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

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 | interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| Type of living accommodation | ordinal |
| Level of Agreement | ratio |
| IQ(Intelligence Scale) | ordinal |
| Sales Figures | Ratio |
| Blood Group | nominal |
| Time Of Day | interval |
| Time on a Clock with Hands | interval |
| Number of Children | interval |
| Religious Preference | Ordinal |
| Barometer Pressure | ratio |
| SAT Scores | interval |
| Years of Education | interval |

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

ANS = 3/8

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

Possibilities to obtain two heads and one tail = {HHT, HTH, THH}

Probability

=N(Possibilities to obtain two heads and one tail) / N(All possible outcomes)

=3/8

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

Solution –

All 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(All Possible Outcomes)

1. Equal to 1 = 0

Probability(a) = N(sum Equal to 1)/ N(All Possible Outcomes)

= 0/36

1. Less than or equal to 4 = 1/6

E(Less than or equal to 4) = {(1,1),(1,2),(1,3), (2,1),(2,2), (3,1)}

Probability(b) = N(Less than or equal to 4)/ N(All Possible Outcomes)

= 6/36 = 1/6

1. Sum is divisible by 2 and 3 = 1/6

E(Sum is divisible by 2 and 3) = {((2,4),(4,2), (3,3), (1,5),(5,1),(6,6)}

Probability(c) = N(Sum is divisible by 2 and 3)/ N(All Possible Outcomes)

= 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?

ANS = 10/21

Probability = 1- 2C2/7C2 – 2C1\*5C1/7C2 = 1-1/21-10/21 = 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) 3

|  |  |  |
| --- | --- | --- |
| 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 = E(x) = SUM{x.P(x)}/SUM(P(x))

SUM(P(x)) = 1

E(x) = 0.015 + 0.28 + 1.95 + 0.025 + 0.06+0.24 = 2.57

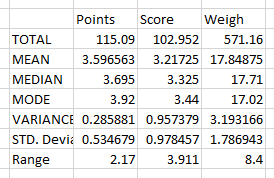
Randomly selected child expected to have 2.57 candies (approx. 3 candies)

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**



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?

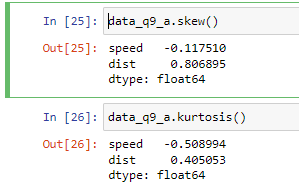
ANS – 145.33 pounds

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

1. Cars speed and distance

Ans -

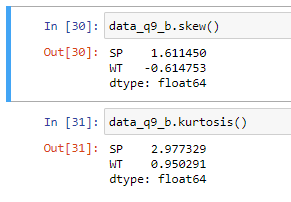


From above data we can infer that,  
a. skew for speed is -0.12(between -0.5 to 0.5) which means speed data is fairly symmetric about its mean. While skew for dist is 0.81(between 0.5 to 1) which indicates the data is positively skewed and moderately symmetrical about its mean

b. kurtosis value for speed is -0.51 and dist is 0.41, from this we can infer that kutosis value is less than 3 thus both speed and dist data have less outliers.

1. SP and Weight(WT)

Ans-



From above data we can infer that,  
a. skew for Sp is 1.6(> 0.5) which means SP data is highly unsymmetric about its mean. While skew for WT is -0.61 (between -1 to -0.5) which indicates the data is negatively skewed and moderately symmetrical about its mean

b. kurtosis value for SP is 2.97 and WT is 0.95, from this we can infer that kutosis value is less than 3 thus both SP and WT data have less outliers.

But the ouliers in SP data are widely spread than th WT data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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





**From histogram we can infer that,**

Frequency range =200-0 =200

Class interval = 50

Max frequency = 200 In 50-100 class and min frequency = 0 in 350-400 class

Data is positively skew, median lies to the left of the mean.

**From boxplt we can infer,**

Data is positively skew as value of median is less than the mean.

Distribution of ouliers are more in positive side of the data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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?

