|  |  |
| --- | --- |
| 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 | Interval |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| 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) P(2 heads and 1 tail) = 3/8 = 0.375

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

1. Equal to 1 = P(Equal to 1) = 0
2. Less than or equal to 4

Total possible outcomes = 62 = 36

P(Less than or equal to 4) = 6/36 = 1/6 = 0.166

1. Sum is divisible by 2 and 3

P(Sum is divisible by 2 and 3) = 6/36 = 1/6 = 0.166

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. of balls = 2+3+2 = 7

Total Number of ways of drawing 2 balls out of 7 balls n(A) = 7 C2 = 7!/ 2!(7-2)!

= 7x6x5!/ 2!5! = 7x3 = 21

Number of ways of drawing 2 balls in which none of them is blue(2red+3green) n(E) = 5C2 = 5!/2!(5-2)! = 5!/2!3! = 5x4 / 2 = 10

P(none of the balls drawn is blue) = n(E)/n(A) = 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

Solution :

Expected number of candies for a randomly selected child = 𐕄 XP(x)

= 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+0.24

= 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.

Solution: Let Points be set A, Score be B and Weigh is C

For Points:

PSW=pd.read\_csv('Q7.csv')

Mean of the set A is = µ(A)= 𐕄(Xi)/n

PSW.Points.mean()

3.5965625

Median of set A is =

PSW.Points.median()

3.6950000000000003

Mode of set A =

PSW.Points.mode()

3.07 and 3.92

(most frequently repeated values in the dataset)

Variance = 𖽚2= 𐕄(Xi-µ)2/n-1

PSW.Points.var()

0.28588135080645166

Standard deviation = 𖽚

PSW.Points.std()

0.5346787360709716

Range =

minP=PSW.Points.min()

maxP=PSW.Points.max()

range1=maxP-minP

print(range1)

2.17

For Score: Mean of the set B is = µ(B) = 𐕄(Xi)/n

PSW.Score.mean()

= 3.217

Median of dataset B =

PSW.Score.median()

= 3.325

Mode of dataset B =

PSW.Score.mode()

=3.44

Variance of dataset B =

PSW.Score.var()

0.9573789677419356

Std deviation of score =

PSW.Score.std()

0.9784574429896967

Range =

minS=PSW.Score.min()

maxS=PSW.Score.max()

maxS-minS

3.9110000000000005

For Weigh: Mean of the set C is = µ(C) = 𐕄(Xi)/n

PSW.Weigh.mean()

= 17.8487500000000003

Median of set C =

PSW.Weigh.median()

= 17.71

Mode of dataset C =

PSW.Weigh.mean()

=17.02 and 18.90

Variance of the Weigh =

PSW.Weigh.var()

= 3.193166129032258

Std Deviation of weigh =

PSW.Weigh.std()

= 1.7869432360968431

Range = minW=PSW.Weigh.min()

maxW=PSW.Weigh.max()

maxW-minW

8.399999999999999

**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?

Solution: Expected value = mean of the given weights

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

=1173/9 = 130.333

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

**Cars speed and distance**

**Use Q9\_a.csv**

Speed\_distances=pd.read\_csv('Q9\_a.csv')

skewness and kurtosis of speed and distances

Speed\_distances.speed.skew()

-0.11750986144663393

Speed\_distances.dist.skew()

0.8068949601674215

Speed\_distances.speed.kurt()

-0.5089944204057617

Speed\_distances.dist.kurt()

0.4050525816795765

Skewness of all the speed values is -0.117 which is a negative value and skewness of all the distances is 0.806 which is a positive value.

Kurtosis of all the speed values is -0.508 which is a negative value and skewness of all the distances is 0.405 which is a positive value.

Both the above values are between -1 to +1 .

Any value of the skewness and kurtosis between -1 and +1 can be represented as Normally distributed data.

Hence, the above given data of the speed and distances of the cars are Normally distributed.

