|  |  |
| --- | --- |
| 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 | Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Nominal |

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 | nominal |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | ordinal |
| Sales Figures | Ratio |
| Blood Group | nominal |
| Time Of Day | interval |
| Time on a Clock with Hands | interval |
| Number of Children | ratio |
| Religious Preference | nominal |
| Barometer Pressure | interval |
| SAT Scores | interval |
| Years of Education | interval |

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

Ans – possibilities = HHH, TTT, HHT, HTH ,THH, THT, TTH , HTT = 8

No.of events happen = HHT, HTH, THH = 3

P (Two heads and one tail) = N (Event (Two heads and one tail)) / N (Event (Three

coins tossed))

= 3/8 = 0.375 = 37.5%

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 2and 3

ANSWERS :

Two dice rolled) =6^2 = 36

1. P (sum is Equal to 1) = ‘0’ no combination with sum is equal to 1
2. P (Sum is less than or equal to 4) =

N (Event (Sum is less than or equal to 4)) / N (Event (Two dice rolled))

= 6 / 36 = 1/6 = 0.166

1. P (Sum is divisible by 2 and 3) = N (Event (Sum is divisible by 2 and 3)) / N

(Event (Two dice rolled))

Possible combinations are – (1,5) (2,4) (3,3) (4,2) (5,1) (6,6) = 6 combinations

= 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?

Ans: Total number of balls =7 balls

N (Event (2 balls are drawn randomly) = 7! / 2! \* 5!

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

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

N (Event (2 balls are drawn randomly) = (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)

Ans = 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

Expected number of candies for a randomly selected child=

0.015+0.2+0.65+0.005+0.01+0.20 = 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**

**Ans :**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **point** | **score** | **weigh** |
| **MEAN** | **3.6** | **3.22** | **17.85** |
| **MODE** | **3.92** | **3.44** | **17.02** |
| **MEDIAN** | **3.7** | **3.33** | **17.71** |
| **SD** | **0.53** | **0.98** | **1.79** |
| **variance** | **0.29** | **0.96** | **3.19** |
| **max** | **4.93** | **5.42** | **22.9** |
| **min** | **2.76** | **1.51** | **14.5** |
| **range** | **2.17** | **3.91** | **8.4** |

**Range( point) = from 4.93 to 2.76**

**Range(score) = from 5.42 to 1.51**

**Range( Weigh) = from 22.9 to 14 .5**

**Here in example mean median and mode values are not differs from each others it means they lie on same point .hence we conclude that data is normally distributed and it follows bell shaped curve**

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 – 108+110+123+134+135+145+167+187+199 / 9 =1308/9

= 145.33 in pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

**q9a.skew()**

**Index 0.000000**

**speed -0.117510**

**dist 0.806895**

**q9a.kurt()**

**Index -1.200000**

**speed -0.508994**

**dist 0.405053**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**q9\_b.skew()**

**Unnamed: 0 0.000000**

**SP 1.611450**

**WT -0.614753**

**dtype: float64**

**q9\_b.kurt()**

**Unnamed: 0 -1.200000**

**SP 2.977329**

**WT 0.950291**

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



Ans: data distribution is rightly skewed and tail is on right.

Mean is greater than Median.

We have heavy chances of getting outliars .

Data is not normally distributed



Ans:

The boxplot has outliers on the maximum side.

Data is not normally distributed , rightly skewed .

Spread of the data is maximum .

**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? 200+-z( 0.67)

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

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 =24.11 and Standard Deviation =4.91

1. What can we say about the student marks?

Ans: here we don’t have any outlialrs and data is slightly right skewed as mean is greater than median.

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

Ans -If mean median of data are equal then there is no skewness , data is normally distributed. Data is symmetric and with zero skewness.

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

Ans - Positive skewness and tailed towards right

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

Ans – negative skewness tailed towards left

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

It indicates more data is present at the tail in positive side

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

Ans: 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 the data?

Ans : data is not normally distributed most of data points are on left side.

What is nature of skewness of the data?

Ans: Data is not normaly distributed and dta is left skewed

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

Ans: 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.

Ans: both the boxplot having same distribution both are symmetric means data for both are normally distributed. Both boxplot are sharing same median 225. Both have almost no or zero skewness.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

Ans:

P(MPG>38) = 0.348

Ans:

P(MPG<40) = 0.729

P (20<MPG<50)

Ans: P(20<MPG<50)= 1.24\*e^ -05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans : By distribution plot we find that data is slightly left skewed .

But approximately we can say it is normally distributed as max value lie between 2 standard deviation.

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

Dataset: wc-at.csv

Ans : – for AT data is not normally distributed(positively skewed)

For Waist data is normally distributed

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

**Solution:**

For 90% confidence interval:

We have the significance level at 5 % ( as it is a two tailed test)

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)

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)

that is:

α =20 % = 0.2

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

z = 0.253

**Therefore, we get that the z score at 90 % confidence interval is 1.645, at 94 % confidence interval is 1.555 and at 60 % confidence interval is 0.253.**

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

Ans: from scipy import stats

from scipy.stats import norm

# t scores of 95% confidence interval for sample size of 25 , we should add 2.5 % above 95%

t\_score\_95=stats.t.ppf(0.975,24) # df = n-1 = 24

t\_score\_95 = 2.0638

t\_score\_96=stats.t.ppf(0.985,24) # df = n-1 = 24

t\_score\_96= 2.30691

t\_score\_99=stats.t.ppf(0.995,24) # df = n-1 = 24

t\_score\_99= 2.796939

**Therefore, we get that the t score at 95 % confidence interval is 2.0638, at 96% confidence interval is 2.30691 and at 99 % confidence interval is 2.796939.**

Q 24**)**A Government companyclaims 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

p\_value=1-stats.t.cdf(abs(-0.4714),df=17)

p\_value = 0.32 = 32%

the probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32 %