|  |  |
| --- | --- |
| 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) | 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 | Internal |
| Weight | Ratio |
| Hair Color | Nomial |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Internal |
| Height | Internal |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Internal |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Internal |
| Number of Children | Ratio |
| Religious Preference | Ordinal |
| Barometer Pressure | Ratio |
| SAT Scores | Internal |
| Years of Education | Ratio |

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

= Probability = Favourable outcomes/Total no of outcomes

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

= 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. Sum is equal to 1= 0
2. Less than equal to 4

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

= 6/36

= 0.17

Sum is divisible by 2 and 3

= {(1,5)(2,4),(3,3),(4,2).(5,2),(6,6)}

= 6/36

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:

= {R,R,G,G,G,B,B}

= Probability of first ball being not blue is 5/7

= Probability of second ball being not blue is 4/6

= 5/7 and 4/6 = 5/7\*4/6

= 20/42

=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:

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count (x) | Probability [p(x)] | x\* P(x) |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.20 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.120 | 0.24 |

= expect number of candies = sum of (X\*p(x)

= sum of ( candies count\* probability)

= 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.

solution:

**Points:**

Mean = 3.596563, Median = 3.695, Mode = 3.92,

Range=2.17

Variance= 0.2858814

Standard deviation= 0.5346787

**Score:**

Mean = 3.21725, Median = 3.325, Mode =3.44,

Range=3.911

Variance= 0.957379,

Standard deviation = 0.9784574

**Weight:**

Mean = 1784875, Median = 17.71, Mode = 17.02,

Range=8.4

Variance= 3.193166,

Standard deviation = 1.78694

1. Central tendency – Here Mean, Median, and Mode are different and closer. Therefore, the Data set is concentric around the mean, slightly asymmetric, and skewness exists. As per the graphs, all are right-skewed.

B)Distribution – The smaller values of Var, SD, and Range confirm that the Data set is concentric around the mean and not widely spread/distributed.

**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 of a Discrete Random Variable is actually the mean of the statements

= Hence Mean is 1308/9

  = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

1. **Cars speed – 1) Skewness: – 0.11 –> Left Skewed**

**2) Kurtosis: ­ – 0.51 – >Platokurtic, Kur < 3**

1. **Cars distance – 1) Skewness: 0.78 –> Right Skewed**

**2) Kurtosis: ­ 0.41 –> Platokurtic, Kur < 3**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Answer:**

1. **SP – 1) Skewness: 1.58 –> Right Skewed**

**2) Kurtosis: ­ 2.98 ~ 3 –> Mesokurtic, Kur = 3**

1. **Weight – 1) Skewness: - 0.60 –> Left Skewed**

**2) Kurtosis: ­ 0.95 –> Platokurtic, Kur < 3**

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



Solution :

Here we can see that the major Chick weights fall in the category of 50-100g(measures in x) as the maximum which is 200.The minimum weights have frequency if less than or equal to 5.

The plot is Right skewed which show that there is lesser concentration of chick weights in the 300-400gram category ,The expected value should be above 46.45



Solution

Median is less than mean right skewed and we have outlier on the upper side of 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:

X (+/-) Z(1-alpha)\*standard\_dev/sqrt(n)

Here

X=sample average or mean

Z(1-alpha)\*standard\_dev/sqrt(n) = Margin of error

Standard dev/sqrt(n)=0.6708

Zvalues

94=1.882

98=2.326

96=2.053

1. 200 (+/-) 1.882 \* 0.6708

=200 (+/-) 1.262

1. 200(+/-)2.326\*0.6708

=200(+/-) 1.560

c)200(+/-) 2.053\*0.6708

= 200(+/-) 1.377

**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.

**Solution:**

Mean = 41

Median= 40.5

Variance= 5.05

Standard deviation= 25.529

1. What can we say about the student marks?

This class contain student that are actually mediocre .most of the student in the class are having an average percentage of 65 and there re only a few student securing value above 90%

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

Answer:

The skewness will be symmetrical .hence both the sides of the plot must be equal in proportion for the data should be normally distributed.

