|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data type |
| Results of rolling a dice | Discrete data type |
| Weight of a person | Continious data type |
| Weight of Gold | Continious data yype |
| Distance between two places | Continious data type |
| Length of a leaf | Continious data type |
| Dog's weight | Continious data type |
| Blue Color | Discrete data type |
| Number of kids | Discrete data type |
| Number of tickets in Indian railways | Discrete data type |
| Number of times married | Discrete data type |
| Gender (Male or Female) | Discrete data type |

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 | Ordinal |
| Socioeconomic Status | Nominal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ordinal |
| Sales Figures | Nominal |
| 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 | Interval |
| Years of Education | Ratio |

===========================================================

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

**A3)** No coins tossed = 3

Total No of Possible Outcomes = 8

{[HHH],[HHT][HTH],[HTT],[TTT],[THH],[THT],[TTH]}

Now we need to Know how many ways we can get Two heads and one tail

{[HHT],[HTH],[THH]}

Probability = No of favarouble Outcome/Total No Of Possible Outcomes

Therefore Probability of getting two heads and one tail when three coins are

tossed Simultaniously:

P(x=2H&1T) = 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

**A4)** No Dice Rolled = 2

Total No Possible Outcomes = 36

{(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. Equal to 1:

The Minimum sum obtained when two dies are rolled simultaniously at a time is 2.So there is no possible way to get sum equals to 1 when two dice are rolled at a time.

Hence P(sum=1) = 0

1. Less than or equal to 4:

No favorable outcomes{(1,1),(1,2),(2,2),(1,3),(3,1)} = 5

Total No Possible outcomes = 36

Probability = No of favarouble Outcome/Total No Of Possible Outcome

Hence P(sum ≤ 4) = 5/36

Therefore the probability of getting sum Less than or equal to 4 is 5/36.

c) Sum is divisible by 2 and 3:

No of favarouble outcomes sum divisible by 2 and 3

{(1,5),(2,4),(3,3),(4,2),(6,6)} = 5

No of possible outcomes = 36

Probability = No of favarouble Outcome/Total No Of Possible Outcome

Hence P(sum divisible by 2 and 3) = 5/36

===========================================================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?

**A5)** Total No of Balls = 2+3+2 = 7

Here we use combinational formula:

nCr = n!/r!\*n-r!

Hence No of ways of drawing 2 balls out of 7 balls = 7C2 =7!/2!\*(7!-2!)

=7\*6/2\*1

=21

Therefore 7C2 = 21

No of ways of 2 balls out of 5 balls(non-blue balls) = 5C2 = 5!/2!\*(5!-2!)

= 5\*4/2\*1

= 10

Therefore 5C2 = 10

P(E) = 5C2/7C2

= 10/21

Therefore the Probability Picking that none of the blue balls is 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

A6) To calculate the Expected number of candies for randomly selected child we need to multiply each candy count by its probability and sum up the results

**Expected Number of candies** = {(Child A candies count \* Child A probability) **+** (Child B candies count \* Child B probability) **+** (Child C candies count \* Child C probability) **+** (Child D candies count \* Child D probability) **+** (Child E candies count \* Child E probability) **+** (Child F candies count \* Child F probability)

Expected Number of candies = {(1\*0.015)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120)}

Expected Number of candies = 0.015 + 0.80 + 0.195 + 0.025 + 0.06 + 0.24

Expected Number of candies = 3.125

Therefore the Expected Number of candies for a randomly selected child is 3.125

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**

**A7)** By using python **Pandas dataframe**,we can solve this problem.

Here is the complete code:

**import numpy as np**

**import pandas as pd**

**df=pd.read\_csv("D:/data science/assignments\_csv/Q7.csv")**

**df**

**####-----------------> Points**

**df['Points'].mean()**

**df['Points'].median()**

**df['Points'].mode()**

**df['Points'].var()**

**df['Points'].std()**

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

**Range**

**#####----------------> Score**

**df['Score'].mean()**

**df['Score'].median()**

**df['Score'].mode()**

**df['Score'].var()**

**df['Score'].std()**

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

**Range**

**####----------------> Weigh**

**df['Weigh'].mean()**

**df['Weigh'].median()**

**df['Weigh'].mode()**

**df['Weigh'].var()**

**df['Weigh'].std()**

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

**Range**

**POINTS:**

Mean: 3.596 Variance: 0.285

Median: 3.695 Standard Deviation: 0.534

Mode: 3.07,3.92 Range: 2.17

**Score:**

Mean: 3.217 Variance: 0.957

Median: 3.325 Standard Deviation: 0.978

Mode: 3.44 Range: 3.911

**Weigh:**

Mean: 17.848 Variance: 3.193

Median: 17.71 Standard Deviation: 1.786

Mode: 17.02,18.90 Range: 8.39

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?

**A8)** Expected Value is known as Average Value or Mean Value

Expected Values = sum of weights of patient/No of weights

Sum of weights of patient = 108+110+123+134, 135+145+167+187+199

No of weights = 9

Expected Values = 108+110+123+134, 135+145+167+187+199 / 9

Expected Values = 1308 / 9

Expected Values = 145.33

Therefore ,the Expected Value of the Weight of a randomly choosen patient at a clinic is 145.33 Pounds

##################################################################

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

**Cars speed and distance**

**Use Q9\_a.csv**

**A9\_a)** For calculating skewness and kurtosis we use the skew() and kurtosis()

Functions .Here is the code of implemntation using pandas dataframe

**import pandas as pd**

**df=pd.read\_csv('D:/data science/assignments\_csv/Q9\_a.csv')**

**df**

**#####------------->skewness for speed and dist**

**df['speed'].skew()**

**df['dist'].skew()**

**#####-------------->kortosis() for speed and dist**

**df['speed'].kurtosis()**

**df['dist'].kurtosis()**

Outputs:

Skew: kurtosis:

Speed: -0.11750986144663393 Speed:-0.5089944204057617

Dist: 0.8068949601674215 Dist: 0.4050525816795765

**Inferences:**

Skewness:

1. Speed data is having negative skewness and it also follow slightly symmetric

Which is near to zero

1. Distance data is positively skewed indicating a right skewed distribution

Kurtosis:

1.Speed data is having negative kurtosis,indicating that it has lighter tails and less peaked when compared to normal distribution

2. Speed data is having negative kurtosis,indicating that it has heavier tails and more peaked when compared to normal distribution

**SP and Weight(WT)**

**Use Q9\_b.csv**

**A9\_b)** For calculating skewness and kurtosis we use the skew() and kurtosis()

Functions .Here is the code of implemntation using pandas dataframe

**import pandas as pd**

**df=pd.read\_csv('D:/data science/assignments\_csv/Q9\_b.csv')**

**df**

**#####------------->skewness for SP and Weight**

**df['SP'].skew()**

**df['WT'].skew()**

**#####-------------->kortosis for SP and Weight**

**df['SP'].kurtosis()**

**df['WT'].kurtosis()**

**Outputs:**

**Skew: kurtosis:**

**SP: 1.6114501961773586 SP: 2.9773289437871835**

**WT: -0.6147533255357768 WT: 0.9502914910300326**

**Inferences:**

**Skewness:**

**1.** The 'SP' variable has positive skewness, indicating a right-skewed distribution. This means there are more data points on the lower end of the scale.

**2.** The 'WT' variable has negative skewness, indicating a left-skewed distribution. This means there are more data points on the higher end of the scale.

**Kurtosis:**

**1.** The 'SP' variable has a kurtosis value of 2.977, indicating leptokurtic distribution. It means the distribution has heavier tails and is more peaked compared to a normal distribution.

