Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| 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 |
|  |  |

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) | Ratio |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nomina |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

ANS: Timesthree coins are Tossed.

(HHH),(TTT),(TTH),(THT),(THH),(HHT),(HTH),(HTT)

Therefore Total No Of Events (N)=8

Condition Is Two Heads And One Tail

(THH),(HHT),(HTH)

Interested Events = 3

probability=3/8

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

* Equal to 1
* ANS:ZERO

Is There is no Case That’s Sum IS= 1

* Less than or equal to 4

ANS:Possible cases are (1,1)(1,2)(1,3)(2,1)(2,2)(3,1)

Total Interesed Events 6

Probability : 6/36=1/6=0.16666

16.66%

* Sum is divisible by 2 and 3

ANS:Possible cases are (1,5)(2,4)(3.3)(4,2)(5,1)(6,6)

Total Interested Events = 6

Probability : 6/36=1/6=0.1666

16.66%

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

THEN NUMBER OF WAY OF DRAWING 2 BALLS OUT OF OF 7

7C2 =7\*6/2\*1=42/2=21

the probability that none of the balls drawn is blue=5C2=5\*4/2\*1=20/2=10

10/21=0.47

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 = x\*P(x)=3.09

Expected number of candies = (1\*0.015) + (4\*0.2) + (3\*0.65) + (5\*0.005) +

(6\*0.01) + (2\*0.12)

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**: **POINTS**

**MEAN**:R Code

> mean(Q7\_csv$Points)

[1] 3.596563

**MEDIAN**:> > median(Q7\_csv$Points)

[1] 3.695

**variance**:> > var(Q7\_csv$Points)

[1] 0.2858814

**Mode:> #Mode**

**> nummode<-function(X){**

+ uniquv<-unique(X)

**+ uniquv[which.max(tabulate(match(X,uniquv)))]}**

**> mode(Q7\_csv$Points)**

**[1] "numeric"**

**> nummode(Q7\_csv$Points)**

**[1] 3.92**

**STANDERD DEVIATION**> sd(Q7\_csv$Points)

[1] 0.5346787

**RANGE:> range(Q7\_csv$Points)**

**[1] 2.76 4.93,> rangevalue<-function(X){max(X)-min(X)}**

**> rangevalue(Q7\_csv$Points)**

**[1] 2.17**

**WEIGH:-**

**MEAN:-> mean(Q7\_csv$Weigh)**

**[1] 17.84875**

**MEDIAN:**> median(Q7\_csv$Weigh)

[1] 17.71

**MODE:**> nummode(Q7\_csv$Weigh)

[1] 17.02

**RANGE:**> range(Q7\_csv$Weigh)

[1] 14.5 22.9

**SCORE:-**

**MEAN:> mean(Q7\_csv$Score)**

**[1] 3.21725**

**MEDIAN:**

**> median(Q7\_csv$Score)**

**[1] 3.325**

**MODE:**

**> nummode(Q7\_csv$Score)**

**[1] 3.44**

**RANGE:**

**> range(Q7\_csv$Score)**

**[1] 1.513 5.42**

**PYTHON CODES:-Q7['Points'].median**

**Out[13]:**

**<bound method Series.median of 0 3.90**

**Q7['Score'].median**

**Out[14]:**

**<bound method Series.median of 0 2.620**

**Q7['Weigh'].median**

**Out[15]:**

**<bound method Series.median of 0 16.46**

**Q7['Score'].mode**

**Out[17]:**

**<bound method Series.mode of 0 2.620**

**Q7['Weigh'].mode**

**Out[18]:**

**<bound method Series.mode of 0 16.46**

**Q7['Points'].var()**

**Out[19]: 0.28588135080645166**

**Q7['Score'].var()**

**Out[20]: 0.9573789677419354**

**Q7['Weigh'].var()**

**Out[21]: 3.193166129032258**

**Q7['Points'].std**

**Out[24]:**

**<bound method Series.std of 0 3.90**

**Q7['Score'].std**

**Out[25]:**

**<bound method Series.std of 0 2.620**

**Q7['Weigh'].std**

**Out[26]:**

**<bound method Series.std of 0 16.46**

**Q7['Points'].var()**

**Out[27]: 0.28588135080645166**

**Q7['Score'].std()**

**Out[28]: 0.9784574429896966**

**# calculating the population standard deviation and variance**

**np.var(Q7.Points) # population variance**

**Traceback (most recent call last)**

**import numpy as np**

**np.var(Q7.Points) # population variance**

**Out[32]: 0.27694755859375003**

**np.std(Q7.gmat) # population standard deviation**

**Traceback (most recent call last):**

**np.std(Q7.Points) # population standard deviation**

**Out[34]: 0.5262580722361891**

**np.var(Q7['Points'])**

**Out[35]: 0.27694755859375003**

**np.std(Q7['Score'])**

**Out[36]: 0.9630477013107918**

**np.var(Q7['Weigh'])**

**Out[37]: 3.0933796874999997**

**np.var(Q7['Score'])**

**Out[38]: 0.927460875**

**range = max(Q7['Points'])-min(Q7['Points']) # max(mba.gmat)-min(mba.gmat)**

**range**

**Out[40]: 2.17**

**Q7['Points'].std**

**Out[41]:**

**<bound method Series.std of 0 3.90**

**np.std(Q7['Score'])**

**Out[42]: 0.9630477013107918**



Q8) Calculate Expected Value for the problem below

* 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**:Expected Value=(108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)/9 = 145.333. It is not mandatory for the expected value to be present in sample space. As calculated above Expected Value of 145.33 is not present in our sample space

