|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continuos |
| Weight of Gold | Continuos |
| Distance between two places | Continuos |
| Length of a leaf | Continous |
| Dog's weight | Continous |
| 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 | Intervals |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Intervals |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Nominal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Intervals |
| Number of Children | Ratio |
| 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?

SOL:-C=COIN

|  |  |  |
| --- | --- | --- |
| C1 | C2 | C3 |
| H | H | H |
| T | T | T |
| H | T | T |
| T | H | T |
| T | T | H |
| T | H | H |
| H | T | H |
| H | H | T |

SO, WE GOT TOTAL 24 OUTCOME FROM THREE COINS AND 9 ARE THOSE WHICH HAVE TWO HEAD AND ONE TAIL. SO, THE PROBABILITY IS 9/24

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

1. Equal to 1

SOL:- THERE IS IMPOSSIBLE IN THIS CASE THAT THE SUM OF TWO ROLLED DICE IS 1.

1. Less than or equal to 4

SOL:-(1,1)(1,2)(1,3)(2,1)(2,2)(3,1)

PROBABILITY =6/36 =1/6

1. Sum is divisible by 2 and 3

SOL:-(1,5)(2,4)(3,3)(4,2)(5,1)(6,6)(1,5)(2,4)(3,3)(4,2)(5,1)(6,6)

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:-2RED=R1&R2 3GREEN=G1,G2&G3 2BLUE=B1&B2

PROBABILITY=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

SOL:- E(X)=SUM(X\*P(X)) X=CANDIES

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

E(X)=3.09

MEAN=3.09/6=0.515

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

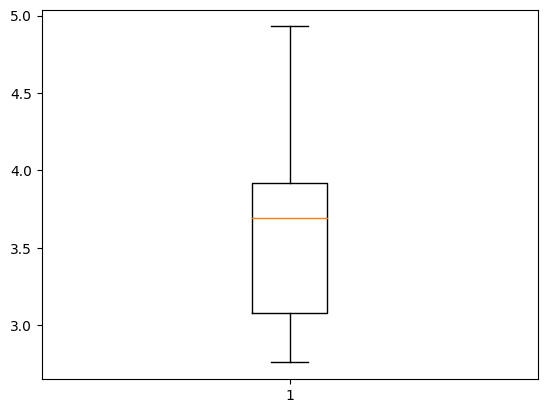
* For Points,Score,Weigh>

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

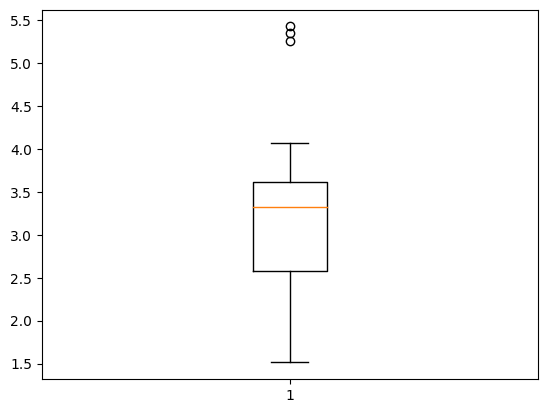
**Use Q7.csv file**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **MEAN** | **MEDIAN** | **MODE** | **VARIANCE** | **STD.DEV** | **RANGE** |
| **POINTS** | 3.596563 | 3.695 | 3.07 | 0.285881 | 0.534679 | 2 |
| **SCORE** | 3.21725 | 3.325 | 3.44 | 0.986138 | 0.978457 | 4 |
| **WEIGH** | 17.84875 | 17.71 | 17.02 | 3.193166 | 1.786943 | 8 |

