|  |  |
| --- | --- |
| 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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Ordinal |
| 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?

Outcomes of 1 coin = H, T

Outcomes of 2 coin = HH, TT, TH, HT

Outcomes of 3 coin = HHH, HHT, THH, HTH, TTT, TTH, HTT, THT

Probability that 2 heads and 1 tail are obtained = 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 2and 3

Probability of 2 dice rolled = (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. **Equal to 1**

= 0 (because when we add to dice outcomes no value would be less than 2, as it is the minimum value )

1. **Less than or equal to 4**

= (1,1) , (1,2), (1,3) , (2,1), (2,2), (3,1) = 6

= 6 / 36

= 1/6

= 0.166

1. **Sum is divisible by 2 and 3**

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

= 6/36

= 1/6

= 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?

Number of red balls = 2

Number of green balls = 3

Number of blue balls = 2

Total number of balls = 7

Probability that first ball is not blue = 2/ 7 + 3/7 = 5 /7

This leaves 6 balls with 2 blue. Assuming that second ball is not blue = 4/6

Probability that neither ball drawn is blue = 5/7 \* 4/6

= 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

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)

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.09

It can be say that randomly selected child has the probability of getting candies is 3.

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**

**For data values please look into the file**

**For points:**

1. **Mean**

115.09/ 32 = 6.975

1. **Median =** 3 + 3.23

= 6.23/2

=3.115

1. **Mode =** (3.92) and (3.07) both repeated 3 times.
2. **Variance =**

**=** 374.1052 / 31

= 12.067

1. **Standard deviation** = 3.473
2. **Range =** 4.93 – 2.76

= 2.17

**For score:**

1. **Mean**

102.952/ 32 = 3.217

2. **Median =** 5.424+ 5.345

= 10.76/2

=5.38

1. **Mode =** (3.44) repeated 3 times.
2. **Variance =**

**=** 29.678 / 31

= 0.957

1. **Standard deviation** = 0.978
2. **Range =** 5.424-1.513

= 3.911

**For weigh:**

1. **Mean =** 571.16/ 32

= 17.84

**2**. **Median =** 17.82+17.42

= 35.24/2

=17.62

1. **Mode =** (18.6) repeated 3 times.
2. **Variance =**

**=** 84.1726 / 31

= 2.715

1. **Standard deviation** =1.647
2. **Range =** 22.9 - 15.41

= 7.49

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?

Total no. of patients = 9

One patient chosen at random = 1/9

Expected value of weight = (1/9)\* 108 + (1/9)\* 110 + (1/9)\* 123 + (1/9)\* 134 + (1/9)\* 135 + (1/9)\* 145 + (1/9)\* 167 + (1/9)\* 187 + (1/9)\* 199

= 12 + 12.22 + 13.66 + 14.88 + 15 + 16.11+ 18.55 + 20.77 +22.11

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**import** numpy **as** np

**import** pandas **as** pd

data**=**pd**.**read\_csv("C:/Users/purti/Downloads/Q9\_a.csv")

Q9\_a**.**skew ()

Index 0.000000

speed -0.117510

dist 0.806895

dtype: float64

Q9\_a**.**kurtosis ()

Index -1.200000

speed -0.508994

dist 0.405053

dtype: float64

**SP and Weight (WT)**

**Use Q9\_b.csv**

Q9\_b **=** pd**.**read\_csv("C:/Users/purti/Downloads /Q9\_b.csv”)

Q9\_b**.**skew ()

Unnamed: 0 0.000000

SP 1.611450

WT -0.614753

dtype: float64

Q9\_b**.**kurtosis ()

Unnamed: 0 -1.200000

SP 2.977329

WT 0.950291

dtype: float64

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



|  |  |
| --- | --- |
| **Chicks weight** | **Frequency** |
| 0 – 50 | 75 |
| 50 - 100 | 200 |
| 100 - 150 | 125 |
| 150 - 200 | 100 |
| 200 - 250 | 50 |
| 250 – 300 | 25 |
| 300 – 350 | Less than 25 approx 12.5 |
| 350 – 400 | Outliers |

The histogram shows that with the increasing weight of the chicks, the frequency is decreasing and led to outliers. Only the maximum frequency is 200, which shows that the ideal weight for chicks is between 50 to 100 to produce maximum growth. And the data is right skewed or positively distributed.



The boxplot shows that the data is right skewed and the outliers are at upper side.

