|  |  |
| --- | --- |
| 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 | Nominal |
| 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 | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| 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?

**Ans:** 3/8

Sample Space n(S) ={HHH, TTT, HHT, THH, HTT, TTH, HTH, THT }

P(E) = n(E)/ n(S) = 3/8 i.e. 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

**Ans:** a) 0

b) 6/36 = 1/6 i.e. 0.166

c) 6/36= 1/6 i.e. 0.166

Sample Space n(S) = {(1,1), (1,2), (1,3),(1,4),(1,5),(1,6)

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

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

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

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

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

1. P(E) = n(E)/n(S) = 0/36 =0
2. n(E)= (1,1), (1,2), (1,3), (2,1), (2,2), (3,1)

P(E) = 6/36 = 1/6

1. n(E) = (1,5), (2,4), (3,3), (4,2), (5,1), (6,6)

P(E) = 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?

**Ans:** 10/21

Sample Space n(S) = 7C2

Event n(E) =2 red, 3 green i.e. two balls drawn out of 5 = 5C2

P(E)= 5C2/ 7C2 = 5x4 / 7x6 = 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

**Ans:** 3.09

Expected number of candies = Average no. of candies = Summation(no. of candies x probability) = (1x.015) + (4x.20) + (3x.65) + (5x.005) + (6x.01) + (2x.12)

= 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**

**Ans: Points**

Mean: 3.596563 (average value)

Median: 3.695 (middle most value)

Mode: 3.92 (most repeated value)

Variance: 0.276948 (population)

Standard Deviation: 0.526258 (population)

Range: 2.17 (max value - min value)

Mode>Median>Mean: left skewed data (negative skewed data)

Std = 0.526258=> maximum no of data points lies within the range of (3.596563- 0.526258 to 3.596563 + 0.526258)

**Score**

Mean: 3.21725

Median: 3.325

Mode: 3.44

Variance: 0.927461 (population)

Standard Deviation: 0.963048 (population)

Range: 3.911

Mode>Median>Mean: left skewed data (negative skewed data)

maximum no of data points lies within the range of (3.21725- 0.963048 to 3.21725+ 0.963048)

**Weigh**

Mean: 17.84875

Median: 17.71

Mode: 17.02

Variance: 3.09338 (population)

Standard Deviation: 1.758801 (population)

Range: 8.4

Mean> Median> Mode: right skewed data (positive skewed data)

maximum no of data points lies within the range of (17.84875 - 1.758801 to 17.84875 + 1.758801)

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?

**Ans:** Expected Value = average value = (108+110+……..+199)/9 = 1308/9=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**

**Ans: 9 a)**

**Speed**

Mean: 15.4

Median: 15

Mode: 20

Mode>Mean>Median: left skewed data (negative skewed data)

Skewness: -0.11751 (left tail, majority of the data lies in the right side of the mean)

Kurtosis: -0.50899 (negative kurtosis which means graph has flat tail, also known as Platy Kurtic; std is more and data points are far more spread about the mean)

**Distance**

Mean: 42.98

Median: 36

Mode: 26

Mean>Median>Mode: right skewed data (positive skewed data)

Skewness: 0.806895 (right tail, majority of data lies in the left side of the mean)

Kurtosis: 0.405053 (positive kurtosis which means graph has long tail, also known as Lapto Kurtic; std is low and less data points spread about the mean)

**9b)**

**SP**

Mean: 121.5403

Median: 118.2087

Mode: 118.289

Mean>Mode>Median: right skewed data (positive skewed data)

Skewness: 1.61145 (right tail, majority of data lies in the left side of the mean)

Kurtosis: 2.977329 (positive kurtosis which means graph has long tail, also known as Lapto Kurtic; std is low and less data points spread about the mean)

**WT**

Mean: 32.41258

Median: 32.73452

Mode: NA (there is no repeated value)

Skewness: -0.61475 (negative skewed data/left skewed data which means graph has left tail, majority of data lies in the right side of the mean)

Kurtosis: 0.950291 (positive kurtosis which means graph has long tail, also known as Lapto Kurtic; std is low and less data points spread about the mean).