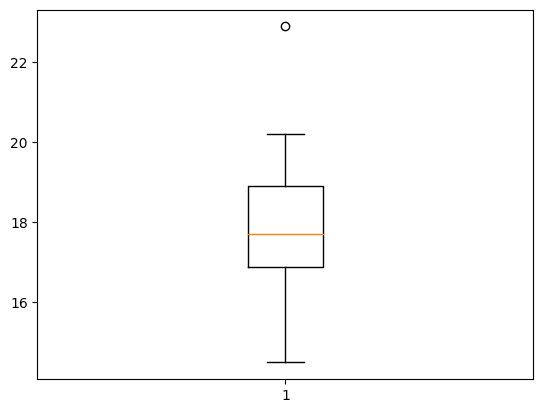
**SOL:-**



THIS IS THE BOXPLOT FOR POINTS COLUMN IN DATA.HERE WE CAN SEE THAT THIS IS LEFT SKEWED DISTRIBUTION BECAUSE THE GAP BETWEEN MEDIAN LINE AND LOWER EDGE IS MORE THAN OTHERS .

****

THIS IS THE BOXPLOT FOR SCORE COLUMN IN DATA.HERE WE CAN SEE THAT THIS IS LEFT SKEWED DISTRIBUTION BECAUSE THE GAP BETWEEN MEDIAN LINE AND LOWER EDGE IS MORE THAN OTHERS AND THIS COLUMNS DATA HAVE SOME OUTLIERS ON THE UPPER SIDE.



THIS IS THE BOXPLOT FOR WEIGH COLUMN IN DATA.HERE WE CAN SEE THAT THIS IS RIGHT SKEWED DISTRIBUTION BECAUSE THE GAP BETWEEN MEDIAN LINE AND UPPER EDGE IS MORE THAN OTHERS AND THIS COLUMNS DATA HAVE SOME OUTLIERS ON THE UPPER SIDE.

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**:- MEAN=SUM OF ALL ELEMENTS/NO. OF ELEMENTS

MEAN=145.33

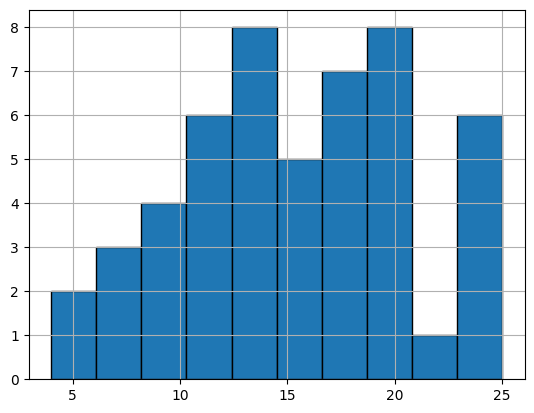
ACTUALLY WE DON’T KNOW WHICH PATIENT HAS ACTUALLY SELECTED,SO WE CALCULATE THE MEAN(AVERAGE) OF THE ENTIRE DATA SO THAT WE GET AN APPROXIMATION IDEA.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SOL:- FOR SPEED**

****

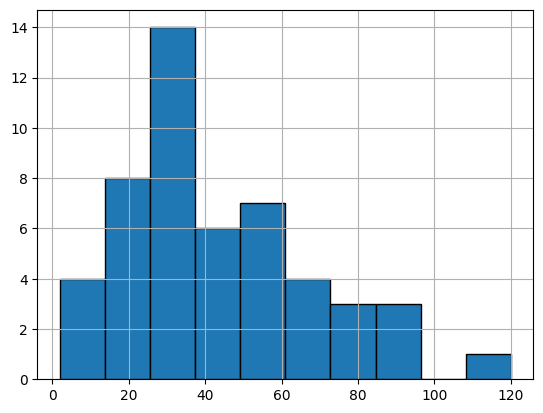
1.THE SKEWNESS VALUE IS -0.1175086 WHICH IS BETWEEN -0.5 & 0.5 WHICH INDICATES THAT THIS IS SYMMETRIC

2. THE KURTOSIS VALUE IS -0.5089944, THIS IS LESS THAN 3 SO THIS IS PLATYKURTIC.

**-FOR DISTANCE**

1.THE SKEWNESS VALUE IS 0.806894 WHICH IS BETWEEN 0.5 & 1 WHICH INDICATES THAT THIS IS POSITIVELY OR RIGHT SKEWED.

