# VRUSHABH RODE

## ASSIGNMENT NO. 1

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 | Interval |
| Celsius Temperature | Interval |
| Weight | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |

| | |
|---|---|
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Ordinal |
| 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 = 3/8 = 0.375

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 2and 3

➔ Answer 1) 0
➔ Answer 2) 0.16
➔ Answer 3) 0.66

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 = 0.47
➔ = 5C2 / 7C2 = 10 / 21 =0.47

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 = 3.09
➔ Expected value = sum of (probability * value)
➔ = 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**

➔ **SOLVE IN PYTHON FILE**

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?

➔ Expected value = sum of (probability * value)
➔ Probability for each Person is (1/9)
➔ so that Expected Value = (108*1/9)+……..+(199*1/9)
➔ = 145.33
➔ Answer = 145.33

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

**Cars speed and distance**

**Use Q9_a.csv**

➔ **SOLVE IN PYTHON FILE**

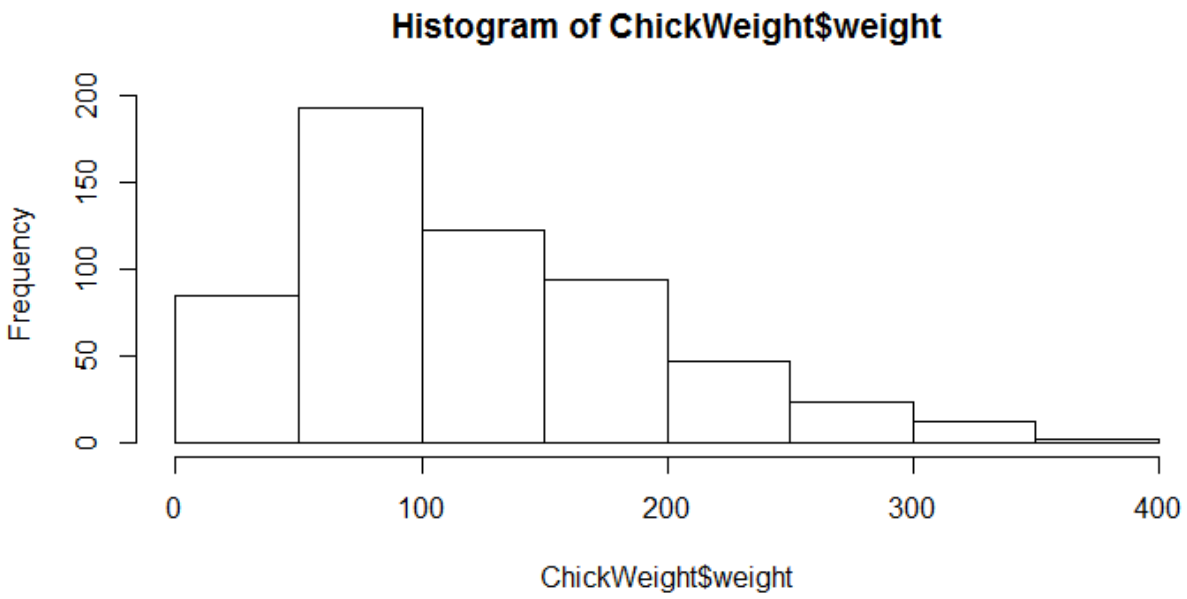
**SP and Weight(WT)**

**Use Q9_b.csv**

➔ **SOLVE IN PYTHON FILE**

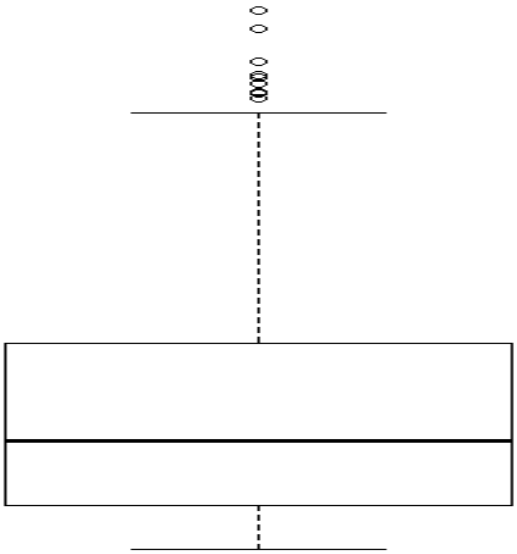
**Q10) Draw inferences about the following boxplot & histogram**

---




➔ Above figure shows graph of Histogram where chickweights$weight weight is on X-axis and the frequency is on Y-axis.
As we can see 200 chicks weight is lying under the bin from the interval of 50-100 that is most chicks weight in bin 50-100.

We also can observe that as the weight of the chick is increasing the number of chicks are decreasing.
As most of the data is started interval from 0-250 there will be mean present in between 50-100 as peakness is there.



