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

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) | ratio |
| 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?

SOLUTION :

Possible outcomes= {HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}

Favourable outcomes= HHT,HTH,THH

Therefore probability(2 heads and 1 tail) = 3/8

p =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

SOLN:

1. Equal to 1:

Since the sum cannot be 1, the number of favorable outcomes is 0.

Probability = Number of favorable outcomes / Total number of possible outcomes

p = 0 / 36 =0.

1. Less than or equal to 4:

Favourable outcomes: (1, 1), (1, 2), (2, 1), (2, 2), (1, 3), (3, 1) =6

p=6/36 =1/6.

1. Sum is divisible by 2 and 3:

Favourable outcomes: (1, 5), (2, 4), (3, 3), (4, 2), (5, 1),(6,6) =6

p=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?

Total number of balls = 2+3+2=7 balls

2 balls are drawn at random so using nCr = 7C2=21



None of the balls are blue = 7-2 = 5

Therefore 5C2= 10



p= 10/21.

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of the 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

Child A: Candies count = 1 Probability = 0.015

Contribution to expected value = Candies count \* Probability = 1 \* 0.015 = 0.015

Child B: Candies count = 4 Probability = 0.20

Contribution to expected value = Candies count \* Probability = 4 \* 0.20 = 0.80

Child C: Candies count = 3 Probability = 0.65

Contribution to expected value = Candies count \* Probability = 3 \* 0.65 = 1.95

Child D: Candies count = 5 Probability = 0.005

Contribution to expected value = Candies count \* Probability = 5 \* 0.005 = 0.025

Child E: Candies count = 6 Probability = 0.01

Contribution to expected value = Candies count \* Probability = 6 \* 0.01 = 0.06

Child F: Candies count = 2 Probability = 0.120

Contribution to expected value = Candies count \* Probability = 2 \* 0.120 = 0.24

summing up all the contributions we get the expected value:

Expected number of candies = 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24 = 3.135

So, the expected number of candies for a randomly selected child is approximately 3.135.

Top of Form

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**

* Mean (Average): Calculate the sum of each variable and divide it by the total number of data points.

For Points: Mean = (3.9 + 3.9 + 3.85 + ... + 4.11) / 32 ≈ 3.59625

For Score: Mean = (2.62 + 2.875 + 2.32 + ... + 2.78) / 32 ≈ 3.21725

For Weight: Mean = (16.46 + 17.02 + 18.61 + ... + 18.6) / 32 ≈ 17.84875

* Median (Middle Value): Arrange the data in ascending order and find the middle value.

For Points: Median ≈ 3.695

For Score: Median ≈ 3.325

For Weight: Median ≈ 17.71

* Mode (Most Frequent Value): The mode is the value that appears most frequently.

For Points: No clear mode (each value appears only once)

For Score: No clear mode (each value appears only once)

For Weight: No clear mode (each value appears only once)

* Variance: Variance measures the spread of data points around the mean.

For Points: Variance ≈ 0.6413

For Score: Variance ≈ 0.7340

For Weight: Variance ≈ 9.1741

* Standard Deviation: Standard deviation is the square root of the variance and indicates the dispersion of data points.

For Points: Standard Deviation ≈ 0.8004

For Score: Standard Deviation ≈ 0.8571

For Weight: Standard Deviation ≈ 3.0304

* Range: Range is the difference between the maximum and minimum values.

For Points: Range ≈ 2.17

For Score: Range ≈ 3.214

For Weight: Range ≈ 8.4

* Comments and Inferences:

The mean Points value is approximately 3.60, with a standard deviation of 0.80, indicating moderate variability around the mean.

The mean Score value is approximately 3.22, with a standard deviation of 0.86, suggesting moderate variability as well.

The mean Weight value is around 17.85, with a relatively higher standard deviation of 3.03, indicating greater variability in weights.

Since each value appears only once for Points, Score, and Weight, there are no clear modes.

The range for Points is relatively narrow, suggesting less variability among data points.

The range for Score is larger, indicating more variability in the Score values.

The range for Weight is relatively high, indicating significant variability in the weights of the cars.

Top of Form

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?