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

**Cars speed and distance**

**ANS:-**> skewness(Q9\_csv$Index)

[1] 0

> skewness(Q9\_csv$speed)

[1] -0.1139548

> skewness(Q9\_csv$dist)

[1] 0.7824835

> kurtosis(Q9\_csv$Index)

[1] 1.79904

> kurtosis(Q9\_csv$speed)

[1] 2.422853

> kurtosis(Q9\_csv$dist)

[1] 3.248019

>

PYTHON:-Q9\_csv.skew()

Out[9]:

Index 0.000000

speed -0.117510

dist 0.806895

Q9\_csv.kurt()

Out[10]:

Index -1.200000

speed -0.508994

dist 0.405053



**ANS:> skewness(Q9\_b$SP)**

**[1] 1.581454**

**> kurtosis(Q9\_b$WT)**

**[1] 3.819466**

**Left Skewed for SP and positive kurtosisb)Left Skewed for WT and Negative kurtosis**



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



**Ans :**  1) The most of the points lies between 50-100 range having maximum frequency 200.

2)positive skewness because long tail is towards right side.



* Negatively Skewed data ,because long tail is towards left side.
* Median < Mean , and outlier is on upper side of box plot.
* Less no. of points lies 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 ?

**ANS:- FORMULA:X+/- Z(1-α) \*(σ/√n)**

Degrees of freedom= 2000-1=1999

Confidence Interval= **94% = 1.89**

Confidence Interval= **98% = 2.33**

Confidence Interval= **96% = 2.06**

**#to calculate z score**

**> qnorm(0.970)#94%**

**[1] 1.880794**

**> qnorm(0.990)#98%**

**[1] 2.326348**

**> qnorm(0.980)#96%**

**[1] 2.053749**

**>**

**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**

* Find mean, median, variance, standard deviation.
* What can we say about the student marks?

**ANS:-**In Python

p = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

**Mean**:

np.mean(p)

Out[23]: 41.0

**Median:**

np.median(p)

Out[24]: 40.5

**Varience:**

np.var(p)

Out[25]: 24.11111111111111

**Standard Deviation :**

np.std(p)

Out[26]: 4.910306620885412

Students gets marks between 34 and 56.

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

**ANS:perfectly symmetrical distribution**

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

**ANS:-If the mean is greater than the median, the distribution is positively skewed**

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

**ANS:If the mean is less than the median, the distribution is negatively skewed.**

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

**ANS:A distribution with a positive kurtosis value indicates that the distribution has heavier tails and a sharper peak than the normal distribution.**

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

**ANS:A distribution with a negative kurtosis value indicates that the distribution has lighter tails and a flatter peak than the normal distribution**

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)?

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

**ANS:-**The above box plot represents the information about Ages of students in School.

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

50% of students are above 10 yrs of old and remaining are less. And the students having age above 15yrs are approximately 40%.

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

Nature of skewness is negative because long tail towards left side.

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

Lower Quartile(Q1)=10, Upper Quartile(Q3)=18

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.

**Ans:**

In both Boxplots, Mean, Median and Mode are same and hence the distribution is

Symmetrical in nature.

the only difference is that Boxplot 2 having high Wisker level as compare toBoxplot 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

* P(MPG>38)
* P(MPG<40)

c. P (20<MPG<50)

**ANS:> (MPG>38)**

**[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE**

**[14] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE**

**[27] TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE**

**[40] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE**

**[53] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE**

**[66] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE**

**[79] FALSE FALSE FALSE**

**> (MPG<40)**

**[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE**

**[14] FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE TRUE FALSE TRUE TRUE**

**[27] TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE**

**[40] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE**

**[53] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE**

**[66] FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE**

**[79] TRUE TRUE TRUE**

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

ANS: It is not normal distribution

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

Dataset: wc-at.csv

ANS:-NOT FOLLOWS.

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

to calculate z score

> qnorm(0.950)#90%

[1] 1.644854

> qnorm(0.970)#94%

[1] 1.880794

> qnorm(0.8)#60%

[1] 0.8416212

**ANS**: At 90% z score is 1.65, at 94% z = 1.88, at 60% z = 0.85

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

**Ans**: At 95% T score = 2.060, at 96% t score is almost 2.2, at 99% t score = 2.787

> qt(0.975,24)=95%

[1] 2.063899

> qt(0.98,24)=96%

[1] 2.171545

> qt(0.995,24)=99%

[1] 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