|  |  |
| --- | --- |
| 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) | Interval |
| 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 |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

Answer:

The possible outcomes will be (HHH,TTT,HTT,TTH,THT,THH,HTH,HHT).

So the favorable outcomes = THH,HTH,HHT

Number of favorable outcomes = 3

Probability of event to happen = number of favorable outcomes / total number of outcomes

Probability=3/8= 0.375

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

Answer:

1. Equal to 1 is 0
2. The sum is equal to 4 the possible outcomes are (1,3),(2,2),(3,1) therefore n(b) = 3/36 = 1/12
3. Their sum is divisible by 2 and 3 are (1,5),(2,4),(3,3),(4,2),(5,1),(6,6) = 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?

Answer:

Total number of balls = 2 + 3 + 2 = 7

Number of ways of drawing 2 balls out of 7C2 = (7 x 6)/(2x1) = 42/2 = 21

* Number of balls other than blue = 5

Number of ways of drawing 2 balls out of 5 = 5C2 = (5x4)/(2x1) = 20/2 = 10

* Required 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

Answer:

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(0.120)

= 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**

Answer:

* Mean for Points = 3.59, Score = 3.21, Weigh = 17.84
* Median for Points = 3.69, Score = 3.32, Weigh = 17.71
* Mode for Points = 3.07, Score = 3.44, Weigh = 17.02
* Variance for Points = 0.28, Score = 0.95, Weigh = 3.19
* Standard deviation for Points = 0.53, Score = 0.97, Weigh = 1.78

Inferences:

* Here in this case of different models of cars data, most type of cars have average points of 3.596563, score of 3.217150 and weigh of 17.848750. Also here in this scenario the standard deviation is very low in points and score so chances of presence of outliers in both the case is very low and comparing to weigh there is little bit higher standard deviation so may be some outliers are present.
* Somehow data points in every case have less spread so most of the data points lie near to the median.

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?

Answer:

Expected value = sum(x\*probability of x)

= (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**

**Answer:**

* **Car speed and distance**

**Use Q9\_a.csv**

**Import pandas as pd**

**Q9\_a = pd.read\_csv(“Q9\_a.csv”)**

**Q9\_a**

**Q9\_a.skew()**

**Q9\_a.kurt()**

Skewness for speed = -0.117510

Skewness for distance = 0.806895

Kurtosis for speed = -0.508994

Kurtosis for distance = 0.405053

**Inferences:**

* **For car speed skewness is negative and also the kurtosis is negative, which suggests that the distribution is more towards left. It means the distribution is left skewed or negative skewed. Here in negative skewed mean is less than median. As taking kurtosis into consideration it shows that the distribution has broad peak and thin tail.**
* **For the distance travel by the car skewness is positive and also the kurtosis is positive, which suggests that the distribution is more towards right. It means the distribution is right skewed or positive skewed. Here in positive skewed mean is greater than median. As taking kurtosis into consideration it shows that the distribution has pointed peak and wide tail.**
* **SP and Weight(WT)**

**Use Q9\_b.csv**

**Import pandas as pd**

**Q9\_b = pd.read\_csv(“Q9\_b.csv”)**

**Q9\_b**

**Q9\_b.skew()**

**Q9\_b.kurt()**

Skewness of SP = 1.611450

Skewness of Weight(WT) = -0.614753

Kurtosis of SP = 2.977329

Kurtosis of Weight(WT) = 0.950291

**Inferences:**

* For SP skewness is positive and also the kurtosis is positive, which suggests that the distribution is more towards right. It means the distribution is right skewed or positive skewed. Here in positive skewed mean is greater than median. As taking kurtosis into consideration it shows that the distribution has pointed peak and wide tail.
* For WT skewness is negative and the kurtosis is positive, which suggests that the distribution is more towards left. It means the distribution is left skewed or negative skewed. Here in negative skewed mean is less than median.

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



Answer:

* The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side.
* The boxplot has outliers on the maximum 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?