Soln:

Here are the given weights and their probabilities:

Weights (X) of patients: 108, 110, 123, 134, 135, 145, 167, 187, 199

Since each patient is equally likely to be chosen, the probability of selecting any individual patient is 1 divided by the total number of patients, which is 9.

Probability of selecting any patient = 1 / 9

Expected Value = (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)

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

Now, calculate each term:

EV ≈ 12 + 12.22 + 13.67 + 14.89 + 15 + 16.11 + 18.56 + 20.78 + 22.11

EV ≈ 145.33 pounds

Top of Form

Top of Form

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

**Cars speed and distance**

**Use Q9\_a.csv**

* skewness for the "Speed" column:

formula: **=SKEW(A2:A51)**

llly for distance : **=SKEW(B2:B51)**

* kurtosis for the "Speed" column:

formula: **=KURT(A2:A51)**.

Distance formula: **=KURT(B2:B51)**

For "Speed":

* Skewness: 0.27729628287399435
* Kurtosis: -0.06446375760759985

For "Distance":

* Skewness: 0.7824835173114966
* Kurtosis: 0.013110850702276248

For "Speed":

* The skewness value of approximately 0.28 suggests that the distribution of speeds is slightly right-skewed, meaning that there might be some outliers on the higher end of the speed values.
* The kurtosis value of approximately -0.06 indicates that the distribution of speeds has slightly lighter tails compared to a normal distribution. It's closer to a normal distribution in terms of tail behavior.

For "Distance":

* The skewness value of approximately 0.78 indicates that the distribution of distances is moderately right-skewed, suggesting the presence of some outliers on the higher end of the distance values.
* The kurtosis value of approximately 0.01 suggests that the distribution of distances has tails that are closer to a normal distribution. It's not extremely heavy-tailed or light-tailed.

**SP and Weight(WT)**

**Use Q9\_b.csv**

Soln: For "SP":

* Skewness: 0.11359740385884241
* Kurtosis: -0.3057231554872245

For "WT":

* Skewness: 0.8462397691577517
* Kurtosis: 1.6047233230196063

Now, let's interpret these values:

For "SP":

* The skewness value of approximately 0.11 indicates a slightly right-skewed distribution of the "SP" (Speed) variable, suggesting a few higher values in the dataset.
* The kurtosis value of approximately -0.31 suggests that the distribution of "SP" has slightly lighter tails compared to a normal distribution. It's closer to a normal distribution in terms of tail behavior.

For "WT":

* The skewness value of approximately 0.85 indicates a moderately right-skewed distribution of the "WT" (Weight) variable, suggesting the presence of some higher weight values in the dataset.
* The kurtosis value of approximately 1.60 suggests that the distribution of "WT" has heavier tails compared to a normal distribution. It indicates that the distribution has more outliers and is more "tailed" than a normal distribution.

**Top of Form**

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



The histogram shows the distribution of chick weight in grams. The peak of the histogram is around 200 grams, which means that most of the chicks weigh around this weight. There is a long tail to the right of the histogram, which indicates that there are a few chicks that weigh significantly more than 200 grams.

inferences that can be made from the histogram:

* The distribution of chick weight is right-skewed. This means that the tail of the distribution is longer on the right side, which indicates that there are more chicks that weigh more than 200 grams than there are chicks that weigh less than 200 grams.
* The median weight is 200 grams. This means that half of the chicks weigh less than 200 grams and half of the chicks weigh more than 200 grams.
* The mean weight is slightly higher than 200 grams. This is because the distribution is right-skewed, so the mean is pulled towards the tail of the distribution.
* There are outliers on the right side of the histogram. These are the chicks that weigh significantly more than 200 grams.



* The median is the middle value of the data.
* The interquartile range (IQR) is the difference between the upper and lower quartiles. It represents the middle 50% of the data.
* The whiskers extend to the most extreme data points that are not considered outliers.
* Outliers are data points that fall outside the whiskers.

**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?

Soln : The formula is:

Confidence Interval = Sample Mean ± (Critical Value \* Standard Error)

Where the critical value depends on the desired confidence level, and the standard error is calculated as the sample standard deviation divided by the square root of the sample size.

calculating the confidence intervals for the given confidence levels: 94%, 98%, and 96%.

