|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete(quantitative) |
| 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 | nominal |
| 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 | Discrete, Nominal |
| High School Class Ranking | Discrete, Ordinal |
| Celsius Temperature | Continuous, Ratio |
| Weight | Continuous, Ratio |
| Hair Color | Discrete, Nominal |
| Socioeconomic Status | Continuous, Interval |
| Fahrenheit Temperature | Continuous, Interval |
| Height | Continuous, Ratio |
| Type of living accommodation | Discrete, Ordinal |
| Level of Agreement | Discrete, Ordinal |
| IQ(Intelligence Scale) | Discrete, Ratio |
| Sales Figures | Discrete, Ratio |
| Blood Group | Discrete, Nominal |
| Time Of Day | Continuous, Ratio |
| Time on a Clock with Hands | Continuous, Ordinal |
| Number of Children | Discrete, Ordinal |
| Religious Preference | Discrete, Nominal |
| Barometer Pressure | Continuous, Interval |
| SAT Scores | Continuous, Ratio |
| Years of Education | Continuous, Interval |

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

Solution :-

Sample space= 2^3=8

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

P(H H T) + P(H T H )+P(T H H)

1/8+1/8+1/8

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

Solution :-

S1= (1,2,3,4,5,6)

S2= (1,2,3,4,5,6)

S=S1\*S2

a). P (getting sum as 1) =0

b). P (less than or equal to 4) = (1,1), (1,2), (1,3), (2,1), (2,2), (3,1) =6

P (less than or equal to 4) =6/36=0.166

c). P(Sum is divisible by 2 & 3)=(1,5),(2,4),(3,3),(5,1),(6,6) =5

P(Sum is divisible by 2 & 3)=5/36=0.138

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. s of ball=2+3+2=7

S=sample space

n (S)=drawing 2 balls out of 7 balls

=7C2

(7\*6)/ (2\*1) =42/2=21

n(S)=drawing 2 balls out of 5 balls

5C2= (5\*4)/ (2\*1) = 20/2 =10

Required probability=5C2/7C2

=10/21

=0.47

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 value= (1\*0.015) +(4\*0.20) +(3\*0.65) +(5\*0.005) +(6\*0.01) +(2\*0.120)

= (0.015+0.8+1.95+0.025+0.06+0.240)

=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 :- referred to excel file name ‘Q7 ans’ ,attached

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 Values= sum (Probability\*Value)

S= (total nos. of Patients) =9

Probability of selecting each Patient=1/9

Values= 108, 110, 123, 134, 135, 145, 167, 187, 199

Expected Value= (1/9\*108) +(1/9\*110) +(1/9\*123) +(1/9\*134)

+(1/9\*135) +(1/9\*145) +(1/9\*167) +(1/9\*187) +(1/9\*199)

=1/9(108+110+123+134+135+145+167+187+ 199)

=0.111(1308)

=145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Solution :-

**Skewness : Kurtosis:**

**speed = -0.117510 speed= -0.508994**

**dist= 0.806895 dist= 0.405053**

**Inference:**

* **Speed is negatively Skewed, and negative Kurtosis (Platykurtic)**
* **dist is positively skewed, with positive kurtosis (leptokurtic)**

**refer to ‘Q9\_a solution’ in excel file, attached**

**SP and Weight (WT)**

**Use Q9\_b.csv**

Solution :-

**Skewness : Kurtosis:**

**SP= 1.611450 SP= 2.977329**

**WT= -0.614753 WT= 0.950291**

**Inference:**

* **SP is positively Skewed, and Positive Kurtosis (Platykurtic)**
* **WT is negatively skewed, with positive kurtosis (leptokurtic)**

**Refer to ‘Q9\_b solution’ in excel file, attached**

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



Solution :-

* Right skewed data
* Positive Kurtosis
* Frequency lies between 0 to 200, range [1 -200]
* Mean=90+200+125+100+50+25+12.5+5 =607.5/8=75.93
* Median=5,12.5,25,50,90,100,125,200

N=8, median= (n+1)/2= (8+1)/2=4.5~ 5th value in row (90).



