Q1) Identify the Data type for the Following:

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 (Categorical)
Number of kids	Discrete
Number of tickets in Indian railways	Discrete
Number of times married	Discrete
Gender (Male or Female)	Discrete (Categorical)

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)	Interval
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Ordinal
Time on a Clock with Hands	Interval
Number of Children	Ratio
Religious Preference	Nominal
Barometer Pressure	Ratio
SAT Scores	Interval
Years of Education	Ratio

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

Answer: Let’s consider three tossed coins A, B & C, we can see total probabilities as shown in table below (H = Head & T = Tail)

Sr. no.	Coin A	Coin B	Coin C
1	H	H	H
2	H	H	T
3	H	T	H
4	H	T	T
5	T	H	H
6	T	H	T
7	T	T	H
8	T	T	T

From table we can see,

Total no. of possibilities = 8

Possibility of getting two heads and one tail = 3 (i.e. Sr. no. 2, 3 & 5)

So, probability that two heads and one tail are obtained = 3/8

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

a) Equal to 1 = **0 probability**
b) Less than or equal to 4 = **1/6**
c) Sum is divisible by 2 and 3 = **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?

Answer: Total no. of balls = 2 (red) + 3 (green) + 2 (blue) = 7
Number of ways drawing 2 random ball out of total 7 = 7_{C}^{2} = (7 x 6) / (2 x1) = 21
Number of ways drawing 2 random ball out of 5 (-2 blue) = 5_{C}^{2} = (5 x 4) / (2 x1) = 10

Probability that none of the balls drawn is blue = 5_{C}^{2} / 7_{C}^{2} = **10 / 21 = 0.4761**

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 number of candies for a randomly selected child = summation of Candies count x Probability of each child (i.e. Child A, B, C, D, E & F)

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

= 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.

Use Q7.csv file

Answer: Please refer the jupyter notebook file 'Assignment -1 .ipynb' for the answer.

Q8) Calculate Expected Value for the problem below

a) 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 can be calculated by below equation

Expected Value = Summation of (Probability * Values)

Probability = 1/total no. of patients = 1/9

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

= 12 + 12.22 + 13.66 + 14.88 + 15 + 16.11 + 18.55+ 20.77 + 22.11

Expected Value = 145.3

Hence, expected Value of the Weight of randomly chosen patient is 145.3

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

Cars speed and distance (Use Q9_a.csv)

SP and Weight(WT) (Use Q9_b.csv)

Answer:

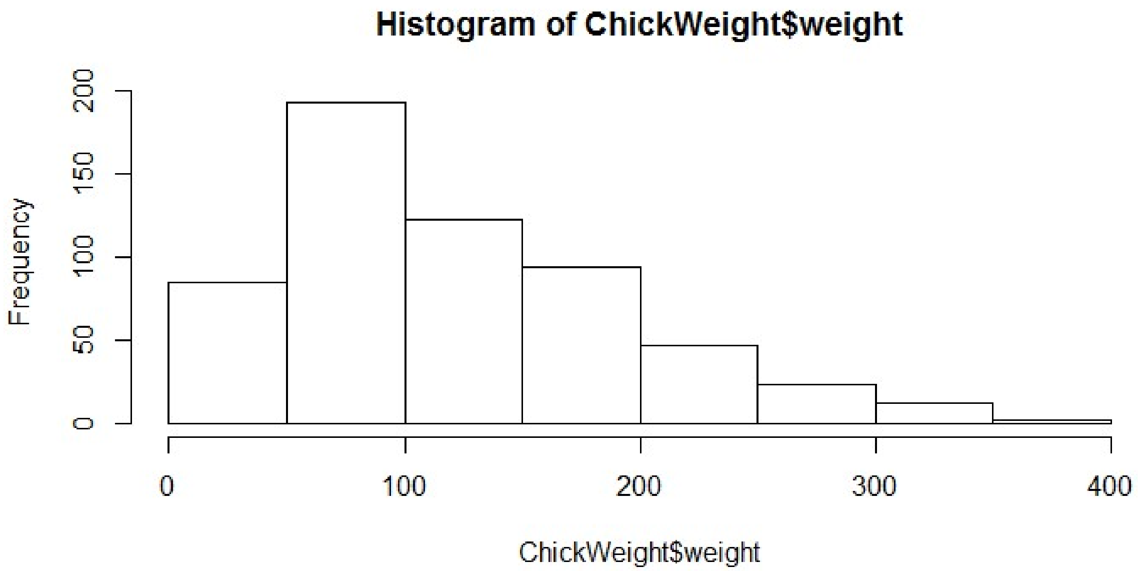
	Cars speed	Distance	SP	Weight(WT)
Skewness	-0.1175	0.8069	1.6115	-0.6148
Kurtosis	-0.509	0.4051	2.9773	0.9503

