1)Identify the Datatypes for the following, whether discrete or continuous?

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

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal - discrete |
| High School Class Ranking | Ordinal - discrete |
| Celsius Temperature | Interval-continuous |
| Weight | Ratio-continuous |
| Hair Color | Nominal-discrete |
| Socioeconomic Status | Ordinal-continuous |
| Fahrenheit Temperature | Interval-continuous |
| Height | Ratio-continuous |
| Type of living accommodation | Nominal-discrete |
| Level of Agreement | Ordinal-discrete |
| IQ(Intelligence Scale) | Interval-discrete |
| Sales Figures | Ratio-discrete |
| Blood Group | Nominal-discrete |
| Time Of Day | Ordinal-continuous |
| Time on a Clock with Hands | Interval-continuous |
| Number of Children | Ratio-discrete |
| Religious Preference | Nominal-discrete |
| Barometer Pressure | Ratio-continuous |
| SAT Scores | Ratio-continuous |
| Years of Education | Ratio-discrete |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

**A)Total no of outcomes=8**

**Favourable outcome=(H,H,T),(T,H,H),(H,T,H)=3**

**SO,probability= 1/8+1/8+1/8=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

Sol**: a) Equal to 1—There is no outcomes which is equivalent sum is one,**

**i.e,0/36=0, probability=0.**

**b) Less than or equal to 4**

**the possible outcomes are (1,3),(3,1)(2,2)(1,2)(1,1)(2,1)**

**probability= 6/36=1/6**

1. **Sum is divisible by 2 and 3**

**The possible outcomes are =(1,5),(5,1)(2,4),(4,2),(3,3)(6,6)=4**

**Probability=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?

**Sol: total no of balls=7**

**Probability=5c2/7c2=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

Sol: **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\*0.120= 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**

**Sol: points:**

**Mean=3.596563,median=3.695,mode=”numeric”, variance=0.285881,Standard Deviation = 0.5346787, Range=2.76 - 4.93.**

**Score:**

**Mean=3.21725, median=3.325, mode=”numeric”, variance=0.957379,**

**Standard Deviation = 0.978457, Range= 1.513 – 5.424.**

**Weigh:**

**Mean= 17.84875, median= 17.71, mode=”numeric”, variance= 3.193166,**

**Standard Deviation = 1.798122 , Range=14.5 – 22.9.**

**Note: Mean value are closer for both points and score.**

**Hence, points and score are closer to each other**

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?

Sol: **Expected Value  =  ∑ ( probability  \* Value )**

**∑ P(x).E(x)**

**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**

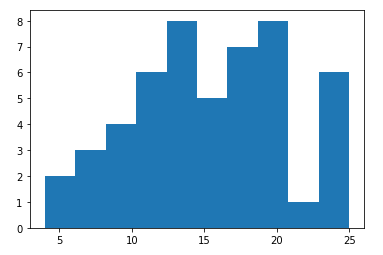
**Use Q9\_a.csv**

**Sol:** For Speed

Skewness for speed= -0.117509861,

Kurtosis for speed= -0.50899442

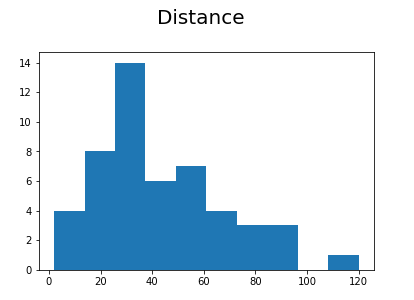
Here, skewness value is negative, so it is left skewed which negatively skewed. Speed



For Distance

Skewness for distance = 0.80689496

Kurtosis for distance = - 0.405052582

For distance skewness is positive, so it is positively skewed. 

