|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continous |
| Weight of Gold | Continous |
| Distance between two places | Continous |
| Length of a leaf | Continous |
| Dog's weight | Continous |
| 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 | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Ordinal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

Sample space S=(HHH,HHT,HTH,THH,THT,TTH,HTT,TTT)

n(S)=8

Event A=Two heads and one tail are obtained

A=(HHT,HTH,THH) n(A)=3

P(A)=n(A)/n(S)= 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
4. There is no any outcome having sum is equal to 1.

So probability of that event should be 0

1. Consider sample space A={(1,1), (1,2), (1,3), (2,1), (2,2),(3,1)}

So, P(A)=6/36 = 1/6

1. Consider Sample space={(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}

So ,P(A)=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 balls in the bag=2+3+2=7 balls

We have to draw 2 balls randomly so this can be done by 7!/2!\*(7-2)! = 21 ways

Now None ball is blue so except that 2 blue balls remaining balls are 5 so they can be selected by 5!/2!\*(5-2)! = 10 ways

Therefore total no of ways are to find none of the ball is blue is = 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

Expected number

E(x)= 1(0.015) + 4(0.20) + 3(0.65) + 5(0.005) + 6(0.01) + 2(0.120)

= 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. Open that file into Excel and apply formulas on separate columns to find mean, median, mode, SD, Variance , Range

For Mean (=AVERAGE(select point column)) + enter

For Median (=Median(select point column)) + enter etc.

**Use Q7.csv file**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | SD | Varience | Range |
| Point | 3.596563 | 3.695 | 3.92 | 0.534679 | 0.285881 | 2.17 |
| Score | 3.21725 | 3.325 | 3.44 | 0.978457 | 0.957379 | 3.911 |
| Weight | 17.84875 | 17.71 | 17.02 | 1.786943 | 3.193166 | 8.4 |

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?

|  |  |  |
| --- | --- | --- |
| x | Nu. of repetitions | P(x) |
| 108 | 1 | 1/9 |
| 110 | 1 | 1/9 |
| 123 | 1 | 1/9 |
| 134 | 1 | 1/9 |
| 135 | 1 | 1/9 |
| 145 | 1 | 1/9 |
| 167 | 1 | 1/9 |
| 187 | 1 | 1/9 |
| 199 | 1 | 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)

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

= 145.3333

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANS. Download file into Excel**

**Apply formulas on separate columns to find skewness and kurtoisis**

**For skewness (=SKEW**(select point column)) + enter)

**For kurtoisis (=KURT**(select point column)) + enter)

|  |  |  |
| --- | --- | --- |
|  | SKEWNESS | KURTOISIS |
| Speed | -0.11395477 | -0.50899 |
| Distance | 0.78248352 | 0.405053 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | SKEWNESS | KURTOSIS |
| SP | 1.581453679 | 2.977328944 |
| WT | -0.603309932 | 0.950291491 |

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



ANS…  **It represents Positive Skewness**



**ANS….. It represents Positive Skewness**

**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?

Consider, Confidence interval=CI

Given, Sample Mean x=200,

SD s=30,

Sample size N=2000

1. CI = stats**.**norm**.**interval(0.94,200,30**/**(2000**\*\***0.5))

print("Confidence Interval at 94% is:",CI))………..(python code)

**Confidence Interval at 94% is: [198.73, 201.27]**

1. CI = stats**.**norm**.**interval(0.98,200,30**/**(2000**\*\***0.5))

print("Confidence Interval at 98% is:",CI)………….(python code)

**Confidence Interval at 98% is: [198.61 201.39]**

In [ ]:

3) CI = stats**.**norm**.**interval(0.96,200,30**/**(2000**\*\***0.5))

print("Confidence Interval at 96% is:",CI)…………(python code)

**Confidence Interval at 96% is: [198.43 201.57]**

In [ ]:

**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?

|  |  |
| --- | --- |
| Mean | 41 |
| Median | 40.5 |
| Varience | 25.52941 |
| SD | 4.910307 |

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

Skewness is 0

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

Positively Skewed

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

Negatively Skewed

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

Distribution is peaked and possesses thick tail

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

Distribution is flat and has thin tail

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



What can we say about the distribution of the data?

**From figure we can see that most of the data can be distributed towards left side of that box plotted.**

What is nature of skewness of the data?

**It represents Negative Skewness**

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.

1. **It represents Positive Skewness**
2. **It represents Symmetric data**

**In Boxplot 2 data is distributed in more region (between 225 to 300) but in boxplot 1 data is distributed in less region (between 250 to 275)**

Draw a

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)

a) 1 - stats.norm.cdf(38,loc=34.422075728024666,scale=9.131444731795982)

**ANS. 0.3475939251582705**

b)stats.norm.cdf(40,loc=34.422075728024666,scale=9.131444731795982)

**ANS 0.7293498762151616**

In [ ]:

c) (stats.norm.cdf(50,loc=34.422075728024666,scale=9.131444731795982)) - (stats.norm.cdf(20,loc=34.422075728024666,scale=9.131444731795982))

**ANS 0.8988689169682046**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Given data follows Normal Distribution if mean=median=mode

**For given data, mean=34.422075728024666**

**Median=35.15272697**

**Mode= 29.629936**

In [16]:

Mean, median, mode are not equal so it is not normal distribution.

In [7]:

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

Dataset: wc-at.csv

For AT

**Mean = 101.89403669724771**

**Median = 96.54**

**Mode = 121.0 and 123.0**

Mean, median, mode are not equal so it is not normal distribution.

For Waist

**Mean = 91.90183486238533**

**Median = 90.8**

**Mode = 94.5, 106.0, 108.5**

Mean, median, mode are not equal so it is not normal distribution.

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

|  |  |  |
| --- | --- | --- |
|  | **from scipy import stats** | **Z score** |
| **90%** | **stats.norm.ppf(0.95)** | **1.6448** |
| **94%** | **stats.norm.ppf(0.97)** | **1.88079** |
| **60%** | **stats.norm.ppf(0.80)** | **0.8416** |

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

|  |  |  |
| --- | --- | --- |
|  | **From scipy import stats** | **T score** |
| **95%** | **stats.t.ppf(0.975,df=24)** | **2.0638** |
| **96%** | **stats.t.ppf(0.98,df=24)** | **2.1715** |
| **99%** | **stats.t.ppf(0.995,df=24)** | **2.7969** |

In [39]:

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

Given, N=270, n=18, x=260, SD=90

t =( x-N)/(s/sqrt(n)) = (260-270)/(90/sqrt18) = -0.4714045208

so average life of no more than 260 days IS

1 - stats.t.cdf(0.4714045208,df=17) …………..(python code)

= 1 - **0.6783274643321531**

**= 0.32167**

**ANS. 0.32167**

In [ ]: