|  |  |
| --- | --- |
| 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) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

**Solution:-**

H = heads

T = tails

Possible events=>

HHH

HHT

HTH

HTT

THH

THT

TTH

TTT

Number with 2 heads: 3

Total number: 8

P (2H, 1T) is 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 2 and 3

**Solution:-**

Total possible outcome =62=36

a) Favorable outcome (sum equal to 1) = 0

Required probability=0/36 =0

b) Favorable outcome (sum less than or equal to 4) = 6 {(1, 1) (1, 2) (1, 3) (2, 1) (2, 2) (3, 1)}

Required probability =6/36 =0.166

c)Favorable outcome (Sum is divisible by 2 and 3) = 6 {(1,5) (2,4) (3,3) (4,2) (5,1) (6,6)}

Required probability =6/36 =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 num of balls= 2+3+2 = 7

Let S be the sample space

N(S)=Num of ways of drawing 2 balls out of 7 = 7C2 =(7\*6)/(2\*1) = 21

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

N(E)= Num of ways of drawing 2 balls out of (2+3) balls = 5C2 =(5\*4)/(2\*1) = 10

Probability P(E) = n(E)/n(S)= 10/21= 0.476

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 = 1\*0.015+ 4\*0.20+ 3\*0.65+ 5\*0.005+ 6\*0.01 +2\*0.120 =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**

**Solution:-**

**Points Score Weigh**

|  |  |  |  |
| --- | --- | --- | --- |
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Standard Dev | 0.534679 | 0.978457 | 1.786943 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Min | 2.76 | 1.513 | 14.5 |
| Max | 4.93 | 5.424 | 22.9 |
| Range | 2.17 | 3.911 | 8.4 |
|  |  |  |  |

To understand the center of the dataset we calculate measure if tendency i.e. mean, median and mode. To understand the measure of how far the data spread apart we calculate measure of dispersion.

Here for e.g.:-

Some cars’ score is more than 3.2 and some car’s score is less than 3.2.

To know how close or how far away car’s score from 3.2. To understand that we calculate measure of dispersion i.e. standard deviation, variance and range.

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 of the Weight of that patient= Mean= 1308

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Solution:-**

Speed Skew : -0.11395477012828319

The speed is fairly symmetrical as the skewness is lies between -0.5 and 0.5. Since the value is negative, the data is skewed to the left— the left tail is slightly longer than the right tail. -> Negatively skewed data

Dist Skew : 0.7824835173114966

The distance is not fairly symmetrical as the skewness is not lies between -0.5 and 0.5. Since the value is positive, the data is skewed to the right— the right tail is slightly longer than the left tail. ->Positively skewed data

Speed Kurtosis : -0.5771474239437371

In speed, as the values are between +1 and -1, the distribution is generally not too peaked and not too flat. Nonetheless, the negative result indicate a slight Platikurty distribution.

Dist Kurtosis : 0.24801865717051808

In Dist, as the value is not between +1 and -1, the distribution is generally too peaked.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Solution:-**

SP Skew : 1.5814536794423764

The SP is not fairly symmetrical as the skewness is not lies between -0.5 and 0.5. Since the value is positive, the data is skewed to the right— the right tail is slightly longer than the left tail. ->Positively skewed data

SP Kurtosis : 2.7235214865269244

In SP, as the value is not between +1 and -1, the distribution is generally too peaked.

WT Skew : -0.6033099322115126

The WT is fairly symmetrical as the skewness is lies between -0.5 and 0.5. Since the value is negative, the data is skewed to the left— the left tail is slightly longer than the right tail. -> Negatively skewed data

WT Kurtosis : 0.8194658792266849

In WT, as the values are between +1 and -1, the distribution is not too peaked and not too flat.

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



**Solution:-**

Histogram =>

It is used to represent frequency of distribution. It plots group of observation within a certain interval. Here, the most of the data points are in the range 50-100 with frequency 200 and least data points are in 0-10 with frequency 400. The data is right skewed, i.e. the tail is more in right.

Boxplot =>

It used to find outliers in the dataset. Any values that are located beyond upper extreme and lower extreme will consider as outliers. Here, we have outliers on the upper side of box plot. There is less data points between lower quartile (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:-**

Confidene interval for 94%= 1.882

Confidence interval for 98%= 2.33

Confidence interval for 96% = 2.05

Using Z-score,

94% =>

[200.82, 198.56]

98% =>

[201.0, 198.98]

96% =>

[200.90, 199.09]

**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.

Mean = 40.94

Median = 40

Variance = 27.05

Std = 5.20

1. What can we say about the student marks?

Some students score is more than 40 and some students score is less than 40.

To know how close or how far away car’s score from 40, we calculated variance and standard deviation.

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

The data is symmetrical

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

The data is right skewed

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

The data is left skewed

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

The data is normally distributed and kurtosis value is 0.

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

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?

50% of the people are above 10 and remaining are less. And above 15 are approx. 40%.

What is nature of skewness of the data?

Left Skewed. Median is greater than mean

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

10-18= -8

Q19) Comment on the below Boxplot visualizations?



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

By observing both the plots the whisker’s level is high in boxplot2, 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)

There are 33 observations in MPG which are greater than 38

* 1. P(MPG<40)

There are 61 observations in MPG which are less than 40

* 1. P (20<MPG<50)

There are 69 observations in MPG which are less than 20 and 50

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Mean = 34.42

Median = 35.15

Median > Mean

So MPG is not a normal distribution. Left skewed data.

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

Dataset: wc-at.csv

Waist =>

Mean= 91.90

Median = 90.8

Mean>Median

Right-Skewed data

AT =>

Mean = 101.89

Median = 96.54

Mean>Median

Right-skewed Data

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

For 90%= 1.96 stats**.**norm**.**ppf(0.95)

For 94%= 2.5 stats**.**norm**.**ppf(0.97)

For 60% = 2.47 stats**.**norm**.**ppf(0.8)

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

np.random.seed(0)

data = np.random.randint(10, 30, 25)

#create 95% confidence interval

st.norm.interval(alpha=0.95, loc=np.mean(data), scale=st.sem(data))

O/p: (17.519022904371603, 22.240977095628395)

#create 96% confidence interval

st.norm.interval(alpha=0.96, loc=np.mean(data), scale=st.sem(data))

O/p: (17.406049368039085, 22.353950631960913)

#create 99% confidence interval

st.norm.interval(alpha=0.99, loc=np.mean(data), scale=st.sem(data))

O/p: (16.777150133422204, 22.982849866577794)

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

Solution:-

*# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days*

*# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days*

*# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))*

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

*# Find P(X>=260) for null hypothesis*

*# p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function*

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

o/p: 0.32167411684460556

*# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function*

p\_value**=**stats**.**t**.**sf(abs(**-**0.4714),df**=**17)

p\_value

o/p: 0.32167411684460556