|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete Data |
| Results of rolling a dice | Discrete Data |
| Weight of a person | Continuous Data |
| Weight of Gold | Continuous Data |
| Distance between two places | Continuous Data |
| Length of a leaf | Continuous Data |
| Dog's weight | Continuous Data |
| Blue Color | Discrete Data |
| Number of kids | Discrete Data |
| Number of tickets in Indian railways | Discrete Data |
| Number of times married | Discrete Data |
| Gender (Male or Female) | Discrete Data |

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 | Nominal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Ratio |
| Socioeconomic Status | Interval |
| Fahrenheit Temperature | Ratio |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Interval |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Ratio |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| 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?**

Let S be the sample space.

n(S) = 8

Let E = Event of two heads and one tail

* P => n(E)/n(S)​ => 3/8 => 0.375

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

Let S be the sample space.

n(S) = 36

1. Equal to 1 => 0
2. Less than or equal to 4 => 6/36 => 0.1667
3. Sum is divisible by 2 and 3 => 6/36 => 0.1667

**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?**

Total number of balls = (2 + 3 + 2) = 7  
Let S be the sample space.

Then, n(S) = Number of ways of drawing 2 balls out of 7  
=> 7C2​ => (2×1)(7×6)​ => 2  
Let E = Event of drawing 2 balls, none of which is blue.  
n(E) = Number of ways of drawing 2 balls out of (2 + 3) balls.  
=> 5C2​ => (2×1)(5×4)​ => 10  
=> P(E) => n(E)/n(S)​ => 10​/21

**Q.6) 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

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

=> 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

=> 3.090

**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**

Mean

Points 3.596563

Score 3.217250

Weigh 17.848750

Median

Points 3.695

Score 3.325

Weigh 17.710

Mode

Car points mode = 0 3.07

1. 3.92

Car weigh mode = 0 17.02

1 18.90

Car Score mode = 0 3.44

Variance

Points 0.285881

Score 0.957379

Weigh 3.193166

Standard Deviation

Points 0.534679

Score 0.978457

Weigh 1.786943

Range

Points\_Range => 2.17

Score\_Range => 3.91

Weigh\_Range => 8.3999

**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  =  ∑ ( probability  \* Value )

 ∑ P(x).E(x)

Total Number of Patients = 9

Probability of selecting one patient = 1/9

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Patient | 108 | 110 | 123 | 134 | 135 | 145 | 167 | 187 | 199 |
| Probability | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 |

Expected Value = (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

Expected Value = 145.33

Expected Value of the Weight of that patient is 145.33 kg

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

**Cars speed and distance**

**Use Q9\_a.csv**

data = pd.read\_csv("Q9\_a.csv")

data.kurt()

data.skew()

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| Speed | -0.11750986144663393 | - 0.5089944204057617 |
| Distance | 0.8068949601674215 | 0.4050525816795765 |

**Interference:**

1. **The value of skewness for speed is negative which indicates that data is left skewed.**
2. **The value of skewness for distance is Positive which indicates that data is right skewed**
3. **The value of kurtosis for speed is negative which indicates that data is not peaked.**
4. **The value of kurtosis for distance is negative which indicates that data is little peaked.**

**SP and Weight(WT)**

**Use Q9\_b.csv**

data1 = pd.read\_csv("Q9\_b.csv")

data1.kurt()

data1.skew()

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| SP | 1.611450 | 2.977329 |
| Weight | -0.614753 | 0.950291 |

**Interference:**

1. **The value of skewness for speed is positive which indicates that data is right skewed.**
2. **The value of skewness for distance is negative which indicates that data is left skewed**
3. **The value of kurtosis for speed is positive which indicates that data is highly peaked.**
4. **The value of kurtosis for distance is negative which indicates that data is little peaked.**

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



The histogram provides the detail as below

1. The data is right skewed which indicates that the frequency of chicks with weight between 300 to 400 is very less.
2. The frequency of chicks with weight 400 is less than or equal to 5 units
3. Maximum frequency of weight is between 50 and 100 units with a value of 200.



In this box plot

1. Median is lower than mean which indicates that data is right skewed.
2. There are outlier present on upper side of the box.

**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?

**import** numpy **as** np

**import** pandas **as** pd

**from** scipy **import** stats

**from** scipy.stats **import** norm

interval\_value = 0.94

weight = 200

sample = 2000

size = 30

con\_int = stats**.**norm**.**interval(interval\_value, weight, size **/**( sample **\*\***0.5))

Average weight of Adult in Mexico with 94% CI => (198.738325292158, 201.261674707842)

Average weight of Adult in Mexico with 98% CI => (198.43943840429978, 201.56056159570022)

Average weight of Adult in Mexico with 96% CI => (198.62230334813333, 201.37769665186667)

**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.**

Mean = 41

Median = 40.5

Variance = 25.529

Standard deviation = 5.0526

1. **What can we say about the student marks?**
2. The average percentage of these student is 65%.
3. Most of the student scored mediocre in the test.
4. Only a single student managed to score above 90%.

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

* If the mean and median along with mode is equal the skewness in zero

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

* If the mean is greater than the median, the distribution is positively skewed

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

* If the mean is greater than the median, the distribution is negatively skewed

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

* Positive values of kurtosis indicate that distribution is peaked and possesses thick tails

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

* Negative excess values of kurtosis (<3) indicate 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?**

The data is no evenly distributed .The distribution is negatively skewed. Also the median doesn’t located at the center of the plot.

**What is nature of skewness of the data?**

Negatively skewed as median is greater than means.

**What will be the IQR of the data (approximately)?**   
  
IQR = 8 unit (approx.)

**Q19) Comment on the below Boxplot visualizations?**



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

1.The value in box plot 2 is more distributed as compared to box plot 1.