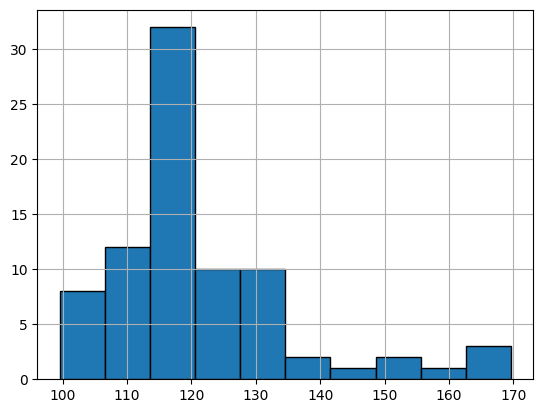
2. THE KURTOSIS VALUE IS 0.405052, THIS IS LESS THAN 3 SO THIS IS PLATYKURTIC.

****

**SP and Weight(WT)**

**Use Q9\_b.csv**

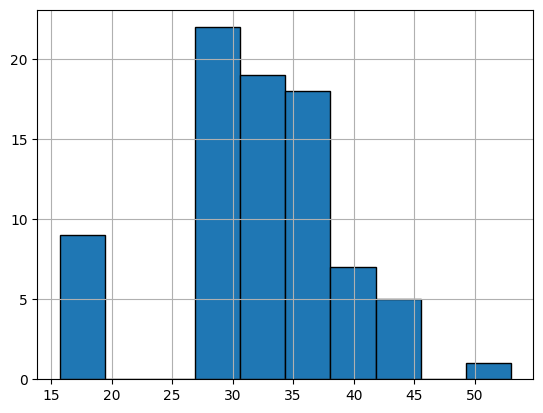
**FOR SP**

****

1. THE SKEWNESS VALUE IS 1.611450 WHICH INDICATTES THAT THE DISTRIBUTION IS HIGHLY POSITIVELY SKEWED OR RIGHT SKEWED.
2. THE KURTOSIS VALUE IS 2.97732 WHICH IS ALMOST EQUAL TO 3,SO THIS IS MESOKURTIC.

**FOR WEIGHT**

1. THE SKEWNESS VALUE IS -0.6147 WHICH INDICATTES THAT THE DISTRIBUTION IS HIGHLY NEGATIVELY SKEWED OR LEFT SKEWED AS IT LIES BETWEEN -1 & -0.5.
2. THE KURTOSIS VALUE IS 0.9502, WHICH IS LESS THAN 3,SO THIS IS PLATYKURTIC.



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



**SOL:-** AS WE CAN SEE, THIS IS A HISTOGRAM WHERE IT IS RIGHT SKEWED OR POSITIVE SKEWED DATA BECAUSE IT HAS MORE DATA ON RIGHT SIDE OF THE TALLEST BAR WHICH MEANS, MEAN>MEDIAN>MODE.WE HAVE OUTLIERS ON THE HIGHER SIDE.



**Sol:-** THIS IS A BOXPLOT WHICH HAVE MANY OUTLIERS IN UPPER SIDE AND WE CAN ALSO SEE THAT THIS IS RIGHT SKEWED BECAUSE THERE IS MORE SPACE BETWEEN THE MEDIAN LINE AND THE UPPER EDGE RELATED TO THE OTHER SIDE ,AND WE HAVE LESSER GAP BETWEEN THE LOWER EDGE AND MEDIAN LINE THAT MEANS WE HAVE MORE DATA WITHIN SMALL RANGE.

