|  |  |
| --- | --- |
| 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) | **RATIO** |
| Sales Figures | **RATIO** |
| Blood Group | **NOMINAL** |
| Time Of Day | **RATIO** |
| Time on a Clock with Hands | **RATIO** |
| Number of Children | **RATIO** |
| Religious Preference | **NOMINAL** |
| Barometer Pressure | **RATIO** |
| SAT Scores | **RATIO** |
| Years of Education | **RATIO** |

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

**Solution**: Total outcomes when 3 coins are tossed:

= (H,H,H) (H,H,T) (H,T,H) (T,H,H) (T,T,T) (T,T,H) (T,H,T) (H,T,T)

Probability of two heads and one tail = Favorable outcomes

Total no. of outcomes

= **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

**Solution**: (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)

Sum of outcomes:

(2,3,4,5,6,7)

(3,4,5,6,7,8)

(4,5,6,7,8,9)

(5,6,7,8,9,10)

(6,7,8,9,10,11)

(7,8,9,10,11,12)

1. Sum equal to 1 = **There is no possibility of sum equal to 1**
2. Less than or equal to 4 = Favorable outcomes

Total no. of outcomes

**= 6**

**36**

1. Sum is divisible by 2 and 3 = Favorable outcomes

Total no. of outcomes

= **6**

**36**

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?

**Solution**:

Total no. of balls = 2+3+2

= 7

No. of ways of drawing 2 balls out of 7 = 7C2

nCr = n!

r!(n-r)!

= 7! = 7\*6\*5! = 7\*6 = 21

2!(7-2)! 2!\*5! 2

Total no.of balls except blue balls = 7-2 = 5balls

Using nCr formula 5C2 = = 5! = 5\*4\*3! = 5\*4 = 10

2!(5-2)! 2!\*3! 2

Probability that none of balls drawn are blue out of 7 balls = Favorable outcomes

Total no. of outcomes

**Probability that none of balls drawn are blue**  =  **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

**Solution**:

Expected no. of candies for a randomly selected child = candies count \* probability

Child A = 1\*0.015

Child B = 4\*0.20

Child C = 3\*0.65

Child D = 5\*0.005

Child E = 6\*0.01

Child F = 2\*0.120

Add all the extracted values to get the expected no. of candies for a randomly selected child.

= 1\*.015 + 4\*.20 + 3\*.65 + 5\*.005 + 6\*.01 + 2\*.120

= **3.09**

**Therefore, the probability of expected candies for a randomly selected child is 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**

**Solution**:

import pandas as pd

df=pd.read\_csv("C:/PYTHON EXCELR/EXCELR ASSIGNMENTS/Assignments data files/Q7.csv")

df

df.describe() # by running this code we can get the following measures for Points,Score,Weigh.

**OUTPUT :**

**Points Score Weigh**

**count 32.000000 32.000000 32.000000**

**mean 3.596563 3.217250 17.848750**

**std 0.534679 0.978457 1.786943**

**min 2.760000 1.513000 14.500000**

**25% 3.080000 2.581250 16.892500**

**50% 3.695000 3.325000 17.710000**

**75% 3.920000 3.610000 18.900000**

**max 4.930000 5.424000 22.900000**

df["Points"].median() **= 3.695**

Range = df["Points"].max() - df["Points"].min()

**Range = 2.17**

df["Points"].var() **= 3.193166129032258**

**df["Points"].mode() = 0 3.07**

**1 3.92**

**# =======================================================**

**df["Score"].median() = 3.25**

Range = df["Score"].max() - df["Score"].min()

**Range = 3.911**

**df["Score"].var() = 0.957**

**df["Score"].mode() = 3.44**

**# ========================================================**

**df["Weigh"].median() =17.71**

Range = df["Weigh"].max() - df["Weigh"].min()

**Range = 8.3999**

**df["Weigh"].var() = 3.1931**

**df["Weigh"].mode() = 1--> 17.02 We got two values for mode both the values**

**2--> 18.90 got repeated for the same no. of times.**

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?

**SOLUTION :**

Expected weight of patient selected at random = sum of all weights

Total no. of patients

= 108+110+123+134+135+145+167+187+199

9

= 1308

9

= 145.33

INFERENCE : We assumed that weight of one of randomly chosen patient will be 145.33 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SOLUTION**:

import pandas as pd

df= pd.read\_csv("C:/PYTHON EXCELR/EXCELR ASSIGNMENTS/Assignments data files/Q9\_a.csv")

df

df['speed'].hist()

df['speed'].skew() = -0.11750986144663393.

df['speed'].kurtosis() = -0.5089944204057617

df['dist'].hist()

df['dist'].skew() = 0.8068949601674215

df['dist'].kurtosis() = 0.4050525816795765

**SP and Weight(WT)**

**Use Q9\_b.csv**

**SOLUTION**:

import pandas as pd

df= pd.read\_csv("C:/PYTHON EXCELR/EXCELR ASSIGNMENTS/Assignments data files/Q9\_b.csv")

df

df['SP'].hist()

