|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Countable |
| Results of rolling a dice | Countable |
| 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 | Nominal |
| Number of kids | Countable |
| Number of tickets in Indian railways | Countable |
| Number of times married | Countable |
| Gender (Male or Female) | Nominal |

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

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

Ans: Probability of 2 heads and 1 tail:

P(2) = (H,H,T)+(H,T,H)+(T,H,H)

= 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

Ans: Total no of events n(d)=36

a) Equal to 1: the probability that the sum is equal to zero

Probability that sum is equal to zero n(A) = 0.

P(A)= n(A)/n(d)= 0/36 = 0.

b) less than or equal to 4:

n(B)=6

P(X<=4) = n(B)/n(d)= 6/36 = 0.16

c) Sum is divisible by 2 and 3:

n(C) = 6

P(X/2&3) = n(C)/n(d) = 6/36 = 0.16

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?

Ans: Total no of balls n(b) = 2+3+2 = 7 balls

n(b) = 7C2

= 7\*6/2\*1 = 42/2= 21.

Remove blue balls: n(A) = 5C2

= 5\*4/2\*1 = 20/2= 10.

Probability P(X)= n(A)/n(b) = 10/21 = 0.47

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

Ans: 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

Expected number of candies = 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**

**Ans:**

**‘Points’**

import pandas as pd

import numpy as np

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

df

df.dtypes

output:

Unnamed: 0 object

Points float64

Score float64

Weigh float64

dtype: object

x1 = df['Points'].mean()

x2 = df['Points'].median()

x3 = df['Points'].mode()

x4 = df[‘Points’].var()

x5 = (df[‘Points’].max())-(df[‘Points’].min())

x6 = df['Points'].std()

print(x1)

print(x2)

print(x3)

print(x4)

print(x5)

print(x6)

Output:

3.5965625000000006

3.6950000000000003

0 3.07

1 3.92

Name: Points, dtype: float64

0.28588135080645166

2.17

0.5346787360709716

**‘Score’**

import pandas as pd

import numpy as np

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

df

df.dtypes

output:

Unnamed: 0 object

Points float64

Score float64

Weigh float64

dtype: object

x1 = df['Score'].mean()

x2 = df['Score'].median()

x3 = df[‘Score’].var()

x4 = (df[‘Score’].max())-(df[‘Score’].min())

x5 = df['Score'].std()

x6 = df['Score'].mode()

print(x1)

print(x2)

print(x3)

print(x4)

print(x5)

print(x6)

Output:

3.2172499999999995

3.325

0.9573789677419356

3.91

0.9784574429896967

0 3.44

**‘Weigh’**

import pandas as pd

import numpy as np

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

df

df.dtypes

output:

Unnamed: 0 object

Points float64

Score float64

Weigh float64

dtype: object

x1 = df['Weigh'].mean()

x2 = df[' Weigh '].median()

x3 = df[‘Weigh’].var()

x4 = (df[‘Weigh’].max())-(df[‘Score’].min())

x5 = df[' Weigh '].std()

x6 = df[' Weigh '].mode()

print(x1)

print(x2)

print(x3)

print(x4)

print(x5)

print(x6)

Output:

17.848750000000003

17.71

3.193166129032258

8.39

1.7869432360968431

0 17.02

1 18.90

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?

Ans: Total number of expected values P(A) = 9

Expected value = Ʃ P(A).E(A)

Sum of expected values E(A)= 108+110+123+134+135+145+167+187+199 = 1308

Expected value = (1/9)\*1308

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

1. For Cars speed **:**

Skewness = -0.1175; Kurtosis = -0.5089

1. For Cars distance :

Skewness = 0.8068; Kurtosis = 0.4050

**SP and Weight(WT)**

**Use Q9\_b.csv**

1. For SP :

Skewness = 1.6114; Kurtosis = 2.9773

1. For WT:

Skewness = -0.6147; Kurtosis = 0.9502

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



Ans: Histogram:- Chick weight data is right skewed or positively skewed according to above data

