|  |  |
| --- | --- |
| 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 | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| 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 | Ratio |

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

If Three Coins are tossed the Sample Space is (S).

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

n(S) = 8

K = {HHT, THH, HTH}

n(K) = 3

P(K) = n(K)/n(S)

P(K) = 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

Ans: Total Number of outcome = 6\*6 = 36

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)

1. Equal to 1

Probability of getting sum equal to less than 1

0/36. Probability is 0

1. Less than or Equal to 4

(1,3), (2,2), (3,1) (1,1) (2,1) (1,2) = 6 outcomes, 6/36 i.e. 1/6

1. Sum is Divisible By 2 And 3

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

=6 / 36 i.e. 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:

Total no. of balls = 2 red + 3 green + 2 blue = 7

Let S be the sample space

Then, n(S) = Number of ways of drawing 2 balls out of 7

N(S) = nCr = 7C2

= (2\*1)(7\*6)

= 21

Let R = Event of Drawing 2 balls, none of which is blue.

n(R) = Number of ways of Drawing 2 balls out of (2 + 3)balls =

5C2

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

P(R) = n(R) / n(S)

=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 number of candies for a randomly selected child = 3.09

Describe:

Excepted 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

=0.0015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.090

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**

**Point:** Mean = 3.596563, Median = 3.659, Mode = numeric

Variance = 0.957379, Standard deviation = 0.5346787.

Score: Mean = 3.217225, Median = 3.325, Mode = “numeric”,

Variance = 0.957379, Standard deviation = 0.9784574

Weight: Mean = 17.84875, Median = 17.71, Mode = “numeric”,

Variance = 3.193166, Standard deviation = 1.786943

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?

**To Find:** Expected Value

**Solution:**

Expected Value = Σ ( probability \* value)

ΣP(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 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)

Mean = (1308)

163.5

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

**Cars speed and distance**

**Use Q9\_a.csv**

Data[‘speed’].skew()

-0.117509

Data[‘speed’].kurtosis()

-0.5089944

Data[‘dist’].skew()

0.806894

Data[‘dist. ’].kurtosis()

0.405052

**SP and Weight(WT)**

**Use Q9\_b.csv**

Data[‘ SP ’].skew()=1.61154

Data[‘ SP ’].kurtosis()=2.977328

Data[‘ WT ’].skew()= -0.61

Data[‘ WT ’].kurtosis()=0.9502914

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



**Solution:**

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.

**Solution:**

Median is less than meaan 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?

**Solution:** Using the t- distribution, it is found = that:

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- t s ^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- t s ^n=200 - 2.3452 30 = 198.43 30^2000

x+ t s ^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**

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

**Solution:** Mean

x.mean () =41.213125

Median

X.median()=40.5

variance

x.var()

28.2990

#Standard Deviation

x.std()

5.3194

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

**Solution:** Symetrical

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

**Solution:** Right Skewed

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

**Solution:** Left Skewed

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

**Solution:** The data is normally distributed and kurtosis value is 0.

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

**Solution:** 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?

**Solution:** Let’s assume above box plot is about age’s of the students in a school.50% of the people are above 10yrs old and remaining are less. And students Who’s age is above 15 are approx. 40%.

What is nature of skewness of the data?

**Solution:** Left skewed, median is greater than mean.

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

Approximately: 8

Q19) Comment on the below Boxplot visualizations?



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

**Solution:** 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

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

**Solution:-**

a = subset(MPG,MPG>38)

show(a)

=33/81

b = subset(MPG,MPG<40)

show (b)

= 67/81

P (20<MPG<50)

Show (c)

=69/81

Step-by-step explanation:

MPG<- Cars$MPG

MPG sample (MPG)

a = subset(MPG,MPG>38)

show(a)

b=subset(MPG,MPG<40)

show(b)

c = subset(MPG,MPG>20 & MPG <50)

show(c)

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

**Solution:** Waist Circumference(Waist) = (91.90183486238533, 90.8)

Adipose Tissue = (101.89403669724771, 96.54)

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

**Explanation:**

From scipy.import stats

From scipy.Stats import norm

#Z-score of 90% confidence interval stats.norm.(0.95)

#Z-Score of 94% confidence interval stats.norm.(0.97)

#Z-Score of 60% confidence interval stats.norm.(0.8)

A 94% confidence interval has two tails of 6/2=3%

So it goes from 3% to 97% which leaves 94% in the middle

So Look up the Z for

P(z<Z) = 0.97

Two closest value in the z- table

P(z<1.88)= 0.96995

P(z<0.89)= 0.97065

Interploting

1.88 +(0.97-0.96995)\*(0.01)/(0.97062-0.96995)

Z critical = approx.1.880746

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

**Solution:**

The formula for a 100(1-a) % confidence interval for µ is

X+ (tα/2,n-1 is the point on the t- distribution

100 (α/2) % of the distribution area to its right

For a 95% confidence interval,α=0.05 because

100(1-0.05)%

100(0.95)%

95%

25.9 + (1.987\*0.2846)

25.9 + (0.5655)

= 25,33,26,47

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

To find:

If the ceo’s claim were true

Ans: Assume H0= An average life of bulb>=260 days

X=260

Using python code

T=(260-270)/(90/18\*0.5)

t-0.4714045207910317

P value = 0.32167411684460556

P0.005 so reject h0

So, an average life of bulb< 260 Days