Answer:

1. Import numpy as np

From scipy import stats

From math import sqrt

ci\_94 = stats.norm.interval(0.94,200,scale = (30/sqrt(2000)))

print("Weight at 94% confidence interval is :",np.round(ci\_94,4))

output : Weight at 94% confidence interval is : [198.7383 - 201.2617]

1. ci\_98 = stats.norm.interval(0.98,200,scale = (30/sqrt(2000)))

print("Weight at 98% confidence interval is :",np.round(ci\_98,4))

output : Weight at 98% confidence interval is : [198.4394 - 201.5606]

1. ci\_96 = stats.norm.interval(0.96,200,scale = (30/sqrt(2000)))

print("Weight at 96% confidence interval is :",np.round(ci\_96,4))

output : Weight at 96% confidence interval is : [198.6223 - 201.3777]

**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.

Answer:

Mean = 41

Median = 40.5

Variance = 25.52

Standard deviation = 5.05

1. What can we say about the student marks?

Answer:

The student score 41 mark most of the time. He scores average 41 mark.

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

Answer:

The nature of skewness is zero.

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

Answer:

Skewness and tail is towards right.

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

Answer:

Skewness and tail is towards left.

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

Answer:

Positive kurtosis means the curve is more peaked and it is leptokurtic.

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

Answer:

Negative kurtosis means the curve will be flatter and broader.

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



* What can we say about the distribution of the data?

**Answer:** The above boxplot is not normally distributed the median is towards the higher value.

* What is nature of skewness of the data?

**Answer:** The data is a skewed towards left. The whisker range of minimum value is greater than maximum.

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

**Answer:** The Inter Quantile Range = q3 upper quanrtile – q1 lower quartile

=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.

Answer:

First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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)

Answer:

Import pandas as pd

Cars[“MPG”].mean()

34.42

Cars[“MPG”].std()

9.13

1. P(MPG>38)

From scipy import stats

1-stats.norm.cdf(38,34.42,9.13)

0.3474

1. P(MPG<40)

stats.norm.cdf(40,34.42,9.13)

0.7294

1. P(20<MPG<50)

stats.norm.cdf(50,34.42,9.13) – stats.norm.cdf(20,34.42,9.13)

0.8989

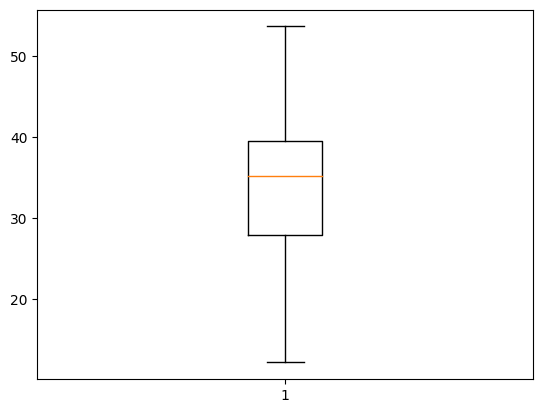
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

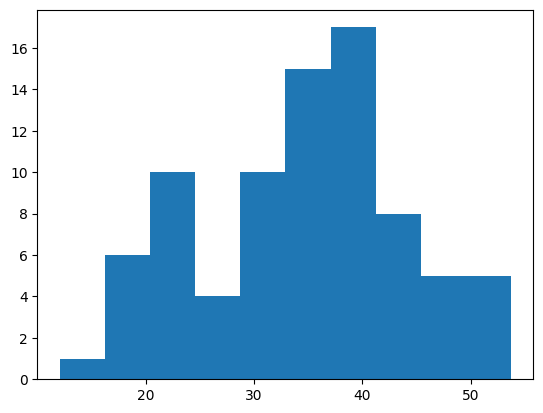
Answer:

Import matplotlib.pyplot as plt

Plt.boxplot(cars[“MPG])



Plt.hist(cars[“MPG”])



From this above boxplot and histogram we can say the MPG of cars follows normal distribution.