More than 50% Chick Weight is between 50 to 150. ---- Yes

Most of the chick weight is between 50 to 100. --- Yes





Ans: Right side skewed and outliers are moved upperside according to above data

**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?

Ans:

from scipy import stats

#alpha = 6

a= stats.norm.interval(0.94, loc=200, scale=30)

print(a)

#alpha = 2

b = stats.norm.interval(0.96, loc = 200, scale=30)

print(b)

#alpha = 4

c = stats.norm.interval(0.92, loc = 200, scale =30)

Print(c)

Output:

(143.57619175546247, 256.42380824453755)

(138.38753268104531, 261.61246731895466)

(130.2095637787748, 269.7904362212252)

**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.

Ans: import numpy as np

df={"Scores":[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]}

f=pd.DataFrame(df)

print(f["Scores"].mean())

print(f["Scores"].median())

print(f["Scores"].var())

print(f["Scores"].std())

**OUTPUT:**

41.0

40.5

25.529411764705884

5.05266382858645

1. What can we say about the student marks?

Ans: Most of the students in class obtained marks in the range between 35 to 41. Only few students obtained marks above 41 and least marks obtained by very few students.

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

Ans: Perfectly symmetric to the center.

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

Ans: Right Skewed data or positively skewed.

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

Ans: Left skewed data or negatively skewed

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

Ans: The peakedness of graph is Sharp(high) *i.e.,* above symmetrical.

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

Ans: The peakedness of graph is low *i.e.,* below symmetrical.

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



What can we say about the distribution of the data?

Ans: The data is distributed between the range of 10-18.

What is nature of skewness of the data?

Ans: Left skewed.

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

Ans: Inter quarlite Range(IOR) = 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.

Ans: 1.There are no outliers in the above data.

2. Both the box plot shares the same median that is in a range between 275 to 250.

3. They are normally distributed with zero and 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)

**Ans:**

**Code:**

**A:**

from scipy.stats import norm

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

a = df["MPG"].mean()

b=df["MPG"].mean()

c=df["MPG"].std()

d= norm(b,c)

e=1-d.cdf(38)

print(e)

**OUTPUT:**

0.3475939251582705

**B:**

from scipy.stats import norm

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

a = df["MPG"].mean()

b=df["MPG"].mean()

c=df["MPG"].std()

d= norm(b,c)

e=d.cdf(40)

print(e)

**OUTPUT:**

0.7293498762151616

**C:**

from scipy.stats import norm

df=pd.read\_csv("D:/Ds Csv files Assignment/Cars.csv")

a = df["MPG"].mean()

b=df["MPG"].mean()

c=df["MPG"].std()

d= norm(b,c)

e=d.cdf(50)

f= d.cdf(20)

g=e-f

print(g)

**OUTPUT:**

0.8988689169682046

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: The MPG of cars follows the 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

Ans: Adipose Tissue(AT) and Waist circumference(Waist) does not follow the normal distribution.

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

Ans: from scipy.stats import norm

from scipy import stats

# z score for 90% confidence interval

Stats.norm.ppf(0.95)

Output:

1.6448536269514722

# z score for 94% confidence interval

Stats.norm.ppf(0.97)

Output:

1.8807936081512509

# z score for 60% confidence interval

Stats.norm.ppf(0.8)

Output:

0.8416212335729143

Q 23) Calculate the t scores of 95% confidence inte From scipy import stats

Ans: From scipy.stats import norm

# t scores of 95% confidence interval for sample size of 25

stats.t.ppf(0.975,24) # df = n-1 = 24

output:

2.0638985616280205

# t scores of 96% confidence interval for sample size of 25

Stats.t.ppf(0.98,24)

Output:

2.1715446760080677

# t scores of 99% confidence interval for sample size of 25

Stats.t.ppf(0.995,24)

Output:

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

Ans: From scipy import stats

From scipy.stats import norm

# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

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

t

# p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function

p\_value=1-stats.t.cdf(abs(-0.4714),df=17)

p\_value

# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function

p\_value=stats.t.sf(abs(-0.4714),df=17)

p\_value

# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

**Output:**

-0.4714045207910317

0.32167411684460556

0.32167411684460556