|  |  |
| --- | --- |
| 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 data |
| High School Class Ranking | Ordinal data |
| Celsius Temperature | Interval data |
| Weight | Ratio data |
| Hair Color | Nominal data |
| Socioeconomic Status | Ordinal data |
| Fahrenheit Temperature | Interval data |
| Height | Ratio data |
| Type of living accommodation | Nominal data |
| Level of Agreement | Ordinal data |
| IQ(Intelligence Scale) | Ordinal data |
| Sales Figures | Ratio data |
| Blood Group | Nominal data |
| Time Of Day | Ordinal data |
| Time on a Clock with Hands | Ordinal data |
| Number of Children | Ratio data |
| Religious Preference | Nominal data |
| Barometer Pressure | Interval data |
| SAT Scores | Interval data |
| Years of Education | Ratio data |

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

Answer:

HHT or HTH or THH

=1/2\*1/2\*1/2 (+) 1/2\*1/2\*1/2 (+) 1/2\*1/2\*1/2

= 1/8 + 1/8 + 1/8

= 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

Answer:

1. Equal to 1

Probability = 0

1. Less than or equal to 4

= (1,1),(1,2)(1,3)(2,1)(2,2)(3,1)=6/36=0.167

Probability=0.16

c)Sum is divisible by 2 and 3

=(1,5)(2,4)(3,3)(4,2)(5,1)(6,6)=6/36=0.167

Probability=0.167

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 balls=7, Red=2, Green=3, Blue=2

Probability that none of the balls drawn is blue =5/7\*4/6=20/42

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 will be

= 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

= 3.09

Expected number of candies for a randomly selected child = 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:

Code:

import pandas as pd

Data=pd.read\_csv('Q7.csv')

X = Data[["Points", "Score", "Weigh"]]

X.describe()

X.mean()

X.median()

X.mode()

X.var()

X.std()

Output:

Points :- Mean = 3.596, Median = 3.659, Mode = 3.92,

Variance =0.285, Standard deviation =0.534.

Score :- Mean = 3.217, Median = 3.325, Mode = 3.44,

Variance =0.957, Standard deviation =0.978.

Weigh :- Mean = 17.848, Median = 17.710, Mode = 18.90,

Variance =3.197, Standard deviation =1.786.

Inferences:

From the data, we observe that mean, median and mode are not equal.Hence, we can say that our data can be skewed .

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:

We have to find mean for the given data,

Code :

import numpy as np

weights\_of\_patients = [108, 110, 123, 134, 135, 145, 167, 187, 199]

np.mean(weights\_of\_patients)

Output : 145.33

Expected Value of the Weight of that patient chosen at random = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Answer:

Code:

Data = pd.read\_csv('Q9\_a.csv')

Data .skew()

Data .kurtosis()

Output: Skewness

speed = -0.117510

dist = 0.806895

Output: Kurtosis

Speed = -0.508994

dist = 0.405053

Inferences:

From the skewness of speed we can say that the data of speed is nearly symmetric, and from the distance we can say that the data is positively skewed.

**SP and Weight(WT)**

**Use Q9\_b.csv**

Answer:

Code:

Data = pd.read\_csv('Q9\_b.csv')

Data.skew()

Data.kurtosis()

Output: Skewness

SP = 1.611450

WT = -0.614753

Output: Kurtosis

SP = 2.977329

WT = 0.950291

Inferences:

From the skewness of SP we can say that the data of SP is positively skewed, and from the skewness of weight we can say that the data is negatively skewed.

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



Answer :

From the above histogram we can say that the data is Positively skewed



Answer:

From the Boxplot we can say that there are some outliers on the upper 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:

Code:

import numpy as np

from scipy import stats

stats.norm.interval(alpha=0.94, loc=200, scale= 30/np.sqrt(2000)) # 94% CI

stats.norm.interval(alpha=0.96, loc=200, scale= 30/np.sqrt(2000)) # 96% CI

stats.norm.interval(alpha=0.98, loc=200, scale= 30/np.sqrt(2000)) # 98% CI

Output:

For 94% confidence interval Range is [ 198.73 – 201.26]

For 98% confidence interval range is [198.43 – 201.56]

For 96% confidence interval 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.

Answer:

Code :

import numpy as np x=[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

np.mean(x)

np.median(x)

np.var(x)

np.std(x)

Output :

Mean = 41

Median = 40.5

Variance = 24.11

Standard Deviation = 4.91

1. What can we say about the student marks?

Answer: Student scores 41 marks as an average and We don’t have

Outliers in the data

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

Answer : The nature of skewness is perfectly symmetric that is it is zero skewed

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

Answer : The nature of skewness is positively skewed.

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

Answer : The nature of skewness is negatively skewed

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 nature of skewness is negatively skewed.

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

Answer : IQR = Q3-Q1

IQR = 18-10

IQR = 8

Q19) Comment on the below Boxplot visualizations?



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

Answer : Here both the Box plots follows normal distribution. The difference is first boxplot have lesser range compared to second boxplot.

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:

Code:

import pandas as pd

from scipy.stats import norm

Data=pd.read\_csv("Cars.csv")

Mean=Data['MPG'].mean()

std=Data['MPG'].std()

N=norm(Mean,std)

#P(MPG<40)

N.cdf(40)

#P(MPG>38)

1-N.cdf(38)

#P(20<MPG<50)

N.cdf(50)-N.cdf(20)

Output :

a) P (MPG>38) = 0.3475

b) P (MPG<40) = 0.7293

c) P (20<MPG<50) = 0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer:

Code:

import pandas as pd

Data=pd.read\_csv("Cars.csv")

Data['MPG'].hist()

Data['MPG'].skew()

Output:

Skewness: -0.17

From the observation of Histogram and skewness of “MPG”,

we can say that the data 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 pandas as pd

Data=pd.read\_csv("wc-at.csv")

Data['AT'].hist()

Data['AT'].skew()

Data['Waist'].hist()

Data['Waist'].skew()

Output:

Skewness of AT : 0.5848

Skewness of Waist : 0.1340

From the observations of Histograms and skewness of the AT and Waist, we can say that, the data is normally distributed for ‘Waist’ and data is positively skewed for ‘AT’.

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

Answer:

Code :

from scipy import stats

stats.norm.ppf(0.95)

stats.norm.ppf(0.97)

stats.norm.ppf(0.80)

Output :

90% confidence interval = 1.64485

94% confidence interval = 1.88079

60% confidence interval = 0.2533

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

Answer:

Code :

from scipy import stats

stats.t.ppf(0.975,df=24)

stats.t.ppf(0.98,df=24)

stats.t.ppf(0.995,df=24)

Output : t scores for

95% confidence interval = 2.0636

96% confidence interval = 2.1715

99% confidence interval = 2.7969

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:

Given,

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

Code:

import numpy as np

Import scipy.stats as stats

t\_score = (x - μ ) / (s/ sqrt(n))

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

t\_score = -0.471

df = n-1 = 18-1 = 17

stats.t.cdf(t\_score, df = 17)

0.32 = 32%