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

Q1) Identify the Data type for the Following: Categorical Numerical

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

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

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

**Ans :-** = P (HHT) + P (HTH) + P (THH) =

(1/2\*1/2\*1/2) + (1/2\*1/2\*1/2\*) + (1/2\*1/2\*1/2\*) = 3/8 = 0.375

Q4) Two Dice are rolled, find the probability that sum is

**Ans :-**

1. Equal to 1 = (Total favorable outcomes / Total possible outcomes) = 0/36 = 0
2. Less than or equal to 4 = possible outcomes = (1,3), (2,2), (3,1). So, = 3/36 = 1/12=0.0833
3. Sum is divisible by 2 and 3

= Possible outcomes = (1,5), (2,4), (3,3), (4,2), (5,1) = 5/36 = 0.1388

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 :-** Possible outcome = (n!/r!(n-r)!) = 5! / 2! (5-2)! = 10

Total Possible outcome = (n! / r! ( n-r)!) = 7! / 2! (7-2)! = 21

Probability = 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 |

**Ans :-**

Child A = 0.015 \* 1 = 0.015

Child B = 0.20 \* 4 = 0.80

Child C = 0.65 \* 3 = 1.95

Child D = 0.005 \* 5 = 0.025

Child E = 0.01 \* 6 = 0.06

Child F = 0.120 \* 2 = 0.240

Mean = (0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.240) / 6 = 0.515

Expected value is 0.515

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:- Mean**= ( of all numerical data) / total no of numerical data

**Median** =if odd then, →[(n + 1)/2]th term, where (n=mid value of data set)

If even then, → [(n/2)th term + ((n/2) + 1)th term]/2**,** where (n=mid values of data set)

**Mode** = Number that appears most frequently in a data set is a Mode

**Variance** = ( (Xi−¯X)^2) / N

**Std. Div.** = √ (( (Xi−¯X)^2) / N)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | 3.60 | 3.22 | 17.85 |
| **Median** | 3.70 | 3.33 | 17.71 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Variance** | 0.28 | 0.96 | 3.19 |
| **Std. Deviation** | 0.53 | 0.98 | 1.78 |
| **Range** | 2.76-4.93 | 1.51-5.42 | 14.50-22.9 |

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:-**

|  |  |
| --- | --- |
| **Weight** | **Probability (Patient)** |
| 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 | 1/9 |

Expected value is =  (1/9)(108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+ (1/9)(145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199) = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:-**

|  |  |  |
| --- | --- | --- |
|  | **Speed** | **Distance** |
| **Skewness** | -0.12 | 0.81 |
| **Kurtosis** | 0.51 | 0.41 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans:-**

|  |  |  |
| --- | --- | --- |
|  | **SP** | **Weight** |
| **Skewness** | 1.61 | -0.61 |
| **Kurtosis** | 2.98 | 0.95 |

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



**Ans:-**

* Chick weight data is right skewed or positively skewed. ---- Yes
* Most of the chick weight is between 50 to 100. --- Yes
* More than 50% Chick Weight is between 50 to 150. ---- Yes



**Ans:-**

* The data is right skewed.
* There are some outliers at the 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?**

**Ans:-** Population count =3000000 men, Sample count = 2000 men

Sample mean = 200 pounds, Sample std. dev. = 30 pounds

Here we have to apply t-test in python so that we can get exact range ---

1. *conf\_94 = stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000))*
2. *print(np.round(conf\_94,0))*
3. *print(conf\_94)*

**--- For 94% confidence interval Range is [ 198.73 – 201.26]**

1. *conf\_96 = stats.t.interval(alpha = 0.96, df=1999, loc=200, scale=30/np.sqrt(2000))*
2. *print(np.round(conf\_96,0))*
3. *print(conf\_96)*

**---For 96% confidence interval range is [198.62 – 201.37]**

1. *conf\_98 = stats.t.interval(alpha = 0.98, df=1999, loc=200, scale=30/np.sqrt(2000))*
2. *print(np.round(conf\_98,0))*
3. *print(conf\_98)*

**--- For 98% confidence interval range is [198.43 – 201.56]**

**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** | **Median** | **Variance** | **Std. Dev.** |
| 41 | 40.5 | 25.53 | 5.05 |

**(2) :-** We can say that the average marks of the student is 41 and there are two outliers (56, 49) in this data. The data is slightly skewed towards right because mean is greater than median.

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

**Ans :-** When mean and median are equal then distribution has zero skewness. We can also say that the distribution is symmetric.

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

**Ans :-** When mean > median then the distribution is positively skewed.

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

**Ans :-** When mean < median then the distribution is negatively skewed.

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

**Ans :-** Positive value of kurtosis indicates that the distribution is peaked and having short tails. This type of distribution is also called leptokurtic.

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

**Ans :-** Positive value of kurtosis indicates that the distribution is peaked and having short tails. This type of distribution is also called leptokurtic.

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



**What can we say about the distribution of the data?**

**Ans :-** The average of data is approx. 15 and it is negatively skewed.

**What is nature of skewness of the data?**

**Ans :-** Data is negatively skewed

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

**Ans :-**  Quartile 1 = 10, Quartile 3 = 18  
so, IQR = Quartile 3 - Quartile 1 = 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.

**Ans :-**

* There are no outliers.
* Both the box plot shares the same median that is approximately in a range between 275 to 250
* All the data points are normally distributed.

**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.)** Total counts of MPG data = **81**,

So, P(MPG>38) = **0.35**

**(b.)** Total counts of MPG data = **81**,

So, P(MPG<40) = **0.72**

**(c.)** Total counts of MPG data = **81**,

So, P(20<MPG<50) = **0.013**

**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 of cars follows normal distribution

**(b)**  Adipose Tissue (AT) and Waist does not follow Normal Distribution

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

**Ans :-**

* If confidence interval is 90% then the significance value is 0.10, for one tail it is 0.05, so the z-score is -1.64.
* If confidence interval is 94% then the significance value is 0.06, for one tail it is 0.03, so the z-score is -1.88.
* If confidence interval is 60% then the significance value is 0.40, for one tail it is 0.20, so the z-score is -0.84.

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

**Ans :-** For exact t-score we have to write a code in python –

1. print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))
2. print('T score for 94% Confidence Interval =',np.round(stats.t.ppf(0.03,df=24),4))
3. print('T score for 99% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))

* **If confidence interval is 95% then the significance value is 0.05, for one tail it is 0.025, so the z-score is -2.064.**
* **If confidence interval is 94% then the significance value is 0.06, for one tail it is 0.03, so the z-score is -1.974.**
* **If confidence interval is 99% then the significance value is 0.01, for one tail it is 0.005, so the z-score is -2.7969.**

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 :- :-** Population mean =270, Sample count = 18

Sample mean = 260, Sample std. dev. = 90

(size) x - population mean) / (sample standard daviation / square root of sample

= (260-270) / (90/√18)

= -0.4714

Now, now we have to write a code for the probability of -0.4714 t-score

1. *import numpy as np*
2. *Import scipy as stats*
3. *stats.t.cdf(t\_score, df = 17)*

So, the probability is 0.32 it means 32%.