**SP and Weight(WT)**

**Use Q9\_b.csv**

SP\_WT=pd.read\_csv('Q9\_b.csv')

skewness and kurtosis of SP and WT

SP\_WT.SP.skew()

1.6114501961773586

SP\_WT.WT.skew()

-0.6147533255357768

SP\_WT.SP.kurt()

2.9773289437871835

SP\_WT.WT.kurt()

0.9502914910300326

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



Histogram Explanation:

The histogram of ChickWeight$weight exhibits skewness in the data . Skewness defines as the measure of asymmetry in the distribution. The data is called a skewed data if the graph does not show the equal tales on the either side.

The graph has the longer tail towards the right side which means the data has the positive Skewness.

Hence, it is written as Skewness > 0 (positive Skewness)

Boxplot Explanation:

Boxplot is generally used to determine or represent the Outliers in the data.

The above Boxplot diagram represents that there are lot of outliers in the data which is denoted in the form of small dots on the other side. The boxplot consists of a median line which is the center line in the diagram and one end of the box is the upper quartile(Q3) and the other one is lower quartile(Q1). The inter quartile range IQR is Q3-Q1. The values in the data which exceeds the value of the upper boundary and lower boundary are called outliers.

Upper boundary = Q1-1.5(IQR)

Lower boundary = Q3+1.5(IQR)

If the values on the either side of the box exceeds the upper boundary or lower boundary values then they are called outliers. In the above boxplot diagram given, some of the values or points exceed the upper boundary. Hence, they are the outliers.

Also, the distance between the upper boundary and median is greater when compared to the distance between the lower boundary and median which concludes that the data is not normally distributed and therefore is a skewed 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: Given total men in sample n = 2000

Mean of the weight of the men (x-)= 200 pounds

Std deviation of weights sample (s)= 30 pounds

Confidence interval for 94%

stats.t.interval(0.94,loc=200,scale=30/np.sqrt(2000),df=1999)

The confidence interval is (198.7376089443071, 201.2623910556929)

Confidence interval for 98%

stats.t.interval(0.98,loc=200,scale=30/np.sqrt(2000),df=1999)

The confidence interval is (198.4381860483216, 201.5618139516784)

Confidence interval for 96%

stats.t.interval(0.96,loc=200,scale=30/np.sqrt(2000),df=1999)

The confidence interval is (198.6214037429732, 201.3785962570268)

**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:

1. Mean : x-- = 𐕄xi/n where n=18

= (34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56)/18

= 738/18 = 41

Mean of the given data is 41

Median: average of the middle most two numbers

= (40 + 41)/2 = 40.5

Variance:

s2 = 𐕄(xi – x-- )2 /n – 1

= 49 +25+25+9+9+4+4+1+1+0+0+0+0+1+1+16+64+225 / 17

= 434 / 17 = 25.529

Standard Deviation:

s = √s2 = √25.529 = 5.05

1. The student marks in the above are not normally distributed as the mean and median of the data is not the same from the data above. There is a slight deviation between the mean and median because of the outlier present in the data. Hence such kind of data is called a Skewed data.

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

Ans) The skewness is zero when mean and median are equal and it represents Normal Distribution.

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

Ans) The skewness is positive if the mean > median because of the outlier in the data, the tale on the probability distribution curve is longer on the right side.

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

Ans) The skewness is negative if the median > mean because of the outlier in the data, the tale on the probability distribution curve is longer on the left side.

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

Ans) Positive kurtosis value for the data indicates narrow peak in the probability distribution curve drawn through plotting of the points in the data.

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

Ans) Negative kurtosis value for the data indicates broader peak in the probability distribution curve drawn through plotting of the points in the data.

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



What can we say about the distribution of the data?

The probability distribution of the data is not uniform and hence it is not a Normal Distribution.

What is nature of skewness of the data?

The nature of the skewness of the data is negative skewness that is skewness<0 as the distance from lower boundary to the median is more than the distance from the outer boundary to the median.

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

From the boxplot Q3 = 18 and Q1 = 10

IQR = Q3 – Q1

IQR = 18 – 10 = 8

IQR = 8

Q19) Comment on the below Boxplot visualizations?



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

In the above distribution of data for boxplot 1 with respect to the boxplot 2, We can say that either side of the median for both the boxplots have the symmetrical space and the probability distribution curve for both the boxplots has the equal number of curve distribution tales with perfect bell shape so it is said to be a Normally Distributed data with the skewness 0.

