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

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

If three coins are tossed,

Total number of possible combinations = 23 = 8

The combinations are HHH, HHT, HTH, THH, TTH, THT, HTT, TTT.

Number of combinations that have two heads and one tail = 3, i.e., HHT, HTH, TTH

The probability of two heads and one tail when three coins are tossed simultaneously are

P (Two heads and One tail) = Number of desired outcomes

= ⅜ or 0.375

**ANS - probability of two heads and one tail is ⅜ 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

Equal to 1 =**0**

Equal to 4 = (1, 3) (2,2) (3,1)

**ANS** - =3/36=**1/12**

Favorable outcomes = sum is divisible by 2 and 3

Sum should be divisible by both 2 and 3

Favorable outcomes = (1, 5), (3, 3), (4, 2), (5, 1), (6, 6)

Therefore,

Number of favorable outcomes = 5

**ANS** - Thus, the probability that sum is divisible by **2 and 3 is **

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?

Sample space =7c2 = 21. Let event E be none of the balls is blue = all balls are either red or green or both. n(E) = 5c2 =10 p(E) =**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 of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

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

=       3.090

**ANS** - = **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**

**(ANS) Points:**

**Mean: 3.32**

**Median: 3.31**

**Mode: 3.07 and 3.92 (both occur twice)**

**Variance: 0.2859**

**Standard Deviation: 0.5347**

**Range: 2.17 (from 2.76 to 4.93)**

**Score:**

**Mean: 3.22**

**Median: 3.215**

**Mode: There is no mode as no value occurs more than once.**

**Variance: 0.2673**

**Standard Deviation: 0.5173**

**Range: 3.21 (from 1.513 to 5.424)**

**Weigh:**

**Mean: 20.09**

**Median: 19.2**

**Mode: There is no mode as no value occurs more than once.**

**Variance: 60.73**

**Standard Deviation: 7.8**

**Range: 10.61 (from 14.5 to 25.11)**

**In terms of Points, the dataset seems to be relatively normally distributed with a mean of 3.32 and a standard deviation of 0.53. The range of points is from 2.76 to 4.93, with a difference of 2.17 between the minimum and maximum values. The mode is not very useful in this case since there are multiple modes with equal frequency.**

**For Score, the distribution is also relatively normal, with a mean of 3.22 and a standard deviation of 0.52. The range of Score is from 1.513 to 5.424, with a difference of 3.21 between the minimum and maximum values. There is no mode in this case as no value occurs more than once.**

**For Weigh, the distribution is not normal, and it is skewed to the right with a mean of 20.09 and a standard deviation of 7.8. The range of Weigh is from 14.5 to 25.11, with a difference of 10.61 between the minimum and maximum values. There is no mode in this case as no value occurs more than once.**

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?

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

**Cars speed and distance**

**Use Q9\_a.csv**

**(ANS) Mean speed = (4+4+7+7+8+9+10+10+10+11+11+12+12+12+12+13+13+13+13+14+14+14+14+15+15+15+16+16+17+17+17+18+18+18+18+19+19+19+20+20+20+20+20+22+23+24+24+24+24+25)/50 = 16.96**

**Mean distance = (2+10+4+22+16+10+18+26+34+17+28+14+20+24+28+26+34+34+46+26+36+60+80+20+26+54+32+40+32+40+50+42+56+76+84+36+46+68+32+48+52+56+64+66+54+70+92+93+120+85)/50 = 42.98**

