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

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

Ans- Sample Space-(HHH,TTT,HHT,THH,HTH,TTH,HTT,THT)(Total-8)

Probability of X i.e. 2H and 1T-(HHT,THH,HTH)(Total-3)

P(X)=3/8

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

a) Equal to 1
b) Less than or equal to 4
c) Sum is divisible by 2 and 3

Ans-Sample Space-(Total-36)

(1,1);(1,2);(1,3);(1,4);(1,5);(1,6)
(2,1);(2,2);(2,3);(2,4);(2,5);(2,6)
(3,1);(3,2);(3,3);(3,4);(3,5);(3,6)
(4,1);(4,2);(4,3);(4,4);(4,5);(4,6)
(5,1);(5,2);(5,3);(5,4);(5,5);(5,6)
(6,1);(6,2);(6,3);(6,4);(6,5);(6,6)

a)Probability of A i.e. sum equal to 1-
P(A)= 0/36 = 0
b)Probability of B i.e. sum ≤ 4 -
sum is 2-[(1,1)]; 3-[(1,2);(2,1)]; 4-[(1,3);(2,2);(3,1)] (Total-6)
P(B)= 6/36 = 1/6
c)Probability of C i.e. sum is divisible by 2&3-
[(1,1);(1,2);(2,1);(1,3);(2,2);(3,1);(1,5);(2,4);(3,3);(4,2);(5,1);(2,6);(3,5);
(4,4);(5,3);(6,2);(3,6);(4,5);(5,4);(6,3);(4,6);(5,5);(6,4);(6,6)] (Total-24)
P(C)= 24/36 = 2/3

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-Sample Space- (Total-21)

(R1,R2);(G1,G2);(G2,G3);(G1,G3);(B1,B2);(R1,G1);(R1,G2);(R1,G3);(R2,G1);
(R2,G2);(R2,G3);(R1,B1);(R1,B2);(R2,B1);(R2,B2);(G1,B1);(G1,B2);(G2,B1);
(G2,B2);(G3,B1);(G3,B2)

Probability of X i.e. drawn two random balls being expect blue-

(R1,R2);(G1,G2);(G2,G3);(G1,G3);(R1,G1);(R1,G2);(R1,G3);(R2,G1);(R2,G2);(R2,G3) (Total-10)

P(X) = 10/21

<u>OR</u>

$ncr = n!/[(n-r)!r!]$

$P(X) = n(X)/n(S) = 5c2 / 7c2$

$= \{5!/[(5-2)!*2!]\} / \{7!/[(7-2)!*2!]\}$

$= [(5*4*3*2*1) / (3*2*1*2*1)] / [(7*6*5*4*3*2*1) / (5*4*3*2*1*2*1)]$

P(X) = 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 A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

<u>Ans</u>-

Child	**Candies count (xi)**	**Probability (P(xi))**	**E(X)=xi*P(xi)**
A	1	0.015	0.015
B	4	0.20	0.8
C	3	0.65	1.95
D	5	0.005	0.025
E	6	0.01	0.06
F	2	0.120	0.24
			Total=3.09

$E(X) = n\Sigma i=1\ \ xi * P(xi)$

∴ The Expected number of candies for a randomly selected child is **3.09**

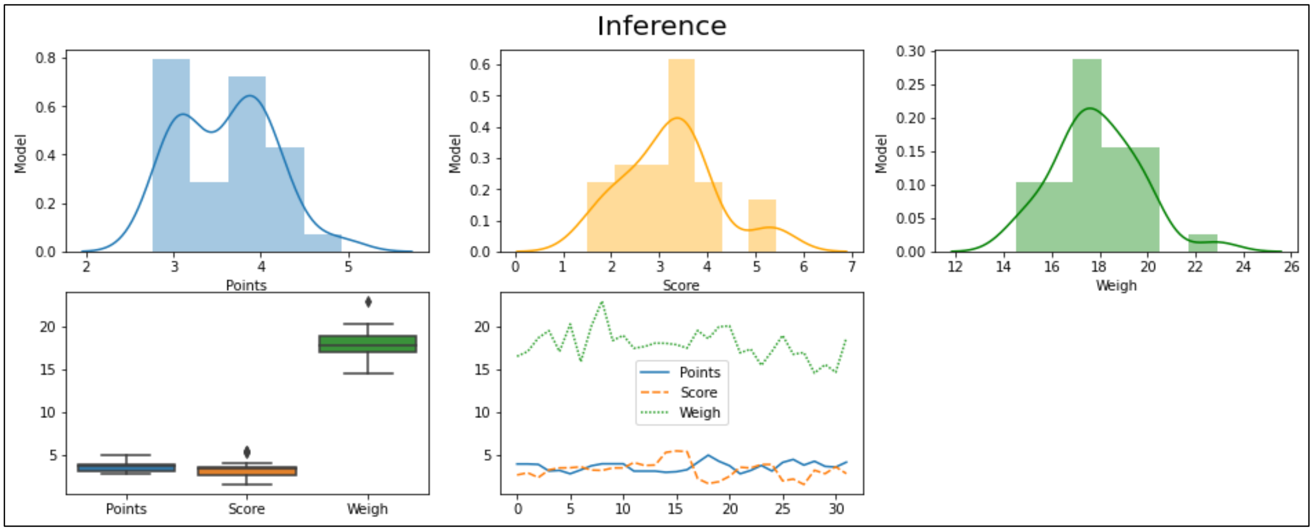
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- 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

Ans-

	Mean	**Median**	**Mode**	**Variance**	**Standard Deviation**	**Range**
Points	3.596563	3.695	3.07**,** 3.92	0.285881	0.534679	2.17
Score	3.217250	3.325	3.44	0.957379	0.978457	3.911
Weigh	17.84875	17.71	17.02, 18.90	3.193166	1.786943	8.399



Inference- We can make an inference by looking at the above graph and Descriptive Statistics that Score and Weigh have Outliers. In the Histogram of Points we can see a Bi-Model. By looking at the Range vales and Variance of Weigh there is more distance compared to others which means the values are far from mean. A positive skewness can be observed in the Weigh Histogram.

Q8) Calculate Expected Value for the problem below-

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-

Weights in Pounds (x)	**Probability (P(x))**	**E(X)=x*P(x)**
108	1/9	12
110	1/9	12.222
123	1/9	13.666
134	1/9	14.888
135	1/9	15
145	1/9	16.111
167	1/9	18.555
187	1/9	20.777
199	1/9	22.111
		Total=145.33

E(X) = nΣi=1 x * P(x)