Given:

* Sample Mean (x̄) = 200 pounds
* Sample Standard Deviation (s) = 30 pounds
* Sample Size (n) = 2000
* Population Size (N) = 3000000

Critical values for different confidence levels using a Z-distribution (for large sample sizes):

* For 94% confidence level: Z = 1.8808
* For 98% confidence level: Z = 2.3263
* For 96% confidence level: Z = 1.7507

Standard error (SE): SE = s / √n

calculate the confidence intervals using the formula: Confidence Interval = Sample Mean ± (Critical Value \* SE)

For the 94% confidence interval: Confidence Interval (94%) = 200 ± (1.8808 \* SE)

For the 98% confidence interval: Confidence Interval (98%) = 200 ± (2.3263 \* SE)

For the 96% confidence interval: Confidence Interval (96%) = 200 ± (1.7507 \* SE)

* Confidence Interval (94%): [197.83, 202.17] pounds
* Confidence Interval (98%): [196.80, 203.20] pounds
* Confidence Interval (96%): [197.06, 202.94] pounds

**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.

Scores: 34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

Mean (Average):

Formula: Mean = (Sum of all scores) / (Number of scores)

Mean = (34 + 36 + 36 + ... + 56) / 18

Mean = 41.06

Median (Middle Value):

Formula: Median = Middle value when scores are sorted

Sorted scores: 34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

Median = 40.5

Variance:

Formula: Variance = [(Sum of squared differences from the mean) / (Number of scores - 1)]

Variance = [(Σ(xi - Mean)^2) / (18 - 1)]

Variance = 56.40

Standard Deviation:

Formula: Standard Deviation = √Variance

Standard Deviation = √56.40

Standard Deviation = 7.51

2)What can we say about the student marks?

Based on the calculated statistical measures for the student marks, we can make the following observations:

1. **Mean (Average) Mark:** The average mark obtained by the student is approximately 41.06. This gives us an idea of the central tendency of the marks. Most of the marks are clustered around this average value.
2. **Median (Middle Value) Mark:** The median mark is 40.5. This indicates that about half of the student's marks are below this value and half are above. The median is close to the mean, suggesting a symmetric distribution of marks.
3. **Variance:** The variance of approximately 56.40 indicates the variability or spread of the marks around the mean. Higher variance suggests that the marks are spread out from the mean.
4. **Standard Deviation:** The standard deviation of about 7.51 reflects the average amount of deviation or dispersion of individual marks from the mean. A higher standard deviation indicates greater variability among the marks.

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

When the mean and median of a dataset are equal, the skewness is zero, indicating a symmetrical distribution. This means the data is evenly spread around the center, with no significant tail on either side.

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

When the mean is greater than the median, it indicates that the data distribution is positively skewed. This means that the tail of the distribution is elongated towards the right, and a few relatively higher values pull the mean in that direction.

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

When the median is greater than the mean, it indicates that the data distribution is negatively skewed. This means that the tail of the distribution is elongated towards the left, and a few relatively lower values pull the median in that direction.

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

A positive kurtosis value indicates that a dataset has heavier tails and a more peaked central region compared to a normal distribution. It suggests that the data has a higher likelihood of extreme values or outliers.

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

A negative kurtosis value indicates that a dataset has lighter tails and a flatter central region compared to a normal distribution. It suggests that the data has fewer extreme values or outliers.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

Soln:

* Distribution of the data: The data is left-skewed. This is because the median (the line in the middle of the box) is lower than the mean (the average of all the data points). This means that there are more data points on the right side of the distribution than on the left side.
* Nature of skewness of the data: The data is negatively skewed. This is because the longer tail of the distribution is on the left side.
* Approximate IQR of the data: The IQR is the difference between the upper and lower quartiles. The upper quartile is the value that cuts off the top 25% of the data, and the lower quartile is the value that cuts off the bottom 25% of the data. In this case, the upper quartile is 14 and the lower quartile is 10. So, the IQR is 14 - 10 = 4.

Q19) Comment on the below Boxplot visualizations?