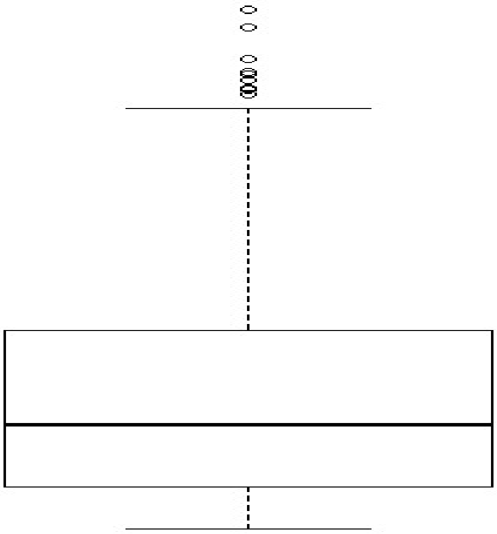
Inferences:

1) *The distribution of speed is slightly negatively skewed and playkurtic*
2) *The distribution of speed is slightly positively skewed and playkurtic.*
3) *The distribution of speed is positively skewed and playkurtic.*
4) *The distribution of speed is slightly negatively skewed and playkurtic.*

Please refer the jupyter notebook file ‘Assignment -1 .ipynb’ for the detail answer.

Q10) Draw inferences about the following boxplot & histogram





Answer:

Histogram

1) By looking at the above histogram it is clear that the data is not normally distributed but it is positively skewed or right skewed.
2) Most of the observations has the weight between 50 and 100.

Boxplot

1) There are some outliers in the data (closer to upper range).
2) Data is not Normally Distributed.
3) We can say that data distribution is 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:

$\bar{x} \pm z_{1-\propto}\frac{\sigma}{\sqrt{n}}$ → If standard deviation of the population is known

$\bar{x} \pm t_{1-\propto,n-1}\frac{s}{\sqrt{n}}$ → If standard deviation of the population is unknown

s = standard deviation of the sample = 30

n = 2000

$\bar{x}$ = average sample weight = 200

α = confidence interval = 94%, 98% & 96%

The average weight of an adult in Mexico with 94% Confidence Interval is: [198.7383 201.2617]

The average weight of an adult in Mexico with 98% Confidence Interval is: [198.4394 201.5606]

The average weight of an adult in Mexico with 96% Confidence Interval is: [198.6223 201.3777]

Please refer the jupyter notebook file ‘Assignment -1 .ipynb’ for the detailed answer.

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:

Mean (**μ**) = (34+36+36+38+38+39+39+40+40+41+41+41+41

+42+42+45+49+56)/18 = 738/18 = **41 = μ**

Median = (40+41) / 2 = **40.5**

Mode = **41**

Variance is calculated by the below formula

Variance = $\sigma^2 = \Sigma\, (X - \mu)^2/N$

where,

- σ^2 = population variance
- Σ = sum of…,
- X = each value
- μ = population mean
- N = number of values in the population

