|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Continuous |
| Results of rolling a dice | 1,2,3,4,5,6 |
| 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 | Continuous |
| Number of tickets in Indian railways | Continuous |
| Number of times married | Continuous |
| 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 | 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 | Nominal |
| Religious Preference | Ordinal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

**Solution:** To find the probability of obtaining two heads and one tail when three coins are tossed, we can consider the different ways this outcomes can occur.

There are three possible scenarios:

1. Head-Head-Tail (HHT)
2. Head-Tail-Head (HTH)
3. Tail-Head-Head (THH)

Each of these scenarios has a probability of (1/2)\*(1/2)\*(1/2) = 1/8, since the probability of getting a head on a fair coin is ½, and the probability of getting a tail is also ½.

Now, we add up the probabilities of these three scenarios to get the total probability of getting two heads and one tail:

Total probability = Probability of HHT + Probability of HTH + Probability of THH

Total probability = (1/8) + (1/8) + (1/8)

Total probability = 3/8

So, the probability of obtaining two heads and one tail when three coins are tossed is 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

**Solution:** When rolling two dice, each die has 6 sides with numbers from 1 to 6. To find the probabilities for the given scenarios, we can use the concept of combinations and sample space.

Total possible outcomes when rolling two dice= 6 (Outcomes for the first die) x 6 (Outcomes for the second die) = 36 outcomes.

Let’s calculate the probabilities for each scenario:

1. Probability that the sum is equal to 1:

There is no possible way where the sum is 1, which is 0.

Probability = (Number of favorable outcomes) / (Total possible outcomes) = 0/36 = 0.

1. Probability that the sum is less than or equal to 4:

Favorable combinations: (1,1),(1,2),(1,3),(2,1),(2,2),(3,1)

Number of favorable outcomes = 6.

Probability = (Number of favorable outcomes) / (Total possible outcomes) = 6/36 = 1/6.

1. Probability that the sum is divisible by 2 and 3:

The only numbers between 2–12 which are divisible by both 2 and 3 are 6 and 12.  
While 12 can only be made 1 way (double 6) 6 can be made by (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) so the probability of throwing a sum of either 12 or 6 is (number of successful outcomes) / (number of all possible outcomes) = (1+5)/36 = 6/36 = 1/6

To summarize:

1. Probability that the sum is equal to 1: 1/36.
2. Probability that the sum is less than or equal to 4: 1/6
3. Probability that the sum is divisible by 2 and 3: 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?

**Solution:**

Given,

A bag contains 2 red, 3 green and 2 blue balls. We need to find the probability that none of the drawn balls is blue.

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

Two balls can be drawn in 7C2 ways = (7!) / (2! (7-2)! ) = 21 ways

Number of ways of drawing 2 balls such that none is blue = Number of ways of drawing 2 balls from 2 red and 3 green balls = 5C2 = (5!) / (2! (5-2)! ) = 10 ways Probability of drawing 2 balls such that none is blue:

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

Probability = 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

**Solution:** To calculate the expected number of candies for a randomly selected child, we need to multiply each candy count by its corresponding probability and then sum up these products for all children. The formula for calculating the expected value (E) in this case would be:

E = (Candies count for Child A x Probability of Child A) +

(Candies count for Child B x Probability of Child B) +

(Candies count for Child C x Probability of Child C) +

(Candies count for Child D x Probability of Child D) +

(Candies count for Child E x Probability of Child E) +

(Candies count for Child F x Probability of Child F)

Let’s calculate it:

E = (1 x 0.015) + (4 x 0.020) + (3 x 0.65) + (5 x 0.005) + (6 x 0.01) + (2 x 0.120)

E = 0.015 + 0.080 + 1.95 + 0.025 + 0.06 + 0.24

E = 3.135

The expected number of candies for a randomly selected child is 3.135.

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**

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Standard Deviation | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.17 | 3.911 | 8.4 |
|  |  |  |  |
| Maximum | 4.93 | 5.424 | 22.9 |
| Minimum | 2.76 | 1.513 | 14.5 |

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?

**Solution:** The expected value (also known as the mean) of a set of number is calculated by summing up all the values and dividing by the total number of values. In this case, you want to find the expected weight of a patient at the clinic.

The weights of the patients are:

108, 110, 123, 134, 135, 145, 167, 187, 199

To calculate the expected value:

1. Add up all the weights:

108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199 = 1318

1. Divide the sum by the total number of patients (which is 9 in this case):

1318 / 9 = 146.44

So, the expected value (average) weight of a patient at the clinic is approximately 146.44 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Solution:**

|  |  |  |
| --- | --- | --- |
|  | speed | distance |
| Skewedness | -0.11751 | 0.806895 |
| Kurtosis | 2.422853 | 3.248019 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Solution:**

|  |  |  |
| --- | --- | --- |
|  | speed | weight |
| Skewedness | 1.451518 | -0.78436 |

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



**Solution:**

Here we can see that the major Chick weights fall in the category of 50-100 g (measures in x) as the maximum which is 200.The minimum weights have a frequency if less than or equal to 5.