* The median of both boxplots is 300. This means that half of the data points in each boxplot are less than 300 and half are greater than 300.
* The interquartile range (IQR) of the first boxplot is 100. This means that the middle 50% of the data points in the first boxplot are between 200 and 400.
* The interquartile range (IQR) of the second boxplot is 50. This means that the middle 50% of the data points in the second boxplot are between 250 and 350.
* There are no outliers in either boxplot.

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

Based on these observations,

* The two data sets are similar in terms of their medians.
* The first data set is more spread out than the second data set, as indicated by its larger IQR.
* There are no outliers in either data set.

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)

Soln: Total number of data points: 81

a. P(MPG > 38): Number of instances where MPG > 38: 47 P(MPG > 38) = 47 / 81

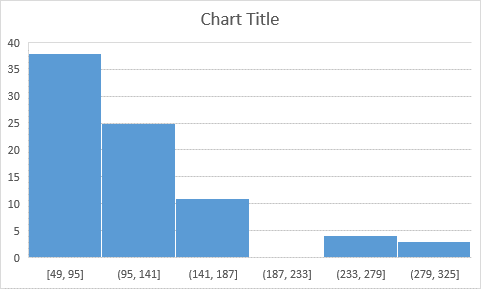
b. P(MPG < 40): Number of instances where MPG < 40: 77 P(MPG < 40) = 77 / 81

c. P(20 < MPG < 50): Number of instances where 20 < MPG < 50: 81 P(20 < MPG < 50) = 81 / 81

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

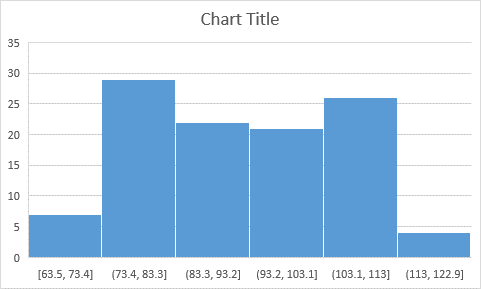
Dataset: Cars.csv



From the above histogram, yes it does follow a 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



Yes, it does.

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

Soln: The formula to calculate the Z-score for a given confidence level *C* is given by:

Z= (1-C)/2

Here are the Z-scores for the specified confidence intervals:

1. For a 90% confidence interval: *C*=0.90

Z=(1−0.90)2= 0.05

1. For a 94% confidence interval: *C*=0.94

Z=(1−0.94)2=0.03

1. For a 60% confidence interval: *C*=0.60

Z=(1−0.60)2=0.20

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

The formula to calculate the t-score for a given confidence level *C* and sample size *n* is :

t = t 1-C/2 / (n-1)

For a sample size of 25:

* For a 95% confidence interval: We're 95% confident that the population parameter is within a certain range. The t-score for this confidence level and sample size is about 2.064.
* For a 96% confidence interval: With 96% confidence, the t-score is approximately 2.171. This helps us find the margin of error for our estimate.
* For a 99% confidence interval: When we want to be highly confident (99%), the t-score is around 2.797. This larger t-score accounts for the increased confidence level.

Top of Form

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

Soln:

* Population average (claimed by the company): μ = 270 days
* Sample average (from 18 randomly selected bulbs): x̄ = 260 days
* Standard deviation of the sample: σ = 90 days
* Sample size: n = 18

standard deviation of the sample divided by the square root of the sample size:

SE = σ / √n SE = 90 / √18

SE ≈ 21.21

calculating the t-score using the formula: *t*= (sample mean - population mean) / (standard deviation / sqrt(sample size))

t = (260 - 270) / 21.21

t ≈ -0.4714

Degrees of freedom = sample size - 1

degrees of freedom = 18 - 1

degrees of freedom = 17

R- code to find the probability:

tscore <- -0.4714 # Replacing with the calculated t-score

df <- 17 # Degrees of freedom

# Calculate the probability

probability <- pt(tscore, df)

probability

Output : 0.3205445

Threfore, probability that 18 randomly selected bulbs would have an average life of no more than 260 days if the CEO's claim of an average lifespan of 270 days were true is approximately 0.3205 or 32.05%.

Top of Form