$\sigma^2 = [(34-41)^2 + (36-41)^2 + (36-41)^2 + (38-41)^2 + (38-41)^2 + (39-41)^2 + (39-41)^2 + (40-41)^2 + (40-41)^2 + (41-41)^2 + (41-41)^2 + (41-41)^2 + (41-41)^2 + (42-41)^2 + (42-41)^2 + (45-41)^2 + (49-41)^2 + (56-41)^2]\ /\ 18$

$= [(-7)^2 + (-5)^2 + (-5)^2 + (-3)^2 + (-3)^2 + (-2)^2 + (-2)^2 + (-1)^2 + (-1)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (1)^2 + (1)^2 + (4)^2 + (8)^2 + (15)^2\ ] / 18$

$= [49 + 25 + 25 + 9 + 9 + 4 + 4 + 1 + 1 + 0 + 0 + 0 + 0 + 1 + 1 + 16 + 64 + 225]\ /\ 18$

$= [434] / 8 = 24.11$

Variance = σ^2 = 24.11

Standard Deviation (**σ**) = Square root of σ2

σ = 4.91

Since, mean =mode=median (approx.) we can say that the students data of scores is normally distributed.

Average score of all the students is **41.**

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

Answer: If mean & median are equal then, there is a Symmetrical Distribution of the data, hence no skewness or zero skewness.

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

Answer: If mean > median then, data is Positively Skewed or Right Skewed.

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

Answer: If median > mean then, data is Negatively Skewed or Left Skewed.

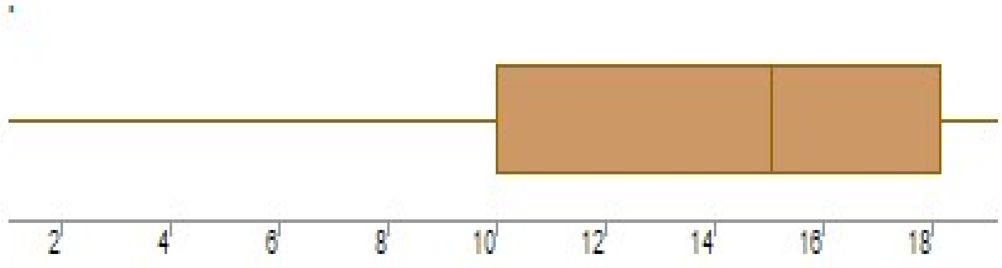
Q16) What does positive kurtosis value indicates for a data?

Answer: Positive kurtosis value indicates that a distribution is longer, tails are fatter. Also we can say that data is heavy-tailed or has outliers. It is also referred as Leptokurtic distribution.

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

Answer: Negative kurtosis value indicates that a distribution is shorter, tails are thinner than the normal distribution. Also we can say that data is light-tailed or has no outliers. It is also referred as Platykurtic 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)?

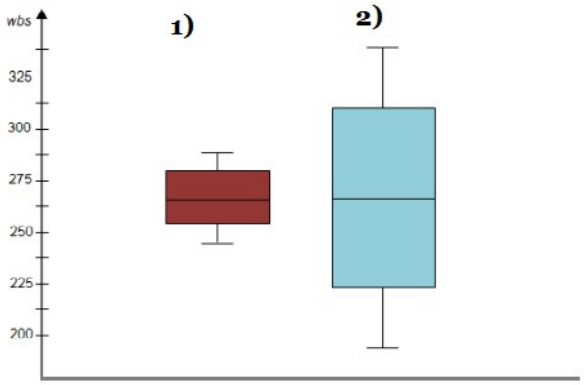
Answer:

The data is not Normally Distributed.

We can say that the data is Left Skewed or Negatively Skewed.

The Interquartile Range of the data is 10 to 18 approximately.

Q19) Comment on the below Boxplot visualizations?



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

Answer:

IQR of Boxplot 1 is less than IQR of Boxplot 2.

Median of both Boxplot 1 and Boxplot 2 is same.

Data of Boxplot 1 has a Positive Kurtosis value (Leptokurtic).

Data of Boxplot 2 follows Normal Distribution (Mesokurtic).