The center line indicates the median and by looking at the diagram we can say that boxplot 1 has the lower IQR (Inter Quartile Range) than that of the IQR of the boxplot 2.

IQR for boxplot 1 = Q3-Q1 = 280 – 250 = 30 wbs

Upper extreme for boxplot 1 = 287.5

Lower extreme for boxplot 1 =240

IQR for boxplot 2 = Q3-Q1 = 312 – 225 = 87 wbs

Upper extreme for boxplot 2 = 325

Lower extreme for boxplot 2 = 200

Median for both boxplots = 265 wbs

Hence, IQR for the boxplot 2 is greater than the IQR for the boxplot 1. Also, it indicates that Q1 and Q3 for boxplot 1 are closer to the median which makes the boxplot have the narrower peak and for boxplot 2 the Q1 and Q3 are far from the median which makes a broader peak.

In boxplot 1, For a narrow peak, the kurtosis value would be greater than 0 which indicates the positive kurtosis.

In boxplot 2, For a broader peak, the kurtosis value would be less than 0 which indicates the negative kurtosis.

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

Mean of MPG =

cars.MPG.mean()

34.42208

Std deviation of MPG =

cars.MPG.()

9.131445

* 1. P(MPG>38) = 1 – P(MPG ≤ 38)

P(MPG ≤ 38)=

stats.norm.cdf(38,34.42208,9.131445)

0.6524058977429089

P(MPG>38) = 1 – P(MPG ≤ 38)

= 1 - 0.6524058977429089

= 0.3475941022570911

* 1. P(MPG<40) =

stats.norm.cdf(40, 34.42208,9.131445)

0.7293497154033285

* 1. P (20<MPG<50) =

stats.norm.cdf(50,34.42208,9.131445) -stats.norm.cdf(20,34.42208,9.131445)

0.8988689170528976

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

MPG of cars is said to be randomly distributed if the mean = median and if the area under the curve is 1.

Mean of MPG = sum of all the MPG / Total number of MPG entries

= 2788.188134/ 81 = 34.422

Median of MPG = 35.152

Hence, the mean and median are not the same, so we can say that MPG of cars are not normally distributed.

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

Dataset: wc-at.csv

Adipose Tissue and Waist Circumference is said to be randomly distributed if the mean = median and if the area under the curve is 1 .

Mean of Adipose tissue = sum of all the Adipose tissue/ Total number of adipose tissues

= 11106.45/ 109 = 101.894

Median of MPG = 96.54

Hence, the mean and median are not the same, so we can say that Adipose tissues are not normally distributed.

Mean of Waist Circumference = sum of all the Waist Circumference / Total number of Waist Circumference

= 10017.3/ 109 = 91.901

Median of Waist Circumference = 90.8

Hence, the mean and median are not the same, so we can say that Waist Circumference are not normally distributed.

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

Z Scores of the 90%, 94% and 60% CI :

90% Confidence Interval:

stats.norm.interval(0.90)

(-1.6448536269514729, 1.6448536269514722)

94% Confidence Interval:

stats.norm.interval(0.94)

(-1.8807936081512509, 1.8807936081512509)

60% Confidence Interval:

stats.norm.interval(0.60)

(-0.8416212335729142, 0.8416212335729143)

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

t Scores of the 95%, 96% and 99% CI:

given n = 25 and df = n-1 = 24

90% Confidence Interval:

stats.t.interval(0.95,df=24)

(-2.0638985616280205, 2.0638985616280205)

94% Confidence Interval:

stats.t.interval(0.96,df=24)

(-2.1715446760080677, 2.1715446760080677)

60% Confidence Interval:

stats.t.interval(0.99,df=24)

(-2.796939504772804, 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 bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

Ans) Given no. of bulbs n = 18 bulbs

Sample average x-- =260 days

Standard deviation s = 90 days

Population mean µ = 270

t score =( - µ )/( s/√n) = (260 – 270)/ (90/√18)

= (-10)/(21.213) =-0.4714

stats.t.cdf(-0.4714,17)

0.32167411684460556