**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?

Sample meanof x = 200

Sample standard deviation of s = 30

Sample **size**of n = 2000

**Confidence interval 94 %** = 200-1

T = **1.8916**

**= 200 – 1.8916 \*(300/√2000) = 198.73**

**= 200 + 1.8916 \*(300/√2000) = 201.27**

**Confidence interval 96 %** = 200-1

T = **2.0673**

**= 200 – 2.0673\*(300/√2000) = 198.61**

**= 200 + 2.0673\*(300/√2000) = 201.39**

**Confidence interval 98 %** = 200-1

T = **2.3452**

**= 200 –2.3452 \*(300/√2000) = 198.43**

**= 200 + 2.3452\*(300/√2000) = 201.57**

**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?

**Mean =** 738/18

= 41

**Median =** 40+41

= 81/2

= 40.5

**Variance =** 434/17

= 25.52

**Standard deviation =** 5.05

The average no. of marks ( mean) gained by students is 41 which is slightly more than median . Most students get marks between 41 – 42 and there are two outliers as 49 , 56.

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

The skewness is zero , when mean and median of data are equal .

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

If the mean is greater than median, the distribution is positively skewed.

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

If the median is more then mean then the data is skewed to the left.

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

The positive values of kurtosis indicate that the graph obtained shows that the distribution is at peak and possess thick tails.

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

The negative values of kurtosis indicate that the graph obtained shows that the distribution is flat and possess thin tails.

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



What can we say about the distribution of the data?

Data is not symmetrically distributed as the median is not divided equally half. Its towards the higher value.

What is nature of skewness of the data?

The data is skewed towards left.

What will be the IQR of the data (approximately)?   
  
Upper quartile range = 18

Lower quartile range = 10

IQR = 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.

In both the boxplots there are no outliers. Secondly, both boxplots shows or lies within the same median i.e 250 – 275 hence the are normally distributed with zero skewness.

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)

**import** pandas **as** pd

**import** numpy **as** np

**from** scipy **import** stats

**from** scipy.stats **import** norm

cars**=**pd**.**read\_csv("C:/Users/purti/Downloads /Cars.csv”)

cars

| **HP** | **MPG** | **VOL** | **SP** | **WT** |
| --- | --- | --- | --- | --- |
| **0** | 49 | 53.700681 | 89 | 104.185353 | 28.762059 |
| **1** | 55 | 50.013401 | 92 | 105.461264 | 30.466833 |
| **2** | 55 | 50.013401 | 92 | 105.461264 | 30.193597 |
| **3** | 70 | 45.696322 | 92 | 113.461264 | 30.632114 |
| **4** | 53 | 50.504232 | 92 | 104.461264 | 29.889149 |
| **...** | ... | ... | ... | ... | ... |
| **76** | 322 | 36.900000 | 50 | 169.598513 | 16.132947 |
| **77** | 238 | 19.197888 | 115 | 150.576579 | 37.923113 |
| **78** | 263 | 34.000000 | 50 | 151.598513 | 15.769625 |
| **79** | 295 | 19.833733 | 119 | 167.944460 | 39.423099 |
| **80** | 236 | 12.101263 | 107 | 139.840817 | 34.948615 |

81 rows × 5 columns

**P(MPG>38)**

1**-**stats**.**norm**.**cdf(38,cars**.**MPG**.**mean(),cars**.**MPG**.**std())

0.3475939251582705

**P(MPG<40)**

stats**.**norm**.**cdf(40,cars**.**MPG**.**mean(),cars**.**MPG**.**std())

0.7293498762151616

**P (20<MPG<50)**

stats**.**norm**.**cdf(0.50,cars**.**MPG**.**mean(),cars**.**MPG**.**std())**-**stats**.**norm**.**cdf(0.20,cars**.**MPG**.**mean(),cars**.**MPG**.**std())

1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Yes, it follows 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

Yes, it follows normal distribution.

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

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

To compute the 95% confidence interval, start by computing the mean and standard error: M = (2 + 3 + 5 + 6 + 9)/5 = 5.

σM = = 1.118. Z.95

The shaded area is 0.95

Confidence Level z

0.90 1.645

0.92 1.75

0.95 1.96

0.96 2.05

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

t = [ x - μ ] / [ s / sqrt( n ) ]

p <- pt((260-270)/(90/sqrt(18)), 18)

[1] 0.3215076