The plot is Right skewed which show that there is lesser concentration of chick weights in the 300-400gram category .

The expected value should be above 46.45



**Solution:**

Median is less than mean right skewed and we have outlier on the upper side of box plot and there is less data points between Q1 and bottom point.

**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?

**Solution:** To calculate confidence intervals for the average of adult males in Mexico based on the sample data we got, we can use the formula for the confidence interval of the mean:

Confidence Interval = Sample Mean ± (Critical value) x (Standard Deviation / √Sample Size)

The critical value is obtained from the Z=distribution, which corresponds to the desired confidence level. For a 94% confidence interval, the critical value is approximately 1.8808, for a 98% confidence interval, it’s approximately 2.3263 and for a 96% confidence interval it’s approximately 1.7507.

Given data:

Sample Mean (x) = 200 pounds

Standard Deviation (σ) = 30 pounds

Sample Size (n) = 2000 men

Let’s calculate the confidence intervals for each confidence level:

1. For a 94% Confidence Interval:

Critical Value = 1.8808

Confidence Interval = 200 ± (1.8808) x (30/√2000)

Confidence Interval = 200 ± 2.6514

Confidence Interval = (197.3486, 202.6514)

1. For a 98% Confidence Interval:

Critical Value = 2.3263

Confidence Interval = 200 ± (2.3263) x (30/√2000)

Confidence Interval = 200 ± 2.6466

Confidence Interval = (197.3534, 202.6466)

1. For a 96% Confidence Interval:

Critical Value = 1.7507

Confidence Interval = 200 ± (1.7507) x (30/√2000)

Confidence Interval = 200 ± 2.4666

Confidence Interval = (197.5334, 202.4666)

So, the calculated Confidence Intervals are as follows:

1. 94% Confidence Interval = (197.3486, 202.6514) pounds
2. 98% Confidence Interval = (197.3534, 202.6466) pounds
3. 96% Confidence Interval = (197.5334, 202.4666) 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.
2. What can we say about the student marks?

**Solution:**

1. Let’sstart by calculating the required statistics for the given set of scores:

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

1. **Calculating Mean:**

Mean (Average) = (Sum of all scores) / (Total number of scores)

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

Mean = 738 / 18

Mean = 41

1. **Calculating Median:**

The median is the value when the scores are arranged in ascending order.

Arranging the scores in ascending order : 34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

Since there are 18 scores, the median is the average of 9th and 10th scores.

Median= (40+41) / 2

Median= 40.5

1. **Calculating Variance:**

Variance = (Sum of squared differences from the mean) / (Total number of scores – 1)

Step 1: Calculate the squared differences from the mean for each score:

(34 - 41)2 + (36 - 41)2 + (36-41)2 + (38-41)2 + (38-41)2 + (39-41)2 + (39-41)2 + (40-41)2 + (40-41)2 + (41-41)2 + (41-41)2 + (41-41)2 + (41-41)2 + (42-41)2 +(42-41)2 +(45-41)2 + (49-41)2 + (56 - 41)2

Step 2: Sum up the squared differences:

Sum = 49+ 25 + 25 + 9 + 9 + 4 + 4 + 1 + 1 + 0 + 0 + 0 + 0 + 1 + 1 + 16 + 64 + 225

Sum = 434

Step 3: Divide the sum by (n-1), where n is the number of scores (18):

Variance = Sum / (18-1)

Variance = 434 / 17

Variance = 25.53 (approximately)

1. **Calculating Standard Deviation:**

Standard Deviation = √Variance

Standard Deviation = √25.53

Standard Deviation = 5.0528 (approximately)

1. Now, regarding what we can say about the student’s marks:
2. The mean (average) score is 41.
3. The median score is 40.5, which indicates that about half of the scores are below this value and half are above it.
4. The variance of 25.53 and the standard deviation of 5.0528 indicate the dispersion or spread of the scores around the mean.
5. The scores are relatively concentrated around the mean, as the standard deviation is not too high.

Overall, the student’s scores show a consistent performance with some variation and the majority of scores are clustered around the middle range of marks.

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

**Answer:** In histogram, mean and median of the data are equal means the distribution of data is symmetric. If the distribution is symmetric the distribution has zero skewness.

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

**Answer:** In histogram, mean is greater than the median of the data means the distribution is positively skewed.

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

**Answer:** In histogram, median is greater than the mean of the data means the distribution is negatively skewed.

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

**Answer:** In histogram, positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.

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

**Answer:** In histogram, a negative value of kurtosis means that the distribution has thinner tails and a flatter peak than the normal distribution.

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



What can we say about the distribution of the data?

**Answer:** This data is skewed, as the median of 15 is much closer to the 75th percentile of 18 than to the 25th percentile of 10.

What is nature of skewness of the data?

**Answer:** Data is negatively skewed

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

**Answer:** Here,

Q3 =18

Q1 = 10

IQR = Q3 – Q1

IQR = 18 - 10  
 IQR = 8

Q19) Comment on the below Boxplot visualizations?