**2.** The 'WT' variable has a kurtosis value of 0.950, which is close to 0. This suggests a distribution that is relatively closer to a normal distribution in terms of peakedness and tail behavior

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



**A10)** : **The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side**.



**Ans: The above box plot has outliers on 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?

**A11)** For calculating Confidence interval we need import scipy stats module.Here is Code for this confidence intervals

**from scipy import stats**

**X\_mean=200**

**X\_std=30**

**######-------->94% confidence interval**

**df\_ci=stats.norm.interval(0.94,loc=X\_mean,scale=X\_std)**

**print('94% confident lies between',df\_ci)**

**######-------->98% confidence interval**

**df\_ci=stats.norm.interval(0.98,loc=X\_mean,scale=X\_std)**

**print('98% confident lies between',df\_ci)**

**######-------->96% confidence interval**

**df\_ci=stats.norm.interval(0.96,loc=X\_mean,scale=X\_std)**

**print('96% confident lies between',df\_ci)**

**Outputs:**

* 94% confident lies under (143.57619175546247,256.42380824453755)
* 98% confident lies between(130.2095637787748,269.7904362212252)
* 96% confident lies between (138.38753268104531 ,261.6124673189)

########################################################

**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?

**A12)**

**1.**By using numpy array we can convert the scores into data frame

Here is the python code:

**import numpy as np**

**scores=np.array([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])**

**scores**

**####------>Mean**

**mean\_scores=np.mean(scores)**

**mean\_scores**

**####------>Median**

**median\_scores=np.median(scores)**

**median\_scores**

**####------>Variance**

**Variance\_scores=np.var(scores)**

**Variance\_scores**

**####------>Standard deviation**

**std\_scores=np.std(scores)**

**std\_scores**

**Outputs:**

**Mean:41.0**

**Median:40.5**

**Variance:24.11**

**Standard deviation:4.91**

**2. Based on the calculated statistics ,what we can say about students marks is:**

* Mean score is 41.0,indicating that on average ,the student scored 41.0 Marks in the tests
* Median score is 40.5,indicating that half of the scores is above 40.5 and remaning half of the scores is below 40.5
* Variance score 24.11 suggests that students scores are relatively spread around mean
* Standard deviation 4.91 suggests that there smaller variation between each score of the student

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

**A13) when the skewness of the mean and median are eqaul then it is in symetric**

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

**A14) when the skewness of mean > median then it is in positive skewness**

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

**A15) when the skewness of median > mean then it is in negative skewness**

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

**A16)** **Positive kurtosis value** indicates that the data has **thicker tails** than a normal distribution. This means that there are more extreme values in the data, both positive and negative.

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

**A17) A negative kurtosis value indicates that the distribution of data has thinner tails than the normal distribution. This means that there are fewer outliers in the data**

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



What can we say about the distribution of the data?

**Ans: The above Boxplot is not normally distributed the median is towards the higher value**

What is nature of skewness of the data?

