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

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

Ans:- When three coins are tossed the total number of possible combinations are 23 = 8.

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

The number of combinations which have two heads and one tail are:

HHT, HTH, TTH which makes them 3 in number.

The Probability of getting two heads and one tails in the toss of three coins simultaneously is 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 2and 3

Ans:-

* (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)
* Total outcome=6×6=36

1. Equal to 1

Probability of getting sum equal to less than

0/36.probability is 0.

1. Less than or equal to 4

(1,3), (2,2), (3,1) (1,1),(1,2),(2,1)=6 outcomes,6/36 i.e. 1/6

1. 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?

Ans:- Total number of balls = (2 + 3 + 2) = 7  
Let S be the sample space.  
Then, n(S) = Number of ways of drawing 2 balls out of 7  
=7C2​  
=(2×1)(7×6)​  
=21  
Let E = Event of drawing 2 balls, none of which is blue.  
∴n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.  
=5C2​  
=(2×1)(5×4)​  
=10  
∴P(E)=n(S)n(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

Ans:- Expected number of cndies 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

=  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** > mean(Q7.assignment$Points)

[1] 3.596563

> mode(Q7.assignment$Points)

[1] "numeric"

> median(Q7.assignment$Points)

[1] 3.695

> sd(Q7.assignment$Points)

[1] 0.5346787

> range(Q7.assignment$Points)

[1] 2.76 4.93

> var(Q7.assignment$Points)

[1] 0.2858814

> mean(Q7.assignment$Score)

[1] 3.21725

> mode(Q7.assignment$Score)

[1] "numeric"

> median(Q7.assignment$Score)

[1] 3.325

> sd(Q7.assignment$Score)

[1] 0.9784574

> range(Q7.assignment$Score)

[1] 1.513 5.424

> var(Q7.assignment$Score)

[1] 0.957379

> mean(Q7.assignment$Weigh)

[1] 17.84875

> mode(Q7.assignment$Weigh)

[1] "numeric"

> median(Q7.assignment$Weigh)

[1] 17.71

> sd(Q7.assignment$Weigh)

[1] 1.786943

> range(Q7.assignment$Weigh)

[1] 14.5 22.9

> var(Q7.assignment$Weigh)

[1] 3.193166

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:-Expected Value  =  ∑ ( probability  \* Value )

 ∑ P(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

Ex  108, 110, 123, 134, 135, 145, 167, 187, 199

P(x)  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

Expected Value  =  (1/9)(108) + (1/9)110  + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

= (1/9) ( 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9)  (  1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Data[‘speed’].skew()

-0.117509

Data[‘speed’].kurtosis()

-0.5089944

Data[‘dist’].skew()

0.806894

Data[‘dist. ’].kurtosis()

0.405052

**SP and Weight(WT)**

**Use Q9\_b.csv**

Data[‘ SP ’].skew()=1.61154

Data[‘ SP’].kurtosis()=2.977328

Data[‘WT’].skew()= -0.61

Data[‘WT’].kurtosis()=0.9502914

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



Ans:- Histogram:- Chick weight data is right skewed or positively skewed.---- Yes-

More than 50% Chick Weight is between 50 to 150. ---- Yes-

Most of the chick weight is between 50 to 100. --- Yes



- The data is right skewed.

- There areoutliers at upper 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:-** Using the t- distribution, it is found= that:

The 94% confidence interval is (198.73,201.27)

The 96% confidence interval is (198.61,201.39)

The 98 % confidence interval is (198.43,201.57)

The interval is

x+t s ^n= 200 – 1.8916 30^2000 =198.73

x+ t s ^n= 200+ 1.816 30 ^2000=201.27

The 94% confidence interval is (198.73,201.27)

x-ts^n=200 – 2.0673 30 ^2000=198.61

x+ t s^n=200 + 2.0673 30-^2000201.39

The 96% confidence interval Is (198.61,201,39).

x-ts^n =200 -2.3452 30 =198.43 30^2000

x+ ts^n 200 + 2.3452 30 = 201.57^2000

The 98% confidence interval is (198.43,201.57)

**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?

**Solution :**Mean

x. mean ()=41.213125

Median

X .median()=40.5

variance

x. var()

28.2990

#Standard Deviation

x.std()

5.3194

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

Ans:- If the mean is equal to the median as well as the mode, hence **the skewness is zero**. If the distribution is symmetric, the mean equals the median, and the skewness of the distribution is zero.