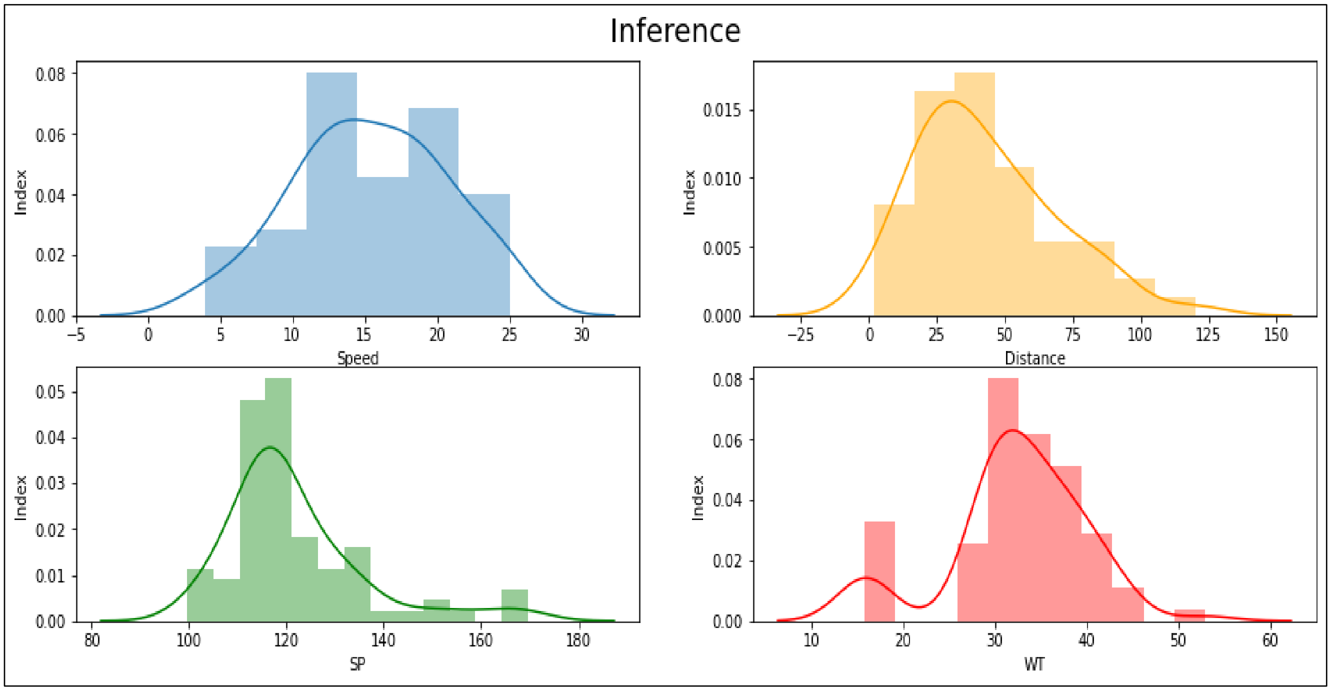
∴ The Expected Value of the Weight of that patient when one of the patients is chosen at random is **145.33 Pounds.**

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

- Car's speed and distance**. Use Q9_a.csv**
- SP and Weight(WT) **Use Q9_b.csv**

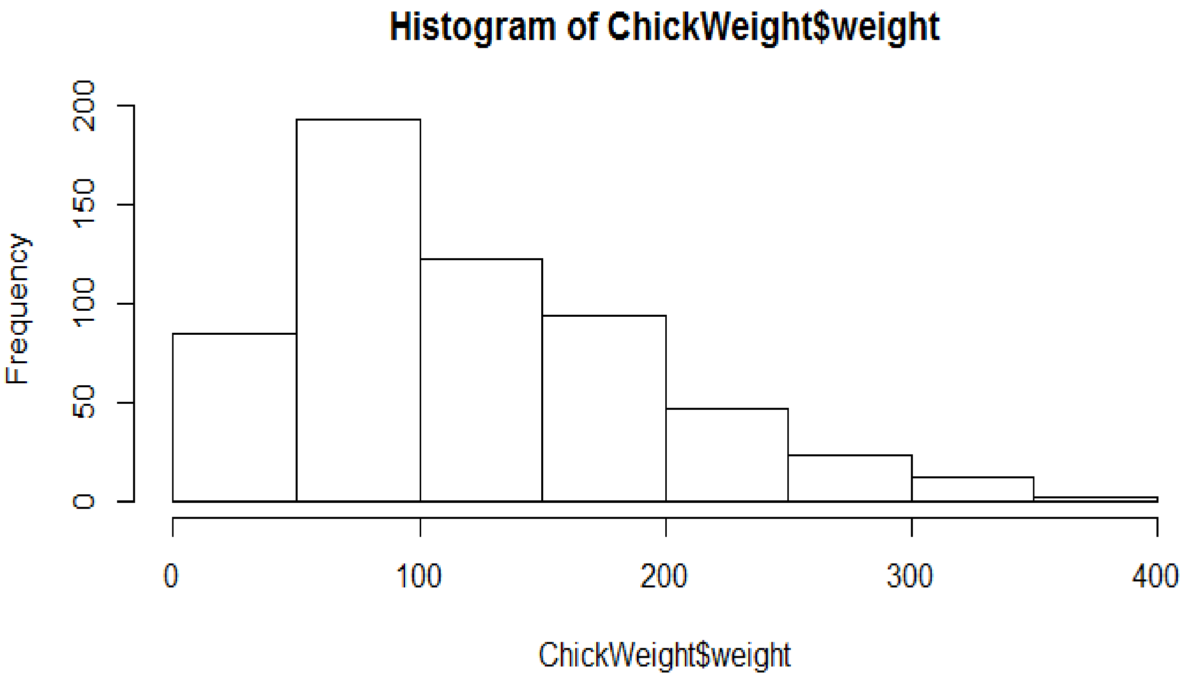
Ans-

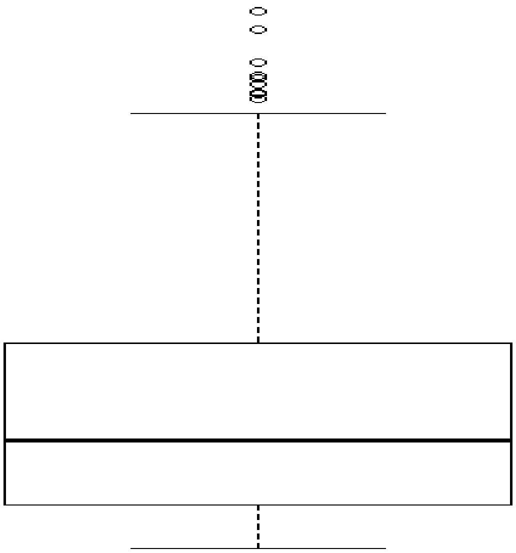
	Car 1		***Car 2***	
	Speed	**Distance**	**SP**	**WT**
Skewness	-0.117510	0.806895	1.611450	-0.614753
Kurtosis	-0.508994	0.405053	2.977329	0.950291



Inference- From the above table we can say that Speed and WT have a Negative Skewness which means that there is a tapering tail on the left-hand side of the distribution. Whereas there is a Negative Kurtosis in Speed which means it's peakness is not that high and is low, flat peak and vice-versa in remaining.

Q10) Draw inferences about the following boxplot & histogram





Inference- From the Histogram of Chick Weight a Positive Skewness is observed showing a tapering tail towards right which may lead to Outliers in the distribution. The Maximum values of the Weight lies between 50-150.This is an Asymmetry distribution.

Whereas the Boxplot has 7 Outliers showing vast distance between the Whisker and Upper-Quartile, which also shows a Positive Skewness with a right tapering tail.

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 30,00,000 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%,96%,98% confidence interval?

Ans-

Confidence Interval	Lower Limit	Upper Limit
94%	198.7376	201.2624
96%	198.6214	201.3786
98%	198.4382	201.5618

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?

