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

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 data – dichotomous |
| High School Class Ranking | ordinal data |
| Celsius Temperature | ratio data |
| Weight | RATIO data |
| Hair Color | nominal data |
| Socioeconomic Status | ordinal data |
| Fahrenheit Temperature | ratio data |
| Height | RATIO data |
| Type of living accommodation | nominal data |
| Level of Agreement | ORDINAL data |
| IQ(Intelligence Scale) | INTERVAL data |
| Sales Figures | nominal data |
| Blood Group | nominal data |
| Time Of Day | INTERVAL data |
| Time on a Clock with Hands | INTERVAL data |
| Number of Children | nominal data |
| Religious Preference | nominal data |
| Barometer Pressure | RATIO data |
| SAT Scores | INTERVAL data |
| Years of Education | RATIO data |

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

Number of possibilities of 3 coins getting tossed = 2^3 = 8

Samples = HHH

HHT

HTH

HTT

THH

THT

TTH

TTT probability = 3/8 = 0.375

Q4) Two Dices 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

Tot sample = 36

Prob pf Sum Equal to 1 = 0

Prob of sum Less than or equal to 4 = 1,1 , 1,2 , 1,3 ,

2,1 , 2,2 ,

3,1 = 6/36 = 1/6

Prob of sum is divisible by 2 and 3 = (1,1), (1,2), (1,3), (2,1), (2,2), (3,1). = 5/36

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?

Total number of balls in the bag = 2 (red) + 3 (green) + 2 (blue) = 7 balls

Total number of ways to draw two balls out of 7 = (7/2)(2/7​) = 7!2!(7−2)!2!(7−2)!7!​ = 7×62×12×17×6​ = 21

the number of ways to draw two balls that are not blue. Since there are 5 balls that are not blue (2 red and 3 green):

Number of ways to draw two balls that are not blue = (5/2)(2/5​) = 5!2!(5−2)!2!(5−2)!5!​ = 5×42×12×15×4​ = 10

Probability = Number of ways to draw two balls that are not blue / Total number of ways to draw two balls

Probability = 10/21

So, the probability that none of the balls drawn is blue is 10/21​.

Top of Form

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

To calculate the expected number of candies for a randomly selected child, we need to multiply each possible candy count by its corresponding probability and then sum up these values.

Let's denote the candy counts as \( X \) and their probabilities as \( P(X) \).

The expected number of candies (\( E[X] \)) is given by the formula:

\[ E[X] = \sum\_{i} X\_i \cdot P(X\_i) \]

Where \( X\_i \) represents each possible candy count and \( P(X\_i) \) represents the probability of each candy count.

Given:

P(1) = 0.015

P(2) = 0.120

P(3) = 0.65

P(4) = 0.20

P(5) = 0.005

P(6) = 0.01

We can calculate the expected number of candies as follows:

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

[ E[X] = 0.015 + 0.24 + 1.95 + 0.80 + 0.025 + 0.06 ]

sum E[X] = 0.015 + 0.24 + 1.95 + 0.80 + 0.025 + 0.06

sum[ E[X] = 3.14

So, the expected number of candies for a randomly selected child is 3.14.

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**

|  |  |  |
| --- | --- | --- |
| **avg** | **3.21725** | **17.84875** |
| **median** | **3.325** | **17.71** |
| **mode** | **3.44** | **17.02** |
| **variance** | **0.957379** | **3.193166** |
| **standard deviation** | **0.978457** | **1.786943** |
| **max** | **5.424** | **22.9** |
| **min** | **1.513** | **14.5** |
| **range** | -3.911 | -8.4 |
|  |  |  |

Variance measures the spread or dispersion of the data points around the mean. weight has a higher variance compared to score, indicating that the values in weight are more spread out from the mean.

Standard deviation measures the average deviation of data points from the mean. Similar to variance, weight has a higher standard deviation compared to score, indicating greater dispersion of values.

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(x) = sum of wt/ num of patients

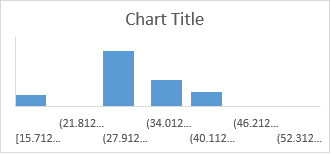
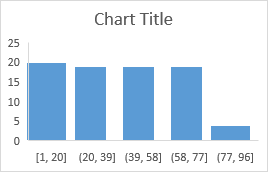
= 1298/9 = 144.22(approx.)

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
| **skewness** | **1.581454** | **-0.60331** |
| **kurtosis** | **2.977329** | **0.950291** |



Skewness:

For the first dataset: Skewness is approximately 1.581.

For the second dataset: Skewness is approximately -0.603.

A positive skewness value (greater than 0), as seen in the first dataset, indicates that the distribution is skewed to the right, meaning it has a longer tail on the right side of the distribution.

A negative skewness value (less than 0), as seen in the second dataset, indicates that the distribution is skewed to the left, meaning it has a longer tail on the left side of the distribution.

A skewness value close to 0 indicates that the distribution is approximately symmetric.

Kurtosis:

For the first dataset: Kurtosis is approximately 2.977.

For the second dataset: Kurtosis is approximately 0.950.

A kurtosis value greater than 3 indicates that the distribution has heavier tails and is more peaked compared to a normal distribution. This is known as leptokurtic.

A kurtosis value less than 3 indicates that the distribution has lighter tails and is less peaked compared to a normal distribution. This is known as platykurtic.

A kurtosis value equal to 3 is considered mesokurtic, meaning it has similar peakedness and tail heaviness as a normal distribution.

In summary, based on the provided skewness and kurtosis values, we can infer the shape and characteristics of the distributions for the two datasets. The first dataset has a right-skewed distribution with heavier tails and more peakedness compared to a normal distribution (leptokurtic), while the second dataset has a left-skewed distribution with lighter tails and less peakedness compared to a normal distribution (platykurtic).

