|  |  |
| --- | --- |
| 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) | Discrete |

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

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

**Ans: -** When three coins are tossed the total number of possible combinations are 2³= 8. These combinations are S = (HHH, HHT, HTH, THH, TTH, THT, HTT, TTT). The number of combinations which have two heads and one tail are: (HHT, HTH, TTH) which makes them 3 in number.

Therefore, the Probability of getting two heads and one tails in the toss of three coins simultaneously is defined as P (Two heads and One Tail) = Number of desired outcomes is 3/8 or 0.375.

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

* Equal to 1
* Less than or equal to 4
* Sum is divisible by 2 and 3

**Ans: -** (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)

Total outcome=6×6=36.

**a)** If two dice are rolled then n(s) = 36, If two dice are rolled then there nothing that we get sum equal 1.

So, n(A)=0

⸫ Probability of get sum of 1= 0/36 = 0.

**b)** Less than or equal to 4

(1,3),(2,2),(3,1),(1,1),(1,2),(2,1) = 6 Outcomes,

6/36 i.e. 1/6

**c)** sum is divisible by 2 and 3. ⸫ 6/36 = 1/6

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: -** A bag contains 2 red, 3 green and 2 blue balls, two balls are drawn at random.

Probability = Favorable outcome/ total outcome

None of the balls drawn is blue, this can only happen when the two balls drawn at random are either red and green or both.

Total Number of balls = 2+3+2 = 7

Number of ways of drawing 2 balls out of 7= 7C2 = (7 x 6) / (2 x 1) = 42/2 = 21

Number of balls other than blue = 5

Number of ways of drawing 2 balls out of 5 = 5C2 = (5 x 4) / (2 x 1)

=20/2 = 10

⸫ Required Probability = 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: -** Expected number of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

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

= 3.090 = 3.09

* Expected number of candies for a randomly selected child = 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: -**  Points :- Mean = 3.596, Median = 3.659, Mode = 3.92,

Variance =0.285, Standard deviation =0.534.

Score :- Mean = 3.217, Median = 3.325, Mode = 3.44,

Variance =0.957, Standard deviation =0.978.

Weigh :- Mean = 17.848, Median = 17.710, Mode = 18.90,

Variance =3.197, Standard deviation =1.786.

Q8) Calculate Expected Value for the problem below

* 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: -**

Expected Value = ∑ (probability \* Value)

∑ P(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

Ex  108, 110, 123, 134, 135, 145, 167, 187, 199

P(x)  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

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)

= (1/9) (1308)

= 145.33

⸫ Expected Value of the Weight of that patient = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:-** Q9\_a.skew(speed)

-0.117509

Q9\_a.kurtosis(speed)

-0.5089944

Q9\_a.skew(dist)

0.806894

Q9\_a.kurtosis(dist)

0.405052

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans:-** Q9\_b.skew(SP)

1.611450

Q9\_b.skew(WT)

-0.614753

Q9\_b.kurtosis(SP)

2.977329

Q9\_b.kurtosis(WT)

0.950291

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





**Ans:-**  The most of the data points are Concerted in the range 50-100 with frequency 200. And least range of weight is 400 Somewhere around 0-10. Around 0-10.

So the expected value the above distribution is 75.

Skewness -we can notice a long tail towards right so it is heavily right skewed.

Median is less than mean right skewed and we have outlier on the upper side of the box plot and there is less data points between Q1 and bottom point.

**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: -**

Sample mean = 20

Standard Deviation of sample = 30

Confidence interval for confidence level = 94%,98%,96%

Standard Error SE= S / √n = 30 / √1000 = 30 / 31.62 = 0.95

α = 1 – 95 / 100 = 0.05

df = n – 1 = 2000 – 1 = 1999

The 94% confidence interval is (198.73,201.27)

The 96% confidence interval is (198.61,201.39)

The 98 % confidence interval is (198.43,201.57)

The interval is

x+t s ^n= 200 – 1.8916 30^2000 =198.73

x+ t s ^n= 200+ 1.816 30 ^2000=201.27

The 94% confidence interval is (198.73,201.27)

x-ts ^n=200 – 2.0673 30 ^2000=198.61

x+ t s^n=200 + 2.0673 30-^2000201.39

The 96% confidence interval Is (198.61,201,39).

x-ts^n =200 -2.3452 30 =198.43 30^2000

x+ ts^n 200 + 2.3452 30 = 201.57^2000

The 98% confidence interval is (198.43,201.57)

**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**

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

**Ans: -**

* import numpy as np

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

print('Mean:', a.mean())

print('Median:', np.median(a))

print('Variance:', a.var())

print('std deviation:',a.std())

Mean: 41.0

Median: 40.5

Variance: 24.11111111111111

std deviation: 4.910306620885412.

**Ans:-** 2)Mass of students marks between 38-42.

Skewness (1.52) is positive because mass of marks in left side of plot.

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

**Ans: -** Data is normalized and there is no skewness. Symmetrical

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

**Ans: -** Negative Skewness implies mass of the Distribution concentrated on right side. Right Skewed.

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

**Ans: -** Positive Skewness implies mass of the Distribution concentrated on left side. Left Skewed.

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

**Ans: -** Positive kurtosis value indicates that thinner peak and wider tails. The data is normally distributed and kurtosis value is 0.

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

**Ans: -** Negative kurtosis value indicates that wider peak and thinner tails. The distribution of the data has lighter tails and a flatter peaks than the normal distribution.

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



What can we say about the distribution of the data?

**Ans:-** Not normally distributed

What is nature of skewness of the data?

**Ans:-**Negative skewness

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

**Ans: -** 18-10=7  
Q19) Comment on the below Boxplot visualizations?



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

**Ans:-**  By observing both the plots whisker’s level is high in boxplot 2.Mean and median are equal hence distribution is symmetrical.

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

* P(MPG>38)
* P(MPG<40)
* P (20<MPG<50)

**Ans: -**

* # P(MPG>38)

1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

0.3475939251582705

b. # P(MPG<40)

stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

0.7293498762151616

c. # P (20<MPG<50)

stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())- stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())

1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:-** cars.MPG.mean(),cars.MPG.mode(),cars.MPG.median()

**(**34.422075728024666, 29.629936, 35.15272697)

b) Check Whether the Adipose Tissue (AT) and Waist

Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

**Ans:-** Waist Circumference (Waist) = (91.90183486238533, 90.8)

Adipose Tissue= (101.89403669724771, 96.54)

Adipose Tissue (AT) and waist cirumference (Waist) both are normal distributed.

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

**Ans:-** from scipy import stats

from scipy.stats import norm

# Z-score of 90% confidence interval

stats.norm.ppf(0.95)

1.6448536269514722

# Z-score of 94% confidence interval

stats.norm.ppf(0.97)

1.8807936081512509

# Z-score of 60% confidence interval

stats.norm.ppf(0.8)

0.8416212335729143

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

**Ans:-** t scores of 95% confidence interval for sample size of 25

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

2.0638985616280205

t scores of 96% confidence interval for sample size of 25

stats.t.ppf(0.98,24)

2.1715446760080677

# t scores of 99% confidence interval for sample size of 25

stats.t.ppf(0.995,24)

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

**Ans:-**

t - statistics for the data is given as follows:

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

t = **-0.471**

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that t < - 0.471 with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.