Ans-

1)

Mean	Median	Variance	Standard Deviation
41	40.50	25.529412	5.052664

2) These are the test scores of 18 students where the average score was 41 and the scores range from 34-56.Majority of the score lies between 39-41 were 41 could also be said as mode. The variance which calculates the distance between the mean and other values also is high i.e. at 25.53 which shows the distribution of scores here will be Positively Skewed.

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

Ans- When mean, median of the data are equal which means it show a normal distribution and there is **Zero Skewness/No Skewness/Normal Skewness.**

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

Ans- When mean > median, the nature of skewness will be **Positive/Right Skewness.**

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

Ans- When median > mean, the nature of skewness will be **Negative/Left Skewness.**

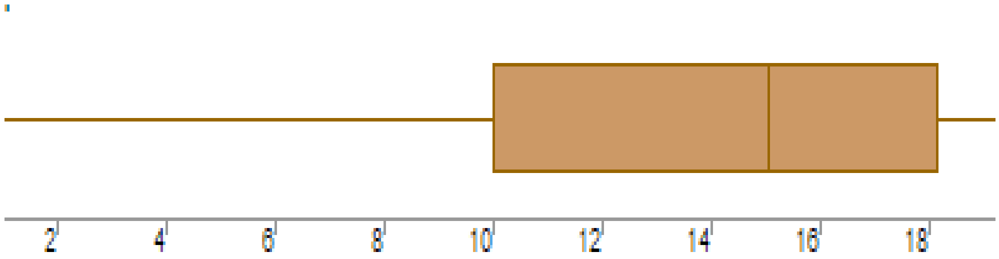
Q16) What does positive kurtosis value indicates for a data ?

Ans-Kurtosis measures the peakness of the distribution. A Positive Kurtosis indicates a **very High, Sharp and Steep Peak.** Which means there is less gap between the two tails and axis. The tails are Narrow.

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

Ans-A Negative Kurtosis indicates a **very Low, Broader and Flat Peak.** Which means there is more gap between the two tails and axis. The Tails are Wide.

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



1)What can we say about the distribution of the data?

2)What is nature of skewness of the data?

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

Ans-

1)Majority of values are mainly covered in the range of Lower Extreme to Lower Quartile so there is **No Normal Distribution**. Here the values range from 0-20,where Q1 i.e. Lower Quartile is 10,Q2 i.e. Median is 15 and Q3 i.e. Upper Quartile is 18.In this Boxplot **No Outliers** are detected.

2)If plotted then the tail with these values will be tapering in the left, which means the distribution is a **Negative/Left Skewness** as the mass of distribution is concentrated on the Right.

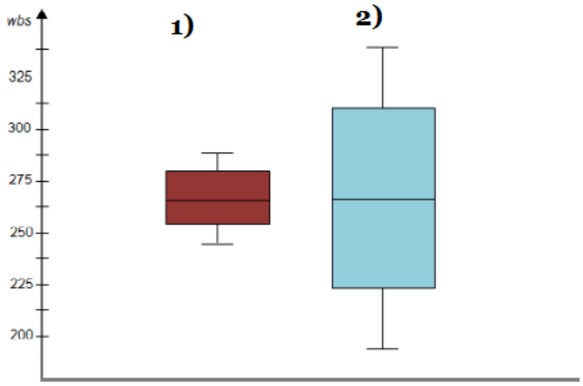
3) Inter Quartile Range = Upper Quartile – Lower Quartile

IQR = Q3 - Q1

= 18-10
= 8

∴ The IQR of the data is **8** Approx.

Q19) Comment on the below Boxplot visualizations?



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

Ans- The range of Boxplot 1 is less comparatively to Boxplot 2.Median of both the Boxplot are similar and both doesn’t have an Outlier. Both Boxplots follow Normal Distribution for which No/Zero Skewness is observed.

Q20) 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)

Ans- The Probability of MPG of Cars are-

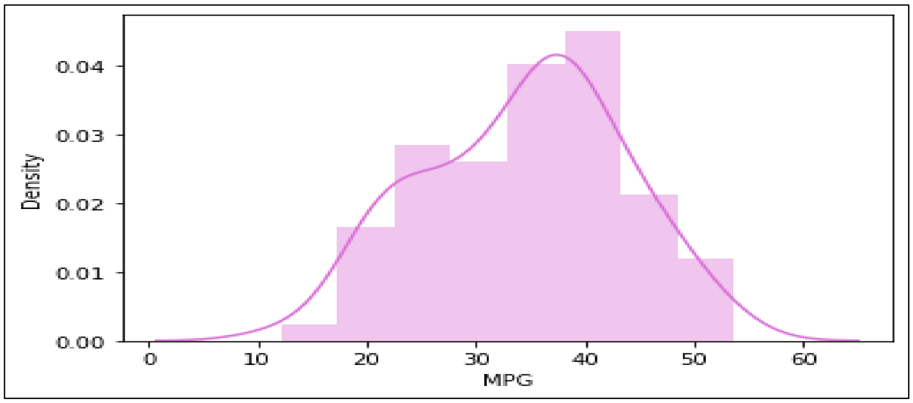
a. P(MPG>38)	**b. P(MPG<40)**	**c. P (20<MPG<50)**
0.34748702501304063	0.7294571279557076	0.8989177824549222

Q21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv
b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution. Dataset: wc-at.csv

Ans-

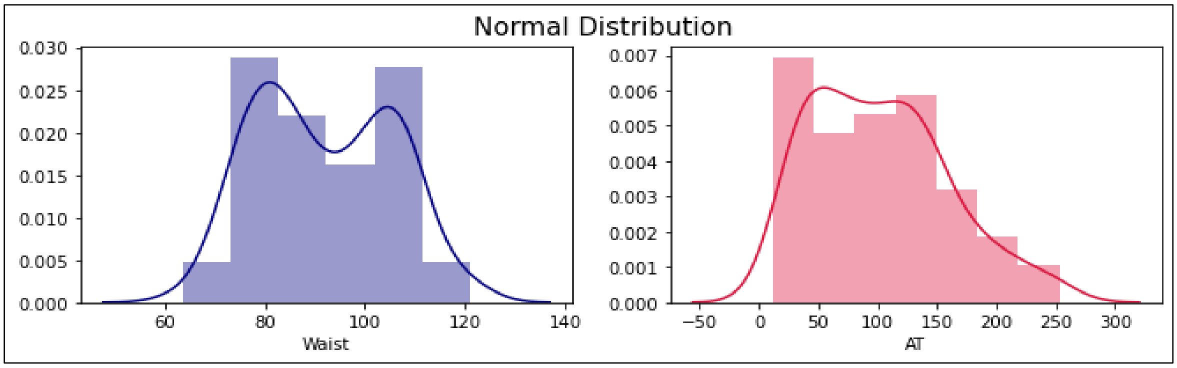
a)



Mean	Median	Mode
34.422076	35.152727	29.629936

For a Normal Distribution the Mean, Median and Mode should be equal, its shape should be Symmetric around the Mean. Normal Distribution are Denser in center and less in Tails. Here the Mean, Median and Mode are not equal nor it is Symmetrical Distribution, even though the values are pretty much closer it could not exactly be said as a Normal Distribution. But meets few criteria of a Normal Distribution.

b)



	Mean	**Median**	**Mode**
Waist	91.901835	90.800000	94.5, 106.0, 108.5
AT	101.894037	96.540000	121.0, 123.0

Here the Mean, Median and Mode are not equal for both and not also a smaller range. The Distribution is also not Symmetric. Therefore, both Waist and AT are not a Normal Distribution.

