|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discreate |
| Results of rolling a dice | Discreate |
| 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 | Discreate |
| Number of kids | Discreate |
| Number of tickets in Indian railways | Discreate |
| Number of times married | Discreate |
| Gender (Male or Female) | Discreate |

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 | Nominal |
| 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 | Interval |
| Years of Education | Ratio |

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

n(S) = {(HHH),(HHT),(HTH),(THH),(TTH),(HTT),(THT),(TTT)}

n(two heads and one tail) = 3

p(two heads and one tail) = 3/8

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

Possible outcomes = {(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) }

n(S) = Total number of events = 36

1. n(A) = 0

Hence p(A) = 0

1. n(A) = 6

p(A) = n(A)/n(S)

= 6/36

= 1/6

1. n(A) = 6

p(A) = n(A)/n(S)

= 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?

n(S) = Total ways of drawing two balls from 7 balls = 7C2 = 21

n(A) = Total ways of drawing two balls from 2 red and 3 green balls = 5C2 = 10

p(A) = 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 value = ∑ p(Xi) \* Xi

= (0.015\*1) + (0.20 \* 4) + (0.65 \* 3) + (0.005 \* 5) + (0.01 \* 6) + (0.120 \* 2)

= 3.09

The expected number of candies for randomly selected child is 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.

**Ans:** Mean = sum of values/ total number of values

Median = if n is odd, n+1/2 th term

If n is even , (n/2)+((n/2)+1)/2 th term

Mode = Value that appeared most in the dataset or which have high frequency

Variance = 1/(n-1) \* (xi – x)2

Where xi = each data point

x = mean

n = number of records

Standard Deviation = square root of variance

Range = Maximum Value – Minimum Value

**Points**

Mean = 3.596

Median = 3.695

Mode = 3.07 and 3.92

Variance = 0.2809

Standard Deviation = 0.53

Range = 4.93 – 2.76 =2.17

**Inference** = 1) The average number of points given to 32 cars lies around 3.5 and there are no outliers present in this column as median is almost equal to mean.

2)Since mean, median and mode are almost equal, the distribution of this column will be normal.

3)All the points for 32 cars are very close to mean as standard deviation is very small.

4)The graph for this column will have two peaks as it have two modes

**Score**

Mean = 3.596

Median = 3.3

Mode = 3.440

Variance = 0.9604

Standard Deviation = 0.98

Range = 5.424 – 1.513 = 3.911

**Inference**: 1) Theaverage number of scores scored by each cars lie around 3.5

2) There are no extreme values present in this column of scores as mean and median are almost similar

3)Scores of 32 cars are clustered around mean as the standard deviation is very small.

**Weigh**

Mean = 17.85

Median = 17.71

Mode = 18.90 and 17.02

Variance = 3.193

Standard Deviation = 1.787

Range = 22.9 – 14.5 = 8.4

**Inference**: 1)The average weight of the cars is about 17.85.

2) There is no outlier or extreme value in this column as mean and median are almost similar

3)All the data points are clustered around mean as standard deviation is very small

**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:** Expected Value = (108+110+123+134+135+145+167+187+199)/9

= 145.33 pounds

So, the expected value of weight is 145 pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

Ans: Skewness for Q9\_a.csv

* **Speed:** The distribution of data for this column is approximately symmetric as it is lying between (-0.5,0.5) and mean is approximately equal to median.
* **Distance:** The distribution of data for this column is moderately skewed as value is between (0.5,1).Also it is right skewed as mean is greater than median.

Kurtosis for Q9\_a.csv

In both the cases of **speed** and **distance,** kurtosis is smaller than 3. This means that the datasets have lighter tails than the normal distribution and can be considered to have a negative kurtosis. In this case, there will be no chances of presence of outliers.

Skewness for Q9\_b.csv

* **SP:** The distribution of data for this column is highly skewed as it is above 1. Also, it is right skewed as mean is greater than median.
* **Weight:** The distribution of data is approximately symmetric or following normal distribution as mean and median are approximately equal.

