|  |  |
| --- | --- |
| 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) | Categorical |

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 | Nominal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio / Interval |
| SAT Scores | Interval |
| Years of Education | Interval |

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

S={‘HHH’,’HHT’,’HTH’,’THH’,’TTH’,’THT’,’HTT’,’TTT’}

S=9

Probability that two heads and one tail={’HHT’,’HTH’,’THH’}

Probability (Event) = Favorable outcomes / Total number of outcomes

=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

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)}

S=36

1. 0
2. S2 = {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)} = 6

= 6/36

= 0.166

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

= 6/36

=1/6 =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?

* S = {R1,R2,G1,G2,G3,B1,B2} = 7
* N (Event (2 balls are drawn randomly from bag) = 7! / 2! \* 5!

= (7\*6\*5\*4\*3\*2\*1) /

(2\*1) \* (5\*4\*3\*2\*1)

* N (Event (2 balls are drawn randomly from bag) = (7\*6)/ (2\*1) = 21
* If none of them drawn 2 balls are blue = 7 – 2 = 5
* N (Event (None of the balls drawn is blue) = 5! / 2! \* 3! = (5\*4) / (2\*1) = 10
* P (None of the balls drawn is blue) = N (Event (None of the balls drawn is blue) / N (Event (2 balls are drawn randomly from bag)
* **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

* 0.015+0.8+1.95+0.025+0.06+0.24 = 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**

* Mean for Points = 3.59, Score = 3.21 and Weigh = 17.84
* Median for Points = 3.69, Score = 3.32 and Weigh = 17.71
* Mode for Points = 3.07, Score = 3.44 and Weigh = 17.02
* Variance for Points = 0.28, Score = 0.95, Weigh = 3.19
* Standard Deviation for Points = 0.53, Score = 0.97, Weigh = 1.78
* Range [Min-Max] for Points [3.59 – 4.93], Score [3.21 – 5.42] and Weigh [17.84 – 22.9]

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?

* Expected value = Sum (X \* Probability of X)

= (1/9)(108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+ (1/9)(145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199)

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

* print('For Cars Speed', "Skewness value=", np.round(q9a.speed.skew(),2), 'and' , 'Kurtosis value=', np.round(q9a.dist.skew(),2))
* For Cars Speed **Skewness value= -0.12 and Kurtosis value= 0.81**
* print('Skewness value =', np.round(q9a.dist.skew(),2),'and', 'Kurtosis value =', np.round(q9a.dist.kurt(),2), 'for Cars Distance')
* **Skewness value = 0.81 and Kurtosis value = 0.41 for Cars Distance**

**SP and Weight(WT)**

**Use Q9\_b.csv**

* Print ('For SP Skewness =', np.round(q9b.SP.skew(),2), 'kurtosis =', np.round(q9b.WT.kurt(),2))
* For SP Skewness = 1.61 kurtosis = 0.95
* Print ('For WT Skewness =', np.round(q9b.SP.skew(),2), 'Kurtosis =', np.round(q9b.WT.kurt(),2))
* For WT Skewness = 1.61 Kurtosis = 0.95

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



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



* 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?

* conf\_94 =stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000))
* print(np.round(conf\_94,0))
* print(conf\_94)
* 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.

* Mean =41, Median =40.5, Variance =25.52, Standard Deviation =5.05

1. What can we say about the student marks?

* 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?

* No skewness is present we have a perfect symmetrical distribution.

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

* Skewness and tail is towards Right.

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

* Skewness and tail is towards Left.

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

* Positive kurtosis means the curve is more peaked and it is Leptokurtic.

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

* 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?

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

What is nature of skewness of the data?

* 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)?

* 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.

* 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 normallydistributed 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.

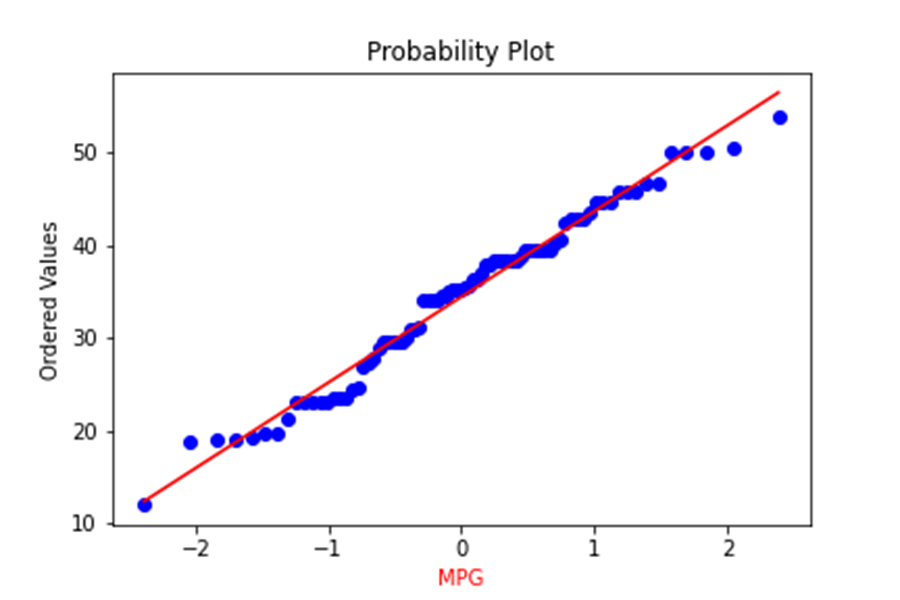
MPG <- Cars$MPG

* 1. P(MPG>38)
* Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3)
* print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)
* **P(MPG>38)= 0.348**
  1. P(MPG<40)
* prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)
* print('P(MPG<40)=',prob\_MPG\_less\_than\_40)
* **P(MPG<40)= 0.729**
  1. P (20<MPG<50)
* prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)
* print('p(MPG>20) =',(prob\_MPG\_greater\_than\_20))
* **p(MPG>20) = 0.943**
* prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)
* print('P(MPG<50)=',(prob\_MPG\_less\_than\_50))
* **P(MPG<50)= 0.956**
* prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20)
* print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))
* P(20<MPG<50)= 0.01300

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



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

Dataset: wc-at.csv



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

* # z value for 90% confidence interval
* print('Z score for 90% =',np.round(stats.norm.ppf(.05),4))
* Z score for 90% = **-1.6449**
* # z value for 94% confidence interval
* print('Z score for 94% =',np.round(stats.norm.ppf(.03),4))
* Z score for 94% **= -1.8808**
* # z value for 60% confidence interval
* print('Z score for 60 =',np.round(stats.norm.ppf(.2),4))
* Z score for 60% = **-0.8416**

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

* # t score for 95% confidence interval
* print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))
* T score for **95% Confidence Interval = -2.0639**
* # t value for 94% confidence interval
* print('T score for 94% Confidence Inteval =',np.round(stats.t.ppf(0.03,df=24),4))
* T score for **94% Confidence Inteval = -1.974**
* # t value for 99% Confidence Interval
* print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))
* T score for **95% Confidence Interval = -2.7969 import numpy as np**
* **import scipy as stats**
* **t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)**
* **= (260-270)/90/np.sqrt(18))**
* **t\_score = -0.471**
* **stats.t.cdf(t\_score, df = 17)**
* **0.32 = 32%**

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

* import numpy as np

import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

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

t**\_score = -0.471**

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

0.32 = 32%