Q22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

Ans-

Z Scores – Confidence Interval		
60%	**90%**	**94%**
0.8416212335729143	1.6448536269514722	1.8807936081512509

Q23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25
Ans-

t Scores – Confidence Interval		
95%	**96%**	**99%**
2.0638985616280205	2.1715446760080677	2.796939504772804

Q24) 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-

H0= $\mu \leq 260$; H1= $\mu > 260$; $\bar{x}=260$; $\mu=270$; s=90; n=18

t-test = -0.4714045207910317

p value = 0.32216394448907903

p > α ; 0.32 > 0.05

Fail to Reject H0 null hypo /Reject H1

Therefore, bulbs would have an average life of no more than 260 days and the CEO’s claim are False.

NOTE: *For reference please check the “1.Basic Statistics-1.ipynb” file attached.*

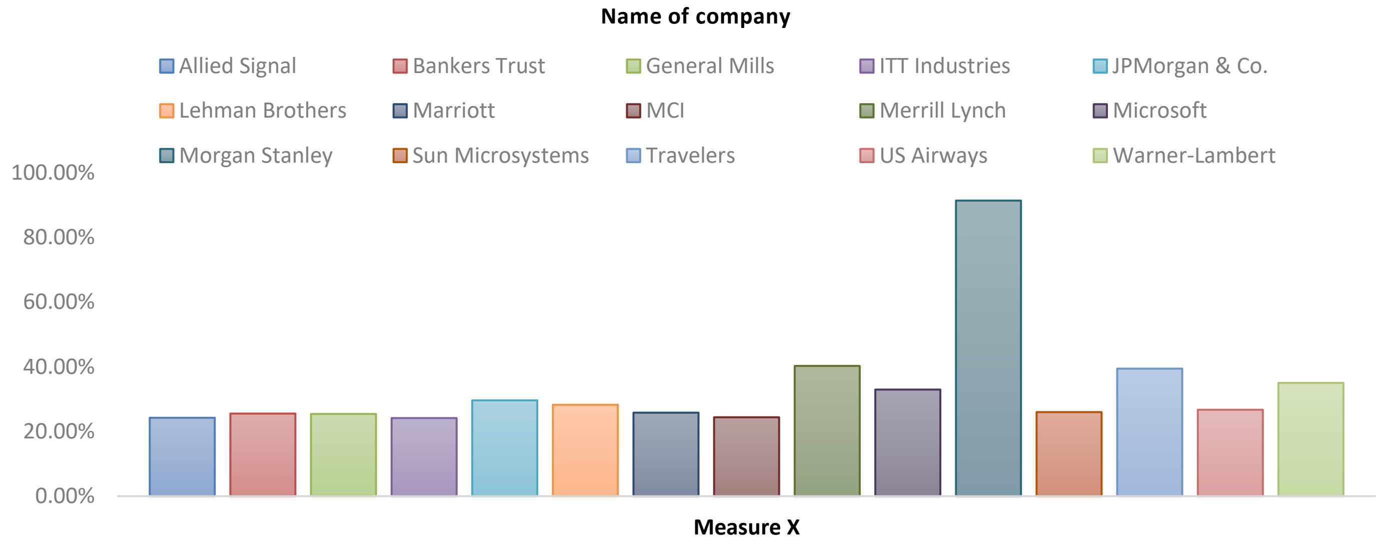
SET-1

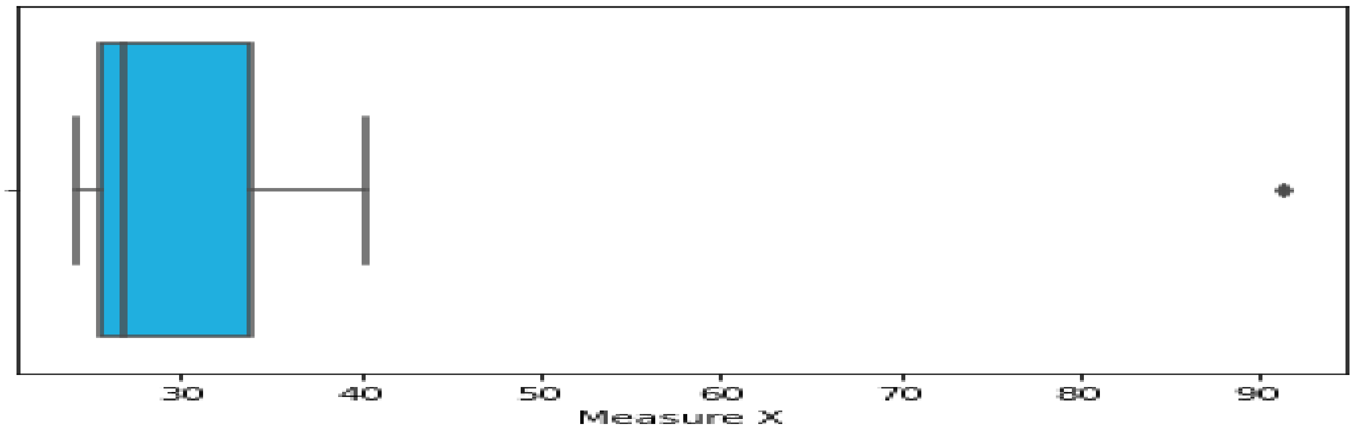
Topics: Descriptive Statistics and Probability

1. Look at the data given below. Plot the data, find the outliers and find out μ, σ, σ^2

Name of company	Measure X
Allied Signal	24.23%
Bankers Trust	25.53%
General Mills	25.41%
ITT Industries	24.14%
JPMorgan & Co.	29.62%
Lehman Brothers	28.25%
Marriott	25.81%
MCI	24.39%
Merrill Lynch	40.26%
Microsoft	32.95%
Morgan Stanley	91.36%
Sun Microsystems	25.99%
Travelers	39.42%
US Airways	26.71%
Warner-Lambert	35.00%