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



**Ans:** **Histogram:** right skewed data.

1-2 % values in the range of 350-400 represents outliers

**Boxplot:** top 3 values are 100% outliers. However, values just below them are very dense, so, they should not be considered as outliers else we may lose some important information if we remove them from the dataset considering them as outliers

**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?

Ans:

Sample mean (x̄) = 200

Sample standard deviation(s) = 30

Sample size (n) = 2000

We have to estimate average weight of an adult weight in population of 3,000,000 men at 94%,98%,96% confidence interval

Average weight at Confidence Interval (CI) = x̄ ± z [s/ sqrt(n)]

Where Z: Z-score corresponding to desired confidence level

**94% confidence Interval:**

CL (Confidence Level) = 0.94, A= (1+CL)/2

A (Area under Normal distribution curve)= (1+0.94)/2 = 0.97

Search 0.97 in Normal distribution Table for the corresponding row & column, where we can see 0.97 is very close to 0.9699 & 0.9706

From this table, Z-score of 0.9699 = 1.88

Z-score of 0.9706 = 1.89

Therefore, Z-score of 0.94 = mean of these values = (1.88+1.89)/2 = 1.885

Hence, Z-score at 94% Confidence Interval = 1.885

**Average weight at Confidence Interval (CI) = x̄ ± z [s/ sqrt(n)]**

**= 200 ± 1.885 [30/sqrt(2000)]**

**= 198.73, 201.26**

**98% confidence Interval:**

Z- score at 98% Confidence Interval = 2.325 (using the same procedure)

**Average weight at Confidence Interval (CI) = x̄ ± z [s/ sqrt(n)]**

**= 200 ± 2.325 [30/sqrt(2000)]**

**= 198.44, 201.56**

**96% confidence Interval:**

Z- score at 96% Confidence Interval = 2.055 (using the same procedure)

**Average weight at Confidence Interval (CI) = x̄ ± z [s/ sqrt(n)]**

**= 200 ± 2.055 [30/sqrt(2000)]**

**= 198.62, 201.38**

**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?

**Ans:** 1)

Mean= (34+36+36+38+……………………..+49+56)/18 = 738/18=41

Median= middle most value= (40+41)/2= 40.5

Variance= [(34-41)2 + (36-41)2  + (36-41)2 + (38-41)2 + (38-41)2 + (39-41)2 + (39-41)2 + (40-41)2 + (40-41)2 + (41-41)2 + (41-41)2 + (41-41)2+ (41-41)2+ (42-41)2 + (42-41)2 + (45-41)2 + (49-41)2 + (56-41)2 ] / 18 = 434/18=24.11

Standard deviation = sqrt(24.11)= 4.9101

2.) The above student get majority of marks in the range of (41-4.9 to 41+4.9) i.e. 36.1 to 45.9

We can see out of 18 scores, 13 of his scores lies from 36.1 to 45.9

Mode = 41 so, here Mean = Mode almost equal to median which means graph is very close to normal distribution.

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

**Ans:** There is no skewness i.e. symmetrically distributed

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

**Ans:** right skewness (positive skewness) i.e. majority of data lies in the left side of the mean

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

**Ans:** left skewness (negative skewness) i.e. majority of data lies in the right side of the mean

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

**Ans:** Graph has a Long tail (Lapto Kurtic) which means standard deviation is low i.e. less data spread about the mean

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

**Ans:** graph has flat tail(Platy Kurtic) means standard deviation is more i.e. more data spread about the mean

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

**Ans:** middle 50% of the data lies from 10 to 18 , which is IQR (Inter Quartile Range)

left skewed data

median = 15.5 (approximately)

IQR = 18-10 = 8 (approximately)

Q19) Comment on the below Boxplot visualizations?



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

**Ans:** Boxplot1: middle 50% data(IQR) lies from 250 to 275 (approx.)

Boxplot2: middle 50% data(IQR) lies from 225 to 300 (approx.)

