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

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

* **Solution :-**

* Total number of outcomes when three coins are tossed

{HHH, HHT, HTH, HTT, TTT, THH, THT, TTH} = 8

1. Probability of getting two heads and one tail

{ HHT, HTH, THH }

P(X) = 3/8

= **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

* **Solution :-**
* Total number of outcomes when two dice are rolled=6\*6=36.

{(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)}

1. Equal to 1 = **0**
2. Less than or equal to 4 :-

{(1,3)(1,2)(1,1)(2,1)(2,2)(3,1)}

P(X<=4) = 6/36

= **0.1666**

1. Sum is divisible by 2 and 3

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

P(X) = 6/36

= **0.1667**

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?

* **Solution :-**
* Total number of events:-

= nCr =7C2 = 42/2 =**21**

* Two balls are drawn at random.

= nCr =5C2 =20/2 =**10**

* Probability that none of the balls drawn is blue:-

P(X) = 10/21

**= 0.4761**

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

* **Solution :-**

E(X)=(0.015\*1)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120) = **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**

* **Solution:-**

df.agg(["mean","var","std","min","max","median"])

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **mean** | 3.596563 | 3.21725 | 17.84875 |
| **var** | 0.285881 | 0.957379 | 3.193166 |
| **std** | 0.534679 | 0.978457 | 1.786943 |
| **min** | 2.76 | 1.513 | 14.5 |
| **max** | 4.93 | 5.424 | 22.9 |
| **median** | 3.695 | 3.325 | 17.71 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **0** | 3.07 | 3.44 | 17.02 |
| **1** | 3.92 | NaN | 18.9 |

df.mode()

range = (df.max()-df.min())

range

**Points 2.170**

**Score 3.911**

**Weigh 8.400**

**dtype: float64**

Conclusion : Here we can see that the average of data is 3.59, 3.21 , 17.84 respectively , Weigh std = 1.78 and var = 3.19 are high as compared to others, Points and Weigh have two modes , Points has a low range = 2.17 as compared to others.

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?

* **Solution :-**

E(X) = 1308/9 = **145.333**

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

**Cars speed and distance**

**Use Q9\_a.csv**

* **Solution :-**

df1.agg(["skew","kurt"])

|  |  |  |
| --- | --- | --- |
|  | **speed** | **dist** |
| **skew** | **-0.11751** | **0.806895** |
| **kurt** | **-0.508994** | **0.405053** |

**SP and Weight(WT)**

**Use Q9\_b.csv**

* **Solution :-**

df2.agg(["skew","kurt"])

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| **skew** | **1.61145** | **-0.614753** |
| **kurt** | **2.977329** | **0.950291** |

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



**Conclusion**: As we can see here, ChickWeight$Weight frequency is very high between 50 and 100, that's frequency close to 200 and on the other hand frequency is very low between 350 and 400 and that's frequency close to 0. From the histogram we can see that the distribution is right skewed.



Interquartile Range

First Quartile(Q1) [Q1 - 1.5\*IQR]

Third Quartile(Q3) [Q3 + 1.5\*IQR]

Upper Quartile

Median

Outliers

Lower Quartile

**Conclusion**: From the box-plot we can see that the distribution is right skewed and outliers are present, the dotted line shows the whisker.

**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?

N=3,000,000 , n=2000 , xbar = 200 , s=30

Confidence Interval Estimate= Z => 200 Z

94% Confidence: ((1-0.94)/2)+0.94) = (0.97) = [1.880794=Z]

200 1.88\*

Lower limit =198.74

Upper limit = 201.26

98% Confidence: ((1-0.98)/2)+0.98) = (0.99) = [1.880794=Z]

200 2.33\*

Lower limit =198.44

Upper limit = 201.56

96% Confidence: ((1-0.96)/2)+0.96) = (0.98) = [1.880794=Z]

200 2.05\*

Lower limit =198.62

Upper limit = 201.38

**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.

* **Solution:-**

df12 = ([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

df12

stats.mean(df12),stats.median(df12),stats.variance(df12),stats.stdev(df12)

**(41, 40.5, 25.529411764705884, 5.05266382858645)**

1. What can we say about the student marks?

* **Solution :-**

Here we can see that most of the students marks are relay around Mean =41 and Mean > Median so that the curve of the data is skewed towards of right.

