|  |  |
| --- | --- |
| 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 | Ratio |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ratio |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Ordinal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Ratio |
| Number of Children | Interval |
| 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 - Sample pace = [HHH, HHT, HTH, THH, TTH, THT, HTT, TTT]**

**Total number of ways = 2\*2\*2=8**

**Fav Cases = 3**

**P(A) = 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

**Ans – Sample Space = [(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. The sum is equal to 1 is zero because they starts with (1,1), likewise other than in the dice we are not having zero.

**P(A) = 0**

1. **Less than or equal to 4 = (1,1), (1,2), (1,3), (2,1), (2,2), (3,1)**

**P(B) = 6/36 = 0.166**

1. **Sum is divisible by 2 and 3 = (1,5), (2,4), (3,3), (4,2), (5,1), (6,6)**

**P(C) = 6/36 = 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?

**Ans.)**

**Total number of possible ways of selecting 2 balls out of 7 = 7C2**

**We know that we must find the probability that 2 balls are selected and none of them are blue.**

**So, after excl–uding blue balls out of total balls we are left with = 2 Red + 3 Green = 5**

**Therefore selecting 2 balls from the above set of 5 balls.**

**Total possible ways of doing so = 5C2**

**Probability = P (None of 2 balls are blue) = 5C2 / 7C2 = 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.)

**Expected number of candies a randomly selected child will get**

**= 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**

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 Comment about the values/ Draw some inferences.

**Use Q7.csv file**

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: Total number of possible ways of selecting 1 patient out of 9 = 9C1**

**Now, we are taking one patient randomly = 1**

**So, Probability = P (One patient at random)**

**= 1/9C1**

**= 1/9**

**= 0.111**

**Expected Weight of the person selected will be**

**= 108 \* 0.111 + 110 \* 0.111 + 123 \* 0.111 +134 \* 0.111 +**

**135 \* 0.111 + 145 \* 0.111 + 167 \* 0.111 + 187 \* 0.111 +**

**199 \* 0.111**

**= 11.988 + 12.221 + 13.653 + 14.874 + 14.985 + 16.095 + 18.537 +**

**20.757 + 22.089**

= **145.99**

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

**Car’s speed and distance**

**Use Q9\_a.csv**

**SP and Weight (WT)**

**Use Q9\_b.csv**

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



**Inference about Histogram: The maximum value is lying in the range between 50 – 100**

**The minimum value is lying in the range between 300-400**

**The trend in the graph shows that, as we move further the**

**Frequency starts decreasing.**



**Inference about Box Plot: There are outliers present in the boxplot.**

**The data is not equally distributed.**

**As, the numerical values is not given, not more**

**inferences can be made.**

**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 Size - 2000**

**Population Size = 3000000**

**Mean = 200**

**SD = 30**

**CI = 94%**

**Step 1.) CHECKING WHETHER DATA IS NORMAL OR NOT?**

**-----------------------------------------------------**

**Sample Size > SD ---> DATA IS NORMAL**

**2000 > 30 ----> The given data is normal.**

**STEP 2.) CALCULATE STANDARD ERROR**

**----------------------------------------**

**Standard Error = SD / sqrt (Sample Size)**

**= 30 / sqrt (2000)**

**= 30/44.72**

**= 0.670**

**Alpha = 1 - CI**

**= 1 - 94%**

**= 1 - 0.94**

**= 0.06**

**Cumulative Probability = 1 - Alpha/2**

**= 1 - 0.06/2**

**= 1 - 0.03**

**= 0.97**

**Degree of Freedom = Sample Size - 1 = 1999**

**t-Score / Critical value = 1.89 (Lookup a table that is t Distribution Table - Degree of Freedom & Cumulative Probability)**

