|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continuous data |
| Weight of Gold | Continuous data |
| Distance between two places | Continuous data |
| Length of a leaf | Continuous data |
| Dog's weight | Continuous data |
| Blue Color | Discrete data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Discrete data |

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

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

**Answer**:

When three coins are tossed the total number of possible combinations are 23 = 8.

These combinations are HHH, HHT, HTH, THH, TTH, THT, HTT, TTT.

The number of combinations which have two heads and one tail are:

HHT, HTH, TTH which makes them 3 in number.

therefore the Probability of getting two heads and one tails in the toss of three coins simultaneously is defined as:

P(Two heads and One tail) = Number of desired outcomes = 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

**Answer**:

The set of possible outcomes when we roll a die are {1,2,3,4,5,6}

So, when we roll two dice, there are 6\*6 = 36 possibilities.

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. Since there are no outcomes which correspond to a sum equal to 1

Therefore in this case i.e, equal to 1 is

Probability = No. Of interested events / Total no of events

Probability = 0/36 = 0

Hence, probability is 0.

1. Less than or equal to 4 {(1,1) (1,2) (1,3) (2,1) (2,2) (3,1)}

Probability = No. of interested events / Total no of events

Probability = 6/36 = 1/6 (0.16666)

1. Sum is divisible by 2 and 3 - 6, 12

Therefore sum of 6: (1,5), (2,4), (3,3), (4,2), (5,1)

Sum of 12: (6,6)

There are total 6 possible outcomes, and we rolling two dice

So 6\*6=36 total outcomes.

The probability (sum divisible by 2 and 3) = 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?

**Answer**:

Total number of balls = 2+3+2 =7 balls.

Let S be the sample space.

Then, n(S) = Number of ways of finding 2 balls out of 7

= 7 C2

We know that nCr = n!/r!(n-r)!

= 7\*6\*5!/2!\*5!. Cancelling 5! from the numerator and denominator, we get

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

= 21

So, there are 21ways of picking 2 balls at random from the bag containing 7 balls.

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

Therefore n(E) = Number of ways of drawing 2 balls out of (2+3) balls.

=5C2

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

= 10

Therefore P(E) = n(E)/n(S) = 10/21

Then

The probability that none of the balls being blue = 10/21

Therefore Required Probability = 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

**Answer**:

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**

**Answer:**

Points Score Weigh

Mean: 3.597 Mean: 3.212 Mean: 17.85

Median: 3.695 Median: 3.325 Median: 17.71

Mode: 3.92 Mode: 3.44 Mode: 17.02

Var: 0.286 Var: 0.933 Var: 3.193

Stdev: 0.535 Stdev: 0.966 Stdev: 1.787

Min: 2.760 Min: 1.513 Min: 14.50

Max: 4.930 Max: 5.345 Max: 22.90

Range: 2.170 Range: 3.832 Range: 8.40

The Mean, Median and Mode are showing very close values for all the three features ‘Points’, ‘Score’ and ‘Weigh’

The Data Type for all the three features ‘Points’, ‘Score’ and ‘Weigh’ are Continuous.

|  |  |  |  |
| --- | --- | --- | --- |
| **Points** | **Score** | **Weigh** |  |
| 3.9 | 2.62 | 16.46 |  |
| 3.9 | 2.875 | 17.02 |  |
| 3.85 | 2.32 | 18.61 |  |
| 3.08 | 3.215 | 19.44 |  |
| 3.15 | 3.44 | 17.02 |  |
| 2.76 | 3.46 | 20.22 |  |
| 3.21 | 3.57 | 15.84 |  |
| 3.69 | 3.19 | 20 |  |
| 3.92 | 3.15 | 22.9 |  |
| 3.92 | 3.44 | 18.3 |  |
| 3.92 | 3.44 | 18.9 |  |
| 3.07 | 4.07 | 17.4 |  |
| 3.07 | 3.73 | 17.6 |  |
| 3.07 | 3.78 | 18 |  |
| 2.93 | 5.25 | 17.98 |  |
| 3 | 5.424 | 17.82 |  |
| 3.23 | 5.345 | 17.42 |  |
| 4.08 | 2.2 | 19.47 |  |
| 4.93 | 1.615 | 18.52 |  |
| 4.22 | 1.835 | 19.9 |  |
| 3.7 | 2.465 | 20.01 |  |
| 2.76 | 3.52 | 16.87 |  |
| 3.15 | 3.435 | 17.3 |  |
| 3.73 | 3.84 | 15.41 |  |
| 3.08 | 3.845 | 17.05 |  |
| 4.08 | 1.935 | 18.9 |  |
| 4.43 | 2.14 | 16.7 |  |
| 3.77 | 1.513 | 16.9 |  |
| 4.22 | 3.17 | 14.5 |  |
| 3.62 | 2.77 | 15.5 |  |
| 3.54 | 3.57 | 14.6 |  |
| 4.11 | 2.78 | 18.6 |  |

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?