**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?

-> Answers :

1.- CI = 94% ( $\mu_0$ - 1,04 < x < $\mu_0$ + 1,04 )

2.- CI = 98 % ( $\mu_0$ - 2,05 < x < $\mu_0$ + 2,05 )

3.- CI = 96 % ( $\mu_0$ - 1,75 < x < $\mu_0$ + 1,75 )

Sample size n = 3000000

Sample mean x = 200

Standard deviation s = 30

From z-table values of z(c):

CI 94 % Confidencial level $\alpha = 6\ \%$ $\alpha = 0{,}06$ $z(c) = 1{,}55$

CI 98 % Confidencial level $\alpha = 2\ \%$ $\alpha = 0{,}02$ $z(c) = 2{,}05$

CI 96 % Confidencial level $\alpha = 4\ \%$ $\alpha = 0{,}04$ $z(c) = 1{,}75$

$MOE = z(c) * \sigma/\sqrt{n}$

1.-MOE = $1{,}55 * 30 / \sqrt{2000}$ MOE = 1,04

2.-MOE = $2{,}05*30/\sqrt{2000}$ MOE = 1,38

3.-MOE = $1{,}75*30/\sqrt{2000}$ MOE = 1,17

Then CI

1.- CI = 94 % $( \mu_0 - MOE < x < \mu_0 - MOE )$

$CI = ( \mu_0 - 1{,}04 < x < \mu_0 + 1{,}04 )$

2.-CI = 98 %

$CI = ( \mu_0 - 2{,}05 < x < \mu_0 + 2{,}05 )$

3.- CI = 96 %

$CI = ( \mu_0 - 1{,}75 < x < \mu_0 + 1{,}75 )$

**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.

➔ Answers :-

➔ Mean=41.0

➔ Median=40.5

➔ Varience=25.529412

➔ Standard Deviation=5.052664

2) What can we say about the student marks?

➔

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

➔ When the mean and median of data are equal then there is no skewness (**ZERO SKEWNESS**)

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

➔ When the mean is greater then median then the nature of skewness is **POSITIVELY SKEWED**

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

➔ When the median is greater then mean then the nature of skewness is **NEGATIVELY SKEWED**

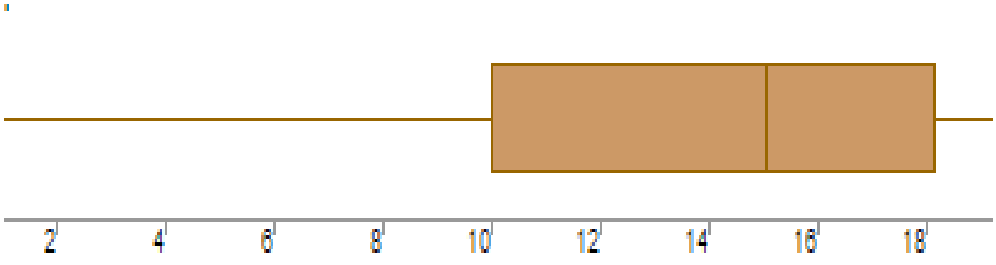
Q16) What does positive kurtosis value indicates for a data ?

➔ Positive values of kurtosis indicates that a distribution is **peaked and possess thick tails.**

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

➔ Negative values of kurtosis indicates 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?

➔ In above boxplot median is 15

➔ Q1=10, Q2=18

➔ Min=1 , Max=19

➔ IQR=8

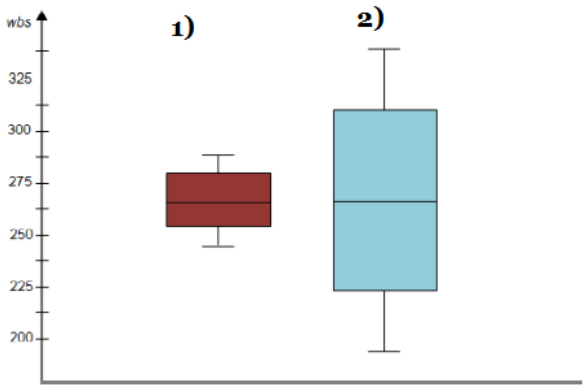
What is nature of skewness of the data?

➔ Above boxplot has negative skewness.

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

➔ IQR is8 or more than that approximately

Q19) Comment on the below Boxplot visualizations?




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

➔ Both median lines lie within the overlap between two boxes. Short boxes mean their data points consistently over around the center values. Taller boxes simply more variable data and both the boxes are without outliers.

Q 20) Calculate probability from the given dataset for the below cases

Data _set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

a. P(MPG>38)

➔ Probability of (MPG>38) = 0.4074074074074074

➔ mpg=(df3['MPG']>38).sum()

➔ total=len(df3)

➔ print("Probability of (MPG>38)= {}".format(mpg/total))

b. P(MPG<40)

➔ Probability of (MPG<40)= 0.7530864197530864

➔ mpg=(df3['MPG']<40).sum()

➔ total=len(df3)

➔ print("Probability of (MPG<40)= {}".format(mpg/total))

c. P (20<MPG<50)

➔ Probability of (20<MPG>50) = 0.8518518518518519

➔ mpg=(((df3['MPG'])>20) & ((df3['MPG'])<50)).sum()

➔ total=len(df3)

➔ print("Probability of (20<MPG>50)= {}".format(mpg/total))

Q 21) Check whether the data follows normal distribution

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

➔ Answer – By ploting histogram for MPG it is clearly seen then this data does not follows Normal distribution.

##thumb rule whether CLT will be applied

##central limit theorem (fairly large sample size)=n=>10*(skewness)^^2

## Internal estimate = points estimate +- margin of error x=sigma/root n

## from scipy import stats

stats.norm.ppf()

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

➔ Answer – By pointing histogram for Adipose Tissue (AT) and waist it is clearly seen then this data does not follows normal distribution.

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

➔ The Z scores of 90% confidence interval = 1.645
➔ The Z scores of 94% confidence interval = 1.8807
➔ The Z scores of 60% confidence interval = 0.85

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

➔ Answer - The T scores of 95% confidence interval = 2.064
➔ The T scores of 96% confidence interval = 2.085
➔ The T scores of 99% confidence interval = 2.797

Q 24)A Government companyclaims 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 – the probability that 18 randomly selected bulbs would have an average life of no more than 260 daysis 0.471