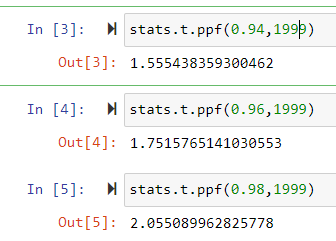
Ans – Population = 3000000

Sample = 2000

Mean = 200

Std dev = 30





94% CI - 200 +/- 1.55\*30/sqrt(2000) = (198.96, 201.04)

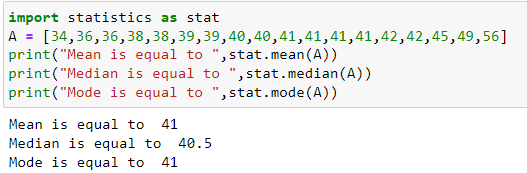
96% CI – 200 +/- 1.75\*30/sqrt(2000) = (198.82, 201.17)

98% CI – 200 +/- 2.05\*30/sqrt(2000) = (198.62, 201.36)

**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.

Ans – 

1. What can we say about the student marks?

Ans – average marks obtain by given set of student is 41. 4 students scored 41 marks which is highest number of student who have scored same marks.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

Ans - when mean and median are for the data then data is symmetrical about its mean.

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

Ans – if mean > median then data is positively skewed

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

Ans – if median > mean then data is negatively skewed.

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

Ans – positive kurtosis indicates that data contributing in tail is higher which indicates that there are more outliers

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

Ans – negative kurtosis indicates that the data contributing for tail is less which means there low outliers present in the data

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



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

Ans – from above boxplot we can say,

25 % data lies in 2 to 10, median is close to 16 (can not find exact from given data),

75 % data lies between 0 to 18, and remaining 25 % data lies above 18.

**What is nature of skewness of the data?**

Ans - Tail of the distribution lies in the left side of the mean which indicates data is negatively skewed

**What will be the IQR of the data (approximately)?**   
ans - the interquartile range(IQR) is measure of the range for middle 50% data.

To find IQR from boxplot, 25 % is at 10 and 75 % is at 18.

Thus IQR = 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.

Ans –

Q2 Second Quartile 50th percentile ,Box plot 1 median = box plot 2 median = 262.5

Q1 first Quartile 25th percentile, Box Plot 1 = 250 and Box Plot 2 = 212.5

Q3 third Quartile 75th Percentile, Box Plot 1 = 275 and Box Plot = 300

1. Extreme ends of boxplpot indicates the maximum and minimum of the data, range = max -min
2. Box plot 1, range 1 = 287.5 – 237.5 = 50
3. Box plot 2 = range 2 = 337.5 - 187.5 = 150

From above we can infer that range first box plot is less than the second which means data of second boxplot is widely dispersed.

1. box plot also gives idea about type of distribution of the data,

for box plot 1, max- median = 287.5 – 262.5 = 25

median – min = 262.5 – 237.5 = 25

for box plot 2, max – median = 337.5 – 262.5 = 75

median – min = 262.5 – 187.5 = 75

from above data we can say both box plots are symmetrically distributed.

--------------------------------------------------------------------------------------------------------

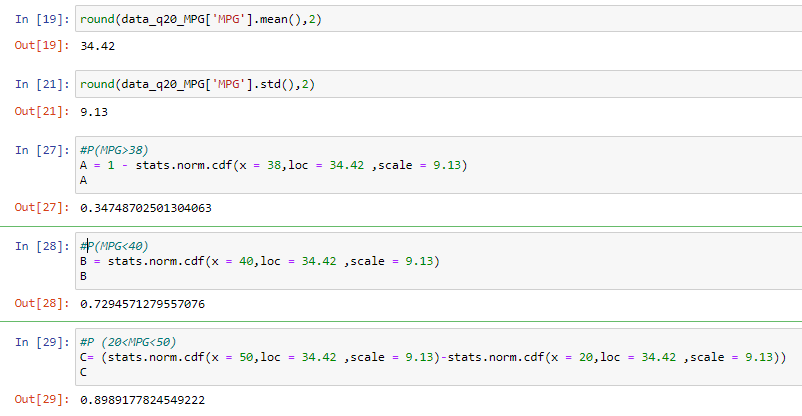
Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

