|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete(count) |
| Results of rolling a dice | Discrete(count) |
| 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(count) |
| Number of tickets in Indian railways | Discrete(count) |
| Number of times married | Discrete(count) |
| 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 | Interval |
| Celsius Temperature | Interval |
| Weight | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| 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 | Nominal |
| 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?

A) All possible occurrences = (HHH, HHT, HTH, THH, TTH, THT, HTT, TTT)

Interested events = (HHT, HTH, THH)

P = 3/8

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1

ans: (0)

1. Less than or equal to 4

(1,1), (1,2), (2,1) (1,3), (2,2), (3,1)

ans: (6/36)

1. Sum is divisible by 2 and 3

The possible numbers that are divisible by both 2 and 3 are 6, 12

(2,4), (4,2), (3,3), (1,5), (5,1), (6,6)

ans: (6/36)

All possible occurrences = {(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)}

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 no. of balls = 2+3+2 = 7

So, there are 7C2 ways of drawing 2 balls randomly (7C2 = 21)

(none. Of the balls drawn is blue) 2 balls can be drawn from 2 red and 3 green in 5c2 ways (10)

P(none of balls drawn is blue) = 10/21 = 0.4761

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.

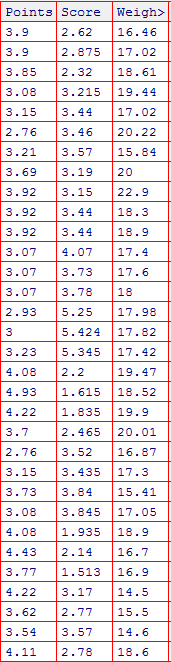
Child B – probability of having 4 candies = 0.20

ans = Expected value = ∑xp(x) = 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:**

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | Points | Score | Weight |
| Mean | 3.597 | 3.217 | 17.849 |
| Median | 3.695 | 3.325 | 17.710 |
| Mode | 3.07 | 3.44 | 17.02 |
| Variance | 0.277 | 0.927 | 3.093 |
| Std. Deviation | 0.526 | 0.936 | 1.759 |
| Range | 2.17 | 3.911 | 8.4 |

**Note: Respective code files: Q7.R, Q7.py**

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: 145.333 as Expected value = ∑xp(x) and p(x) = 1/9

(refer **Q8 Sol** Excel Sheet)

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure** | **speed** | **dist** | **SP** | **WT** |
| Skewness | -0.114 | 0.782 | 1.581 | -0.603 |
| Kurtosis | -0.577 | 0.248 | 2.724 | 0.819 |

**Note: Respective code files: Q9.R, Q9.py**

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



Ans:

* Data is in between 0 and 400. Has the highest peak between 0 and 50.
* As there are no gaps in between, the data is continuous.
* Data is Right Skewed, which means mean>median>mode
* Also there are no outliers



Ans: Data is not normally distributed. Has outliers. As the length of the box is less, dispersion is less.

**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?

**Given, n=2000, N=3,000,000, x̅=200, s=30**

**Confidence Interval =** **x̅ ±Z(s/sqrt(n))**

**@94% confidence,**

**CI = 200** **± 1.88(30/sqrt(2000)) = 198.73 - 201.26**

**@ 98% confidence,**

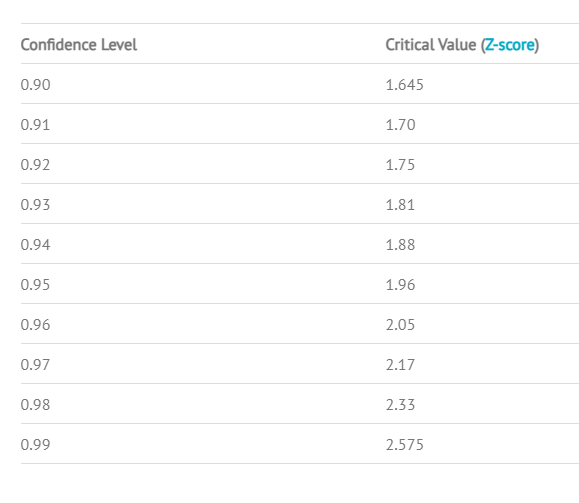
**CI = 200** **± 2.33(30/sqrt(2000)) = 198.43 - 201.56**

**@ 96% confidence,**

**CI = 200** **± 2.05(30/sqrt(2000)) = 198.62 - 201.37**

sample calculation in python: import math

val = 200 + 2.05\*(30/math.sqrt(2000))



**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 | 25.52941176 |
| Standard Deviation | 5.052663829 |

The data is right skewed as mean>median.

1. What can we say about the student marks?

Ans: Doesn’t have any outliers. Also, the data is multimodal.

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

Symmetric distribution so Zero skewed

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

Right Skewed

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

Left Skewed

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

Positive Kurtosis means thin peaks and wider tails

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

Negative Kurtosis means wider peaks and thin tails

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



What can we say about the distribution of the data?

