|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| 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 |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| 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 | Ordinal |
| 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 | Interval |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Nominal |

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

Sol: {H,H,H},{T,T,T},{H,T,T},{H,H,T},{H,T,H},{T,H,H},{T,H,T},{T,T,H}

Probability = Favourable Outcome / Total No. Of Outcome

= 3/8

= 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

Sol: a) Probability = 0/36

= 0

As there are no outcomes which corresponds sum is equal to 1.

b) {1,1},{1,2},{1,3},{2,1},{2,2},{3,1}

Probability = 6/36

= 1/6

= 0.1667

c) {1,5},{2,4},{3,3},{4,2},{5,1},{6,6}

Probability = 6/36

= 1/6

=0.1667

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?

Sol:

Probability =No. of ways of drawing 2 balls out of 5 balls (none is blue = 7-2=5) /

No. of ways of drawing 2 balls out of 7 balls

= 5C2 / 7C2

= 10/21

= 0.476

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

Sol: Expected no. of candies for a randomly selected child

= 1\*0.015 + 4\*0.2 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 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.

**Use Q7.csv file**

Sol: For Points:

1. Mean = 115.09/32

= 3.596

1. Median = 3.695
2. Mode = 3.92
3. Range = 4.93 – 2.76

= 2.17

1. Variance = 0.2858
2. Standard Deviation = 0.5346

For Score:

1. Mean = 102.952/32

= 3.217

1. Median = 3.325
2. Mode = 3.44
3. Range = 5.424 – 1.513

= 3.911

1. Variance = 0.9574
2. Standard Deviation = 0.9785

For Weigh:

1. Mean = 571.16/32

= 17.85

1. Median = 17.71
2. Mode = 17.02
3. Range = 22.9 – 14.5

= 8.4

1. Variance = 3.1392
2. Standard Deviation = 1.7869

Since mean value are close to median for ‘Points’ , ’Score’ and ‘Weigh’ , the data set has symmetrical distribution.

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?

Sol: There are 9 patients.

Probability of 1 patient = 1 / 9

Expected Value = ∑ P(X).E(X)

= 1/9 \* (108 + 110 +123+ 134+ 135+ 145 +167 +187 +199)

= 1308/9

= 145.33

**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**

Sol: Speed:

Skewness = -0.11751 and Kurtosis = - 0.5089

Since skewness is negative , so it is left skewed. Since Kurtosis is negative , the data distribution is flat and light tailed.

Distance:

Skewness = 0.806 and Kurtosis = 0.405

Since skewness is positive , so it is right skewed. Since Kurtosis is positive , the data distribution is peaked and heavy tailed.

SP:

Skewness = 1.611 and Kurtosis = 2.977

Since skewness is positive , so it is right skewed. Since Kurtosis is positive , the data distribution is peaked and heavy tailed.

Weight:

Skewness = -0.6147 and Kurtosis = 0.9502

Since skewness is negative , so it is left skewed. Since Kurtosis is positive , the data distribution is peaked and heavy tailed.

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



From Histogram:

1. The least range of weight is 400 which is about 0-10.
2. Most of the data points are in the range 50-100 and has the frequency of 200
3. From the above plot, we can say that the data is positively or right skewed.



From Boxplot:

1. We can see that there are outliers on the upper side of the box plot.
2. There are 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?

Sol: Confidence Interval (CI) = x̅ ± z\* σ/√n

For 94%:

z = 1.88 n=2000 σ=30 x̅=200

CI = 200 ± 1.88 \* 30/√2000

= 200 ± 1.26

Therefore, we can be 94% confident that the population mean falls within the interval 200 ± 1.26

For 98%:

z = 2.33 n=2000 σ=30 x̅=200

CI = 200 ± 2.33 \* 30/√2000

= 200 ± 1.561

Therefore, we can be 98% confident that the population mean falls within the interval 200 ± 1.561

For 96%:

z = 2.05 n=2000 σ=30 x̅=200

CI = 200 ± 2.05 \* 30/√2000

= 200 ± 1.374

Therefore, we can be 96% confident that the population mean falls within the interval 200 ± 1.374

**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?

Sol: 1) Mean = 738/18

= 41

Median = 40.5

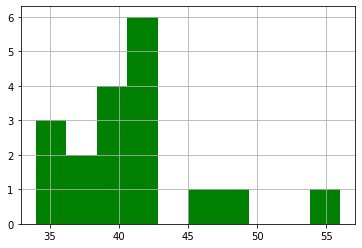
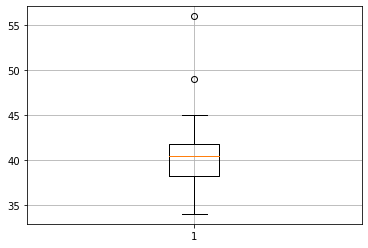
Variance = 25.53

Standard Deviation = 5.05

2) From the below histogram and boxplot, we can say that mean of marks of student is 41 which is slightly greater than median.