**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?

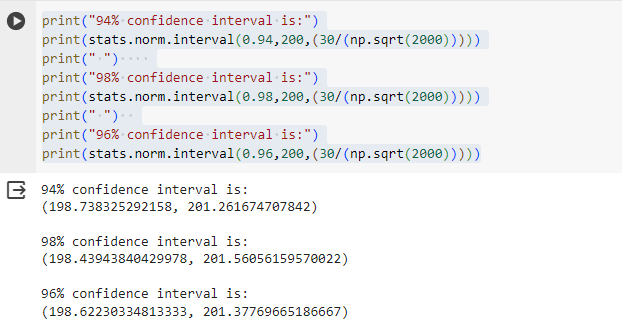
**SOL:-** POPULTION SIZE(N)=3,000,000

RANDOM SAMPLE(n)=2000, SAMPLE MEAN=200 POUNDS, SAMPLE STD DEV-30POUNDS

**CI AT 94%=(LL= 198.68)&(UP=201.31)**

**CI AT 98%=(LL=198.45) & (UP=201.54)**

**CI AT 96%=(LL= 198.68)&(UP=201.31)**

****

**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.

SOL:- MEAN=41

MEDIAN=40.5

VARIANCE=25.52

STD. DEV.=5.05

1. What can we say about the student marks?

SOL:- BY SEEING THE MEAN AND MEDIAN ARE NEARLY CLOSE,SO THIS IS SYMMETRICAL DISTRIBUTION AND THE MOST OF THE MARKS OBTAINED ARE NEARLY 40-42 WHICH IS NEAR THE MEAN VALUE OF THE DATA

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

**SOL:-** THE SKEWNESS WILL BE ZERO IN THIS CASE, THE DATA IS SYMMETRICAL

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

**SOL:-** THE SKEWNESS IS POSITIVELY SKEWED WHEN MEAN>MEDIAN

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

**SOL:-** THE NATURE WILL BE NEGATIVE SKEWED

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

**SOL:-** FOR POSITIVE KURTOSIS, MEANS THAT MORE PEAKED DISTRIBUTION,AND IT IS LEPTOKURTIC(KURTOSIS>3).MORE NUMBERS ARE LOCATED IN TAILS OF DISTRIBUTION INSTEAD OF NEAR THE MEAN

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

**SOL:-** FOR NEGATIVE KURTOSIS AND KURTOSIS<3, MEANS THAT IT IS PLATYKURTIC AND LESS PEAKED, STRETCHED AROUND CENTER.

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



What can we say about the distribution of the data?

**SOL:-** WE CAN SEE THAT ONE PART IS SMALLER AND ONE IS GREATER BEFORE AND AFTER THE MEDIAN LINE,SO IT’S CONFIRMED THIS IS UNEQUAL DISTRIBUTION.THIS IS LEFT SKEWED DATA BECAUSE Q1>Q3.

What is nature of skewness of the data?

**SOL:-** THIS IS LEFT SKEWED DATA BECAUSE Q1>Q3,THE GAP BETWEEN MEDIAN LINE AND THE LOWER EDGE IS MORE THAN THE OTHER SIDE.

What will be the IQR of the data (approximately)?   
  
SOL:- IF Q1=10 , Q3= 18.1

IQR=Q3-Q1=8.1

Q19) Comment on the below Boxplot visualizations?



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

**SOL:-** BOTH ARE THE BOXPLOT HAVING ALMOST SAME MEDIAN AND THERE IS NO OUTLIERS SO THEY ARE NORMALLY DISTRIBUTED BECAUSE WE CAN SEE THE GAP BETWEEN THE UPPER AND LOWER EDGE FROM THE MEDIAN IN BOTH BOXPLOT IS ALMOST SAME.

