|  |  |
| --- | --- |
| 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 | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Interval |

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

**Answer:**

Total number of Event is: 8(HHH,HHT,HTH,THH,HHT,TTH,HTT,THT)

No. of interested events for 2 heads and 1 tail is: 3 (HTH,THH,HHT)

Probability: No. of interested events/total no of events

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:**

S={(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)}

1. Since there are no outcomes which correspond to a sum equal to 1

Therefore in this case i.e., equal to 1 is

Probability = No. of interested events/Total no of events

Probability = 0/36=0

1. Less than or equal to 4 {(1,1) (1,2) (1,3) (2,1) (2,2) (3,1)}

Probability = No. of interested events/Total no of events

Probability = 6/36=1/6 =0.16666

1. Sum is divisible by 2 and 3

Divisible by 2 are 2,4,6,8,10 and 12 {(1,1) (1,3) (1,5) (2,2) (2,4) (2,6) (3,1)(3,3) (3,5) (4,2) (4,4) (4,6) (5,1) (5,3) (5,5) (6,2) (6,4) (6,6)}

Probability = No. of interested events/Total no of events

Probability = 18/36=1/2 (0.5)

Divisible by 3 are 3,6,9 and 12 {(1,2) (1,5) (2,1) (2,4) (3,3) (3,6) (4,2) (4,5)(5,1) (5,4) (6,3) (6,6)}

Probability = No. of interested events/Total no of events

Probability = 12/36=1/3 (0.333)

Therefore sum of divisible by 2 and 3 are

1/2+1/3=5/6 (0.8333)

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 numbers of balls are (2+3+2=7)

Total of ways that the balls can be picked out of 7 balls are 2 randomly [N(s)]

ncr = n!/(n-r)!(r)!

7c2 = (7\*6)/2 = 21

N(s) = 21

Total balls after picking the 2 balls out of 7 balls is 5 [N(e)]

ncr = n!/(n-r)!(r)!

5c2 = (5\*4)/2 = 10

N(e) = 10

Therefore probability = N (e)/N(s)

i.e., 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:**

Child A:

Probability=0.015

Total event =1

Interested event/Expected candies =?

So formula:

Probability = No. of interested events/Total no of events

No. of interested events= Probability\* Total no of events

No. of interested events/ Expected candies(child A)=0.015\*1=0.015

Similarly, for Child B

No. of interested events/ Expected candies(child B) =0.20\*4=0.80

Similarly other

= 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**

**Answer:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

cars = pd.read\_csv(r'C:\Users\Bssuraj\Desktop\Assignments\Q7.csv')

cars

**#mean\_value**

mean\_points\_value = cars['Points'].mean()

print('Mean Value of Column Points: ', mean\_points\_value)

mean\_score\_value = cars['Score'].mean()

print('Mean Value of Column Score: ', mean\_score\_value)

mean\_weigh\_value = cars['Weigh'].mean()

print('Mean Value of Column Weigh: ', mean\_weigh\_value)

**#median\_Value**

median\_points\_value = cars['Points'].median()

print('Median Value of Column Points: ', median\_points\_value)

median\_score\_value = cars['Score'].median()

print('Median Value of Column Score: ', median\_score\_value)

median\_weigh\_value = cars['Weigh'].median()

print('Median Value of Column Weigh: ', median\_weigh\_value)

**#mode\_Value**

mode\_points\_value = cars['Points'].mode()

print('Mode Value of Column Points:\n', mode\_points\_value,'\n')

mode\_score\_value = cars['Score'].mode()

print('Mode Value of Column Score:\n', mode\_score\_value,'\n')

mode\_weigh\_value = cars['Weigh'].mode()

print('Mode Value of Column Weigh:\n', mode\_weigh\_value,'\n')