**Standard deviation speed = 4.147**

**Standard deviation distance = 25.77**

**Skewness speed = -0.117**

**Skewness distance = 0.782**

**Kurtosis speed = -0.508**

**Kurtosis distance = 0.307**

**SP and Weight (WT)**

**Use Q9\_b.csv**

**import numpy as np**

**from scipy.stats import skew, kurtosis**

**data = np.array([[1, 104.1853528, 28.7620589],**

**[2, 105.4612635, 30.46683298],**

**[3, 105.4612635, 30.19359657],**

**[4, 113.4612635, 30.63211391],**

**[5, 104.4612635, 29.88914864],**

**[6, 113.1853528, 29.59176832],**

**[7, 105.4612635, 30.30847957],**

**[8, 102.5985128, 15.84775807],**

**[9, 102.5985128, 16.35948352],**

**[10, 115.6452041, 30.92015417],**

**[11, 111.1853528, 29.36334142],**

**[12, 117.5985128, 15.75353468],**

**[13, 122.1050553, 32.81359241],**

**[14, 111.1853528, 29.3784363],**

**[15, 108.1853528, 29.34727902],**

**[16, 111.1853528, 29.60452658],**

**[17, 114.3692933, 29.5357836],**

**[18, 117.5985128, 16.19412154],**

**[19, 114.3692933, 29.92939368],**

**[20, 118.4729364, 33.51697417],**

**[21, 119.1050553, 32.32464971],**

**[22, 110.8408174, 34.90821127],**

**[23, 120.2889958, 32.6758277],**

**[24, 113.8291446, 31.83712236],**

**[25, 119.1853528, 28.78172789],**

**[26, 114.5985128, 16.04317492],**

**[27, 120.7605199, 38.06282335],**

**[28, 119.1050553, 32.83506939],**

**[29, 99.56490661, 34.4832075],**

**[30, 121.8408174, 35.54935984],**

**[31, 113.4846092, 37.04235003],**

**[32, 112.2889958, 33.23436141],**

**[33, 119.9211148, 31.38004084],**

**[34, 121.3926389, 37.57328965],**

**[35, 111.2889958, 32.701644],**

**[36, 115.0130851, 31.9112234],**

**[37, 114.0933825, 28.75400008],**

**[38, 116.909442, 27.87991549],**

**[39, 116.909442, 28.63050247],**

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



**ANS - The histogram has a right skew and tail is on the Right-Side Mean > Median. We have outliers on the higher side.**

**The Boxplot has outliers on the maximum side.**

**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** - **conf\_94 =stats.t. interval (alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt (2000)) print (np. round(conf\_94,0)) print(conf\_94) For 94% confidence interval Range is [ 198.73 – 201.26] For 98% confidence interval range is [198.43 – 201.56] For 96% confidence interval range is [198.62 – 201.37]**

**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** - **Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05**

1. What can we say about the student marks?

**ANS** - We don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

**ANS** - **No skewness is present we have a perfect symmetrical distribution**

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

**ANS – Skewness and Tail is towards Right side**

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

**ANS – Skewness and tail is towards left**

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

**ANS - Positive kurtosis means the curve is more peaked and it is Leptokurtic**

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

**ANS** - **Negative Kurtosis means the curve will be flatter and broader**

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**

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**

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

**ANS - The Inter Quantile Range = Q3 Upper quartile – Q1 Lower Quartile = 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.**

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)

**ANS – 0.348**

* 1. P(MPG<40)

**ANS – 0.729**

* 1. P (20<MPG<50)

**ANS – prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20)) p(MPG>20)= 0.943**

**prob\_MPG\_less\_than\_50 = np. round (stats.norm.cdf (50, loc = q20.MPG.mean(), scale = q20.MPG.std ()),3) print('P(MPG<50)=',(prob\_MPG\_less\_than\_50)) P(MPG<50)= 0.956**

**prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20) print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50)) P(20<MPG<50)= 0.013000000000000012**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANS - a.) MPG of cars follows 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) and Waist does not follow Normal Distribution**

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

**ANS - print ('Z score for 60% Confidence Interval = ‘, np. round (stats.norm.ppf (.05),4)) Z score for 60% Confidence Interval = -1.6449**

**ANS - print ('Z score for 60% Confidence Interval =',np.round(stats.norm.ppf(.03),4)) Z score for 60% Confidence Interval = -1.8808**

**ANS - print('Z score for 60% confidence interval =',np.round(stats.norm.ppf(.2),4)) Z score for 60% confidence interval = -0.8416**

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

**ANS - print ('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4)) T score for 95% Confidence Interval = -2.0639**

**ANS - print ('T score for 96% Confidence Interval =',np.round(stats.t.ppf(0.03,df=24),4)) T score for 94% Confidence Interval = -1.974**

**ANS - print ('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4)) T score for 95% Confidence Interval = -2.7969**

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 - import numpy as np**

**import scipy as stats t\_score = (x - pop mean) / (sample standard deviation / square root of sample size) (260-270) /90/np.sqrt (18)) t\_score = -0.471 stats.t.cdf (t\_score, df = 17) 0.32 = 32%**