**df["SP"].skew() = 1.6114501961773586**

**df["SP"].kurtosis() = 2.9773289437871835**

df['WT'].hist()

**df["WT"].skew() = -0.6147533255357768**

**df["WT"].kurtosis() = 0.9502914910300326**

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



**SOLUTION: 1) The histogram is positively skewed and it has outliers on the right hand side.**



**SOLUTION: The box plot has the outliers on the higher values 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?

**SOLUTION:**

from scipy import stats

# confidence interval for 94%

df\_ci = stats.norm.interval(0.94,

loc=200,

scale=30)

print ("I am 94% confident that population mean weight lies under:", df\_ci)

**(143.57619175546247, 256.42380824453755)**

# confidence interval for 96%

df\_ci = stats.norm.interval(0.96,

loc=200,

scale=30)

print ("I am 96% confident that population mean weight lies under:", df\_ci)

**(138.38753268104531, 261.61246731895466)**

# confidence interval for 98%

df\_ci = stats.norm.interval(0.98,

loc=200,

scale=30)

print ("I am 98% confident that population mean weight lies under:", df\_ci)

**(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.
2. What can we say about the student marks?

**SOLUTION:**

**1)**

import pandas as pd

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() = 41.0**

**df.median() = 40.5**

**df.var() = 25.529412**

**df.std() = 5.052664**

**2)Here we can say that on an average a student got 41 marks**

**And there was standard deviation of marks from the center.**

**On the large scale the variance was recorded as 25.5 and the median we got here is 40.**

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

**SOLUTION:**

**When the mean and median of data are equal then there is no skewness. Hence, we can say that data is normally distributed.**

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

**SOLUTION: When the mean > median. The nature of skewness will be positive.**

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

**SOLUTION: When the median > mean. The nature of skewness will be negative.**

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

**SOLUTION: Positive kurtosis value indicates that fewer data values are located near the mean and more data values are located on the tails i.e either on the right side or the left side.**

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

**SOLUTION: Negative kurtosis value indicates that more data values are located near the mean and data values located on the tails are less.**

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



What can we say about the distribution of the data?

**SOLUTION: The Q1 is at 10 and the Q3 is at 18, therefore 50% of the data lies between Q1&Q3. The other 50% of the data lies below the Q1 and above the Q3 in equal proportion.**

What is nature of skewness of the data?

**SOLUTION: Here the median of the box plot is closer to the upper side of the box plot and the upper whisker is shorter as compared to the lower whisker therefore we can say that the nature of skewness is negative.**

What will be the IQR of the data (approximately)?   
**SOLUTION: The IQR of the data is between 10 to 18.**

Q19) Comment on the below Boxplot visualizations?



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

**SOLUTION: In both the 1&2 box plots the data is symmetrical as we can see that the median is at the center of the box plots. Hence there is no skewness.**

* **In the box plot 1 the IQR lies between 255 to 280 approximately.**
* **In the box plot 2 the IQR lies between 225 to 315 approximately.**

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)

**SOLUTION:**

import pandas as pd

df= pd.read\_csv("C:/PRACTISE CODING EXCELR/EXCELR ASSIGNMENTS/Assignments data files/Assignment 1 data file/cars.csv")

df

len(df)

**a)**

df[df["MPG"]>38]

df[df["MPG"]>38].count()

(df[df['MPG']>38].count ()/len (df)).round(3)\*100

**P(MPG>38) = 40.7**

**b)**

df[df["MPG"]<40]

df[df["MPG"]<40].count()

(df[df['MPG']<40].count ()/len (df)).round(3)\*100

**P(MPG<40) = 75.3**

**c)**

df[(df['MPG']>20) & (df['MPG']<50)].count()

(df[(df['MPG']>20) & (df['MPG']<50)].count()/len (df)).round(3)\*100

**P (20<MPG<50) = 85.2**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**SOLUTION:**

import pandas as pd

df= pd.read\_csv("C:/PRACTISE CODING EXCELR/EXCELR ASSIGNMENTS/Assignments data files/Assignment 1 data file/cars.csv")

df

df["MPG"].hist()

**df["MPG"].skew() = -0.17794674747025727**

# Check whetherthe MPG of cars is following normal distribution or not

from scipy.stats import shapiro

calc,p = shapiro(df["MPG"])

calc

p

alpha =0.05

if (p<alpha):

print("Ho is rejected and H1 is accepted")

else:

print ("Ho is accepted and H1 is rejected")

# Ho is accepted : Data is normal

# H1 is accepted : Data is not normal

**Output: Ho is accepted and H1 is rejected**

**Therefore, with the above test we can say that the MPG of cars follows normal distribution.**

A graph with blue squares

Description automatically generated

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

Dataset: wc-at.csv

**SOLUTION:**

**import pandas as pd**

df= pd.read\_csv("C:/PRACTISE CODING EXCELR/EXCELR ASSIGNMENTS/Assignments data files/Assignment 1 data file/wc-at.csv")

df

df["AT"].hist()