**#variance**

var\_points\_value = cars['Points'].var()

print('Variance Value of Column Points: ', var\_points\_value)

var\_score\_value = cars['Score'].median()

print('Variance Value of Column Score: ', var\_score\_value)

var\_weigh\_value = cars['Weigh'].median()

print('Variance Value of Column Weigh: ', var\_weigh\_value)

**#Satndard Deviation**

std\_points\_value = cars['Points'].std()

print('Satndard Deviation of Column Points: ', std\_points\_value)

std\_score\_value = cars['Score'].std()

print('Satndard Deviation of Column Score: ', std\_score\_value)

std\_weigh\_value = cars['Weigh'].std()

print('Satndard Deviation of Column Weigh: ', std\_weigh\_value)

**#Range**

cars.describe()

points\_range=cars.Points.max()-cars.Points.min()

print('Points Range:',points\_range)

score\_range=cars.Score.max()-cars.Score.min()

print('Score Range:',points\_range)

weigh\_range=cars.Weigh.max()-cars.Weigh.min()

print('Weigh Range:',points\_range)

**#Boxplot**

f,ax=plt.subplots(figsize=(15,5))

plt.subplot(1,3,1)

plt.boxplot(cars.Points)

plt.title('Points')

plt.subplot(1,3,2)

plt.boxplot(cars.Score)

plt.title('Score')

plt.subplot(1,3,3)

plt.boxplot(cars.Weigh)

plt.title('Weigh')

plt.show()

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:**

**#given data**

**#weights as X in pounds**

X = [108, 110, 123, 134, 135, 145, 167, 187, 199]

**#calculate the no of patients**

N = len(X)

print('No of patients:', N)

**#calculate the expected value**

exp\_value = sum(X) / N

print('Expected Value:', exp\_value)

**Output:**

No of patients: 9

Expected Value: 145.33333333333334

**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:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

data = pd.read\_csv(r'C:\Users\Bssuraj\Desktop\Assignments\Q9\_a.csv')

data

**#Skewness**

data.skew()

**#Kurtosis**

data.kurt()

data1 = pd.read\_csv(r'C:\Users\Bssuraj\Desktop\Assignments\Q9\_b.csv')

data1

**#Skewness**

data1.skew()

**#Kurtosis**

data1.kurt()

**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.



**Answer:**

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:**

import math

from scipy import stats

from scipy.stats import norm

**#Given Data**

x\_barh = 200

sample\_std\_dev = 30

sample\_size = 2000

**#CI = 94% = 0.94**

alpha = 1 - 0.94

stats.norm.ppf(1-alpha/2)

lower\_limit = (200 - (1.88079 \* (30/math.sqrt(2000))))

lower\_limit

upper\_limit = (200 + (1.88079 \* (30/math.sqrt(2000))))

upper\_limit

**#CI = 98% = 0.98**

alpha = 1 - 0.98

stats.norm.ppf(1-alpha/2)

lower\_limit = (200 - (2.32634 \* (30/math.sqrt(2000))))

lower\_limit

upper\_limit = (200 + (2.32634 \* (30/math.sqrt(2000))))

upper\_limit

**#CI = 96% = 0.96**

alpha = 1 - 0.96

stats.norm.ppf(1-alpha/2)

lower\_limit = (200 - (2.053748 \* (30/math.sqrt(2000))))

lower\_limit

upper\_limit = (200 + (2.053748 \* (30/math.sqrt(2000))))

upper\_limit

**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:**

import matplotlib.pyplot as plt

import seaborn as sns

import scipy.stats as stats

import pandas as pd

import numpy as np

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

df = pd.DataFrame(marks)

df

**#MEAN**

df.mean()

**#MEDIAN**

df.median()

**#MODE**

df.mode()

**#Standard Deviation**

df.std()

**#Variance**

df.var()

**#Line plot for representing marks**

plt.plot(df)

**#boxplot for checking outliers**

plt.boxplot(df)

plt.grid()

1. What can we say about the student marks?

**Answer:**

We don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

**Answer:**

If the mean is equal to the median as well as the mode, hence the skewness is zero. If the distribution is symmetric, the mean equals the median, and the skewness of the distribution is zero.