No outliers present in both the 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

a. P (MPG>38)
b. P (MPG<40)
c. P (20<MPG<50)

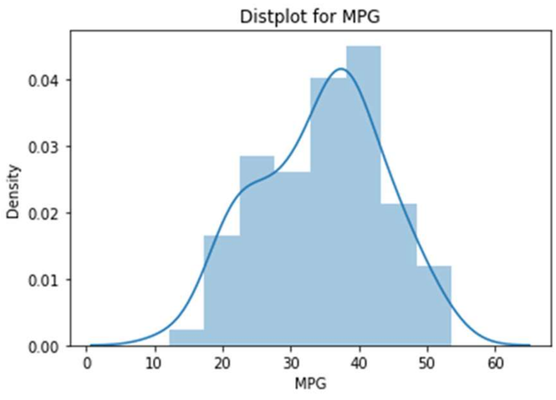
Answer: Please refer the jupyter notebook file ‘Assignment -1 .ipynb’ for the detailed answer.

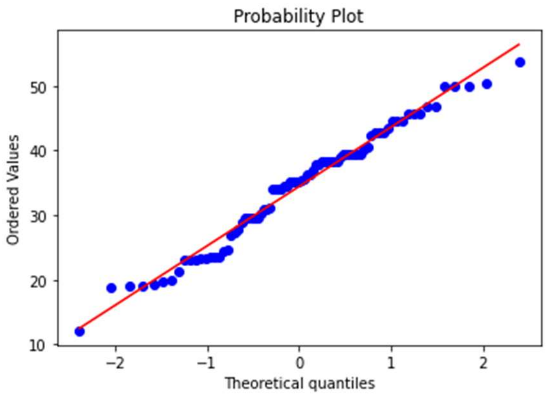
a) Probability of MPG > 38 is 0.3476 (34.76%)
b) Probability of MPG < 40 is 0.7293 (72.93%)
c) Probability of 20 > MPG > 50 is 0.8989 (89.89%)

Q 21) Check whether the data follows normal distribution

d) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

Answer: Since Mean, Median & Mode are approximately same, we can say that MPG of Cars follows Normal Distribution. By looking at distplot and probability plot we can say that MPG of Cars follows Normal Distribution.

