|  |  |
| --- | --- |
| 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 | Categorical |
| 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 | Interval |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

Ans: When Three coins are tossed the probability that we having two heads and one tail are.

HHH, HTH,HHT,THH,TTT,THT,TTH,HTT

There are 3 Outcomes

3/8 =0.374

It is binominal distribution.

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1

Ans If Two dies are rolled there is 0 probablity that sum will be equal to 1, because the highest number we get by rolling two dies is 2.

1. Less than or equal to 4

Ans: The probability of getting less than or equal to 4 is 6.

Because the possible outcome of rolling two dies are 6\*6=36

(1,1),(1,2),(2,1),(2,2),(3,1),(1,3)

1. Sum is divisible by 2 and 3

Ans The probability of rolling dies and getting outcome are

(1,1),(1,2),(2,1),(2,2)(2,4),(3,1),(1,3),(3,3),(4,2),(4,4)(6,2),(6,4)

Therefore the probability of getting the number divisible by 2 and 3 are 13.

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 number of balls =7 balls

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

Ans number of probability and childrens having candies .

Child A has probability of 0.015 of having candies

Child B has probability of 0.20 of having candies

Child C has probability of 0.65 of having candies

Child D has probability of 0.005 of having candies

Child E has probability of 0.01 of having candies

Child F has probability of 0.120 of having candies

Now let’s calculate the Expected number of candies for a randomly selected child

Expected no. of candies = (1\*0.015)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120).

Expected no of candies =0.015+0.80+0.1.95+0.02+0.06+0.24=3.085

3.085 is Expected number of candies for a randomly selected child

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.

Ans – mean for points : 3.5965

Score : 3.21

Weigh : 17.84

Median of : points 3.695

Score 3.325

Weigh 17.710

Mode of : points 3.07

Score 3.44

Weigh 18.90

Variance of cars for : points

0.285881 score

0.9573 weigh

3.193166

Standard Deviation cars for : points

0.534679 score

0.9784

Weigh 1.786943

**Use Q7.csv file**

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 Given weights of patients : (108+110+123+134+135+145+167+187+199)/9

1308/9 =145.333

Therefore the correct expected value of random choosen patients is 145.33 pounds.







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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans - vk=read.csv('C:\\Users\\VISHNU\\Downloads\\Q9\_a.csv')**

> #skewness and kurtosis for speed

> skewness(vk$speed)

[1] -0.1139548

> kurtosis(vk$speed)

[1] 2.422853

>

> #skewness and kurtosis for dist

> skewness(vk$dist)

[1] 0.7824835

> kurtosis(vk$dist)

[1] 3.248019

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans -** Ans :-

> vk=read.csv('C:\\Users\\YASH COMPUTER\\Downloads\\Q9\_b.csv')

> #skewness and kurtosis for SP

> skewness(vk$SP)

[1] 1.581454

> kurtosis(vk$SP)

[1] 5.723521

>

> #skewness and kurtosis for WT

> skewness(vk$WT)

[1] -0.6033099

> kurtosis(vk$WT)

[1] 3.819466

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



Ans: Histogram: - The most of the datapoints are concentrated 50 to 100 with frequency 200.

And least range of weight is 400 is between 0 to 10. So the expected value the above distribution is 75.



Ans: Boxplot median is less than mean rightskewed and outlier on the upper side of boxplot and there is less datapoints between Q1 and bottom.

**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 - 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.

Ans: mean= 41, median=40.5, variance=24.111, standard deviation=4.910

1. What can we say about the student marks?

Ans: The student has a range of scores from 34 to 56, indic\+ates variations in the score across the test. The lowest score of student is 34 and highest is 56.

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

Ans: skewness is measure how asymmetric a distribution of data

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

Ans: Right skewness

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

Ans: Leftskewness

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

Ans: Data is normally distributed an kurtosis value is 0.

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

Ans: The distribution of data has lighter tail and flatter peak than normal distribution.

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 data is skewed towards the left. The whisker range of minimum value is greater than the maximum.

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

Ans: The IQR= Q3 upper quartile-Q1Lower quartile

Q19) Comment on the below Boxplot visualizations?



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

Ans: 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)

Ans - 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)

Ans - 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)

Ans - 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)= 0.013000000000000012

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

1. Ans - MPG of cars follows 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 does not follow Normal Distribution





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

Ans:

# z value for 90% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.05),4))

Z score for 60% Conifidence Intervla = -1.6449

# z value for 94% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.03),4))

Z score for 60% Conifidence Intervla = -1.8808

# z value for 60% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.2),4))

Z score for 60% Conifidence Intervla = -0.8416

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

Ans -# 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

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 - 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%