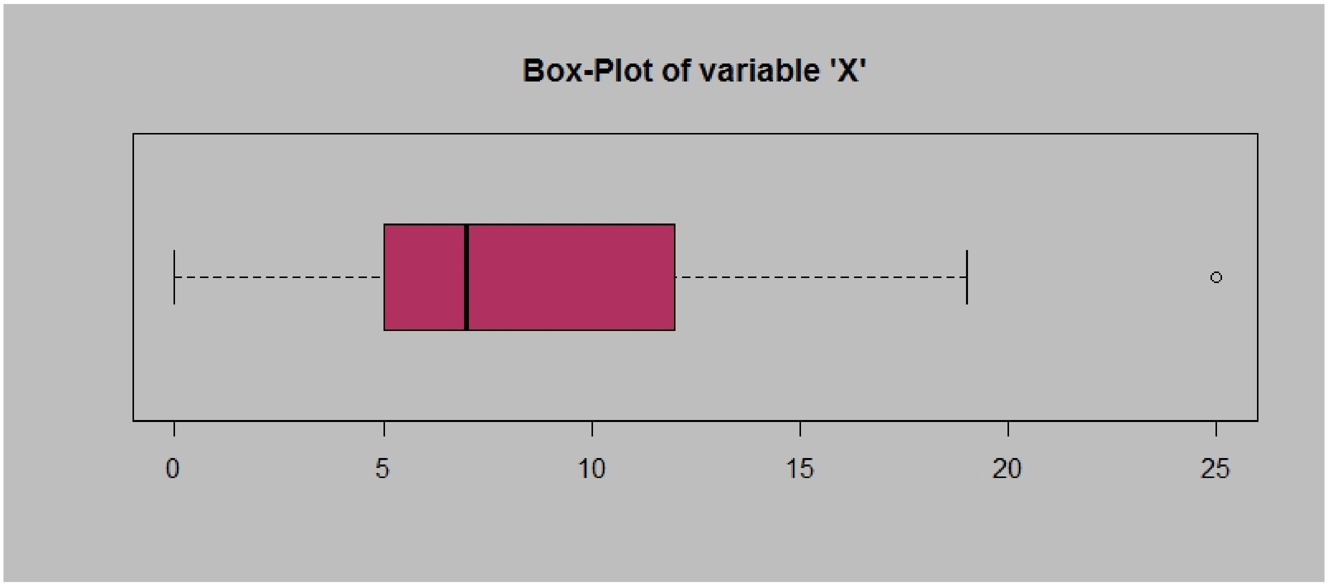
Ans-





Outliers	**Mean (μ)**	**Variance (σ)**	**Std Deviation (σ^2)**
Morgan Stanley 91.36%	33.271333	287.146612	16.945401

2.



Answer the following three questions based on the box-plot above.

(i) What is inter-quartile range of this dataset? (Please approximate the numbers) In one line, explain what this value implies.
(ii) What can we say about the skewness of this dataset?
(iii) If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?

Ans-

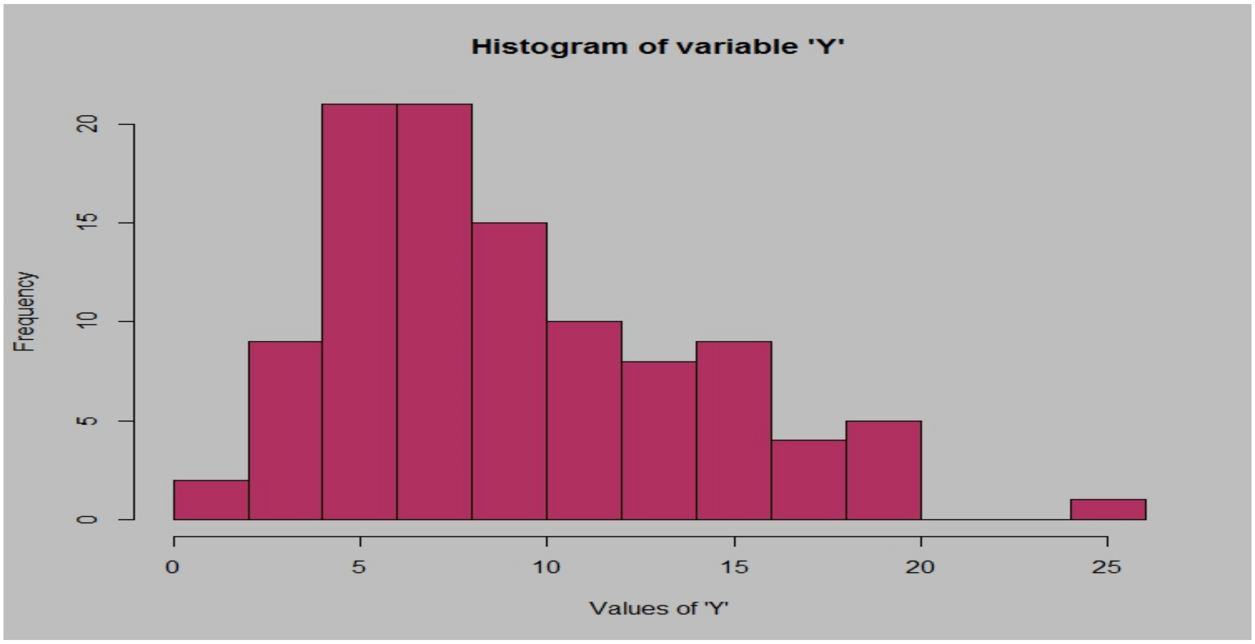
LE =0, Q1 =5, Q2 =7, Q3 =12, UE =19, Outlier =25

(i) IQR = Q3 - Q1 =Upper Quartile – Lower Quartile =12-5 =**7**
IQR is a range where 50% (i.e. 75% - 25%,Q3-Q1) of the data point lies.

(ii) A **Positive Skewness** is observed which may show a tapering tail towards right in a Histogram(Frequency Distribution) as it represents an Outlier after the Upper Extreme value. It shows an Asymmetry Distribution.

(iii) If the data point 25 would be 2.5 the box plot will remain the same with range 0-19, just without an outlier i.e. no Outlier would be detected as the data point 25 is an outlier in this boxplot if it is changed to 2.5 it will be a normal value and hence, No Skewness would be observed. It will be a Normal Distribution.

3.



Answer the following three questions based on the histogram above.

(i) Where would the mode of this dataset lie?

(ii) Comment on the skewness of the dataset.

(iii) Suppose that the above histogram and the box-plot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.

Ans-

(i) The mode of variable "Y" in this data set will lie between **4-8** with a frequency of 20.

(ii) The tail is tapering in the Right, which means the distribution is a **Positive/Right Skewness** as the mass of distribution is concentrated on the Left as few values are stretched till 26 in this dataset.

(iii) Both plots Boxplot and Histogram complement each other by having a common outlier at the data point 25.Both the plots display a Positive Skewness and the datapoints in the dataset ranges from 0-25.The mode or the highest frequency also lies between 4-12,even the IQR where 50% of the data points lie falls in this range.

4. AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that "could happen." Suppose that one in 200 long-distance telephone calls is misdirected. What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

Ans-

Using Bernoulli Trials Formulas-

$\mathbf{P(x) = {}^nC_x\, p^x\, q^{n-x}}$

p = Probability of Success =1/200

q = Probability of Failure =1-p =1-1/200 =199/200

n = Number of Independent trials =5

x = The number of times an event occurred =0

$P(X)= 5C0 * (1/200)^0 *(199/200)^{5-0}$

P(at least 1 in 5 calls reaches wrong no.) $= 1 - P(X)$

$= 1 - 5C0 * (1/200)^0 * (199/200)^5$

$= 1 - (199/200)^5$

$= 1 - 0.9752$

= **0.02475**

∴ The probability that at least one in five attempted telephone calls reaches the wrong number is **0.02475**

5. Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution

x	P(x)
-2,000	0.1
-1,000	0.1
0	0.2
1000	0.2
2000	0.3
3000	0.1

(i) What is the most likely monetary outcome of the business venture?

