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

**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 scale |
| Hair Color | nominal |
| Socioeconomic Status | ordinal |
| Fahrenheit Temperature | Interval scale |
| Height | Ratio scale |
| Type of living accommodation | nominal |
| Level of Agreement | ordinal |
| IQ (Intelligence Scale) | Ratio scale |
| Sales Figures | Ratio scale |
| Blood Group | nominal |
| Time Of Day | ordinal |
| Time on a Clock with Hands | interval |
| Number of Children | nominal |
| 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?**

Ans: - {HHH, HHT, HTT, THH, TTH, TTT, HTH, THT, }

S = 8 and N =3

Probability of two heads and one tail = **P = 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

Ans:- A = no. probability Equal to 1 ,Thus N(a)= 0

B= no. probability Less than or equal to 4 ,Thus N(b) = 6

C = no. probability Sum is divisible by 2 and 3 , Thus N(c) = 6

{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),

(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),

(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),

(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6),

(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),

(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)} Hence,S= 36

probability of A: N(a)/ S = 0/36 = 0

probability of B: N(b)/ S = 6/36 = 1/6

probability of C: N(c)/ S = 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?**

**Answer:-**

Total number of balls =7 balls

No of Red balls = N (R) =2

No of Green balls = N(G) = 3

No of Blue balls = N(B) = 2

Probability of red balls = N(R)/S = 2/7

Probability of Green balls = N(G)/S = 3/7

a) probability that two balls are drawn at random =

b) probability that none of the balls drawn is blue = (2+3)/7 = 5/7

**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:** 1\*0.015+4\*0.20+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**

**Answer:-**

**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?**

**Answer :** Expected value = summation of [X\* P(X)]

As result, the following is the expected value of a patient's weight:

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

Rounding to two decimal places, E(X) equals 144.33pounds.

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

**Cars speed and distance : skewness Use Q9\_a (1).csv**

**Answer :-**

from scipy.stats import skew

print(skew(x, axis=0, bias=True))

[ 0. -0.11395477 0.78248352]

from scipy.stats import kurtosis

print(kurtosis(x, axis=0, bias=True))

[-1.20096038 -0.57714742 0.24801866]

**For car speed**

**skewness value= -0.11 and Kurtosis value= 0.57**

**For car distance**

**skewness value= -0.78 and Kurtosis value= 0.24**

**SP and Weight(WT) Use Q9\_b.csv**

**Answer :-**

from scipy.stats import skew

print(skew(q, axis=0, bias=True))

[ 0. 1.58145368 -0.60330993]

from scipy.stats import kurtosis

(kurtosis(x, axis=0, bias=True))

[-1.20096038 -0.57714742 0.24801866]

**For car sp**

**skewness value = 1.58 and Kurtosis value = -0.57**

**For car wt**

**skewness value = -0.60 and Kurtosis value = 0.24**

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



The histograms peak has right skew and tail is on right. 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?

Answer:

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?**

As mean and median are equal there is zero skew.

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

Right skew: mean > median

A right-skewed distribution has a long tail on its right side.

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

Left skew: mean < median

A left-skewed distribution has a long tail on its left side.

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

Positive kurtosis value indicates that a data distribution has heavy tails and a peaked center compared to a normal distribution (which has a kurtosis of 0).

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

Negative kurtosis value indicates that the distribution of a dataset has thinner tails and a flatter peak compared to the normal distribution

Negative kurtosis between -1 and -2: The distribution has moderately thinner tails and a moderately flatter peak than the normal distribution

Negative kurtosis less than -2: The distribution has very thin tails and an extremely flat peak compared to the normal distribution.

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?**

The nature of skewness is Left Skew.

**What will be the IQR of the data (approximately)?**IQR = Q3 - Q1

Where:

* Q1 is the value below which 25% of the data falls.
* Q3 is the value below which 75% of the data falls

IQR = Q3 – Q1

= 4.5 -2.5 = 1.5 (approx..)

**Q19) Comment on the below Boxplot visualizations?**



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

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

Ans: Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)

# P(MPG>38)

stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

0.6524060748417286

* 1. P(MPG<40)

# P(MPG<40)

stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

0.7293498762151609

* 1. P (20<MPG<50)

# P (20<MPG<50)

stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())

1.2430968797327491e-05

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

Ans: 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: # Z-score of 90% confidence interval

stats.norm.ppf(0.95)

1.6448536269514722

# Z-score of 94% confidence interval

stats.norm.ppf(0.97)

1.8807936081512509

# Z-score of 60% confidence interval

stats.norm.ppf(0.8)

0.8416212335729143

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

**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:**

*# find 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

*# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function*

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

p\_value

0.32167411684460556