Most of the students have marks in between 41-42.

From box plot, we can see that there are two outliers at 49 and 56.

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

Sol: When mean, median of data are equal then there is Zero Skewness i.e. the data distribution is symmetrical.

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

Sol: When mean > median, then the data is Positively Skewed or Right Skewed.

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

Sol: When median > mean, then the data is Negatively Skewed or Left Skewed.

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

Sol: Kurtosis is the degree of presence of outliers in the distribution.

Positive kurtosis value indicates that distribution is peaked and possesses thick tails.

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

Sol: Negative kurtosis indicates that distribution is stretched around center tails means most of the data points are present in high proximity with mean and possesses lower tails.

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)?   
  
Sol: 1) From the above boxplot, we can say that

1. 50% of data is above 10 units.
2. Out of which, 60% of data lies between 10 – 15.
3. 40% of data lies above 15.

2) Data is left skewed since median is greater than mean.

3) IQR = Q3 – Q1

= 18 – 10

= 8 (Approximately).

Q19) Comment on the below Boxplot visualizations?



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

Sol: From both the box plot, mean and median are equal. Hence the distribution is symmetrical. Also we can say that the mean and median values lies at 262.5 (approximately) and the whiskers value of boxplot 2 is higher than boxplot 1.

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)

Sol: a) *μ=34.422 σ = 9.131*

P(MPG>38) = 1-pnorm (38,34.422,9.131)

=0.3475908

b) *μ=34.422 σ = 9.131*

P(MPG<40) = pnorm (40,34.422,9.131)

= 0.7293527

c) *μ=34.422 σ = 9.131*

P(MPG<40) = pnorm (50,34.422,9.131) – (1-pnorm (20,34.422,9.131))

= 0.01311818

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Sol: Mean = 34.422

Median = 35.1527

Since median is greater than mean, data is left skewed.

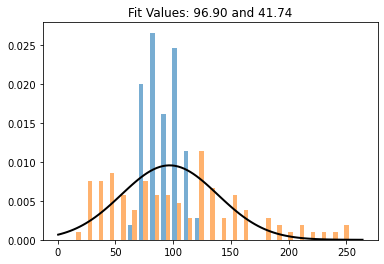
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Sol: Adipose Tissue (AT)

Mean= 101.89

Median = 96.54



Since mean > median in case of AT, data is positively skewed.

Waist Circumference(WC)

Mean = 91.90

Median = 90.8

Since mean is close to median in case of WC, data is symmetrically distributed.

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

Soln: For 90% confidence interval:

We have the significance level at 5 % (as it is a two tailed test)

that is:

α = 5 % = 0.05

z at α = 0.05 from the z table will be:

z = 1.645.

For 94 % confidence interval, we get:

We have the significance level at 3 % (as it is a two tailed test)

that is:

α = 3 % = 0.03

z at α = 0.03 from the z table will be:

z = 1.555.

For 60 % confidence interval, we get:

We have the significance level at 20 % (as it is a two tailed test)

that is:

α =20 % = 0.2

z at α = 0.2 from the z table will be:

z = 0.253

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

Soln: From Python code, we can get

# t scores of 95% confidence interval for sample size of 25 is

stats**.**t**.**ppf(0.975,24) (df = n-1 = 24)

= 2.0638985616280205

# t scores of 96% confidence interval for sample size of 25

stats**.**t**.**ppf(0.98,24)

= 2.1715446760080677

# t scores of 99% confidence interval for sample size of 25

stats**.**t**.**ppf(0.995,24)

=2.796939504772804

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:

Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

t-scores at x=260;

t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

t**=** (260**-**270)**/**(90**/**18**\*\***0.5)

t= -0.4714045207910317

p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function

p\_value**=**1**-**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)

p\_value= 0.32167411684460556

Since p\_value > 0.05, we say we failed to reject the null hypothesis. So we can say that the average life of bulb is greater than or equal to 260 days.