(ii) Is the venture likely to be successful? Explain

(iii) What is the long-term average earning of business ventures of this kind? Explain

(iv) What is the good measure of the risk involved in a venture of this kind? Compute this measure

Ans-

(i) Since the probability of P(2000) = 0.3 which is the highest among others so it is most likely that **$2,000** will be the monetary outcome of the business venture.

(ii) **Yes,** the venture is likely to be successful P(x>0) =0.2+0.3+0.1 =0.6 i.e. the probability of success or earning profit is **60%** whereas failure or suffering loss is 40%.

(iii) The long-term average earning of the business venture can be calculated using Expected Value-

E(X) = xi * P(xi)

xi	**P(xi)**	**E(X) = xi * P(xi)**
-2,000	0.1	-200
-1,000	0.1	-100
0	0.2	0
1000	0.2	200
2000	0.3	600
3000	0.1	300
Total		**800**

∴The long-term average earning of the business venture is **$800.**

(iv) The good measure of the risk involved in a venture of this kind can be calculated by the standard deviation of E(**x^2)-**

xi	**x^2**	**P(xi)**	**E(x^2) = x^2 * P(xi)**
-2,000	40,00,000	0.1	4,00,000
-1,000	10,00,000	0.1	1,00,000
0	0	0.2	0
1000	10,00,000	0.2	2,00,000
2000	40,00,000	0.3	12,00,000
3000	90,00,000	0.1	9,00,000
Total			**28,00,000**

∴ The good measure of the risk involved in a venture of this kind is **28,00,000**

Questions referred to from *Aczel A., Sounder Pandian J., Complete Business Statistics (7ed.)*

SET-2

Topics: Normal distribution, Functions of Random Variables

1. The time required for servicing transmissions is normally distributed with μ = 45 minutes and σ = 8 minutes. The service manager plans to have work begin on the transmission of a customer's car 10 minutes after the car is dropped off and the customer is told that the car will be ready within 1 hour from drop-off. What is the probability that the service manager cannot meet his commitment?

A. 0.3875
B. 0.2676
C. 0.5
D. 0.6987

Ans-

To find: **P(x ≥ 60)** as car should be ready in 1 hour (60 mins)

As the work will begin after 10 mins of drop, so the μ =45+10=55

Using jupyter notebook –

x=60, μ =55, σ = 8 (being Normal Distribution as specified)

P(x ≥ 60) = 1 - P(x ≤ 60)

=1-stats.norm.cdf (x, μ,σ)

=1-stats.norm.cdf(60,55,8)

= 0.26598552904870054

∴ The probability that the service manager cannot meet his commitment is **0.2659**

2. The current age (in years) of 400 clerical employees at an insurance claims processing center is normally distributed with mean μ = 38 and Standard deviation σ =6. For each statement below, please specify True/False. If false, briefly explain why.

A. More employees at the processing center are older than 44 than between 38 and 44.
B. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

Ans-

A. To find: **P(38 ≤ x ≤ 44)**

Using jupyter notebook –

x1=38, x2=44, μ =38, σ = 6 (being Normal Distribution as specified)

P(38 ≤ x ≤ 44) = P(x ≤ 44) - P(x ≥ 38)

= stats.norm.cdf (x_1, μ,σ) - stats.norm.cdf (x_2, μ,σ)

= stats.norm.cdf (44,38,6) - stats.norm.cdf (38,38,6)

= 0.8413447460685429 – 0.5

= 0.3413447460685429

Since the **P(x ≤ 44) = 0.84** which is **True** but **P(38 ≤ x ≤ 44) = 0.34** which is **False** as 400*0.34 is 136 clerical employees which is very less.

So **True & False** makes the statement **False.**

∴ More employees at the processing center are older than 44 than between 38 and 44 is a **False** statement.

B. To find: **$P(x \leq 30)$ & $N^* P(x \leq 30) \approx 36$**

Using jupyter notebook –

x=30, μ =38, σ = 6 (being Normal Distribution as specified)

$P(x \leq 30)$

= stats.norm.cdf (x, μ,σ)

= stats.norm.cdf (30,38,6)

= 0.09121121972586788

&

$N^* P(x \leq 30)$

= 400 * 0.09

= 36

∴ A training program for employees under the age of 30 at the center would be expected to attract about 36 employees is a **True** statement.

3. If $X_1 \sim N(\mu, \sigma^2)$ and $X_2 \sim N(\mu, \sigma^2)$ are *iid* normal random variables, then what is the difference between 2 X_1 and $X_1 + X_2$? Discuss both their distributions and parameters.

Ans-

iid (independent and identically distributed) normal random variable means that the distribution of both the variables are normal and identical, and both have independent outcomes which means outcome of one individual independent outcome does not affect the other.

$2X_1 \sim N(2\mu ,2 \sigma^2)$ & $X_1+X_2 \sim N(\mu, \sigma^2)+ N(\mu, \sigma^2)$

$2X_1$ will multiply the whole random variable with a change in its distribution which will affect it largely.

Whereas, X_1+X_2 will add both the random variable and its distribution.it will sum up its samples.

As it is random so the value of the parameter will not remain constant it will vary. Both the variables have same parameters like N which could be the total sample size ,μ which is mean / average and σ^2 which is variance. If we input values in these parameters their distribution will completely change.

∴ $2X_1$ & X_1+X_2 are **different** not by formula but by input values it will differ.

4. Let X ~ N(100, 20^2). Find two values, *a* and *b*, symmetric about the mean, such that the probability of the random variable taking a value between them is 0.99.

A. 90.5, 105.9
B. 80.2, 119.8
C. 22, 78
D. 48.5, 151.5
E. 90.1, 109.9

Ans-

To find: **a ,b**

$P(a \leq x \leq b) = 0.99$

$X \sim N(\mu, \sigma^2)$ = $X \sim N(100, 20^2)$

x=0.99, μ =100, σ = 20

x is also the confidence interval at 99%

Using jupyter notebook –

= stats.norm.interval(x, μ, σ)

= stats.norm.interval(0.99,100,20)

= 48.48341392902199, 151.516586070978

OR

Using Z-scores formula-

$Z =(x_i - \mu) / \sigma$

$x_i = \mu \pm \sigma * Z$

Using jupyter notebook –

Z = stats.norm.ppf(x)

Z = stats.norm.ppf(0.995)

Z = 2.5758293035489004

$x_i = 100 \pm 2 * 2.57$

x_i = 48.48341392902199, 151.516586070978

∴ **a = 48.5** and **b = 151.5**

∴ **Option D** is the correct answer.

5. Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions $Profit_1 \sim N(5, 3^2)$ and $Profit_2 \sim N(7, 4^2)$ respectively. Both the profits are in $ Million. Answer the following questions about the total profit of the company in Rupees. Assume that $1 = Rs. 45

A. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.