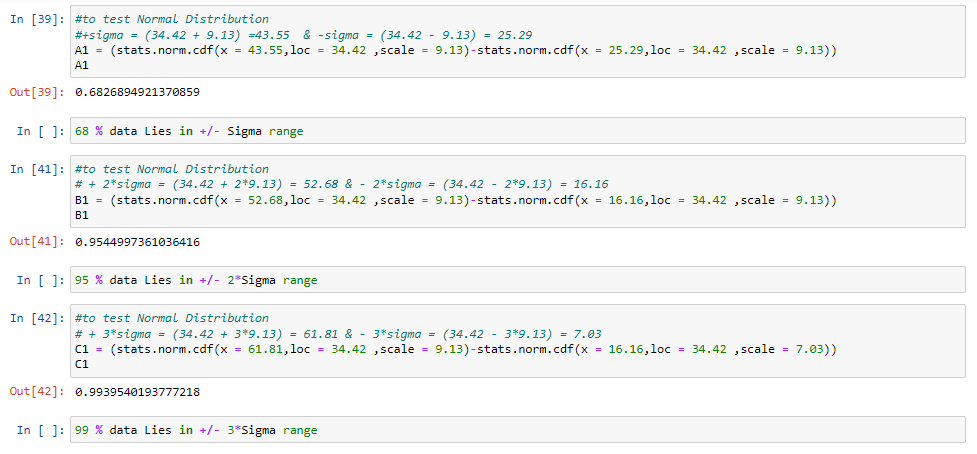
* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