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

Answer:

For data which produces mean>median the skeness will be a +ve skewness or the data will be right skewed. Most of the data will be lying on the lift side of the plot . mean always tends to go towards the most skewed part since skewness influences the mean

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

Answer:

For the data which produces mean> median the skewness will be a -ve skewness or the data will be left skewed . most of the data will be lying on the right side of the plot mean always tends to go towards the most skewed part since skewness influences the mean.

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

Answer:

A distribution with a positive kurtosis value indicates that the distribution has heavier tails than the normal distribution. For example ,data that follow a t distribution have a positive kurtosis value .

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

Answer:

A distribution with negative kurtosis value indicates that the distribution has lighter tails 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:**

The data is not actually equally distributed across the plane. The might be outliers influencing the data median of the data is 14.7(app x)

25 percentage of the data lies between 0-10

50 percentage of the data lies between 10-18

25 percentage of the data lies after 18-20 appx

What is nature of skewness of the data?

The data will be left skewed since whisker length on the upper quadrant is higher than the data on the lower quadrant .median will be greater than the mean since data is left skewed

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

IQR is the inter quartile range

Here Q1=10

Q2= 14.7

Q3 = 18

IQR =Q3-Q1 =8(approx)

Q19) Comment on the below Boxplot visualizations?

Here there is a representation of 2 box plots in which box plot 2) is highlydistributed across the plane and 1) is slightly less distributed.(variance)

Whiskers in these diagrams also show this.100% of the data is spread acrossvalues from 350 in 2 whereas its spread in range 250-290 app x in 1



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

Here when we compare box plot 1 with box plot 2 we can say that the data in box plot 1 is widely spread. Here the main inference is that since the data range varies high in box plot 2 it is hard to make a prediction in box plot 2. The median in the 2box plots are equal. And the data spread in both of them are 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)

c. P (20<MPG<50)

Solution

Ans – Total nos of observations = 81

|  |  |  |
| --- | --- | --- |
| Criteria | Count | Probability |
| MPG > 38 | 33 | 0.41 |
| MPG < 40 | 61 | 0.75 |
| 20 < MPG < 50 | 69 | 0.85 |

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer:

All data values of MPG lie between 3 Standard deviation levels. However, there is slight skewness in the data set, and the mean is almost the same as the median. Therefore, the MPG of cars **follows Normal distribution** with slightly left skewness.

|  |  |
| --- | --- |
| **MPG descriptive stats** | |
| Mean | 34.42 |
| Median | 35.15 |
| SD | 9.07 |
| 3Sigma lower | 7.20 |
| 3Sigma upper | 61.65 |
| Maximum | 53.70 |
| Minimum | 12.10 |
| Skewness | -0.17 |

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

Dataset: wc-at.csv

Answer:

i) All data values of Waist lie between 3 Standard deviation levels. However, there is slight skewness in the data set, and the mean is almost the same as the median. Therefore, the Waist **follows Normal distribution** with slightly Right skewness.

|  |  |
| --- | --- |
| **Waist descriptive stats** | |
| Mean | 91.90 |
| Median | 90.80 |
| SD | 13.50 |
| 3Sigma lower | 51.41 |
| 3Sigma upper | 132.39 |
| Maximum | 121.00 |
| Minimum | 63.50 |
| Skewness | 0.13 |

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

Z scores

=90%

=95+2.5

=97.5

=qnorm(0.975)

=1.96

=94 %

=94 +4

=qnorm(0.97)

=1.88

60%

=60 +20

=80

=qnorm(0.80)

=0.841

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

T Score calculation

T((1, alpha ), (n-1))

Here n=25

n-1 =24

Here t score values will be :

95%

= qt (0.975,24)

=2.063899

96%

=qt(0.98,24)

=2.171545

99%

=qt(0.995,24)

=2.79694

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:

Samplesize = 18 =n

Sample mean = 260 days =X

Sample standard deviation =s = 90days

T score = (x – mu) / (s / sqrt(n)) = (260 – 270) / (90 / sqrt(18)) = (-- 0.471)

Probability value of the above T score for left-tailed with df 17 = 0.3218

Hence, there is a 32.18% probability that 18 nos randomly selected bulbs would have an average life of no more than 260 days.