**SP and Weight(WT)**

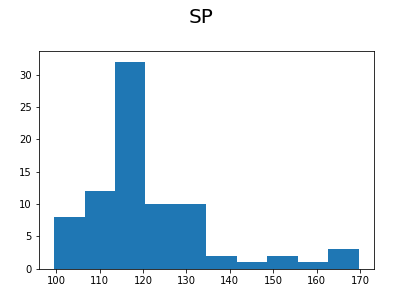
**Use Q9\_b.csv**

**Sol:** For SP

Skewness for SP= 1.61145

Kurtosis for SP= 2.977329

Here, skewness value is positive, so it is right skewed which positively skewed.

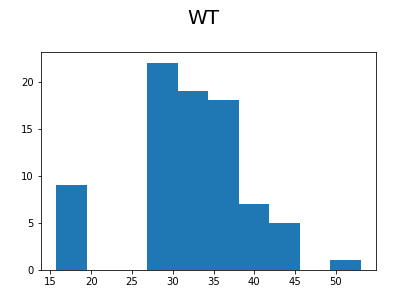


For Weight

Skewness for Weight = -0.61475

Kurtosis for Weight= 0.950291

For distance skewness is negative, so it is negatively skewed.



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



**Sol:** The histogram represents right skewed data, so it is positively skewed data.

And the Boxplot representes

There is no outlier in the lower extreme whereas more outliers on the upper extreme and lower whisker contains less data whereas upper whisker contains more data and Lower quartile is closer to median whereas upper quartile is little far if compared with each other.

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?

**Sol:**

**1)** **Make sure the sample is normal and is a simple random sample.**

 Is **n(the sample size) > 30**? i.,e 2000 > 30 yes

2) **Identify a sample statistic**. Since we are trying to estimate the mean weight in the population, we choose the mean weight in our sample (200) as the sample statistic.

3)confidence interval: 94%

Standard Error of the mean(SE)=s/sqrt(n)

SE=30/sqrt(2000)=0.67

Finding Critical value:

Compute α= (1 – Confidence level)/100=0.06

the critical propability:1-α/2=1-0.06/2=0.97

degrees of freedom(df)=n-1=2000-1=1999

a) **For (94%)**

The critical value of the ‘t’ score having 1999 degrees of freedom and cumulative probability is 0.97. From the T distribution Calculator we find the critical value is 1.88

**margin of error (ME)**: ME = critical value \* standard error

= 1.88 \* 0.67 = 1.2596 =1.26

Therefore,

The confidence interval for 94 %= sample statistics±Margin of Error

= 200±1.26

b)  **For (96%)**

The critical value of the ‘t’ score having 1999 degrees of freedom and cumulative probability is 0.97. From the T distribution Calculator we find the critical value is 2.05

**margin of error (ME)**: ME = critical value \* standard error

= 2.05 \* 0.67 = 1.3735 =1.38

Therefore,

The confidence interval for 94 %= sample statistics±Margin of Error

= 200±1.38

c) **For (98%)**

The critical value of the ‘t’ score having 1999 degrees of freedom and cumulative probability is 0.97. From the T distribution Calculator we find the critical value is 2.33

**margin of error (ME)**: ME = critical value \* standard error

= 2.33 \* 0.67 = 1.5611 =1.57

Therefore,

The confidence interval for 94 %= sample statistics±Margin of Error

= 200±1.57

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?

Sol: Mean = 41

Median = 40.5

Variance =25.5294

Standard Deviation =5.05266

Mean and median values of student scores are very closer. And it is positively skewed .

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

Sol: When mean and median of data are equal, then the distribution is symmetric, and the distribution has zero **skewness**. If the distribution is both symmetric and unimodal, then the **mean** = **median** = mode.

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

Sol: If the **mean is greater than** the **median**, the distribution is positively **skewed.**

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

Sol: If the **median is greater than** the **mean**, the distribution is negatively **skewed.**

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

Sol: **Positive values** of **kurtosis indicate** that a distribution is peaked and possess thick tails.