**Answer:**

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

= 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)**

|  |  |
| --- | --- |
| **Speed** | **Dist** |
| 4 | 2 |
| 4 | 10 |
| 7 | 4 |
| 7 | 22 |
| 8 | 16 |
| 9 | 10 |
| 10 | 18 |
| 10 | 26 |
| 10 | 34 |
| 11 | 17 |
| 11 | 28 |
| 12 | 14 |
| 12 | 20 |
| 12 | 24 |
| 12 | 28 |
| 13 | 26 |
| 13 | 34 |
| 13 | 34 |
| 13 | 46 |
| 14 | 26 |
| 14 | 36 |
| 14 | 60 |
| 14 | 80 |
| 15 | 20 |
| 15 | 26 |
| 15 | 54 |
| 16 | 32 |

**Use Q9\_b.csv**

|  |  |
| --- | --- |
| **Speed** | **Weight** |
| 104.1854 | 28.76206 |
| 105.4613 | 30.46683 |
| 105.4613 | 30.1936 |
| 113.4613 | 30.63211 |
| 104.4613 | 29.88915 |
| 113.1854 | 29.59177 |
| 105.4613 | 30.30848 |
| 102.5985 | 15.84776 |
| 102.5985 | 16.35948 |
| 115.6452 | 30.92015 |
| 111.1854 | 29.36334 |
| 117.5985 | 15.75353 |
| 122.1051 | 32.81359 |
| 111.1854 | 29.37844 |
| 108.1854 | 29.34728 |
| 111.1854 | 29.60453 |
| 114.3693 | 29.53578 |
| 117.5985 | 16.19412 |
| 114.3693 | 29.92939 |
| 118.4729 | 33.51697 |
| 119.1051 | 32.32465 |
| 110.8408 | 34.90821 |
| 120.289 | 32.67583 |
| 113.8291 | 31.83712 |
| 119.1854 | 28.78173 |
| 114.5985 | 16.04317 |
| 120.7605 | 38.06282 |
| 119.1051 | 32.83507 |
| 99.56491 | 34.48321 |
| 121.8408 | 35.54936 |
| 113.4846 | 37.04235 |
| 112.289 | 33.23436 |
| 119.9211 | 31.38004 |
| 121.3926 | 37.57329 |

**Answer:** Skewness for speed = -0.1139548, skewness value is negative so it is left skewed. Since magnitude is slightly greater than 0 it is slightly left skewed.

And for distance = 0.7824835, right skewed (Positive) slight magnitude to right.

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



**Answer:** The most of the data points are concerated in the range 50 -100 with frequency 200.

And least range of weight is 400 somewhere around 0 - 10.

So the expected value the above distribution is 75.

Skewness - We can notice a long tail towards right so it is heavily right skewed.



**Answer:** 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?

**Answer:**

X+/-(Z1-α.σ/sqrt(n)

Degrees of freedom = 2000-1 = 1999

Confidence interval = 94%

(1-σ/2) = 1-0.03) = 0.97

Confidence interval for 94% is 1.882

Confidence interval for 98% is 2.33

Confidence interval for 96% is 2.05.

**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?

**Answer:**

Mean = 41, Median = 40, Variance = 24.111, Standard deviation = 4.910.

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

**Answer:** Symmetrical.

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

**Answer:** Right skewed.

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

**Answer:** Left skewed.

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

**Answer:** The data is normally distributed and kurtosis value is 0.

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

**Answer:** The distribution of the data has lighter tails and a flatter peaks than the normal distribution.

Q18) Answer the below questions using the below box plot visualization.



What can we say about the distribution of the data?

**Answer:** Let’s assume above box plot is about age’s of the students in a school.

50% of the people are above 10 yrs old remaining are less.

And student whose age is above 15 are approx 40%.

What is nature of skewness of the data?

**Answer:** Left skewed, median is greater than mean.

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

**Answer:** Approximately = -8.

Q19) Comment on the below Boxplot visualizations?



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

**Answer:** By observing both the plots whisker’s level is high in boxplot 2, 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)
  2. P(MPG<40)
  3. P (20<MPG<50

**Answer:** By using filter command and installing the dplyr package into the ‘R’.

1. There are 33 observations in MPG which are greater than 38.
2. 61 observations in MPG which are lesser than 40

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

**Answer:**

1. MPG of cars not following the normal distribution.
2. Variable Waist circumference(waist) does not follow normal distribution.

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

**Answer:** Z score of 90% confidence interval is 1.65

Z score of 94% confidence interval is 1.55

Z score of 60% confidence interval is 0.85

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

**Answer:** t score of 95% confidence interval= 1.96

t score of 96% confidence interval= 2.5

t score of 99% confidence interval = 2.47

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

**Answer:** A government company claims that an average light bulb last 270 days. A researcher randomly selects a18 bulbs for testing.

The sample bulbs last an average of 260 days with a standard deviation of 90 days.

t - statistics for the data is given as follows:

t=x-μ/8/√ n

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=260-270/90√ 18

t= -10/90/3√ 2

t=-10/30/√ 2

t=-1\*√ 2/3

Therefore,

t = - 0.47

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that t < - 0.471 with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean life of the bulbs is 300 days.