Data is not normally distributed.

What is nature of skewness of the data?

Left Skewed

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

(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.

Both 1) and 2) data sets are normally distributed and has the same median.

Box 2 is longer than box 1, which means data set 2 has high dispersion compared to data set 1.

Neither of the data sets has outliers.

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) = 0.3475939
  2. P(MPG<40) = 0.7293499
  3. P (20<MPG<50) = 0.8988689

**Code file: Q20\_21.R**

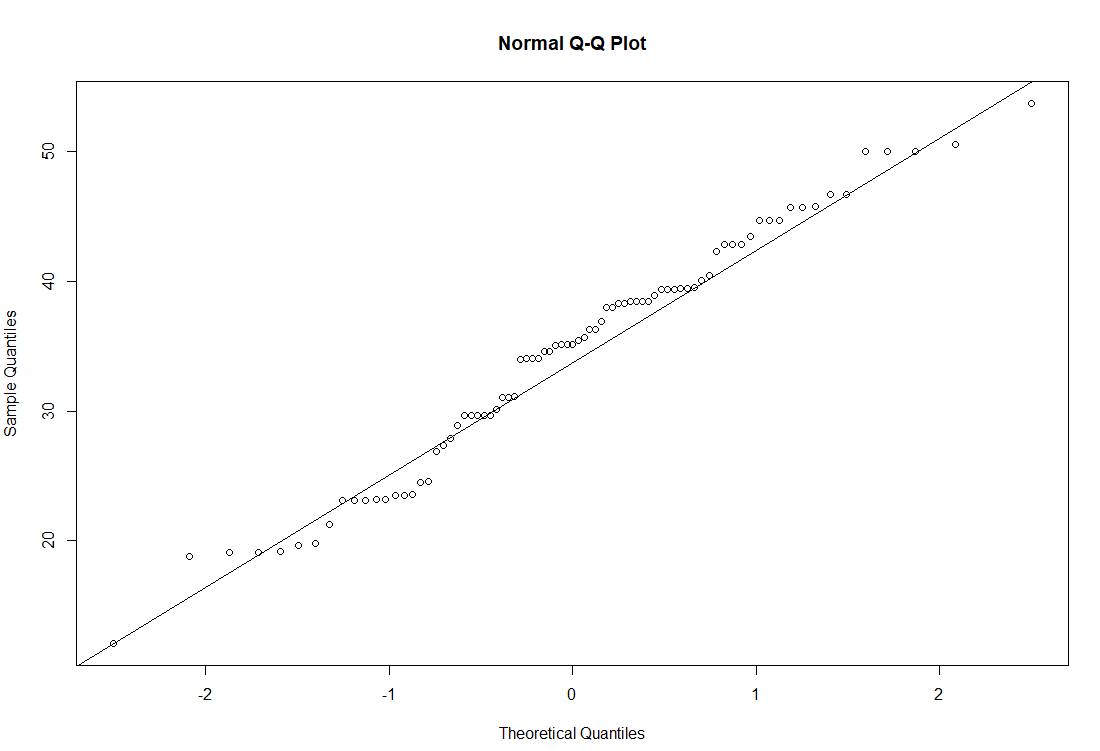
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

ans: **qqnorm(MPG)**

**qqline(MPG)**



From Shapiro Test: p-value = 0.1764 which is higher than the significance level. So we accept the Null Hypothesis. That means Data is following 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:

|  |  |  |
| --- | --- | --- |
|  | P Value from Shapiro Test | Result |
| AT | 0.000654 (<0.05) | Not Normal |
| Waist | 0.00117 (<0.05) | Not Normal |

P>0.05 => Null Hypothesis => Data is Normal

P<0.05 => Alternative Hypothesis => Data is not Normal

|  |  |  |
| --- | --- | --- |
|  | AT | Waist |
| QQ PLot |  |  |

Both AT, Waist are not following Normal Distribution

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

qnorm(0.95)

[1] 1.644854

qnorm(0.97)

[1] 1.880794

qnorm(0.80)

[1] 0.8416212

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

Ans: **T1-𝞪, n-1** **can be calculated from the formula qt(p,df)**

**T0.95,24**  **= qt(0.975,24) = 2.063899**

**T0.96,24 = qt(0.98,24) = 2.171545**

**T0.99,24** **= 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

Ans: **0.3216725**

Sol: Given, Population mean (μ)= 270, Sample size(n) = 18, sample mean(x̅) = 260, sample Std. dev.(s) = 90.

Therefore, df= n-1 = 17.

We know that,  t = (x̅-μ)/(s/sqrt(n)), let us take the confidence interval to be 95% as per the business standards.

t = (260-270)/(90/sqrt(18)) = -0.4714045

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days = pt(tscore,df)  = pt(-0.4714045,17) = **0.3216725**