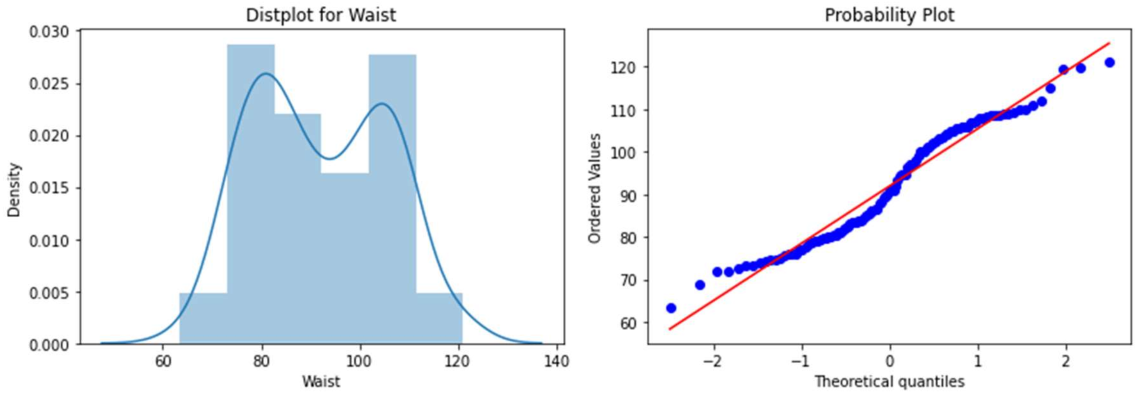


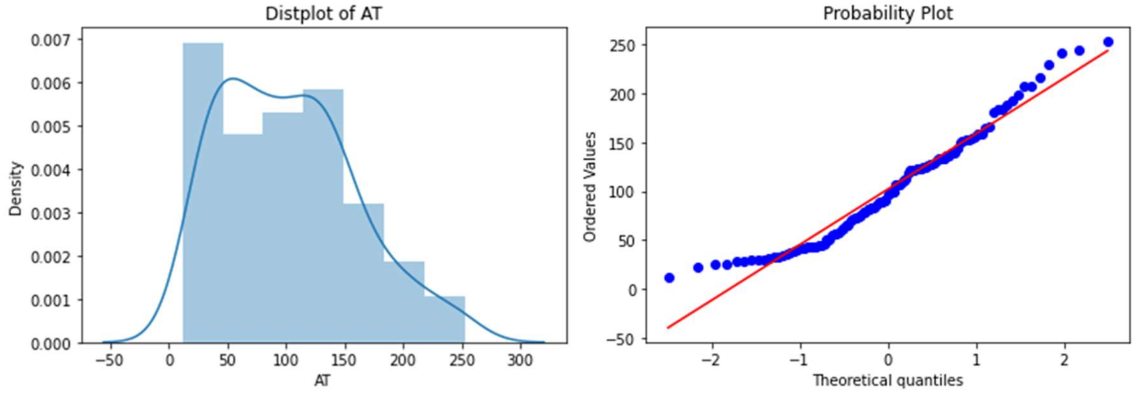


Please refer the jupyter notebook file ‘Assignment -1 .ipynb’ for the detailed answer.

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

Answer: By looking at the distplot and probability plot below it is clear that AT is fairly symmetrical whereas Waist from wc-at data set does not follow a Normal Distribution.





Please refer the jupyter notebook file ‘Assignment -1 .ipynb’ for the detailed answer.

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

Answer: For normal distribution, 90% confidence interval has two tails of

(100-90 = 10) → 10/2 = 5%

So, considering from 5% to **95%** leaves 90% in the middle. In the Similar way,

94% → (100-94 = 6) → 6/2 = 3% → **97%**

60% → (100-60=40) → 40/2 = 20% → **80%**

Let’s calculate Z score with the help of python

Z score of **90%** confidence interval is: (**-1.6449, 1.6449)**

Z score of **94%** confidence interval is: (**-1.8808, 1.8808)**

Z score of **60%** confidence interval is: (**-0.8416, 0.8416)**

Please refer the jupyter notebook file ‘Assignment -1 .ipynb’ for the detailed answer.

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

Answer: For normal distribution, 95% confidence interval has two tails of

(100-95 = 5) → 5/2 = 2.5%

So, considering from 2.5% to **97.5%** leaves 95% in the middle. In the Similar way,

96% → (100-96 = 4) → 4/2 = 2% → **98%**

99% → (100-99 = 1) → 1/2 = 0.5% → **99.5%**

Degree of Freedom = sample size – 1 = 25 – 1 = 24

Let’s calculate Z score with the help of python

By stats.t.ppf

t score of **95%** confidence interval is: **2.0639** (stats.t.ppf(0.975,df=24)

t score of **96%** confidence interval is: **2.1715** (stats.t.ppf(0.98,df=24)

t score of **99%** confidence interval is: **2.7969** (stats.t.ppf(0.995,df=24)

By stats.t.interval(CI,0,1)

t score of **95%** confidence interval is: stats.t.interval(0.95,25,loc=0,scale=1) =
(-2.059538552753294, 2.059538552753294)
t score of 96% confidence interval is: stats.t.interval(0.96,25,loc=0,scale=1) =
(-2.1665866344527562, 2.1665866344527562)
t score of 99% confidence interval is: stats.t.interval(0.99,25,loc=0,scale=1)
(-2.787435813675851, 2.787435813675851)

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:

t =(x – μ) / (s / √n)

x = mean of the sampled bulbs = 260 days

μ = population mean = 270 days

s = standard deviation = 90 days

n = randomly selected bulbs = 18

t = (260 -270) / (90 / √18) = -0.471

Degrees of freedom = n – 1 = 18 – 1 = 17