|  |  |
| --- | --- |
| 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 | 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 | 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?

Total number of events= {HHH,HHT,HTH,THH,TTT,TTH,THT,HTT} = 8

Interested events = 3(HHT, HTH, THH)

Probability = 3/8

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

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

Number of possible outcomes = 6 x 6 = 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)

Equal to 1: Possible events = 0, hence the probability = 0/36 = 0

Less than or equal to 4: Possible events = 6, hence the probability = 6/36 = 1/6

Sum is divisible by 2 and 3: Possible events = 6, hence the 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?

Number of events = nCr = 7C2 = 7!/2! (7-2)! = 7!/2!5! = 5040/240 = 21

Interested events = nCr = 5C2 = 5!/2! (5-2)! = 5!/2! 3! = 120/12 = 10

Probability = 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

Expected number of candies for a randomly selected child = Summation of all probabilities

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

= 0.015 + 0.80 + 1.95 +0.025 + 0.06 + 0.240

= 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.

data = pd.read\_csv('C:/Users/Dell/Desktop/Assignments/Q7.csv')

data.describe() to find the Mean, Standard Deviation and Range. Median and Variance can be found using individual commands.

**For** **Points**:

Mean: **3.596563**

Median: data["Points"].median() = **3.695**

Mode: data["Points"].mode() = **3.92**

Standard Deviation = **0.534**

Range: **2.76**(min) **4.93**(max)

Variance: np.var(data[“Points”]) = **0.27**

**For Score**:

Mean: **3.217**

Median: data[“Score”].median() = **3.325**

Mode: data[“Score”].mode() = **3.44**

Standard Deviation: **0.978**

Range: **1.513**(min) and **5.424**(max)

Variance: np.var(data[“Score”]) = **0.927**

**For Weigh**:

Mean: **17.848**

Median: data[“Weigh”].median() = **17.71**

Mode: data[“Weigh”].mode() = **18.90**

Standard Deviation: **1.786**

Range: **14.50**(min) and **22.90**(max)

Variance: np.var(data[“Weigh”]) = **3.093**

Inference:

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?

Expected value of the weight of a patient chosen in random = sum of all weights/number of patients

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

Expected Value = 1308/9

Expected Value = 145.33

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

**Cars speed and distance**

print("Skewness")

print(df.skew(axis=0))

print("Kurtosis")

print(df.kurtosis())

Output:

Skewness

Index 0.000000

speed -0.117510

dist 0.806895

dtype: float64

Kurtosis

Index -1.200000

speed -0.508994

dist 0.405053

dtype: float64

**SP and Weight(WT)**

print("Skewness")

print(sf.skew(axis=0))

print("Kurtosis")

print(sf.kurtosis())

Output:

Skewness

Unnamed: 0 0.000000

SP 1.611450

WT -0.614753

dtype: float64

Kurtosis

Unnamed: 0 -1.200000

SP 2.977329

WT 0.950291

dtype: float64

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



Inference: The above Boxplot suggests that the distribution of the data has plenty of outliers(dots above the upper extreme)

**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: n=2000

= 200

s= 30

Confidence Interval Estimate= Z => 200 Z

1**.For 94% Confidence interval**

α =1-0.94=0.06

(t-value for 1999 by using t-table)

t 1-α,n-1=t 0.94,1999=,

+ t 0.94,1999\*s/=200+1.881861\*0.671=200+1.26272873=201.262729

- t 0.94,1999\*s/=200- 1.881861\*0.671=200-1.26272873=198.737271

[198.737271, 201.104466]

2**.For 98% confidence interval**

α =1-0.98=0.02

t 1-α,n-1=t 0.98,1999= 2.328215

+ t 0.98,1999\*s/=200+2.328215\*0.671=200+1.56223227=201.562232

- t 0.98,1999\*s/=200-2.328215\*0.671=200-1.56223227=198.437768

[198.437768, 201.562232]

**For 96% confidence interval**

α =1-0.96=0.04

t 1-α,n-1=t 0.96,1999=2.05509

+ t 0.96,1999\*s/=200+2.05509\*0.671=200+1.37896539=201.378965

- t 0.96,1999\*s/=200-2.05509\*0.671=200-1.37896539=198.621035

[198.621035, 201.378965]

**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?

data = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

Mean: np.mean(data) = 41

Median: np.median(data) = 40.5

Variance: np.var(data) = 24.11

Standard Deviation: np.std(data) = 4.91

Inference: We can see that the Mean is slightly greater than the mode and hence we can say that the distribution is skewed to the right.