**Ans: The above Boxplot is not normally distributed the median is towards the higher value**

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

**Ans**: The Inter Quantile Range = Q3 Upper quartile – 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.

**A19)** 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)

**A20)** Here is the python code for the following data:

**import pandas as pd**

**import numpy as np**

**df=pd.read\_csv('D:/data science/assignments\_csv/Cars.csv')**

**df**

**#####---------->MPG>38**

**from scipy.stats import norm**

**k=1-norm.cdf(38,df.MPG.mean(),df.MPG.std())**

**k.round()**

**#####---------->MPG<40**

**from scipy.stats import norm**

**k=norm.cdf(40,df.MPG.mean(),df.MPG.std())**

**k.round()**

**#####---------->20<MPG<50**

**from scipy.stats import norm**

**k1=norm.cdf(50,df.MPG.mean(),df.MPG.std())**

**k1**

**k2=1-norm.cdf(20,df.MPG.mean(),df.MPG.std())**

**k2**

**k3=k1-k2**

**k3.round()**

**Outputs:**

**a.** P(MPG>38): The probability of getting (MPG>38) MPG greater than 38 is 0.347

**b.** p(MPG<40): The probability of getting (MPG<40) MPG less than 40 is 0.729

**c**.P(20<MPG<50): The probability of getting (20<MPG<50) MPG between 20 and 50 is 0.013

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**A21)**

**a).**MPG of cars follows normal distribution.Here is the Python code for it

**import pandas as pd**

**import numpy as np**

**df=pd.read\_csv('D:/data science/assignments\_csv/Cars.csv')**

**df**

**df['MPG'].mean()**

**df['MPG'].hist()**

**df['MPG'].skew()**

**### Normality Test**

**##--->Ho:Data is Normal**

**##--->H1:Data is not Normal**

**from scipy.stats import shapiro**

**calc,p = shapiro(df['MPG'])**

**calc**

**p**

**alpha=0.05**

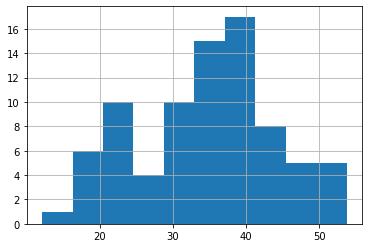
**if (p < alpha):**

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

**else:**

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

**graph:**



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 Normal distribution.here is the python code for it

**import pandas as pd**

**df=pd.read\_csv('D:/data science/assignments\_csv/wc-at.csv')**

**df**

**#############------------>Waist**

**df['Waist'].hist()**

**### Normality Test**

**##--->Ho:Data is Normal**

**##--->H1:Data is not Normal**

**from scipy.stats import shapiro**

**calc,p = shapiro(df['Waist'])**

**calc**

**p**

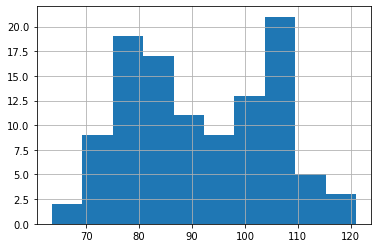
**alpha=0.05**

**if (p < alpha):**

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

**else:**

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



**#############------------>AT**

**df['AT'].hist()**

**### Normality Test**

**##--->Ho:Data is Normal**

**##--->H1:Data is not Normal**

**from scipy.stats import shapiro**

**calc,p = shapiro(df['AT'])**

**calc**

**p**

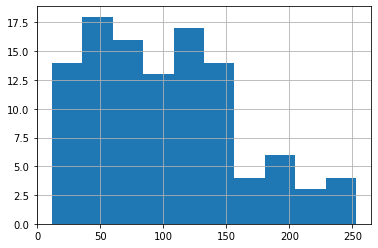
**alpha=0.05**

**if (p < alpha):**

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

**else:**

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



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

**A22)** Z scores confidence interval will be in the given below code

**from scipy import stats**

**from scipy.stats import norm**

**####------>90%**

**confidence\_interval = 0.90**

**z=(1-confidence\_interval)/2**

**stats.norm.ppf(z)**

**####------>94%**

**confidence\_interval = 0.94**

**z=(1-confidence\_interval)/2**

**stats.norm.ppf(z)**

**####------>60%**

**confidence\_interval = 0.60**

**z=(1-confidence\_interval)/2**

**stats.norm.ppf(z)**

**Outputs:**

* Z score of 90% confidence interval is -1.6448536269514729
* Z score of 94% confidence interval is -1.8807936081512509
* Z score of 60% confidence interval is -0.8416212335729142

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

**A23) from scipy import stats**

**from scipy.stats import norm**

**####------>95%**

**confidence\_interval = 0.95**

**n = 25**

**z=(1-confidence\_interval)/2**

**stats.t.ppf(z,n-1)**

**####------>96%**

**confidence\_interval = 0.96**

**n = 25**

**z=(1-confidence\_interval)/2**

**stats.t.ppf(z,n-1)**

**####------>99%**

**confidence\_interval = 0.99**

**n = 25**

**z=(1-confidence\_interval)/2**

**stats.t.ppf(z,n-1)**

**Outputs:**

* **t score for 95% confidence interval is -2.0638985616280205**
* **t score for 95% confidence interval is -2.1715446760080677**
* **t score for 95% confidence interval is -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

**A24)**

**from scipy import stats**

**from scipy.stats import norm**

**x = 260**

**pop\_mean = 270**

**std = 90**

**n = 18**

**df = (x-pop\_mean)/(std/np.sqrt(n))**

**df**

**stats.t.cdf(df,17)**

**Output:** The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.321