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

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 | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ordinal |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans:-

Sample Space are

X={HHH,THH,HTH,HHT,TTH,THT,HTT,TTT}

* The probability that two heads and one tail are 3/8 or 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

Ans:-

Sample Space are

X={HH,HT,TH,TT}

|  |  |
| --- | --- |
| Out Come | Probability |
| 0 | 1/4 |
| 1 | 1/2 |
| 2 | 1/4 |

Two dice thrown=36

1. Zero
2. Less than or equal to 4 is(1,3),(2,2),(3,1) and n(b)=3/36 or 1/12
3. Sum is divisible by 2 is 1

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:- There are 7 balls originally with 2 of them blue so the probability of the first ball not being blue is 5/7. This leaves 6 balls with 2 blue. The probability of the second ball not being blue assuming that the first wasn't is 4/6. The probability that neither ball drawn was blue is

**(5/7)\*(4/6)=20/42**

**=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

Ans:-

Total Candies are 21

Expected Value

E(x) = ∑xi.P(x=xi) = 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120

E(x) = 3.09

So we say that 2100 Candies for 309 Childrens.

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**

**Ans:-**

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **mode** | 3.07, 3.92 | 3.44 | 17.02, 18.90 |
| **mean** | 3.596563 | 3.217250 | 17.848750 |
| **std** | 0.534679 | 0.978457 | 1.786943 |
| **var** | 0.285881 | 0.957379 | 3.193166 |
| **median** | 3.695000 | 3.325000 | 17.710000 |
| **range** | 2.17 | 3.911 | 8.4 |
|  |  |  |  |

Inference:- The value of mean and median are nearly equal so we say that the cure is nearly **symmetric.**

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:-

Total Patients are 9

So, Probability of selecting each patient is 1/9

Expected Value

E(x) = ∑xi.P(x=xi) = (1/9)\*(108+110+123+134+135+145+167+187+199)

E(x)= 145.33

Expected Value of the weight of that patient is 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Ans:-

Skewness

Speed:- -0.117510

distance:- 0.806895

Kurtosis

Speed:- -0.508994

Dist:- 0.405053

Fromthe graph speed are linearly increasingbut in distance there are lots of variances.

**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans:-

Skewness

speed -0.117510

dist 0.806895

Kurtosis

speed -0.508994

dist 0.405053

Fromthe graph SP are linearly increasingbut in WT there are lots of variances.

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



Ans:- From the above histrogram we conclude that in data many or more than one **outlier**s are present. It no symmetric or normal Distribution and maximum points lies between 50-100 range.



Ans:- From the above Boxplot or Wisker plot we conclude that in data many **outliers** are present.In given Boxplot mean and median are not same.

**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:-

Given

n=2000, xbar=200, sigma=30

For 94% of Confidence interval

[198.73 , 201.26]

For 96% of Confidence interval

[198.62 , 201.37]

For 98% of Confidence interval

[198.43 , 201.56]

**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?

Ans:-



|  |  |
| --- | --- |
| **var** | 24.1 |
| **mean** | 41 |
| **std** | 4.9 |
| **median** | 40.5 |
|  |  |

1. Mean and median is similar so we say marks are equally distributed and 4 students get 41 marks so it’s a mode.

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

Ans:-Symmetric (Bell Curve) or Normal Curve.

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

Ans:- Positive or Right Tail Skewness

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

Ans:- Negative or Left Tail Skewness

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

Ans:- Positive kurtosis is mean the curve is more peaked and its Leptokurtic. Large number of sample point present in data which is greature than median.

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

Ans:- Negative kurtosis means the curve will be flatter and broader. Large number of sample point present in data which is less than median.

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



What can we say about the distribution of the data?

Ans:- The above Boxplot is not normally distributed the median is towards the higher value. Median of data is 15(appro.) and Asymmetric.

What is nature of skewness of the data?

Ans: The data is a skewed towards left. The whisker range of minimum value is greater than maximum .Negative or Left Tail Skewness.

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

Ans:-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:- First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range. In above graph, two Boxplot are drawn of differents data points. Median of both Boxplot is same 265(appro.) . In 1st Boxplot the data set or sample are less than 2nd data set but both are symmetric . so we said that mean and median are same value.

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

Code:- stats.norm.cdf(n,mean(),std())

* 1. P(MPG>38) = 0.348
  2. P(MPG<40)= 0.729

c. P (20<MPG<50)= 0.013

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:- mean =34.422076

median= 35.152727

mode=29.62

All values are nearly equal so MPG of cars follow 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:-

Adipose Tissue (AT)

Mean = 91.901835, median= 90.800000, mode= 94.5

Waist Circumference(Waist)

Mean = 101.894037, median= 96.540000, mode= 120

Adipose Tissue (AT) and Waist Circumference(Waist) does not follow Normal distribution.

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

Ans:

Code: stats.norm.ppf()

Z score for 90% confidence interval is 1.645

Z score for 94% confidence interval is 1.88

Z score for 60% confidence interval is 0.84

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

Ans:-

t score for 95% confidence interval is 2.064

t score for 96% confidence interval is 2.11

t score for 99% confidence interval is 2.796

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:-

Given

mu=270, SD=90, n=18, xbar=260

df=n-1=18-1=17

Code:- t\_score=(x-pop mean)/(sample std/sqrt of sample size))

(260-270)/(90/np.sqrt(18))

t\_score=-0.4714045207910317

stats.t.cdf(-0.4714045207910317,17)

=0.32

=32%