Ans - 

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

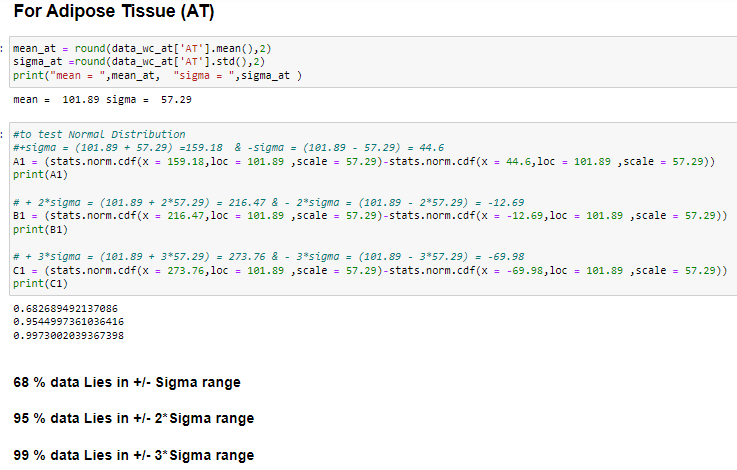
Dataset: Cars.csv

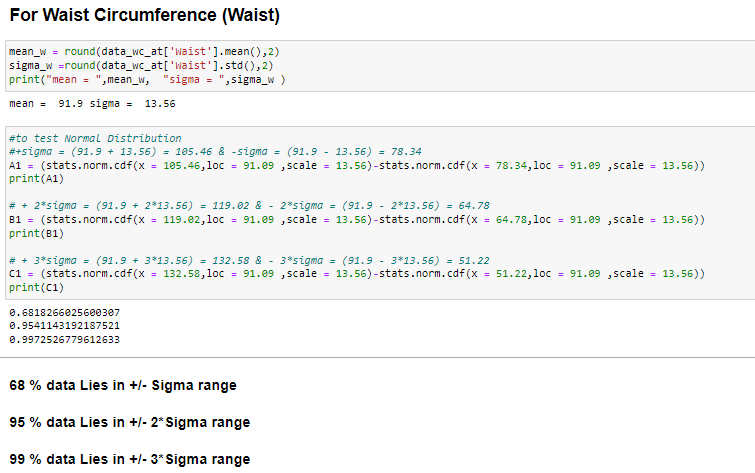


From above we can confirm the MPG of Cars data is Normally Distributed

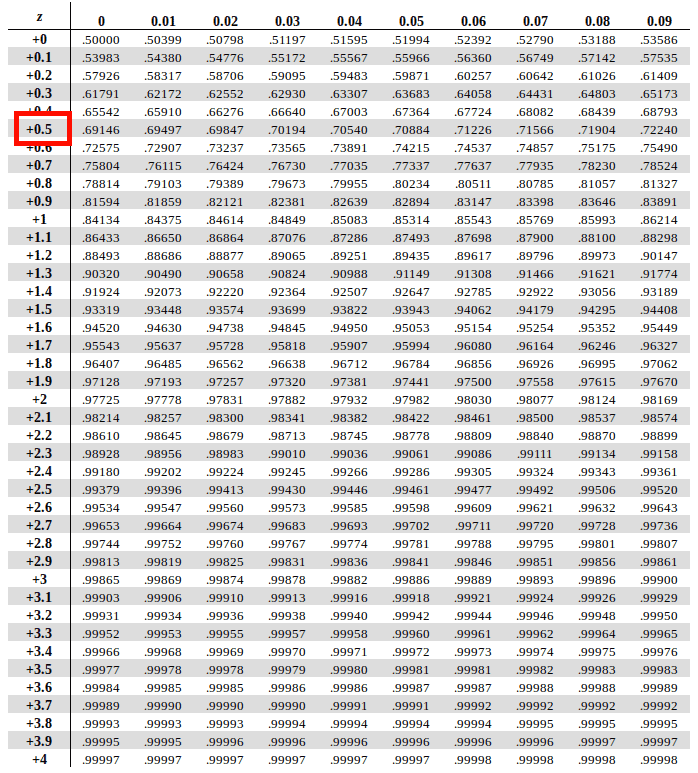
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv





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



Z score for 90 % -

Z score for 90 % lies between 1.28 to 1.29

By interpolating,

(x-1.28)/(0.9-0.89973) = (1.29-1.28)/(0.901474-0.89973)

X = 1.2816 ………. (Z score for 90%)

Z score for 94 % -

Z score for 94 % lies between 1.55 to 1.56

By interpolating,

(x-1.55)/(0.94-0.93943) = (1.55-1.56)/(0.94062-0.93943)

X = 1.5548 ………. (Z score for 94%)

Z score for 60 % -

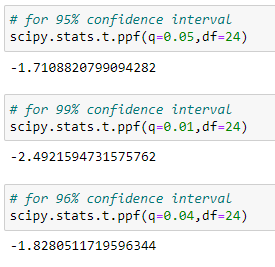
Z score for 60 % lies between 0.25 to 0.26

By interpolating,

(x-0.25)/(0.6-0.59871) = (0.25-0.26)/(0.60257-0.59871)

X = 0.2533 ………. (Z score for 60%)

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



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 – Population Mean = 270 days

Sample mean = 260 days

Sample std deviation = 90 days

Sample size = 18

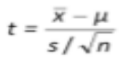
TO FIND – probability for bulbs life more than 260 days

Hypothesis –

Null – the probability that 18 randomly selected bulbs would have an average life of 270 Days

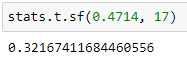
Alternate -the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

From above we can say that distribution is one tailed and right tailed as alternate hypothesis states greater than statement.

t = ((260 -270 )\*sqrt(18)) / 90

T = 0.4714

Degree of Freedom = df = 18-1 = 17



Therefor probability = 0.3217