**Ans:** Here there is a representation of 2 box plots in which box plot 2 is highly distributed across the plane and 1 is slightly less distributed.(variance)Whiskers in these diagrams also show this. 100% of the data is spread across values from 350 in 2 whereas its spread in range 250-290 app x in 1



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

**Ans:**

Here when we compare box plot 1 with box plot 2 we can say that the data in boxplot 1 is widely spread. Here the main inference is that since the data range varies high in box plot 2 it is hard to make a prediction in box plot 2. The median in the 2box plots are equal. And the data spread in both of them are symmetrical

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)

**Solution:**

P (MPG>38)

= mean (MPG) = 34.42208

= sd (MPG) = 9.131445

= 1**–** pnorm (38, mean(MPG), sd(MPG))

= 0.330

= 33%

P (MPG<40)

=pnorm (40, mean(MPG), sd(MPG))

=0.7293499

=72.3%

P (20<MPG<50)

= pnorm (50, mean(MPG), sd (MPG))**–** pnorm (20,mean(MPG), sd (MPG))

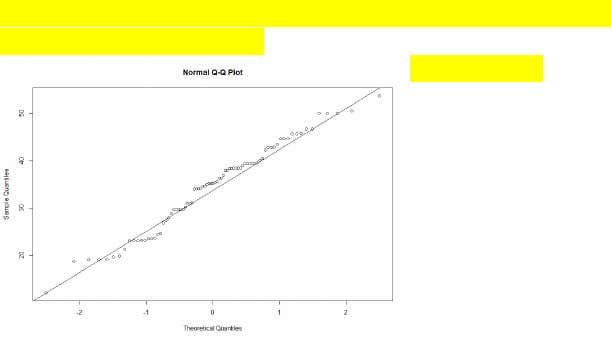
=0.955 -0.057

=0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

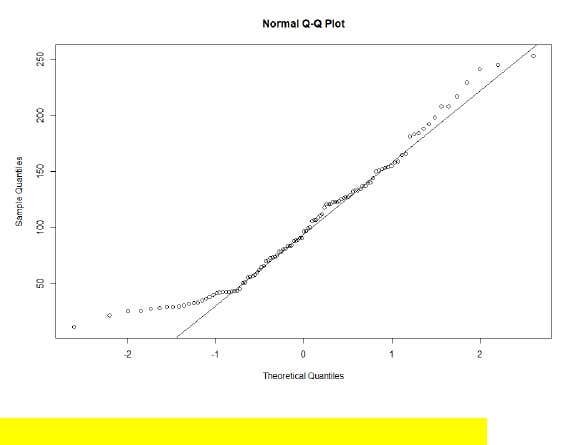
Dataset: Cars.csv



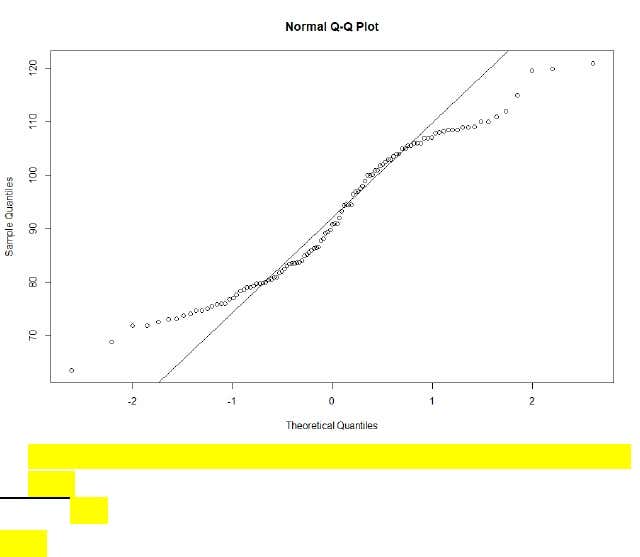
When we plot check the qqnorm and qqline we can almost get a straightline thus the data is normalized.

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

Dataset: wc-at.csv



Majority of the data points lie on the qqline hence normal.



This data set is not normal because the data points follows an abnormal curve.

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

**Solution:**

**Z Scores**

=90%

= 95+2.5

=97.5

=qnorm(0.975)

=1.96

94%

= 94+4=97

=qnorm(0.97)

 =1.88

60%

= 60 + 20= 80

= qnorm (0.80)

= 0.841

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

**Solution:**

TSCORE CALCULATION

T((1,alpha),(n-1))

Here n = 25

n-1 = 24

Hence t score values will be:

95%

= qt(0.975,24)

= 2.063899

96%

= qt(0.98,24)

=2.17154599%

= qt(0.995,24)

= 2.79694

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

**Solution:**

t - Statistics for the data is given as follows:

t = (x - μ) / (s / √n)

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

t = (260 - 270) / (90 / √18)

t = (-10) / (90 / 3√2)

t = -10 / (30 / √2)

t = (-1 \* √2) / 3

t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here we need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.