|  |  |
| --- | --- |
| 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 | Categorical (Nominal) |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Categorical (Nominal) |

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 | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Ratio |
| 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 | Ordinal |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

Solution: Probability is 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

Solution:

1. Probability of sum equal to 1 is 0
2. Probability of less than or equal to 4 is 6/36 = 0.1667
3. Probability of sum is divisible by 2 and 3 is 5/36 = 0.1389

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: The probability of drawing 2 balls that are not blue is

5C2=10; 7C2=21

P(E) = 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 |

Solution:

Child A – probability of having 1 candy = 0.015 = 1\*0.015 = 0.015

Child B – probability of having 4 candies = 0.20 = 4\*0.20 = 0.8

Child C – probability of having 3 candies = 0.65 = 3\*0.65 = 1.95

Child D – probability of having 5 candies = 0.005 = 5\*0.005 = 0.025

Child E – probability of having 6 candies = 0.01 = 6\*0.01 = 0.06

Child F – probability of having 2 candies = 0.120 = 2\*0.120 = 0.24

The expected number of candies a randomly selected child getting is 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:

Mean, Median, Mode, Variance, Standard Deviation, and Range for Points, Score, Weigh is

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| mean | 3.596563 | 3.217250 | 17.848750 |
| median | 3.695000 | 3.325000 | 17.710000 |
| mode | 3.92 | 3.44 | 17.02 |
| variance | 0.285881 | 0.957379 | 3.193166 |
| std | 0.534679 | 0.978457 | 1.786943 |
| range | 2.17 | 3.911 | 8.399999 |

1. Most of the data lies around the median value itself.
2. From the box plot, ‘Score’ and “Weigh’ have got outliers.
3. From the above table we can tell ‘Points’ has got very less variance and ‘Weigh’ has got most dispersion of data.
4. ‘Points’ and ‘Score’ have negative skewness and ‘Weigh’ is positive skewed.
5. ‘Weigh’ has got more spread of data by looking at the 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: Assuming one of the patients is chosen at random and the expected value of the patient weight is the mean of all the given data. i.e.,

(108+110+123+134+135+145+167+187+199)

9

=145.33

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

**Car’s speed and distance**

**Use Q9\_a.csv**

Solution:

*Skewness*

speed -0.117510

dist 0.806895

* **Inference**

1. speed is negatively skewed (Left skewed)
2. dist is positively skewed (Right skewed)

*Kurtosis*

speed -0.508994

dist 0.405053

* **Inference**

1. speed has got negative kurtosis (Platykurtic)
2. dist has got positive kurtosis (Leptokurtic)

**SP and Weight (WT)**

**Use Q9\_b.csv**

Solution:

*Skewness*

SP 1.611450

WT -0.614753

* **Inference**

1. SP is positively skewed (Right skewed)
2. WT is negatively skewed (Left skewed)

*Kurtosis*

SP 2.977329

WT 0.950291

* **Inference**

Both SP and WT have Positive Kurtosis (Leptokurtic)

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



Solution:

* The data is Right skewed (Positive Skewed)
* Data has got Positive Kurtosis (Leptokurtic)
* Most of the data lies in between 50 to 100
* Not a symmetric data



Solution:

* We can clearly see the outliers in the data
* The data has got Positive skewness (Right Skewed)
* The spread of data is more by looking at the length of Whiskers
* Not a symmetric distribution of 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:

Here total no of sample mean (n) = 2000

The average weight of person in sample (X̅) = 200

Standard deviation of sample () = 30

Confidence Interval =

#Z Score values

Z score for 94% is 1.5547735945968535

Z score for 98% is 2.0537489106318225

Z score for 96% is 1.7506860712521692

#Confidence interval for 94%

Upper Limit 🡪 201.0429738341421

Lower Limit 🡪 198.9570261658579

#Confidence interval for 98%

Upper Limit 🡪 201.37769665186667

Lower Limit 🡪 198.62230334813333

#Confidence interval for 96%

Upper Limit 🡪 201.17439591877456

Lower Limit 🡪 198.82560408122544

**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:

| **DATA** | **VALUE** |
| --- | --- |
| MEAN | 41.000000 |
| MEDIAN | 40.500000 |
| VARIANCE | 25.529412 |
| STANDARD DEVIATION | 5.052664 |
|  |  |

1. Most of the students scored between 35 to 45 marks

2. Most of the times students score 41. So mean is 41

3. Not normally distributed

4. Data has got outliers

5. Skewness of data Left Skewed (Negative skewness)

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

Solution: Symmetrical Distribution (Zero skewness).