As mean = median in both plots , so, data is symmetrically distributed in both the plots

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)

**Ans:** a. P(E) = n(E)/n(S) = 33/81 = 0.407

b.P(E) = 61/81 = 0.753

c.P(E) = 69/81 = 0.852

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

Ans: **a) MPG:**

Mean= 34.42208, Median= 35.15273, Mode= 29.62994.

Mean ≠ Median ≠ Mode . So, MPG of cars does not follow Normal Distribution

**b.) AT:**

Mean= 101.894, Median= 96.54, Mode = 121

**Waist:**

Mean= 91.90183, Median= 90.8, Mode = 94.5

Mean ≠ Median ≠ Mode . So, AT, Waist do not follow Normal Distribution

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

**Ans:** A = (1+CL)/2 where CL: Confidence Level, A: Area under normal

distribution curve for the given Z-score

**90% confidence interval:**

CL = 0.9

A= (1+0.9)/2 = 0.95

Search 0.95 in Normal distribution Table for the corresponding row & column, where we can see 0.95 is very close to 0.9495 & 0.9505

From this table, Z-score of 0.9495 = 1.64

Z-score of 0.9505 = 1.65

Therefore, Z-score of 0.95 = mean of these values = (1.64+1.65)/2 = 1.645

**Hence, Z-score at 90% Confidence Interval = 1.645**

**94% confidence interval:**

CL = 0.94

A= (1+0.94)/2 = 0.97

Search 0.97 in Normal distribution Table for the corresponding row & column, where we can see 0.97 is very close to 0.9699 & 0.9706

From this table, Z-score of 0.9699 = 1.88

Z-score of 0.9706 = 1.89

Therefore, Z-score of 0.94 = mean of these values = (1.88+1.89)/2 = 1.885

**Hence, Z-score at 94% Confidence Interval = 1.885**

**60% confidence interval:**

CL = 0.6

A= (1+0.6)/2 = 0.8

Search 0.8 in Normal distribution Table for the corresponding row & column, where we can see 0.8 is very close to 0.7995 & 0.8023

From this table, Z-score of 0.7995 = 0.84

Z-score of 0.8023 = 0.85

Therefore, Z-score of 0.8 = mean of these values = (0.84+0.85)/2 = 0.845

**Hence, Z-score at 60% Confidence Interval = 0.845**

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

Ans:

**95% confidence interval:**

Express confidence level as a number i.e. c= 0.95

Significance level α = 1-c = .05

Sample size = 25 i.e. n= 25

Degrees of freedom = n-1 = 24

tα/2  = t.025 ( use t-table) = 2.06390

**96% confidence interval:**

Express confidence level as a number i.e. c= 0.96

Significance level α = 1-c = .04

Sample size = 25 i.e. n= 25

Degrees of freedom = n-1 = 24

tα/2  = t.02 ( use t-table) = 2.49216

**99% confidence interval:**

Express confidence level as a number i.e. c= 0.99

Significance level α = 1-c = .01

Sample size = 25 i.e. n= 25

Degrees of freedom = n-1 = 24

tα/2  = t.005 ( use t-table) = 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

**Ans:** average life of a bulb (μ) = 270 days

Sample size , n = 18

Average life of a bulb from sample i.e sample mean (x̄) = 260 days

Standard deviation of population (σ) = 90 days

If CEO’s claim is true which means Null hypothesis is : A Government company claims that an average light bulb lasts 270 days.

P (X <=260)= ?

Since, sample size is below 30 (small sample size) , therefore, we can use t-distribution table to find the probability of sampled data

Formula of t-statistics: t = [x̄ - μ] / [σ/sqrt(n)]

= [260-270] / [90/sqrt(18)]

= -0.47

Degrees of freedom = n-1 = 17

Refer to t-distribution table, search t= -0.47 with degree of freedom =17,

We get Probability = 0.3192 i.e. 31.92 %