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

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 | Ratio |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Ratio |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal (e.g. Morning -> afternoon ->) |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans : Total Possible Events : 8 , No of desired events: 3

P = 3/8

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

1. Equal to 1 Ans : 0
2. Less than or equal to 4 Ans : 1/6
3. Sum is divisible by 2 and 3 Ans : 5/36

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 : 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 : 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:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| MEAN | 3.59375 | 3.21725 | 17.84875 |
| MEDIAN | 3.695 | 3.325 | 17.71 |
| MODE | 3.92 | 3.44 | 17.02 |
| Variance | 0.2858814 | 0.957379 | 3.193166 |
| Std. Dev | 0.534678 | 0.978457 | 1.786943 |
| Range | 2.17 | 3.911 | 8.4 |

* In none of the variable Mean = Median = Mode so we can say.
* Two of the columns i.e. “Score” and “Weigh”in the data set contains outliers.
* In “Points” and “Scores” Mean<Median<Mode so we can say that the data may be negatively skewed. In “Weigh” Mean>Median>Mode so we can say that the data is positively skewed.

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: 145.333

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

**Cars speed and distance**



Ans : Skewness Kurtosis

Column(speed) : -0.110 -0.6730924

Column(dist) : 0.7591268 0.1193971

* Both the columns(variables) are skewed in nature i.e. non symmetric
* Skewness : “speed” is negatively skewed and “dist” is positively skewed, i.e
  + - * In column “speed” the mass of the distribution is concentrated on the right ;
      * In column “dist” the mass of the distribution is concentrated on the left ;
* Kurtosis : “speed” and “dist” are “-ve” and “+ve” ly distributed correspondingly.i.e.
  + - * In column “speed” peaks are wider.
      * In column “dist” peaks are basically larger and we get short visible tails on both side as compare to the column “speed”

**SP and Weight(WT)**



**Ans :**  Skewness Kurtosis

Column(SP) : 1.552258 5.723521

Column(WT) : -0.5921721 3.819466

* Both the columns(variables) are skewed in nature i.e. non symmetric
* Skewness: “SP” is positively skewed and “WT” is negatively skewed, i.e.
  + - * In column “SP” the mass of the distribution is concentrated on the left;
      * In column “WT” the mass of the distribution is concentrated on the right;
* Kurtosis: “SP” and “WT” both are positively distributed. i.e. in both the columns we can see higher peaks in smaller width.

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



* This data is positively skewed.
* Many chickensare having weights in between 50-100 gm (unit)
* Very rare no of chickens are having weight more than 300gm
* After a certain point i.e. (100gm) the frequency of the chicken decreases with increase in their weight.



* This boxplot represents the data is positively skewed
* The data contains outliers.

**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 :

|  |  |  |  |
| --- | --- | --- | --- |
|  | 94% | 98% | 96% |
| Upper | 201.2617 | 201.5606 | 201.3777 |
| Lower | 198.7383 | 198.4394 | 198.6223 |

**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 | Median | Variance | Std Deviation |
| 41 | 40.5 | 25.529 | 5.052664 |

import statistics

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

mean = statistics.mean(data)

median = statistics.median(data)

stddev = statistics.stdev(data)

variance = statistics.variance(data)

print("mean of dataset is = ", mean)

print("median of dataset is = ", median)

print("stddev of dataset is = ", stddev)

print("variance of dataset is = ", variance)

mean of dataset is = 41

median of dataset is = 40.5

stddev of dataset is = 5.05266382858645

variance of dataset is = 25.529411764705884

1. What can we say about the student marks?

Ans: Marks are not normally distributed.

Person with mark 56 can be the outlier in our data.

Many of the students are having 41 marki.e. the modal value of our data.

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

Ans: Skewness don’t exists in those case i.e. skewness = 0. The curve is called as a perfectly symmetric bell shaped curve.

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

Ans : The curve is positively skewed

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

Ans : The curve is negatively skewed

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

Ans : We can see high peakedness in the central part of the data .

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

Ans : We can see wider peaks.

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



What can we say about the distribution of the data?

Ans : The data is not symmetric. Mass of the data is concentrated towards right side.

What is nature of skewness of the data?

Ans : The Data is negatively skewed.

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

Ans: 18-10 = 8 . The Inter Quartile Range of the data is 8

Q19) Comment on the below Boxplot visualizations?



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

Ans : Looking at both the box plots we can say that:

Data in both cases are symmetrically distributed,(Normally distributed ), no outliers. Both seems to have same median value around 262.5 .

Comparing Boxplot\_1 (Red) has less variability, less variation, less standard deviation, less range, less Inter-quartile-range value as compared to Boxplot\_2(blue)

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:0.4074074
  2. P(MPG<40) Ans: 0.7530

c. P (20<MPG<50) Ans : 0.8518

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans : Yes 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 : None of the data follows Normal distribution

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

|  |  |  |
| --- | --- | --- |
| 90% | 94% | 60% |
| ± 1.644854 | ± 1.880794 | ± 0.8416212 |

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

|  |  |  |
| --- | --- | --- |
| 95% | 96% | 99% |
| ± 2.063899 | ± 2.171545 | ± 2.79694 |

**# for small samples (<50) we use t-statistics**

**# n = 25, degree of freedom = 25-1 = 24**

**# for 95% confidence interval, alpha = 5% = 0.05 and alpha/2 = 0.025**

**from scipy import stats**

**ci = 95**

**n = 24**

**t = stats.t.ppf(1- ((100-ci)/2/100), n-1) # 90% CI, t24, 0.025**

**print(t) # 2.0686576104190406**

* t scores of 95% confidence interval = **2.0686576104190406**
* t scores of 96% confidence interval =

(for 96% confidence interval, alpha = 5% = 0.04 and alpha/2 = 0.02)

**= 2.1769581113153715**

* t scores of 99% confidence interval =

(for 99% confidence interval, alpha = 1% = 0.01 and alpha/2 = 0.005)

**= 2.8073356837675227**

Q 24**)**A Government companyclaims 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 : t-score = -0.4714, and Degree of freedom = 17

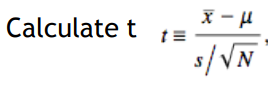
P(t) = 0.3216725

mean = 270 days

sample size = 18

sample mean = 260

deviation sample = 90 days



T = (260 – 270)/(90/srt(18))

= (-10)/ (90/4.24264068))

= - 0.471404

Degree of freedom = n-1 = (18-1) = 17

From T gables find 0.471404 for df(Degree of freedom) 17 =

from scipy import stats

print (stats.t.ppf(-0.47, 17))