|  |  |
| --- | --- |
| 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) | Interval |
| 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 | Ratio |

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

* Sample space = [HHH,HHT,HTH,HTT,THH,THT,TTH,TTT]

P(obtaining two head and one tail) = 3/8

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

1. Equal to 1

* P(sum = 1) = 0/36 = 0 # lowest possible sum is 2

1. Less than or equal to 4

* P(sum <= 4) = 6/36 = 1/6 # 6 samples have sum <= 4

1. Sum is divisible by 2 and 3

* P(sum divisible by 2 and 3) = 6/36 = 1/6 # 6 samples have sum divisible by 2 and 3

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?

* NumOfBalls = 7

P(first ball drawn is not blue) = 1 – P(first ball drawn is blue)

= 1 – (2/7)

= 5/7

P(second ball drawn is also not blue) = 1 – P(second ball drawn is blue)

= 1- (2/6)

= 2/3

P(none of the balls drawn is blue) =

P(first ball drawn is not blue)\* P(second ball drawn is also not blue)

= (5/7)\*(2/3)

= 10/21

probability that none of the balls drawn is blue is 10/21 or ~0.48

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

* E(num of candies for randomly selected child) =

1\*0.015 + 4\*0.2 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.12 = 3.09

Expected number of candies for a randomly selected child ~ 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**

* **Refer ‘A1] solution.ipynb’ for calculations .**
* **1 > Score column have mode > median > mean , implies**

**Distribution might be left skewed**

**2 > span of distribution is largest for Weigh column.**

**3 > Points column data is distributed closer to its mean.**

**4 > In case of Weigh data is distributed far from its mean .**

**5 > Weigh column have mode < median < mean, implies**

**Distribution might be right skewed**

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?

* E(weight of patient chosen at random) = E(X)

= (sum of weights of all patients)/(total number of patients)

= (108+110+123+134+135+145+167+187+199)/9 = 145.33 pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

* **Refer ‘A1] solution.ipynb’ for calculations .**
* **1 > For speed column its skewness is between**

**-0.5 to +0.5 , implies its distribution is normal.**

**Ie, its mode , median ,mean are almost equal.**

**2 > In case of dist column its skewness is > 0.5,**

**Implies its distribution in right skewed. Ie,**

**Dist column have higher values in its data.**

**Also its mode < median < mean**

**3 >both speed and dist columns have kurtosis close to 0,**

**Implies distributions are mesokurtic distributions. And are**

**Normal distribution.**

**SP and Weight(WT)**

**Use Q9\_b.csv**

* **Refer ‘A1] solution.ipynb’ for calculations .**
* **1 > For SP column its skewness is > 0.5, implies its**

**Distribution is right skewed .ie, its mode<median<mean.**

**2 > In case of WT column its skewness is < -0.5, implies its**

**Distribution is left skewed. Ie, its mode>median>mean.**

**3 > For SP column its kurtosis < 3 means it have 0 excess kurtosis,**

**Implies distribution is mesokurtic. And have normal distribution.**

**4 > In case of WT its kurtosis < 3 means it have 0 excess kurtosis ,**

**Implies distribution is mesokurtic. And have normal distribution.**

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



* 1 > distribution have mean<median<mode, implies it is right skewed.

2 > lower outliers are present in distribution

3 > IQR is less suggesting more data points are distributed closer to mean.

**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?

* Refer ‘A1] solution.ipynb’ for calculation.

**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.

* Refer ‘A1] solution.ipynb’ for calculation.

1. What can we say about the student marks?

* 1 > Expected score of any student chosen randomly is 41

2 > since standard deviation is 4.91 , it can be inferred that

68% students have obtained scores between [36 , 46]

95% students have obtained scores between [31, 51]

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

* If mean = median then it generally gives symmetrical distribution

with ZERO skewness

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

* If mean > median then it generally gives assymmetrical distribution

with POSITIVE skewness

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

* If mean < median then it generally gives assymetrical distribution with

NEGATIVE skewness

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

* 1> majority of data is distributed near mean ,ie distribution

Have sharp peak

2> data may have more outliers because of heavy tails

Than normal distribution

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

* 1> majority of data is not concentrated near mean ,ie

Distribution have flat peak

2> data may have fewer outliers because of thin tails

Than normal distribution

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



What can we say about the distribution of the data?

* 1> data is not normally distributed.

What is nature of skewness of the data?

* 1> data distributed in LEFT (NEGATIVELY) skewed

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

* IQR will be approximately equal to (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.

* 1> both distribution don’t have ouliers

2> both are normal distributions

3> distribution 1 (might have ) have higher kurtosis than distribution 2

4> variance of distribution 1 is less than distribution 2

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)

* Refer ‘A1] solution.ipynb’ for calculations.

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

* Refer ‘A1] solution.ipynb’ for calculations

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

Dataset: wc-at.csv

* Refer ‘A1] solution.ipynb’ for calculations

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

* Refer ‘A1] solution.ipynb’ for calculations

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

* Refer ‘A1] solution.ipynb’ for calculations

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

* Refer ‘A1] solution.ipynb’ for calculations