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

Solution: Positive skewness (Most of the data is distributed towards the right side of the median).

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

Solution: Negative skewness (Most of the data is distributed towards the left side of the median).

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

Solution: Positive kurtosis, also known as Leptokurtic indicates that the data is peaked and has got a thicker tail.

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

Solution: Negative kurtosis, also known as Platykurtic indicates that the data distribution is flat and has got thin tails.

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



What can we say about the distribution of data?

Solution: Distribution of data is concentrated more towards the right side. We can tell that the data is not symmetrical in nature.

What is the nature of skewness of the data?

Solution: The data is negatively skewed.

What will be the IQR of the data (approximately)?   
Solution: IQR(Approx.) = 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.

Solution:

1. There are no outliers in both boxplots.
2. The data is normally distributed in both plots.
3. The median of both boxplots lies approximately around 265.5.
4. The second box plot has got more range compared to first box plot.

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)

Solution:

1. P(MPG>38)

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

= 0.3475939251582705

1. P(MPG<40)

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

= 0.7293498762151616

1. P(20<MPG<50)

(stats.norm.cdf(50, cars['MPG'].mean(), cars['MPG'].std())) - (stats.norm.cdf(20, cars['MPG'].mean(), cars['MPG'].std()))

= 0.8988689169682046

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Solution:

Mean of MPG: 34.422075728024666

Median of MPG: 35.15272697

Mode of MPG: 0 29.629936

Name: MPG, dtype: float64

Std. Deviation of MPG: 9.131444731795982

Skewness of MPG: -0.17794674747025727

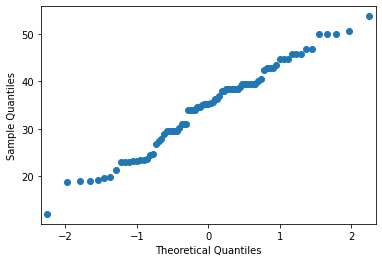
Kurtosis of MPG: -0.6116786559430913

AxesSubplot(0.125,0.125;0.775x0.755)

Figure(432x288)

Chart, histogram

Description automatically generated



***MPG of Cars follow 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

Solution:

Mean of MPG: 91.90183486238533

Median of MPG: 90.8

Mode of MPG: 0 94.5

1 106.0

2 108.5

Name: Waist, dtype: float64

Std. Deviation of MPG: 13.559115982678826

Skewness of MPG: 0.1340560824786468

Kurtosis of MPG: -1.1026666011768886

AxesSubplot(0.125,0.125;0.775x0.755)

Figure(432x288)

Chart, histogram

Description automatically generated

A picture containing chart

Description automatically generated

Mean of MPG: 101.89403669724771

Median of MPG: 96.54

Mode of MPG: 0 121.0

1 123.0

Name: AT, dtype: float64

Std. Deviation of MPG: 57.29476272231215

Skewness of MPG: 0.584869324127853

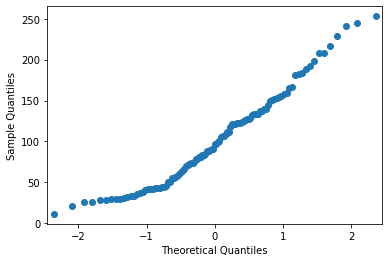
Kurtosis of MPG: -0.28557567504584425

AxesSubplot(0.125,0.125;0.775x0.755)

Figure(432x288)

Chart, histogram

Description automatically generated



***Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution***

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

Solution:

#for 90% confidence interval

stats.norm.ppf(0.95)

1.6448536269514722

#for 94% confidence interval

stats.norm.ppf(0.97)

1.8807936081512509

#for 60% confidence interval

stats.norm.ppf(0.80)

0.8416212335729143

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

Solution:

#for 95% confidence interval for sample size of 25.

stats.t.ppf(0.975,24)

2.0638985616280205

#for 96% confidence interval for sample size of 25.

stats.t.ppf(0.98,24)

2.1715446760080677

#for 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

Solution:

Population mean, µ = 270

Sample size, n = 18

Sample mean, x̅ = 260

Standard deviation, s =90

t score = (x̅- µ)/(s/sqrt(n))

= -0.4714045207910317

Df = degrees of freedom = n-1 = 18-1 = 17

Probability = 0.32167253567098364