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)

**SOL:- 0.3476(BY USING NORM DIST)**

* 1. P(MPG<40)

**SOL:- 0.7293**

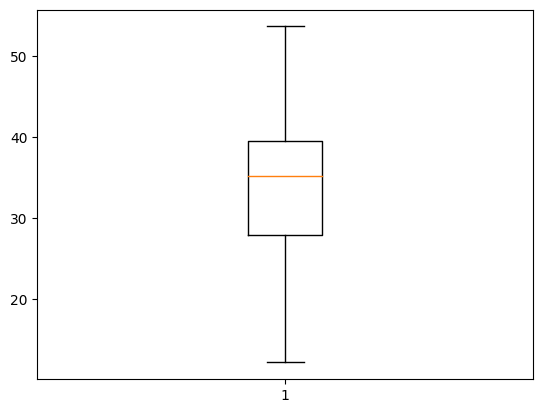
* 1. P (20<MPG<50)

**SOL:-0.0131**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

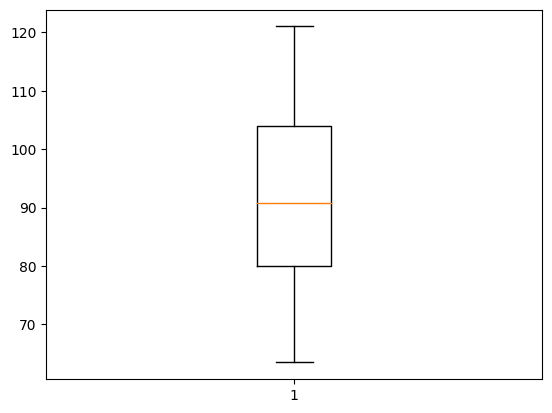


THIS IS NOT NORMAL DISTRIBUTION AS WE CAN SEE THE GAP BETWEEN UPPER EDGE AND LOWER EDGE FROM THE MEDIAN IS NOT SAME.THIS IS LEFT SKEWED AS THE GAP BETWEEN MEDIAN AND LOWER EDGE IS MORE.

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

Dataset: wc-at.csv

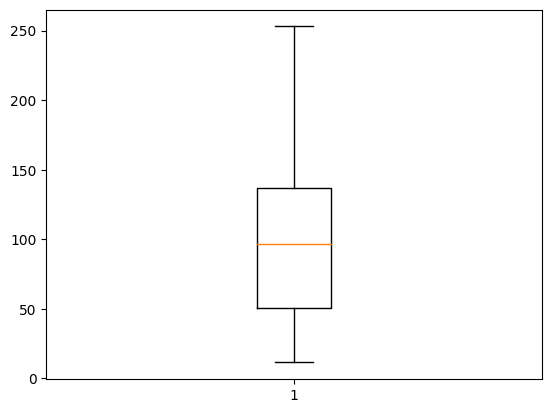
**FOR WAIST**



THIS IS NORMAL DISTRIBUTION AS WE CAN SEE THE GAP BETWEEN UPPER EDGE AND LOWER EDGE FROM THE MEDIAN IS SAME ALONG WITH THE WHISKER OF THE BOXPLOT

**FOR AT**

THIS IS NOT NORMAL DISTRIBUTION AS WE CAN SEE THE GAP BETWEEN UPPER EDGE AND LOWER EDGE FROM THE MEDIAN IS NOT SAME. THIS IS RIGHT SKEWED OR POSITIVE SKEWED.



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

**SOL:- Z score of 90% CI=1.645**

**Z score of 94% CI=1.5540**

**Z score of 60% CI= 0.253**

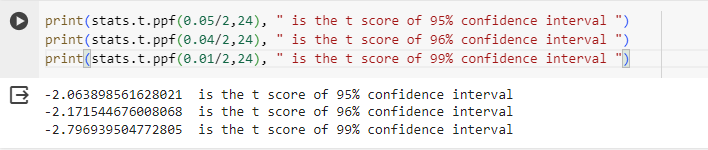
Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

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

**T\_SCORES AT 95% CI=2.064**

**T\_SCORES AT 96% CI=2.171**

**T\_SCORES AT 99% CI=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

NULL HYPOTHESIS:-THE CLAIM THAT BULB WILL NOT LAST 270 DAYS

ALTERNATE HYPOTHESIS:-THE CEO’S CLAIM THAT BULB LASTS 270 DAYS

n=18

population mean=270

sample mean=260

SAMPLE STD DEV=90DAYS

