|  |  |
| --- | --- |
| **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 |

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

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

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

probability that two heads and one tail= 3/8

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

1. Equal to 1

Sol: As the sum always >1 when two dice are rolled, the answer is ‘0’

1. Less than or equal to 4

Sol: {(1,3),(3,1),(2,2),(1,1),(1,2),(2,1)}

P=6/36=1/6

1. Sum is divisible by 2 and 3

Sol:{ (1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6)}

P=5/36

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?

Sol:

the probability that none of the balls drawn is blue =

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: 1 \* 0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12=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**

**Sol:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | **3.59** | **3.21** | **17.84** |
| Median | **3.69** | **3.32** | **17.7** |
| Mode | **3.07** | **3.44** | **17.02** |
| Variance | **0.28** | **0.95** | **3.19** |
| Standard Deviation | **0.53** | **0.97** | **1.78** |
| Range | 3.59 – 4.93 | 3.21 – 5.42 | 17.84 – 22.9 |

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: Expected Value of the Weight of that patient =

 (1/9)(108) + (1/9)110  + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199 = 145.33

**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**

**Sol:** For Cars Speed : Skewness value= -0.12 and Kurtosis value= 0.81

For Cars Distance: Skewness value = 0.81 and Kurtosis value = 0.41

For SP Skewness = 1.61 kurtosis = 0.95

For WT Skewness = 1.61 Kurtosis = 0.95

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



Sol: The histogram is right skewed. We have outliers on the higher side.

**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:

Let’s solve using Z

CI =

Given ,

Find Z value from python using *stats.norm.ppf(CI value)*

|  |  |  |
| --- | --- | --- |
| CI% | CI value= (1+CI%)/2 | *Z= stats.norm.ppf(CI value)* |
| 94% =0.94 | (1+0.94)/2= 0.97 | *Z= stats.norm.ppf(0.97) = 1.88* |
| 96%=0.96 | (1+0.96)/2= 0.98 | *Z= stats.norm.ppf(0.98) = 2.05* |
| 98%=0.98 | (1+0.98)/2= 0.99 | *Z= stats.norm.ppf(0.99) = 2.32* |

Substitute every value in the above expression

CI =

For 94% CI= the range is 198.73 – 201.26

For 98% CI= the range is 198.43 – 201.56

For 98% CI= the range is 198.62 – 201.37

**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?

Sol: 1) Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05

2)No outliers present, right skewed as mean > median

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

Sol: symmetrical distribution and will have no skewness

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

Sol: Right skewed

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

Sol: Left skewed

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

Sol: Positive kurtosis means the curve is more peaked

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

Sol: Negative Kurtosis means the curve will be flatter

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



What can we say about the distribution of the data?

Sol: The above Boxplot is not normally distributed the median is towards the higher value

What is nature of skewness of the data?

Sol: The data is a skewed towards left

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

Sol: The Inter Quantile Range= Q3-Q1=18-10=8

Q19) Comment on the below Boxplot visualizations?



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

Sol: There are no outliers. The box plots are both normally distributed, with zero to no skewness at either the lowest or maximum whisker range, and they both share the same median, which is about in the range of 275 to 250.

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:

Mean of MPG column is 34.4

Standard Deviation is 9.13

P(MPG>38)= 1-stats.norm.cdf(38,34.4,9.13)=0.348

* 1. P(MPG<40)

Sol: P(MPG<40)= stats.norm.cdf(40,34.4,9.13)=0.729

* 1. P (20<MPG<50)

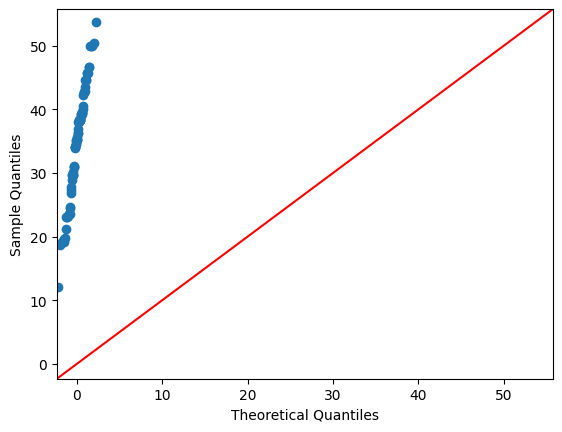
Sol: P(MPG<50) - P(MPG>20) = 0.956 - 0.943=0.013

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

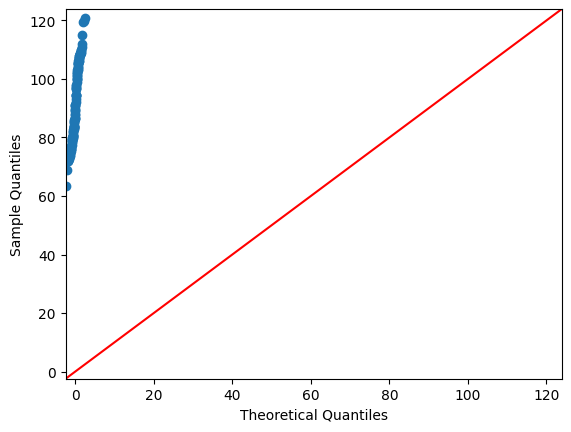
Sol: Q-Q Plot for MPG: As the data is not lying on the line it is not normally distributed



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

Dataset: wc-at.csv

Sol: Q-Q Plot: Not a normal distribution



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

Sol:

Z=stats.norm.ppf((1+CI)/2)

from scipy import stats

z1=np.abs(stats.norm.ppf((1+0.94)/2))

z2=np.abs(stats.norm.ppf((1+0.90)/2))

z3=np.abs(stats.norm.ppf((1+0.60)/2))

print(z1,z2,z3)

1.8807936081512509 1.6448536269514722 0.8416212335729143

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

Sol:

t\_crit = np.abs(t.ppf((1-confidence)/2,dof))

t1=np.abs(stats.t.ppf((1-0.95)/2,df=24))

t2=np.abs(stats.t.ppf((1-0.96)/2,df=24))

t3=np.abs(stats.t.ppf((1-0.99)/2,df=24))

print(t1,t2,t3)

2.0638985616280205 2.1715446760080677 2.796939504772804

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

Sol:

As population mean is mentioned we can go with ‘t’ rather than ‘z’

We know,

t=(sample\_mean-Population\_mean)/(sample\_SD/sqrt(n))

Given sample\_mean =260, Population\_mean=270, sample\_SD=90, n=18

Find probability that average life **of no more than 260** days

P(Average life < =260) = ?

t= ((260-270)/90)\*18^0.5 = - 0.471

df= n-1 =17

P(Average life < =260) **=**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)

P(Average life <= 260) = 0.678