|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| 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 data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Discrete 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 | Interval |
| Celsius Temperature | Interval |
| Weight | Interval |
| Hair Color | nominal |
| Socioeconomic Status | nominal |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| Type of living accommodation | Ordinal |
| Level of Agreement | Nominal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Interval |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ordinal |
| Years of Education | Ordinal |

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

Sample space= {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Probability= (no of favorable outcomes)/(total no: of outputs)

P (two heads and one tail) =3/8

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

1. Equal to 1

P(sum=1)=0

1. Less than or equal to 4

P(less than or equal to 4)=6/36=1/6

1. Sum is divisible by 2 and 3

P(sum is divisible 2 and 3)=10/36

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?

P(none of the balls drawn is blue)=20/21

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Child | A | F | C | B | D | E |
| Candies count(x=) | 1 | 2 | 3 | 4 | 5 | 6 |
| probability | .015 | .120 | .65 | .20 | .005 | .01 |
| p\*x | .015 | .240 | 1.95 | .80 | .025 | .06 |

*Expected number of candies for random selected child=.015+.24+1.95+.8+.025+.06=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**

***Points***

*Mean=*3.596563

*Mode=* 3.92

*Median=* 3.695

*Range=* 2.17

***Score***

*Mean=* 3.21725

*Mode=* 3.44

*Median=* 3.325

*Range=* 3.911

***Weight***

*Mean=* 17.84875

*Mode=* 17.02

*Median=* 17.71

*Range=* 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?

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weights(x) | 108 | 110 | 123 | 134 | 135 | 145 | 167 | 187 | 199 |
| P(Weights) | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 |
| p\*x | 12 | 12.22 | 13.66 | 14.88 | 15 | 16.11 | 18.55 | 20.77 | 22.11 |

*Expected value=∑p\*x=12+12.22+13.66+14.88+15+16.11+18.55+20.77+22.11 =145.3*

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Cars speed :**

*Skewness(sample)= -0.11751*

*Skewness(population)= -0.11395*

*Kurtosis= -0.50899*

***Skewness*** *helps to find the distribution of data point across the range of speed .for case of car speed ,The skewness is negative value indicating that most of the data point are located to the right of the distribution curve. It is also known as left skewed or negative skewed.*

*Kurtosis helps to find whether the Distribution plot is pointed or flat (ie: deviation from normal distribution curve ) . Here in the case of speed of the car we get negative kurtosis which indicate that the distribution plot is flat and has thin tail or light tailed compared to normal distribution plot.*

***Distance:***

*Skewness(sample)= 0.806895*

*Skewness(population)= 0.782484*

*Kurtosis= 0.405053*

*In case of distance the* ***skewness*** *is observed to be positive value indicating that the data points are concentrated to the left of distribution plot compared to normal distribution plot.*

*Kurtosis is also a positive value in the case of distance which indicate that the distribution is peaky or peaked. that there is a high frequency of data points getting concentrated at or close to the mean value*

**SP and Weight(WT)**

**SP:**

*Skewness(sample)= 1.61145*

*skewness(population)= 1.581454*

*Kurtosis= 2.977329*

*Here skewness is positive or right skewed which indicate that data points are concentrated towards left side of the distribution curve.*

*Kurtosis is also positive value which indicate that the distribution curve is highly pointy or heavily pointed. (i.e.: high frequency of data point is observed for particular range of interval.*

**WT:**

*Skewness(sample)= -0.61475*

*Skewness(population)= -0.60331*

*Kurtosis= 0.950291*

*Here skewness is negative indicating that data points are concentrated toward right side of distribution curve. It is a left skewed curve.*

*Here kurtosis is a positive value which indicates that the distribution is pointed compared to normal distribution curve.*

**Use Q9\_b.csv**

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



*The histogram shows that chick weight range of 50-100 has highest frequency,it helps to determine the frequency distribution of the weight across certain intervales.*

*From the box plot we are able to find the median as also the maximum and minimum value of chick weight.the box plot also gives details about chick weight and its distribution across the certain range using q1,q2,q3 which are quartile 1(25% percentile) ,quartile 2(median ) ,quartile 3(75% percentile)*

**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?

For 94% confidence interval=[198.68-201.31]

For 98% confidence interval = [198.34 – 201.65]

For 96% confidence interval = [198.55 – 201.44]

**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.

*Mean=41*

*Median=40.5*

*Variance=24.11111*

*Standard deviation=4.910307*

1. What can we say about the student marks?

*The students of class have scored an average of 41 marks in the test. The highest mark scored is 56 and the lowest mark reported is 34.*

*Range is observed to be 22.*

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

*We will get a normal distribution curve, when mean, median and mode becomes equal, that is ,skewness will be zero or we can say no skewness*.

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

*When mean is greater than median we will get positive skewness.*

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

*When median is greater than mean we will get negative skewness*

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

*Positive kurtosis indicate that there is a high frequency for particular range variable .the graph will contain a pointy peak in case of positive kurtosis.it shows the variation of height of normal distribution*

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

*In case of negative kurtosis, it show how heavily tailed the data is along the distribution curve. the graph will be having a lower peak than the normal distribution curve.*

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



What can we say about the distribution of the data?