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

**Answer:**

If the mean is greater than the median, the distribution is positively skewed. If the mean is less than the median, the distribution is negatively skewed.

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 indicates heavier tails and a more peaked distribution

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

**Answer:**

Negative kurtosis indicates lighter tails and a flatter distribution.

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:**

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 when we compare box plot 1 with box plot 2 we can say that the data in boxplot 1 is widely spread. Here the main inference is that since the data range varies high in box plot 2 it is hard to make a prediction in box plot 2. The median in the 2box plots are equal. And the data spread in both of them are symmetrical

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)

**Answer:**

import numpy as np

import pandas as pd

import scipy.stats as stats

from scipy.stats import norm

cars = pd.read\_csv(r'C:\Users\Bssuraj\OneDrive\Desktop\Assignments\Cars.csv')

cars

cars.describe()

#P(MPG>38)

p\_38 = 1 - norm.cdf(38,loc=34.422076,scale=9.131445)

p\_38

#P(MPG<40)

p\_40 = stats.norm.cdf(40,loc=34.422076,scale=9.131445)

p\_40

p\_20\_mpg\_50 = (stats.norm.cdf(0.50,loc=34.422076,scale=9.131445)) - (stats.norm.cdf(0.20,loc=34.422076,scale=9.131445))

p\_20\_mpg\_50

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Answer:**

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

cars=pd.read\_csv(r'C:\Users\Bssuraj\OneDrive\Desktop\Assignments\Cars.csv')

cars

cars.describe()

cars['MPG'].hist()

sns.distplot(cars['MPG'])

plt.grid(True)

plt.show()

cars['MPG'].skew()

cars['MPG'].kurt()

#From above plot and values we can say that data is fairly symmetrical, i.e fairly normally distributed.

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

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

wcat = pd.read\_csv(r'C:\Users\Bssuraj\OneDrive\Desktop\Assignments\wc-at.csv')

wcat

# plotting distribution for Waist Circumference (Waist)

sns.distplot(wcat.Waist)

plt.ylabel('density')

# plotting distribution for Adipose Tissue (AT)

sns.distplot(wcat.AT)

plt.ylabel('density')

# WC

wcat.Waist.mean() , wcat.Waist.median()

# AT

wcat.AT.mean() , wcat.AT.median()

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

**Answer:**

# Z-score of 90% confidence interval

stats.norm.ppf(0.95)

# Z-score of 94% confidence interval

stats.norm.ppf(0.97)

# Z-score of 60% confidence interval

stats.norm.ppf(0.8)

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

**Answer:**

import scipy.stats as stats

n = 25

df = n -1

df

#t\_scores of 95% confidence interval

CI = 0.95

alpha = 1 - CI

alpha

#alpha = 0.050

stats.t.ppf(1 - alpha/2, df)

#t\_scores of 96% confidence interval

CI = 0.96

alpha = 1 - CI

alpha

#alpha = 0.096

stats.t.ppf(1 - alpha/2, df)

#t\_scores of 99% confidence interval

CI = 0.99

alpha = 1 - CI

alpha

#alpha = 0.010

stats.t.ppf(1 - alpha/2, df)

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:**

import scipy.stats as stats

import math

#Given Data

x = 260 #sample\_mean

s = 90 #sample\_std\_dev

n = 18 #sample\_size

p = 270 #population\_mean

df = n - 1 #degree of freedom

#finding t\_stats

tstats = (x - p) / (s / math.sqrt(n))

tstats

#Calculate probability using cumulative distribution function (CDF)

pvalue = 1 - stats.t.cdf(abs(tstats), df)

pvalue

print(f"The t-score is: {tstats:.3f}")

print(f"The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is: {pvalue:.4f}")