B. Specify the 5th percentile of profit (in Rupees) for the company

C. Which of the two divisions has a larger probability of making a loss in a given year?

Ans-

Profit ~$N(\mu, \sigma^2)$

A. Combining profit parameters of both the division x=0.95,μ = 5+7 =12*45=540 , σ= 3+4 = 7*45=315

Using jupyter notebook –

= stats.norm.interval(x, μ, σ)

= stats.norm.interval(0.95,540,315)

= -77.38865513011706 ,1157.388655130117

∴ The Rupee range for the 95% probability for the annual profit of the company is **-77.39 to 1157.39 Rs. Million**

B. μ=540, σ = 315,x = 0.05

Using jupyter notebook –

= stats.norm.ppf(x, μ, σ)

= stats.norm.ppf(0.05,540,315)

=21.871107510286038

OR

$x_i = \mu \pm \sigma * Z$

=540+315*(-1.647)

= 21.871107510286038

∴ the 5th percentile of profit for the company is **21.87 Rs. Million**

C. To find: $\mathbf{P_1(x \leq 0)}$ **&** $\mathbf{P_2(x \leq 0)}$

Using jupyter notebook –

x=0, μ_1 =5, σ_1 = 3, μ_2 =7, σ_2 = 4

$\mathbf{P(x \leq 0)}$

= stats.norm.cdf (x, μ,σ)

$\mathbf{P_1(x \leq 0)}$ **=** stats.norm.cdf (0, 5,3) =0.0477903522728147 = **0.048**

$\mathbf{P_2(x \leq 0)}$ **=** stats.norm.cdf (0, 7,4) = 0.040059156863817086 = **0.040**

Both the divisions are making loss but $\mathbf{P_1 > P_2}$ (0.048 > 0.040)

∴ the larger probability of making a loss in a given year is **Division 1** with a Probability of **0.048**

SET-3

Topics: Confidence Intervals

1. For each of the following statements, indicate whether it is True/False. If false, explain why.

I. The sample size of the survey should at least be a fixed percentage of the population size in order to produce representative results.
II. The sampling frame is a list of every item that appears in a survey sample, including those that did not respond to questions.
III. Larger surveys convey a more accurate impression of the population than smaller surveys.

Ans-

I. True- as the results are made on the basis of sample size and it will make an inference to it.
II. False- Sampling Frame is simply a list of items from which to draw a sample not including those who did not respond to questions.
III. True- Larger surveys convey a more accurate impression of the population than smaller surveys.as larger surveys involve large sample size which reduces the chances of error with less standard deviation.

2. *PC Magazine* asked all of its readers to participate in a survey of their satisfaction with different brands of electronics. In the 2004 survey, which was included in an issue of the magazine that year, more than 9000 readers rated the products on a scale from 1 to 10. The magazine reported that the average rating assigned by 225 readers to a Kodak compact digital camera was 7.5. For this product, identify the following:

A. The population
B. The parameter of interest
C. The sampling frame
D. The sample size
E. The sampling design
F. Any potential sources of bias or other problems with the survey or sample

Ans-

A. Population- Readers of PC Magazine
B. Parameter of interest- Sample Size ,Average ,Rating Scale
C. Sampling Frame- More than 9000
D. Sample Size- 225
E. Sampling Design- Judgmental / Voluntary Sampling Design
F. The survey might be bias as it was a rating scale for an electronic product and only readers of PC Magazine have rated due to which other magazine readers or non-magazine readers would not have been considered. As well as the readers those who have voted might have not given accurate ratings due to which the results of survey may differ and so a conclusion might not be accurate.

3. For each of the following statements, indicate whether it is True/False. If false, explain why.

I. If the 95% confidence interval for the average purchase of customers at a department store is \$50 to \$110, then \$100 is a plausible value for the population mean at this level of confidence.

II. If the 95% confidence interval for the number of moviegoers who purchase concessions is 30% to 45%, this means that fewer than half of all moviegoer's purchase concessions.

III. The 95% Confidence-Interval for μ only applies if the sample data are nearly normally distributed.

<u>Ans</u>-

I. True- Confidence interval identifies the collection of values for the population parameter that are consistent with the observed sample. It displays the probability that a parameter will fall between a pair of values around the mean.so \$100 is a plausible value for the population mean.

II. True-At 95% confidence interval of moviegoers, who purchase concession is between 30% and 45%, so the lower and upper limit for it are 30% and 45%.So at 95% confidence interval for the minimum proportion of moviegoers, who do not purchase from the concession is (100−45)%=55%.
The proportion of moviegoers, who do not like to purchase from the concession is more than 50%.

III. False- The distribution of sample mean will be approximately normal even if the distribution of population is not normal.

4. What are the chances that $\overline{X} > \mu$?

A. ¼

B. ½

C. ¾

D. 1

<u>Ans</u>-

The expected value of $\bar{x}$ is equal to µ.Therefore, the chances that $\bar{x} > \mu$ is **1.**
Option- D is correct.

5. In January 2005, a company that monitors Internet traffic (WebSideStory) reported that its sampling revealed that the Mozilla Firefox browser launched in 2004 had grabbed a 4.6% share of the market.

I. If the sample were based on 2,000 users, could Microsoft conclude that Mozilla has a less than 5% share of the market?

II. WebSideStory claims that its sample includes all the daily Internet users. If that's the case, then can Microsoft conclude that Mozilla has a less than 5% share of the market?

Ans-

I. Defining the hypothesis and calculating probability:
Using One-sample z-test for proportions-
$H_0 = \mu \leq 5\%$,$H_1 = \mu > 5\%$,$\mu = 5\%$,s = 4.6% ,n=2000
$Z = (s- \mu) / \sqrt{} ((\mu * (1- \mu)) / n)$ **OR** $Z= (\hat{P} - P_0) / \sqrt{} ((P_0 * (1- P_0)) / n)$ where $P_0 = 0.05$ & $\hat{P} = 0.046$
= (0.046-0.05) / √ ((0.05*(1-0.05)) / 2000)
= -0.820782681668124
= **-0.821**
-0.0821< 0.05
Z < 0.05 i.e. α
∴ we reject the H_0 and accept H_1.
$H_1 = \mu > 5\%$ so, Microsoft can conclude that **Mozilla has more than 5%** share of the market.

II. The sample includes all the daily Internet users, which means that the 4.6% share of the market represents the whole population. So, Microsoft can conclude that **Mozilla has a less than 5%** share of the market.

6. A book publisher monitors the size of shipments of its textbooks to university bookstores. For a sample of texts used at various schools, the 95% confidence interval for the size of the shipment was 250 ± 45 books. Which, if any, of the following interpretations of this interval are correct?

A. All shipments are between 205 and 295 books.
B. 95% of shipments are between 205 and 295 books.
C. The procedure that produced this interval generates ranges that hold the population mean for 95% of samples.
D. If we get another sample, then we can be 95% sure that the mean of this second sample is between 205 and 295.
E. We can be 95% confident that the range 160 to 340 holds the population mean.

Ans-

The size of the shipment was 250 ± 45 books.
So 250-45 = **205** and 250+45 = **295**
At 95% Confidence Interval the shipment lies between 205 to 295.
∴ We can conclude that the interpretations of **B** and **C** are **Correct.**

7. Which is shorter: a 95% *z*-interval or a 95% *t*-interval for μ if we know that σ =s?

A. The z-interval is shorter
B. The t-interval is shorter
C. Both are equal
D. We cannot say

Ans-

At 95% Confidence Interval-
z value =**1.96.**
t value = **2.262**.
z value < t value
∴ The Statement **A** that the z-interval is shorter is Correct.