**Margin of Error = Critical Value \* Standard Error = 1.89 \* 0.670 = 1.2663**

**94% Confidence Interval --> 200-1.2663 to 200+1.2663 ---> 198.7 to 201.26**

**For 98 % Confidence Interval**

**Alpha = 1 – CI**

**= 1 – 98%**

**= 1 – 0.98**

**= 0.02**

**C.P. = 1 – Alpha/2**

**= 1 – 0.02/2**

**= 0.99**

**D.O.F = 1999**

**t-score/Critical Value = 2.32812**

**Margin Of Error = 2.3812 \* 0.670**

**= 1.595**

**98% Confidence Interval 🡪 200-1.595 to 200 + 1.595 🡪198.405 to 201.595**

**For 96% Confidence Interval**

**Alpha = 1 – CI**

**= 1 – 96%**

**= 1 – 0.96**

**= 0.04**

**C.P. = 1 – Alpha/2**

**= 1 – 0.04/2**

**= 1 – 0.02**

**= 0.98**

**D.O.F. = 1999**

**tScore = 2.05469**

**Margin Of Error = 2.05469 \* 0.670**

**= 1.3766**

**96 % Confidence Interval 🡪 200 – 1.3766 to 200 + 1.3766 🡪198.6234 to 201.3766**

**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: Q1) Mean X = (Sum of all the observations/Total number of observations)**

**= 34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56/18**

**= 738/18**

**= 41**

**Median: Here n is even so,**

**Med = {(n/2) + (n/2+1)/2}th**

**= {18/2 + (18/2 +1)/2}th**

**= {(9 + 10) / 2}th**

**= {(40 +41)/2}**

**= 40.5**

**For Standard Deviation,**

**Mean = 41**

**(Xn - X) = 34 – 41 = -7**

**= 36 - 41 = -5**

**= 36 – 41 = -5**

**= 38 – 41 = -3**

**= 38 - 41 = -3**

**= 39 – 41 = -2**

**= 39 – 41 = -2**

**= 40 – 41 = -1**

**= 40 – 41 = -1**

**= 41 – 41 = 0**

**= 41 – 41 = 0**

**= 41 – 41 = 0**

**= 41 – 41 = 0**

**= 42 – 41 = 1**

**= 42 – 41 = 1**

**= 45 – 41 = 4**

**= 49 – 41 = 8**

**= 56 – 41 = 15**

∑ **(Xn - X)2 = 434**

**S.D. = sqrt (**∑ **(Xn - X)2 / (n-1))**

**= sqrt (434/ (18 – 1))**

**= sqrt (434/ (17))**

**= sqrt (25.529)**

**= 5.0526**

**Variance = (5.0526)2**

**= 25.5288**

**Q2) The average marks scored by the student is 41.**

**The standard deviation is high as the difference is high from mean.**

**The middle value of the scores obtained by students is 40.5**

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

**Ans: Nature of skewness: - When the value of mean, median of data are equal then there will be no skewness i.e., the data is symmetrical.**

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

**Ans: When mean>median, the nature of skewness is positively skewed.**

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

**Ans: When median>mean, the nature of skewness is negatively skewed**.

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

**Ans: Positive values of kurtosis indicate that distribution is peaked and possesses thick tails**.

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

**Ans: Negative excess values of kurtosis (<3) indicate that a distribution is flat and has thin tails. Platykurtic distributions have negative kurtosis values**.

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



What can we say about the distribution of the data?

**Ans: The distribution of the data is not symmetrical.**

What is nature of skewness of the data?

**Ans: The nature of skewness of the data is negatively skewed as the whisker and half-box are longer on the left side of the median than on the right side.**

What will be the IQR of the data (approximately)?   
  
**Ans: IQR = Q3 – Q1**

**= 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: The Boxplot 1 and Boxplot 2 has the symmetrical distribution.**

**The median of the Boxplot 1 and Boxplot 2 is nearly equal.**

**The range of the Boxplot 2 is much higher than the Boxplot 1.**

**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)**

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

Q 22) Calculate the Z scores of 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

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