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

Ans:-The mean = median = mode if the distribution is both symmetric and unimodal. Skewness can be positive, negative, or zero in nature. There is no skewness when the mean,median,and mode values are all the same

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

Ans:- **Left skewness** The mean, mode and median can be used to figure out if you have a positively or negatively skewed distribution. If the mean is greater than the median, the distribution is positively skewed.

Q16) What does positive kurtosis value indicates for adata ?

Ans:- A distribution with a positive kurtosis value indicates that the distribution has heavier tails than the normal distribution.

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

Ans:- A negative kurtosis means that your distribution is flatter than a normal curve with the same mean and standard deviation. The easiest way to visualise this is to plot a histogram with a fitted normal curve.

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



What can we say about the distribution of the data?

Ans:-The distribution of a data set is the shape of graph when all possible value are plotted on a frequency graph (showing how often they occur). Usually, we are not able to collect all the data for our variable of interest. Therefore we take a sample. This sample is used to make conclusions about the whole data set.

What is nature of skewness of the data?

Ans:- Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve,or normal distribution ,in a set of data.If the curve is shifted to the left or to the right, it is said to be skewed.

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

Ans:- The formula for finding the interquartile range takes the third quartile value and subtracts the first quartile value. Equivalently, the interquartile range is the region between the 75th and 25th percentile (75 – 25 = 50% of the data).

Q19) Comment on the below Boxplot visualizations?



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

**Solution** :By observing both the plots whisker’s level is high in boxplot 2.Mean and median are equal hence distribution is symmetrical.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

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

c. P (20<MPG<50)

**Ans:-import** pandas **as** pd

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

**from** scipy **import** stats

**from** scipy.stats **import** norm

cars**=**pd**.**read\_csv('Database/Cars.csv')

cars

|  | **HP** | **MPG** | **VOL** | **SP** | **WT** |
| --- | --- | --- | --- | --- | --- |
| **0** | 49 | 53.700681 | 89 | 104.185353 | 28.762059 |
| **1** | 55 | 50.013401 | 92 | 105.461264 | 30.466833 |
| **2** | 55 | 50.013401 | 92 | 105.461264 | 30.193597 |
| **3** | 70 | 45.696322 | 92 | 113.461264 | 30.632114 |
| **4** | 53 | 50.504232 | 92 | 104.461264 | 29.889149 |
| **...** | ... | ... | ... | ... | ... |
| **76** | 322 | 36.900000 | 50 | 169.598513 | 16.132947 |
| **77** | 238 | 19.197888 | 115 | 150.576579 | 37.923113 |
| **78** | 263 | 34.000000 | 50 | 151.598513 | 15.769625 |
| **79** | 295 | 19.833733 | 119 | 167.944460 | 39.423099 |
| **80** | 236 | 12.101263 | 107 | 139.840817 | 34.948615 |

81 rows × 5 columns

sns**.**boxplot(cars**.**MPG)

<matplotlib.axes.\_subplots.AxesSubplot at 0x21fb6a63e80>

*# P(MPG>38)*

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

0.3475939251582705

*# P(MPG<40)*

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

0.7293498762151616

*# 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())

1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Car MPG mean 34.422075728024666

Car MPG median 35.15272697

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

Dataset: wc-at.csv

Waist Circumference(Waist) = (91.90183486238533, 90.8)

Adipose Tissue= (101.89403669724771, 96.54)

Out[11]:

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

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

Ans:- A 94 % confidence interval has two tails of 6/2 = 3 %

so it goes from 3% to 97 % which leaves 94 % in the middle

so look up the Z for

P(z<Z) = 0.97

two closest values in the z-table

P(z<1.88 ) = 0.96995

P(z< 1.89) = 0.97062

interpolating

1.88 + ( 0.97- 0.96995)\*(0.01)/ (0.97062- 0.96995)

Z\_critical = approx. 1.880746

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

Ans:- The formula for a 100 (1-a) % confidence interval for µ is

X+ (tα/2,n-1 is the point on the t- distribution

100 (α/2) % of the distribution area to its right

For a 95% confidence interval,α=0.05, because

100(1-0.05) %

100(0.95)%

95%

25.9 + (1.987× 0.2846)

25.9 +(0.5655)

(25,33,26,47)

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:-from** scipy **import** stats

**from** scipy.stats **import** norm

*# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days*

*# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days*

*# 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

*# Find P(X>=260) for null hypothesis*

*# 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