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

No Skewness

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

Right Skewed/Positive Skewed

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

Left Skewed/Negative Skewed

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

Distribution has a sharp peak and heavy tails than the normal distribution

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

Distribution has lighter tails and flat peak than the normal distribution

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



What can we say about the distribution of the data?

As skewness exists, this is not a Normal Distribution

What is nature of skewness of the data?

Left Skewed

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

IQR = 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.

Inferences:

No skewness in both Boxplots

Median is almost or if not the same

No outliers seen

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)
  2. P(MPG<40)
  3. P (20<MPG<50)

1. data = pd.read\_csv('C:/Users/Dell/Desktop/Assignments/Cars.csv')

df = pd.DataFrame(data)

criteria = df[df.iloc[:,1] > 38]

print(criteria) and criteria.shape to get the number of entries with MPG>38

Probability = 33/81

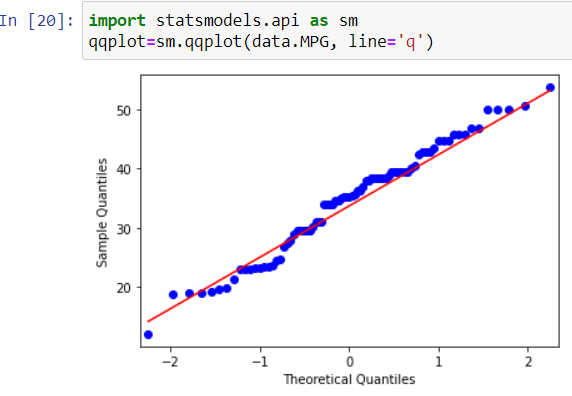
1. Similarly, criteria.shape for entries with MPG<40 is 61 and hence the probability is 61/81
2. 69/81

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

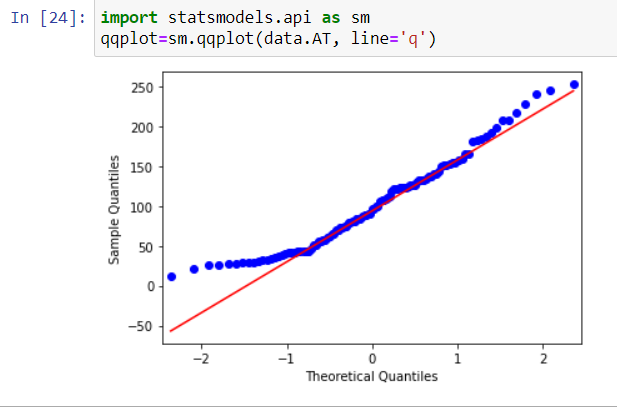
Yes, it does follow Normal Distribution.



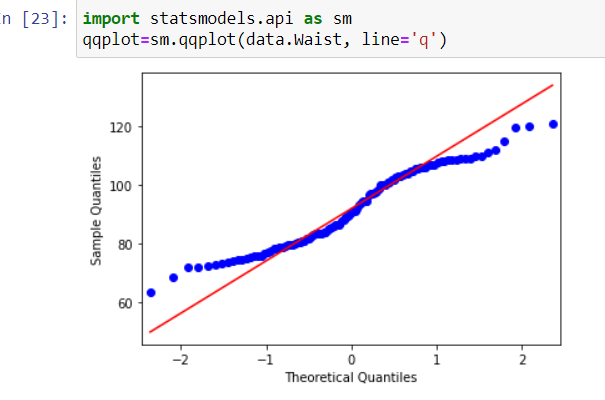
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

QQ Plot for AT can be found below and it follows Normal Distribution



QQ Plot for Waist can be found below and it follows Normal Distribution too.



The following are said to follow Normal Distribution since most of the data points fall on the red line.

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

stats.norm.ppf(.95) = 1.6448536269514722

stats.norm.ppf(.97) = 1.8807936081512509

stats.norm.ppf(.80) = 0.8416212335729143

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

stats.t.ppf(.975,25) = 2.059538552753294

stats.t.ppf(.98,25) = 2.1665866344527562

stats.t.ppf(.995,25) = 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

µ=270  
n=18  
Xbar=260  
S=90 We will use T-distribution -n<30 -ơ is Unknown t=( Xbar - µ ) / ( s / √n ) t= ( 260 -270 )/ ( 90 / √18 ) t=-0.471

stats.t.cdf(-.471,17) = 0.3218140331685075