**df["AT"].skew() = 0.584869324127853**

# Check whether the adipose tissue data follows normal distribution or not

from scipy.stats import shapiro

calc,p = shapiro(df["AT"])

calc

p

alpha = 0.05

if (p < alpha):

print ("Ho is rejected and H1 is accpeted")

else:

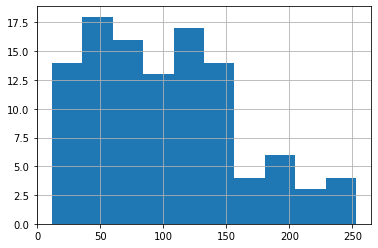
print ("Ho is accepted and H1 is rejected")

# Ho is accepted : Data is normal

# H1 is accepted : Data is not normal

**Output:** **Ho is rejected and H1 is accepted**

**Therefore, with the above test we can say that the adipose tissue does not follow the normal distribution.**



df["Waist"].hist()

**df["Waist"].skew() = 0.1340560824786468**

# Check whether the waist data follows normal distribution or not

from scipy.stats import shapiro

calc,p = shapiro(df["Waist"])

calc

p

alpha = 0.05

if (p < alpha):

print ("Ho is rejected and H1 is accpeted")

else:

print ("Ho is accepted and H1 is rejected")

# Ho is accepted : Data is normal

# H1 is accepted : Data is not normal

**Output : Ho is rejected and H1 is accepted**

**Therefore, with the above test we can say that the waist data does not follow the normal distribution.**

A graph with blue squares

Description automatically generated

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

interval, 60% confidence interval

**SOLUTION:**

from scipy.stats import norm

# Z-score for 90% confidence interval

confidence\_level = 0.90

z\_score = norm.ppf(1 - (1 - confidence\_level) / 2)

print(f"Z-score for a {int(confidence\_level \* 100)}% confidence interval: {z\_score:.4f}")

**Output : Z-score for a 90% confidence interval: 1.6449**

# Z-score for 94% confidence interval

confidence\_level = 0.94

z\_score = norm.ppf(1 - (1 - confidence\_level) / 2)

print(f"Z-score for a {int(confidence\_level \* 100)}% confidence interval: {z\_score:.4f}")

**Output : Z-score for a 94% confidence interval: 1.8808**

# Z-score for 60% confidence interval

confidence\_level = 0.60

z\_score = norm.ppf(1 - (1 - confidence\_level) / 2)

print(f"Z-score for a {int(confidence\_level \* 100)}% confidence interval: {z\_score:.4f}")

**Output : Z-score for a 60% confidence interval: 0.8416**

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

**SOLUTION:**

from scipy.stats import t

# t score for 95% confidence interval

confidence\_level = 0.95

sample\_size = 25

degrees\_of\_freedom = sample\_size - 1

# Calculate t-score for the given confidence level and degrees of freedom

t\_score = t.ppf(1 - (1 - confidence\_level) / 2, df=degrees\_of\_freedom)

print(f"t-score for a {int(confidence\_level \* 100)}% confidence interval with {sample\_size} sample size: {t\_score:.4f}")

**Output : t-score for a 95% confidence interval with 25 sample size: 2.0639**

# t score for 96% confidence interval

confidence\_level = 0.96

sample\_size = 25

degrees\_of\_freedom = sample\_size - 1

# Calculate t-score for the given confidence level and degrees of freedom

t\_score = t.ppf(1 - (1 - confidence\_level) / 2, df=degrees\_of\_freedom)

print(f"t-score for a {int(confidence\_level \* 100)}% confidence interval with {sample\_size} sample size: {t\_score:.4f}")

**Output : t-score for a 96% confidence interval with 25 sample size: 2.1715**

# t score for 99% confidence interval

confidence\_level = 0.99

sample\_size = 25

degrees\_of\_freedom = sample\_size – 1

# Calculate t-score for the given confidence level and degrees of freedom

t\_score = t.ppf(1 - (1 - confidence\_level) / 2, df=degrees\_of\_freedom)

print(f"t-score for a {int(confidence\_level \* 100)}% confidence interval with {sample\_size} sample size: {t\_score:.4f}")

**Output : t-score for a 99% confidence interval with 25 sample size: 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

**SOLUTION:**

import math

import scipy.stats as stats

x\_bar = 260

mu = 270

s = 90

n = 18

t\_score = (x\_bar - mu) / (s / math.sqrt(n))

df = n - 1

p\_value = stats.t.cdf(t\_score, df)

print("The probability that 18 randomly selected bulbs would have an average life of no more than 260 days:", p\_value)

**Output : The probability that 18 randomly selected bulbs would have an average life of no more than 260 days: 0.32167253567098364**