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



Average chicken wt is between 50 to 100 by seeing seeing graph with skewed towards right side.

Seeing box plot we see that, data is concentrated towards q1 and q2 and there are a few outliers outside the range.

Hence avg chicken weight in between 50 to 100 grms

**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?

A confidence interval is a range of values calculated from a sample of data that is likely to contain the true population parameter with a certain level of confidence.

For example, let's say you're trying to estimate the average height of adult males in a city. You take a sample of 100 adult males and calculate their average height, which turns out to be 175 centimeters. However, you know that this sample average might not perfectly represent the true average height of all adult males in the city.

A confidence interval provides a range of values around the sample statistic (in this case, the sample mean) within which you are reasonably confident that the true population parameter (in this case, the population mean) lies.

For instance, you might calculate a 95% confidence interval for the average height of adult males in the city as 170 to 180 centimeters. This means that you are 95% confident that the true average height of adult males in the city falls within this range.

First z score is cal for all 3, then applied to confidence interval formula

Confidence Interval=(198.7382,201.2618)

Confidence Interval=(198.6856,201.3144)

Confidence Interval=(198.4376,201.5624)

**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.

|  |  |  |
| --- | --- | --- |
| mean | 41 |  |
| median | 40.5 |  |
| variance | 25.52941 |  |
| std var | 4.910307 |  |
|  |  |  |

1. What can we say about the student marks?

A variance of 25.52941 and a standard deviation of approximately 4.91 indicate that the student marks vary around the mean by an average of approximately 4.91 marks.

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

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

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

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

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

a14) Right skewness.

a15) Left skewness.

A16) Heavy tails, peaked distribution.

a17) Light tails, flat distribution.

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



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

What is nature of skewness of the data? Right side skewed

What will be the IQR of the data (approximately)? Towards q4

Q19) Comment on the below Boxplot visualizations?



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

ans

Data 1 has smaller range lying bt/ 235-285 and data 2 has larger range bt/100=350 approx. both data are symmetric and have normal peakedness.

No outliers in both.

Whiskers are equal for both

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) = 33/81 = 0.407
  2. P(MPG<40) = 61/81 = 0.753

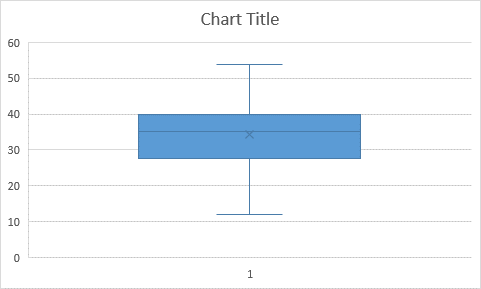
c. P (20<MPG<50) = 69

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

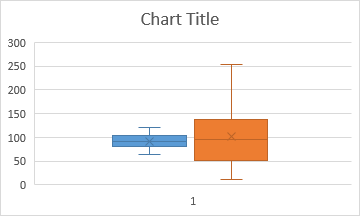
Dataset: Cars.csv

Asymmetric



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

Dataset: wc-at.csv



Assymetic, left skewed and median between 100-120

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

|  |  |  |
| --- | --- | --- |
| **mean** | 92.02275 | 129.8934 |
| **std dev** | 17.53972 | 304.5225 |
| **Score** | 90 | 90 |
|  |  |  |

Z = (interval – mean)/standard dev

For AT:

At 90: -0.115

At 94: 0.113

At 64: -0.216

For circumference:

At 90: -0.131

At 94: -0.118

At 64: -0.216

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

|  |  |  |
| --- | --- | --- |
| mean | 76.564 | 36.0768 |
| popluation mean | 92.02275 | 129.834 |
| std dev | 5.437975 | 10.53019971 |
| n | 25 | 25 |

Degrees of freedom (df) for both sets: n−1= 25−1 =24 ,

n−1=25−1=24

t-scores for each confidence interval:

For a 95% confidence interval:

apha =1−Confidence level=1−0.95=0.05α=1−Confidence level=1−0.95=0.05

For a 96% confidence interval:

alpha =1−Confidence level=1−0.96=0.04α=1−Confidence level=1−0.96=0.04

For a 99% confidence interval:

=1−Confidence level=1−0.99=0.01α=1−Confidence level=1−0.99=0.01

|  |  |  |
| --- | --- | --- |
| **95 percent t score** | 0.031666 | 0.03166636 |
| **96 percent t score** | 0.025332 | 0.0253315 |
| **99 percent t score** | 0.006332 | 0.006332214 |

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= (mean of sample – mean of population)/((standard dev of sample/ (under root of n))

mean of sample= 260

mean of population=270

n= 18

s= 90

degrees of freedom df= n-1 = 18- 1= 17

rcode 🡪 pt(tscore,df)

*t*≈−0.9487

*P*(*t*≤−0.9487)=pt(−0.9487,17)

If the probability is sufficiently low (usually below a significance level, such as 0.05), it suggests that the observed sample mean is significantly lower than what would be expected if the company's claim were true. This could lead to rejecting the company's claim about the average lifespan of the light bulbs.

However, if the probability is not very low, it suggests that observing such a sample mean is reasonably likely even if the company's claim were true. In this case, we wouldn't have strong evidence to reject the company's claim.

So, to draw a conclusion, you would compare the calculated probability to a predetermined significance level (e.g., 0.05). If the probability is less than or equal to the significance level, you may conclude that the observed sample mean is significantly different from the claimed population mean. Otherwise, you may not have sufficient evidence to reject the company's claim.