Questions 8 and 9 are based on the following: To prepare a report on the economy, analysts need to estimate the percentage of businesses that plan to hire additional employees in the next 60 days.

8. How many randomly selected employers (minimum number) must we contact in order to guarantee a margin of error of no more than 4% (at 95% confidence)?

A. 600
B. 400
C. 550
D. 1000

Ans-

Assume p=0.5,MOE=4%=0.04,z at 95% CI =1.96
$\text{MOE}= z * \sqrt{(p(1-p))} / \sqrt{n}$
$0.04=1.96* \sqrt{(0.5-0.25)} / \sqrt{n}$
n=**600.25 ~ 600**
∴ **A. 600** employers must be contacted in order to guarantee MOE less than 4%

9. Suppose we want the above margin of error to be based on a 98% confidence level. What sample size (minimum) must we now use?

A. 1000
B. 757
C. 848
D. 543

Ans-

Assume p=0.5,MOE=4%=0.04,z at 98% CI =2.326
$\text{MOE}= z * \sqrt{(p(1-p))} / \sqrt{n}$
$0.04=2.326* \sqrt{(0.5-0.25)} / \sqrt{n}$
n=**845.36 ~ 848**
∴ **C. 848** employers must be contacted.

SET-4

CBA: Practice Problem Set-2

Topics: Sampling Distributions and Central Limit Theorem

1. Examine the following normal Quantile plots carefully. Which of these plots indicates that the data …?

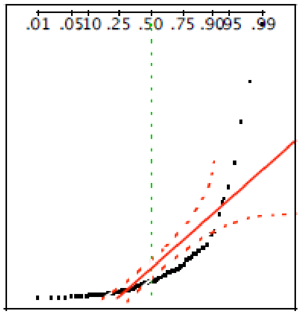
I. Are nearly normal?

II. Have a bimodal distribution? (One way to recognize a bi-modal shape is a "gap" in the spacing of adjacent data values.)

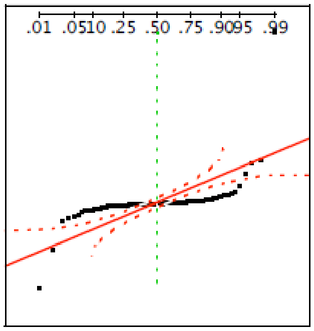
III. Are skewed (i.e. not symmetric) ?

IV. Have outliers on both sides of the center?

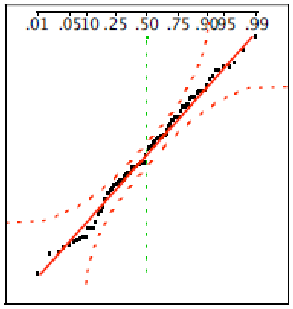
A



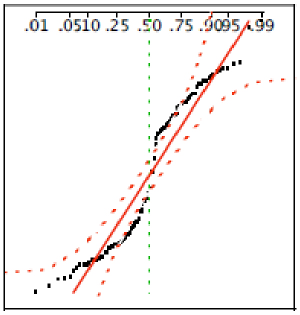
B



C



D



Ans-

I. The plot **C** has a normal distribution and plot **D** is nearly normal.

II. The plot **D** has a bi-modal distribution forming a S-shape curve.

III. The plot **A** is **Right-**Skewed.

IV. The plot **A** have outliers on both the sides of center.

2. For each of the following statements, indicate whether it is True/False. If false, explain why.
The manager of a warehouse monitors the volume of shipments made by the delivery team. The automated tracking system tracks every package as it moves through the facility. A sample of 25 packages is selected and weighed every day. Based on current contracts with customers, the weights should have μ = 22 lbs. and σ = 5 lbs.

(i) Before using a normal model for the sampling distribution of the average package weights, the manager must confirm that weights of individual packages are normally distributed.

(ii) The standard error of the daily average SE($\bar{x}$) = 1.

Ans-

(i) **True -** based on central limit theorem the distribution of sample mean will be approximately normal even if the distribution of data in population is not normal but the sample size should be fairly large. Here the sample size is very small i.e. < 30.

(ii) **True -** SE($\overline{x}$) = σ / √n = 5/√25 = **1**

3. Auditors at a small community bank randomly sample 100 withdrawal transactions made during the week at an ATM machine located near the bank's main branch. Over the past 2 years, the average withdrawal amount has been $50 with a standard deviation of $40. Since audit investigations are typically expensive, the auditors decide to not initiate further investigations if the mean transaction amount of the sample is between $45 and $55. What is the probability that in any given week, there will be an investigation?

A. 1.25%
B. 2.5%
C. 10.55%
D. 21.1%
E. 50%

Ans-

To find: **1-P(45 ≤ x ≤ 55)**
x_1=45, x_2=55, n=100, μ =50, σ = 40, df = n-1 = 100-1 = 99

P(45 ≤ x ≤ 55) = P(x ≤ 45) - P(x ≥ 55)
Calculating z values-
z value = ($\overline{x}$ - μ) / (σ /√n)
Using jupyter notebook –
p1=(50-45)/(40/np.sqrt(100)) = 1.25
p2=(50-55)/(40/np.sqrt(100)) = -1.25
= stats.t.cdf (p_1,df) - stats.t.cdf (p_2, df)
= stats.t.cdf (1.25,99) - stats.t.cdf (-1.25,99)
= 0.7857536624316135

P(45 ≤ x ≤ 55) = 0.7858
1- P(45 ≤ x ≤ 55) = 1- 0.7858 = 0.2142463*100= **21.43%**
∴ The probability that in any given week, there will be an investigation is **21.1%**

4. The auditors from the above example would like to maintain the probability of investigation to 5%. Which of the following represents the minimum number transactions that they should sample if they do not want to change the thresholds of 45 and 55? Assume that the sample statistics remain unchanged.

A. 144
B. 150
C. 196
D. 250
E. Not enough information

Ans-

The Probability of investigation is 5% so z-score = 100-5=95

The z-score at 95% Confidence Interval is 1.95996398454005

z value = ($\bar{x}$ - μ) / (σ /√n)

1.95996398454005 = (50-45)/(40/√n))

1.95996398454005 = √n / 8

√n = 15.6797118763204

∴ n = 245.853364524423 ~ 250

∴The minimum number of transactions that they should sample is **250.**

5. An educational startup that helps MBA aspirants write their essays is targeting individuals who have taken GMAT in 2012 and have expressed interest in applying to FT top 20 b-schools. There are 40,000 such individuals with an average GMAT score of 720 and a standard deviation of 120. The scores are distributed between 650 and 790 with a very long and thin tail towards the higher end resulting in substantial skewness. Which of the following is likely to be true for randomly chosen samples of aspirants?

A. The standard deviation of the scores within any sample will be 120.

B. The standard deviation of the mean of across several samples will be 120.

C. The mean score in any sample will be 720.

D. The average of the mean across several samples will be 720.

E. The standard deviation of the mean across several samples will be 0.60

Ans-

x_1=650, x_2=790, n=40,000, μ =720, σ = 120

SE($\bar{x}$) = σ / √n = 120/√40000 = **0.60**

Option **E** is likely to be true for randomly chosen samples of aspirants.

NOTE: *For reference please check the "2.Basic Statistics-2.ipynb" file attached.*