2. The values in plot 1 lies between 225 to 275(approx.)

3.The values in plot 2 lies between 200 to 350(approx.)

4. The median in plot 2 seems to be at the center which indicates symmetric distribution of data. The median in plot 1 seems to be slightly below the mean which indicates the left skewed data.

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)

c. P (20<MPG<50)

a. mean(MPG>38)= 34.42208

sd(MPG>38)= 9.131445

P = 1 - pnorm(38,mean(MPG),sd(MPG)) = 0.33 = 33%

b. mean(MPG<40)= 34.42208

sd(MPG<40)= 9.131445

P = pnorm(40,mean(MPG),sd(MPG)) = 0.7293499 = 72.93%

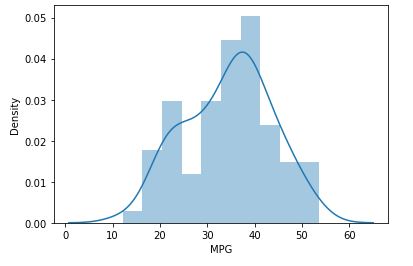
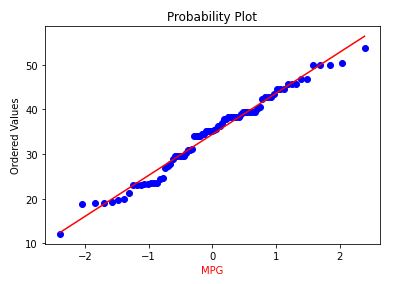
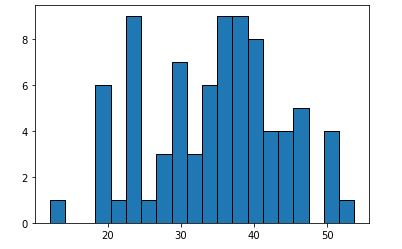
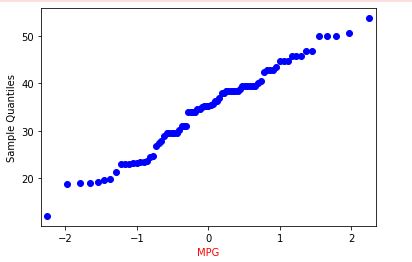
c. P = pnorm(50,mean(MPG),sd(MPG)) – pnorm(20,mean(MPG),sd(MPG))

P = 0.955 -0.057 = 0.8988689 =89.88%

Q 21) Check whether the data follows normal distribution

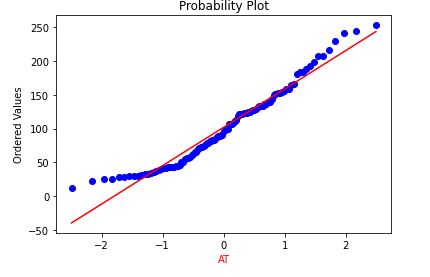
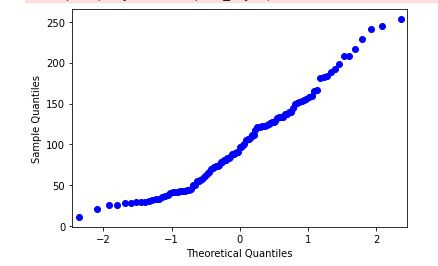
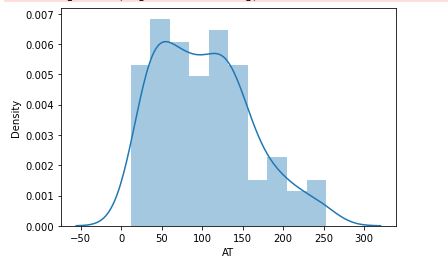
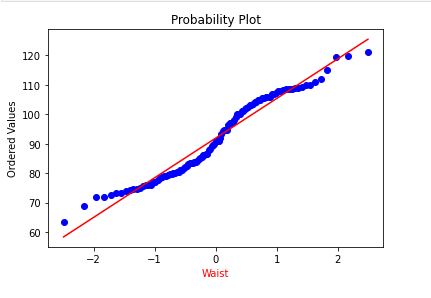
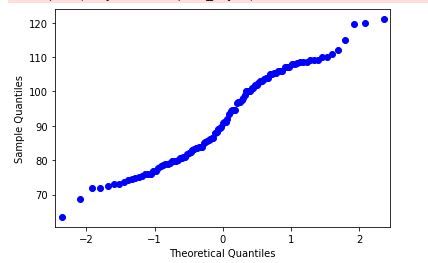
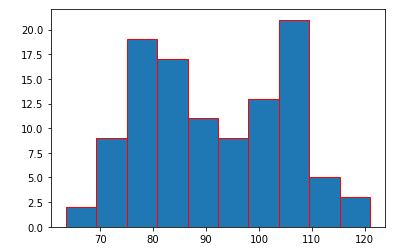
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



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

Dataset: wc-at.csv



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

from scipy import stats

from scipy.stats import norm

#z-score of 90% confidence interval

stats.norm.ppf(0.95) => 1.6448536269514722

#z-score of 94% confidence interval

stats.norm.ppf(0.97) => 1.8807936081512509

#z-score of 60% confidence interval

stats.norm.ppf(0.80) => 0.8416212335729143

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

import numpy as np

from scipy.stats import t

x = np.random.normal(size=25)

dof = len(x)-1

confidence = 0.95,0.96,0.99

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

t\_score(0.95) = 2.0638985616280205

t\_score(0.96) = 2.1715446760080677

t\_score(0.99) = 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

Solution:-

import numpy as np Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

= (260-270)/90/np.sqrt(18)) t\_score

P = -0.471 stats.t.cdf(t\_score, df = 17) 0.32

= 32%