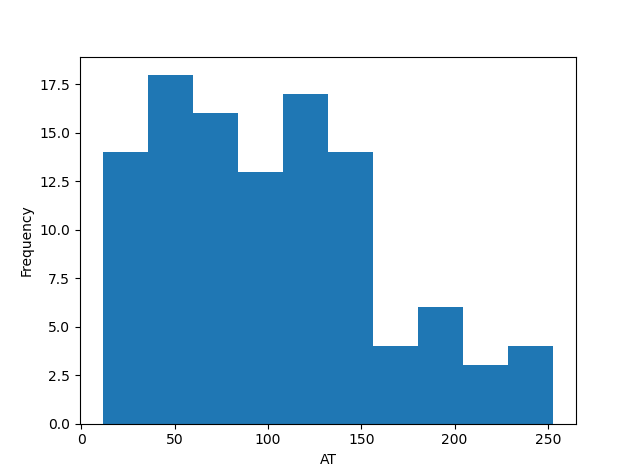
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Ans:



The data is not normally distributed for Waist data



The data is not normally distributed for Adipose Tissue data

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

Ans: Z Scores for 90% confidence interval = (1+0.90)/2 = 0.95

from Z Score table for 0.95, Z Score = 1.64

Z Scores for 94% confidence interval = (1+0.94)/2 = 0.97

from Z Score table for 0.97, Z Score = 1.88

Z Scores for 60% confidence interval = (1+0.6)/2 = 0.8

From Z Score table for 0.8, Z Score = 0.84

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

Ans:

* t scores for 95% confidence interval = (1-0.95)/2 = 0.025

To find the degree of freedom(df) = sample size - 1 = 25-1=24,

from t score table for 0.025 one tail and df of 24, Z Score = 2.064

* t scores for 96% confidence interval = (1-0.96)/2 = 0.02

To find the degree of freedom(df) = sample size - 1 = 25-1=24,

from t score table for 0.02 two tail and df of 24, Z Score = 2.492

* t scores for 99% confidence interval = (1-0.99)/2 = 0.005

To find the degree of freedom(df) = sample size - 1 = 25-1=24,

from t score table for 0.005 one tail and df of 24, Z Score = 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:

sample mean(x) = 260

population mean(μ) = 270

standard deviation of the sample(s) = 90

sample size(n) = 18

t statistic (t)= [ x - μ ] / [ s / sqrt( n ) ]  
t = ( 260 - 270 ) / [ 90 / sqrt( 18) ]  
t = -10 / 21.21 = - 0.47147

The degrees of freedom = 18 - 1 = 17.

t statistic = - 0.47147.

According to T Distribution Calculator, the cumulative probability = 0.3217. Hence, if the true bulb life were 270 days, there is a 32.17% chance that the average bulb life for 18 randomly selected bulbs would be less than or equal to 260 days.