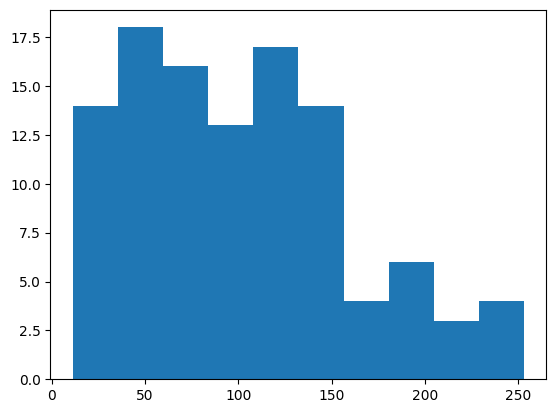
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

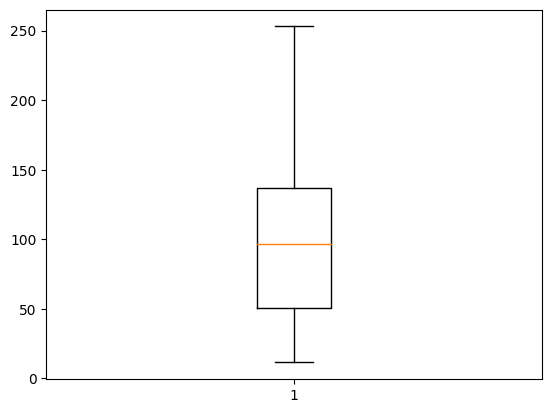
Answer:

Import matplotlib.pyplot as plt

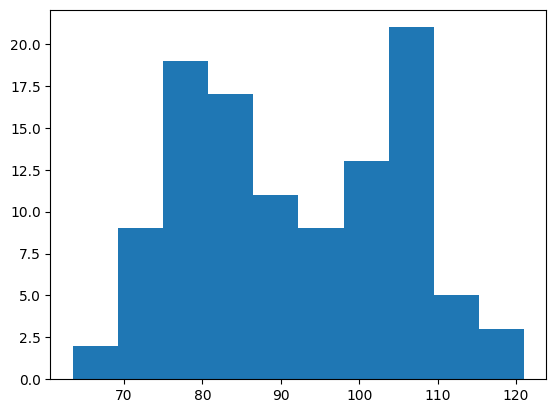
Plt.hist(wc-at[“AT”])

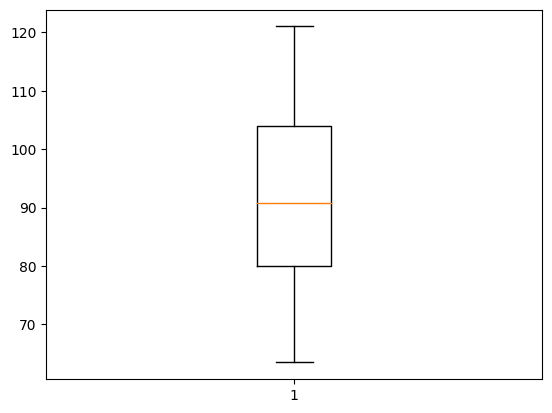


Plt.boxplot(wc\_at[“AT”]



Plt.hist(wc\_at[“Waist”])



Plt.boxplot(wc\_at[“Waist”]

From the above histogram and boxplot for both AT &Waist of wc-at dataset, it shows that both AT & Waist follows normal distribution.

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

Answer:

from scipy import stats

stats.norm.ppf(0.95)

Z scores of 90% confidence interval = 1.65

from scipy import stats

stats.norm.ppf(0.97)

Z scores of 94% confidence interval = 1.89

from scipy import stats

stats.norm.ppf(0.80)

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:

from scipy import stats

stats.t.ppf(0.975,24)

t score of 95% confidence interval for sample size of 25 = 2.064

from scipy import stats

stats.t.ppf(0.98,24)

t score of 96% confidence interval for sample size of 25 = 2.172

from scipy import stats

stats.t.ppf(0.995,24)

t score of 99% confidence interval for sample size of 25 = 2.797

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

Answer:

Population mean = 270

Sample size, n = 18

Sample mean,x = 260

Standard deviation, s =90

t score = (x – population mean) / (s / sqrt(n))

=(260 – 270) / (90 / sqrt(18))

= -10 / 21.23

= -0.47

df = degree of freedom = n – 1 = 18 – 1 = 17

Probability

Stats.t.cdf( -0.47, 17 )

Ans = 0.3221639