Kurtosis for Q9\_b.csv

In both the cases of **SP** and **weight,** kurtosis is smaller than 3. This means that the datasets have lighter tails than the normal distribution and can be considered to have a negative kurtosis. In this case, there will be no chances of presence of outliers.

**SP and Weight(WT)**

**Use Q9\_b.csv**

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



**Ans:10.a**

1. The distribution of data is right skewed or positively skewed and most of the data lie on the left side of the graph. Also, it’s a unimodal distribution.
2. Majority of chicks are having weight less than 200
3. The frequency of chicks having weight between 50 to 100 is maximum
4. The mean in this case of distribution will be greater than the median



**Ans:10.b** 1. Theboxplot is skewed right and most of the data is concentrated in the lower quartile.

2. There are seven outliers present in the data records which can affect the mean. Hence, a correct way to measure the central tendency will be median.

**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 11:**

For 94% = (134.85077199775026, 265.14922800224974)

For 98% = (122.6512694143423, 277.3487305856577)

For 96% = (130.15355671679083, 269.84644328320917)

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

Mean = 41

Median = 40.5

Variance = 1/(n-1) \* (xi – x)2

= 156/17

= 9.17

S.D = square root of variance

= (9.17) 1/2

= 3.02

**2)**The average marks obtained by student in tests is 41 and we can say that student is maintaining his/her marks around 41 as deviation is very less.

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

**Ans**: When mean and median are equal, then there is no skewness and there will be symmetric distribution or normal distribution.

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

**Ans:** Whenmean> median, then the distribution will be right skewed or positively skewed.

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

**Ans:** When median > mean, then distribution will be left skewed or negatively skewed.

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

**Ans:** Kurtosis is the measure use to describe the degree to which the data cluster in the tails. It is actually measure of outliers present in the distribution.

Positive kurtosis indicates data possess thick or heavy tails. Due to this, it has chances of having outliers.

**Positive kurtosis = means distribution is larger = means tails are fatter = means peak is higher**

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

**Ans:** Negative kurtosis indicates that the distribution of data is flat and has thin tails. There is no chance of presence of outliers.

**Negative kurtosis = means distribution is shorter = means tails are thinner = means peak is lower**

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



What can we say about the distribution of the data?

**Ans:** We can say that the distribution of data is not symmetrical. In other words, most of the data lies on the right side or in the upper quartile.

What is nature of skewness of the data?

**Ans: Nature:** Data is left skewed or negatively skewed.

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

**Ans:** IQR =Q3 – Q1

= 18 –10

= 8Q19) Comment on the below Boxplot visualizations?



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

**Ans:** 1.Both the box plots have uniform distribution and same medians.

2. The IQR for boxplot 1 is lower than the IQR for boxplot 2. Hence, the data points in boxplot 1 are consistently hovered around the central value than in boxplot 2.

3.The range of data of boxplot 1 is smaller than the range of data of boxplot 2. Hence, the overall data is widely distributed in boxplot 2 as compared to boxplot 1.

4.There are no outliers present in both the boxplots.

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)

**Ans: 0.32**

* 1. P(MPG<40)

**Ans: 0.74**

* 1. P (20<MPG<50)

**Ans: 0.90**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:** 1. It is not following normal distribution as the mean, median and mode for each of the column in data is not equal.

2.Also the graphical distribution of each column proves that either it is left skewed or right skewed and does not follow 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

**Ans:** The data is not following normal distribution as the mean, median and mode are not equal for the columns in data.

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

**Ans:**

|  |  |
| --- | --- |
| **C.I** | **Z-score** |
| 90% | 1.644 |
| 94% | 1.88 |
| 60% | 0.841 |

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

**Ans:**

|  |  |
| --- | --- |
| **C.I** | **t-score** |
| 95% | 2.06 |
| 96% | 2.171 |
| 99% | 2.79 |

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) p

df 🡪 degrees of freedom

**Ans: t = difference in sample mean/(s.d of sample/√n)**

**(Find the attached script )**

The probabilitythat 18 randomly selected bulbs would have an average life of no more than 260 days is 0.32