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

Sol: **Negative values** of **kurtosis indicate** that a distribution is flat and has thin tails.

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)?

Sol: 1) By observing the data it is left skewed, so it the **median is greater than** the **mean.**

**2) And the boxplot is negatively skewed.**

**3)The IQR of the data will be -8 apporox.**

Q19) Comment on the below Boxplot visualizations?



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

Sol: : By observing both the plots whisker’s level is high in boxplot 2, mean and

median are equal hence distribution is symetrical.

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) b.P(MPG<40) c. P (20<MPG<50)

Sol: By the Given data for MPG µ = 34.42 and sigma=9.13

a)P(MPG>38), z-score corresponding to 38=34.42-38/9.13=0.39,

Calculate the probability using Z table P(Z> -0.39) =1 - 0.6517=0.3483

b)P(MPG<40), z-score corresponding to 40=34.42-40/9.13=-0.61,

Calculate the probability using Z table P(Z< 40) = 0.25

c)P(20<MPG<50), z-score corresponding to P(20 < MPG < 50)

=P(MPG<50) – P(MPG <20)

P(MPG<50)=34.42-50/9.13= - 1.70 = 0.0446

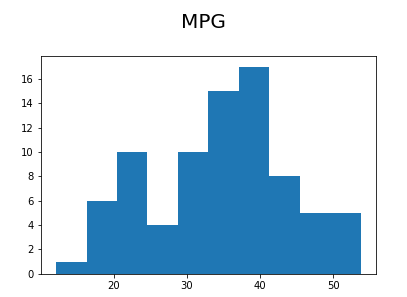
P(MPG < 20) = 34.42-20/9.13= 1.57 = 0.9418

Calculate the probability using Z table P(20<MPG<50) = 0.0446 – 0.9418 = -0.89

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

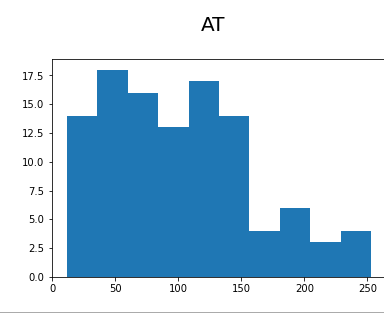
Sol:As per the histogram chart, it is not symmertric and it is not equally distributed.So it doesn’t follows Normal Distribution.



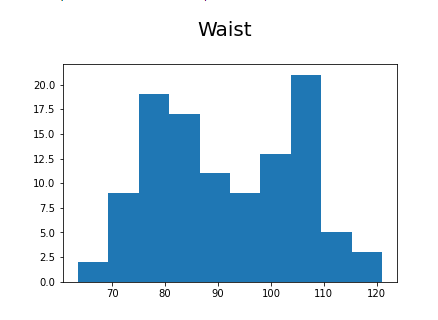
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Sol:



By observing the above graph, it is not symmetric , and Adipose Tissue (AT) doesn’t follows Normal Distribution .



By observing the above graph, it is not symmetric , and Waist Circumference(Waist) doesn’t follows Normal Distribution .

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

Sol: a) Z scores of 90% confidence interval

1 – α = 90% , α = 10%=5%+5%

So,90+5 = 95 i.,e 0.95, Z score of 0.95 =1.645.

b) Z scores of 94% confidence interval

1 – α = 94% , α = 6%=3%+3%

So,94+3 = 95 i.,e 0.97, Z score of 0.97 =1.88.

c) Z scores of 60% confidence interval

1 – α = 60% , α = 40%=20%+20%

So,60+20 = 80 i.,e 0.80, Z score of 0.80 =0.85

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

Sol: Sample size =25, therefore degrees of freedom = 25-1=24.

Confidence interval = 95%=0.975. Then, the t-score is 2.064

Confidence interval = 96%=0.98. Then, the t-score is 2.172

Confidence interval = 99%=0.995. Then, the t-score is 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

Sol: 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.