Solution :-

* Presence of outliers
* Positively skewed
* More numbers of data into the whiskers part
* Median is likely to fall near left of the data

**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 :-

Sample size= 2000

Avg wt (X̅) =200

Standard deviation=30

Confidence Interval for mean: X̅E

E=Confidence Interval

Confidence interval=

Find margin of error=z(alpha/2)\*s/

E= 30/= 30/44.72 =0.67

Find critical value:

Alpha(a) for 94% confidence interval=1-(confidence level/100)

a=1-(94/100)= 1-0.94 =0.06

a/2 (each tail value)=0.06/2=0.03

check the corresponding z-score -1.88 for left tail and due to symmetry 1.88 for right tail

margin of error=1.88\*0.67

MOE=1.2596

Lower limit=mean-MOE

200-1.2595=198.7405

Upper Limit=mean +MOE

=200+1.2595= 201.2595

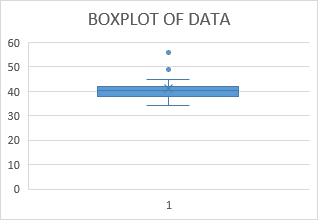
**Q12)** Below are the scores obtained by a student in tests

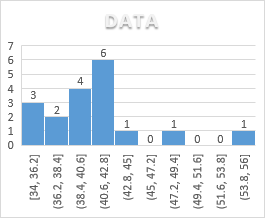
**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

Solution :-

|  |  |
| --- | --- |
| MEAN | 41.00 |
| MEDIAN | 40.50 |
| VARIANCE | 25.529 |
| STANDARD DEVIATION | 5.025 |





* This data got outliers
* Most of the student have scored around 41
* This data is negatively skewed
* Range of data is 22 (max-min)

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

Solution :-

Symmetrical Distribution

Zero skewness

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

Solution :-

Positive skewness

Most data is distributed towards the right of the median

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

Solution :-

Negative Skewness

Most data toward left side of the median

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

Solution :-

Kurtosis measures the ‘tailedness’ of the probability of the data relative to normal distribution of the data

Positive kurtosis: Distribution of the data is peaked and posses thick tail.

Where more numbers of data are located on the tail side than the normal distribution.

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

Solution :-

Negative Kurtosis: Distribution of the data on the tail side are lightly distributed than the normal distribution.

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



What can we say about the distribution of the data?

Solution :-

* Median of the data is between 14 & 16
* Negatively skewed data
* Q1= 10 (first Quartile) , values under first quartile holds 25% of the data
* Q3=18 (Third Quartile), value under third quartile holds upto 75% of the data

What is nature of skewness of the data?

Solution :-

* Negatively Skewed

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

Solution :-

IQR=q3-q1=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.

Solution :-

* Data are normally distributed in both the boxplot
* Median of the both boxplot is same as 265 approx.
* Boxplot 2 have more no.s of data and range than the boxplot 1
* Boxplot1: q1=260, q3=275, median=265.5
* Boxplot2: q1=225, q3=312.5, median=265.5

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 :-

* Done in python notebook

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Solution :-

* Excel file named “Cars solution”, attached.

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

Dataset: wc-at.csv

* Done in python notebook

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

Solution :-

* Done in python notebook

For 90% confidence interval:

We have the significance level at 5 % ( as it is a two tailed test), 10/2=5%

that is:

α = 5 % = 0.05

z at α = 0.05 from the z table will be:

z = 1.645.

For 94 % confidence interval, we get:

We have the significance level at 3 % ( as it is a two tailed test), 6/2=3%

that is:

α = 3 % = 0.03

z at α = 0.03 from the z table will be:

z = 1.555.

For 60 % confidence interval, we get:

We have the significance level at 20 % ( as it is a two tailed test), 40/2=20%

that is:

α =20 % = 0.2

z at α = 0.2 from the z table will be:

z = 0.253

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

Solution :-

#for 95% of the confidence interval of sample size=25

100-95=5%

5%/2=2.5%

1-2.5%

1-0.025=0.975

df=n-1; 25-1=24

code: stats.t.ppf(0.975,24)

from t-table t\_0.975 and df at24, the value is 2.0638

#for 96% of the confidence interval of sample 25

Code: stats.t.ppf(0.98,24)

100-96=4

4/2=2%

1-0.02=0.98

Check from the t-table, t\_0.98 and df=24, value is 2.1715

#for 99% of the confidence interval

100-99=1%

For two tail test, 1%/2=0.5% =0.05

1-0.05=0.995

Code: stats.t.ppf(0.095,24)

Check from t-table, t\_0.995 with df=24, value is 2.796

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 :-

Population mean, µ = 270

Sample size, n = 18

Sample mean, x̅ = 260

Standard deviation, s =90

t score = (x̅- µ)/(std/sqrt(n))

= -0.4714045207910317

df=n-1,18-1=17

Probability = 0.32167253567098364