|  |  |
| --- | --- |
| 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 | interval |
| Fahrenheit Temperature | ratio |
| 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 | interval |
| Religious Preference | ordinal |
| Barometer Pressure | interval |
| SAT Scores | ratio |
| Years of Education | nominal |

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

= three coin are tossed the total number of possible combinations = 8

(HHH, HHT, HTH, THH, TTH, THT, HTT, TTT)

Combination having two heads and one tail are:

HHT, HTH, TTH

which makes them 3 in no.

P (two heads and one tails) = 3

= 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

=total possible outcome

= 36

1. sum equal to 1 = 0

0/36 = 0

1. sum equal to 4 = 3

{(1,1) , (1,2) ,(1,3) , (2,2) , (2,3) , (3,3)}

3/36 = 12

1. sum is divisible by 2 & 3 = 6

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

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

(3,3) (3,4) (3,5) (3,6)

(4,4) (4,5) (4,6)

(5,5) (5,6)

(6,6)

Probability = 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?

= total number of balls = (2+3+2) =7

Let s be the sample space.

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

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

=21

Let E= event of drawing 2 balls, none of which is blue.

=5 c2

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

=10

P (E) = n (E)/n(S)

= 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 cadies 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

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**

= cars.mean ():- points 3.596563, Score 3.217250

Weight 17.848750, dtype: - float 64

Cars.median ():- point 3.695, score 3.325, weight 17.710, dtype: - float 64

Cars.point.mode ():- 0 3.07

1. 3.92

Name: points, dtype: float64

Cars.score.mode ():-0 3.44

Name: score, dtype: float64

Cars.weight.mode ():- 0 17.02

1. 18.90

Name: weight, dtype: float64

Cars. Variance ():- point 0.285881

Score 0.957379

Weight 3.193166

dtype: float64

cars.stad ():- point 0.534679

Score 0.978457

Weight 1.786943

dtype: float64

#range

Point\_Range = cars.Points.max () - cars.Points.min ()

Point\_Range

: -2.17

Score\_Range = cars.Score.max () - cars.Score.min ()

Score\_Range

: - 3.911000000000005

Weigh\_Range = cars.Weigh.max () - cars.Weigh.min ()

Weigh\_Range

: - 8.399999999999999

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?

= Expected value ∑ (probability\*value)

∑P x).E(x)

there are 9 patients

Probability of each patient = 1/9

= 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**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**=**

**Q9\_a:-** # skewness

Data.Skew ():- index 0.000000

Speed -0.117510

dist 0.806895

dtype: float 64

# Kurtosis

Data.kurt ():- index -1.200000

Speed -0.508994

dist 0.405053

dtype: float 64

**Q9\_b :-**

#skewness

Data2.skew ():- sp 1.611450

Wt -0.614753

dtype: float 64

#kurtosis

Data2.kurt():- sp 2.977329

Wt 0.950291

dtype: float 64

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



= the histograms peak has right skewness and positive skewed, tail is on right. Mean > Median. We have outliers on the higher side.



= the boxplot has outliers on the maximum side.

**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?

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

Print (np.round(conf\_94,0)) 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.
2. What can we say about the student marks?
3. = mean = 41,

Median = 40.5,

variance = 25.52

standard deviation = 5.05

1. = we don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

= nature of skewness. Skewness can be positive, negative or zero

When the value of mean, median and more are equal there is no skewness.

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

= the mean of positively skewed data will be greater than the median.

Skewness and tail is towards right.

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

= skewness and tail is towards left.

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

= positive kurtosis means the curve is more peaked and it is leptokurtic.

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

= 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 of the data?

= the above boxplot is not normally distributed the median is toward the higher value.

What is nature of skewness of the data?

= the data is a skewed towards left the whisker range of minimum value is greater than maximum.

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

= the inter quantile range= Q3 upper quartile – Q1 lower quartile

= 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.

= first there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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)

1. = Prob\_MPG\_greater\_than\_38=np.round(1-stats.norm.cdf(38, loc=q20.MPG.mean(), scale= q20.MPG.std()),3)print(‘P(MPG>38)=’,prob\_MPG\_greater\_than\_38

P(MPG>38)= 0.348

1. = prob\_MPG\_less\_than\_40= np.round(starts.norm.cdf(40,loc = q20.MPG.mean(), scal = q20.MPG.std()),3)

print(‘P(MPG<40)=’,prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

1. = prob\_MPG\_greater\_than\_20= np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(),scale= q20.MPG.std()),3)print(.p(MPG>20)=’,

(prob\_MPG\_greater\_than\_20)) p(MPG>20)= 0.943

Prob\_MPG\_less\_than\_50=

np.round(stats.norm.cdf(50,loc = q20.MPG.mean(),

scale= q20.MPG.std()),3) print(‘P(MPG<50)=’,

(prob\_MPG\_less\_than\_50)) P(MPG<50)= 0.956

Prob\_MPG\_greaterthan20\_and\_lessthan50=

(prob\_MPG\_less\_than\_50) –

(prob\_MPG\_greater\_than\_20) print(‘P(20<MPG<50)=’,

(prob\_MPG\_greaterthan20\_and\_lessthan50))

P(20<MPG<50=0.013000000000000012

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 follows normal distribution.

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 (AT) and waist does not follow normal distribution.

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

= 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

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

= To compute the 95% confidence interval,

M = (2 + 3 + 5 + 6 + 9)/5

= 5.

σ M = 1.118. Z.95

specifying that the shaded area is 0.95

Confidence Level z

0.90 1.645

0.92 1.75

0.95 1.96

0.96 2.05

A 99 percent confidence interval would be wider than a 95 percent confidence interval

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

= 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 =

t = - 0.471