*From the given boxplot we can infer that the majority of data points are concentrated between 10-18 and the mean of the data points is at 15 .Here the q1 or quartile 1 shows that 25%percentile of the data points lie below 10 and q2 or the second quartile depicts the mean of the data set and q3 or the 3rd quartile shows that 75% percentile of the data points lie below 18.and finally the iqr or inter quartile range shows that most of the data points lies between 10 and 18.*

What is nature of skewness of the data?

*Nature of skewness helps to determine symmetry of the distribution curve, weather the data is concentrated along the right sided or left oriented from the normal distribution curve.*

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

*Inter quartile range it is the difference between the 25th and 75th percentile of the data.in simple language it depicts the 50 percent of the data, it also helps to find the outliners.it tells about the spread of middle half of the data points.*  
  
  
Q19) Comment on the below Boxplot visualizations?



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

*Here comparing the 2 box plots we can infer that :-*

*Box plot 1 is having more consistent values*

* *Box plot 2 is having the greatest no of variety or unique data points compared to box plot 1 as it is having the biggest range value. Box plot 1 is having less no of unique points.*
* *The range of box plot 1 is between (287-237)=50 which is lesser when compared to box plot 2 range (337-187)=150 which is 3 time the range of plot 1.*
* *Box plot 2 is having the highest and the lowest data point values.*
* *Box plot 2 is having greater no of data point below 25% percentile as it has a broader whisker similar for the case of 75% percentile also.*
* *Even though they are having almost the same mean but still the box plot show a larger Deviation.*

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.

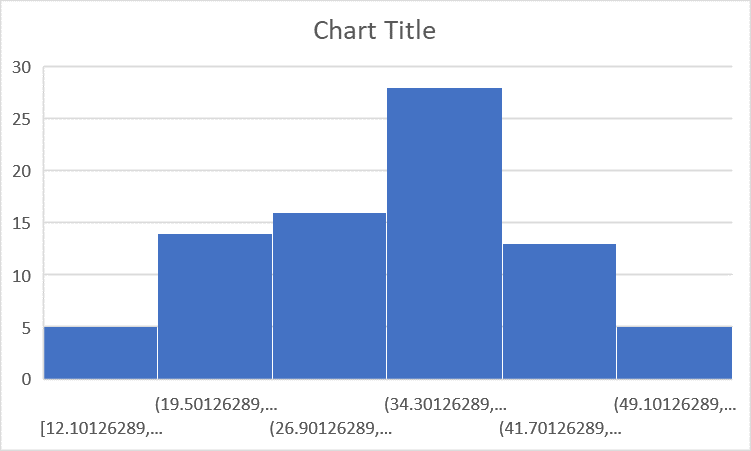
MPG <- Cars$MPG

* 1. P(MPG>38)= 33/81=.407
  2. P(MPG<40) = 61/81=.753
  3. P (20<MPG<50)=69/81=.851

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

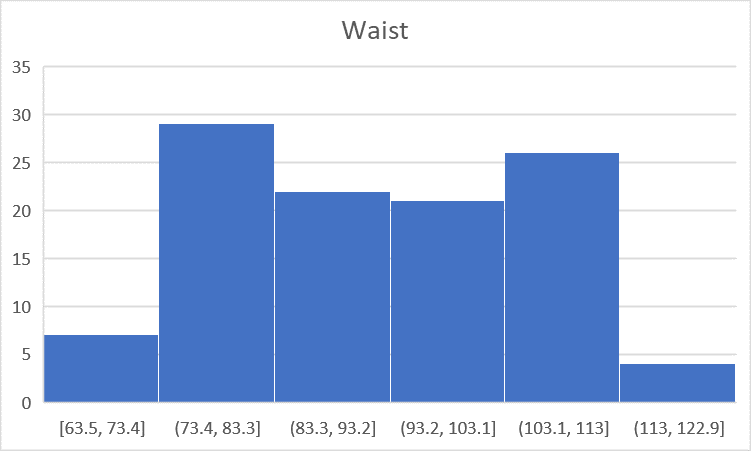
Dataset: Cars.csv

Here from the skewness and kurtosis value we can infer that they lie within the acceptable range to be almost normally distributed (accepted skewness to be normal distribution=, kurtosis = .the mean mode and the median are almost equal with mode a bit less.

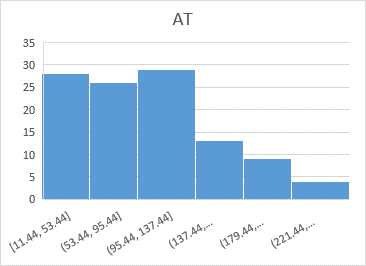
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

From the above study show that the data set of Waist Follows Almost normal distribution curve. With skewness lying within the range of and kurtosis also in the range of .

And also the mean median and mode are almost equal.



Even though the skewness and kurtosis are within the range the mean ,median and mode are having greater deviation so we can infer that the data set does not follows normal distribution.

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

*We can find the Z-score using the formula area=(1+confidence interveal)/2 and finding the currosponding value in the z table*

*Z-score for 90% confidence interval =1.645*

*Z-score for 94% confidence inerval=1.88*

*z-score for 90% confidence interval=.84*

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

*t-score of 95%=2.063*

*t-score of 96%=2.171*

*t-score of 99%=2.796*

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

*we Know t=*

*t=-.47*

*p(average life not more than 260)=stats.t.cdf(-.47,df=17)*

*= 0.32*