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

* **Solution :-**

There is no skewness and distribution is symmetric.

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

* **Solution :-**

The nature of skewness is Right skewed or Right Tail

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

* **Solution :-**

The nature of skewness is Left skewed or Left tail

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

* **Solution :-**

In a positive skew, the tail of a distribution curve is longer on the right side.

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

* **Solution :-**

In a Negative skew, the tail of a distribution curve is longer on the left side.

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



What can we say about the distribution of the data?

* **Solution :-**

The tail of a distribution curve is longer on the left side.

What is nature of skewness of the data?

* **Solution :-**

The nature of skewness is Left skewed or Left Tail

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

IQR(inter quartile range)

Q1=10, Q3=18

IQR=(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.

* **Solution :-**

Both the box plot follows the normal distribution and both of having approximately same median, there are no outliers in both the plot, 1st plot having less range than the 2nd plot, both of having different IQR;

IQR(1st box plot): Q3-Q1 = 278-253 = 25

IQR(2nd box plot): Q3-Q1 = 310-224 = 86

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)
* **Solution :-**

1-norm.cdf(38,loc=34.422076,scale=9.131445)

0.34759394041453007

* 1. P(MPG<40)
* **Solution :-**

norm.cdf(40,loc=34.422076,scale=9.131445)

0.7293498604157946

* 1. P (20<MPG<50)
* **Solution :-**

norm.cdf(50,loc=34.422076,scale=9.131445) -

norm.cdf(20,loc=34.422076,scale=9.131445)

0.8988689076273199

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

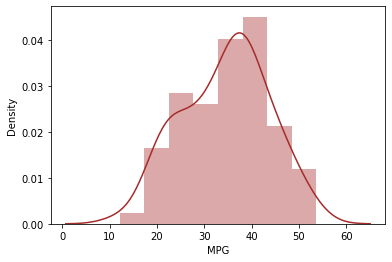
Dataset: Cars.csv

* **Solution :- ( Kolmogorov test for normality (**N < 5000**)**

H0 : The data is normal.

H1 : The data is not normal.

sns.distplot(df20["MPG"], kde=True, color = "brown" )



stats.kstest(df20["MPG"],'norm')

KstestResult(statistic=1.0, pvalue=0.0)

Conclusion : here the p-value is less than 0.05, so we reject the null

hypothesis and conclude that the data is not normal.

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

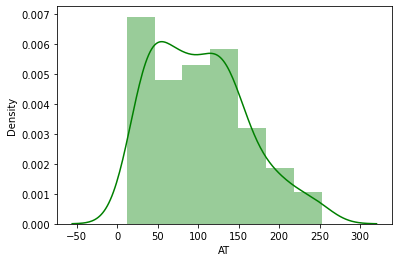
Dataset: wc-at.csv

* **Solution :- ( Kolmogorov test for normality (**N < 5000**)**

H0 : The data is normal.

H1 : The data is not normal.

sns.distplot(df21["AT"], kde=True, color = "green" )



stats.kstest(df21["AT"],'norm')

KstestResult(statistic=1.0, pvalue=0.0)

Conclusion : here the p-value is less than 0.05, so we reject the null

hypothesis and conclude that the data is not normal.

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

* **Solution:-**

Computer\_Data.csv

#price 90% CI

norm.interval(0.90,loc = 2219.576610,scale = 580.803956)

(1264.2391164256364, 3174.9141035743633)

#speed 94% CI

norm.interval(0.94,loc = 52.011024,scale = 21.157735)

(12.217691249041998, 91.804356750958)

#hd 60% CI

norm.interval(0.60,loc = 416.601694,scale = 258.548445)

(199.00183278074124, 634.2015552192588)

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

* **Solution:-**

stats.t.interval(0.975,df=24)

**(-2.3909493151293866, 2.390949315129389)**

stats.t.interval(0.98,df=24)

**(-2.4921594731575762, 2.4921594731575762)**

stats.t.interval(0.995,df=24)

**(-3.0905135487060877, 3.090513548706097)**

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

* **Solution:-**

H0 : average life of bulbs are 260 days

H0 : average life of bulbs are not more than 260 days

µ=270, n = 18, =260, s=90

tscore= = = -10/21.23= -0.47

> pt(-0.47,17)

[1] 0.3221639

Conclusion : here the p-value is greater than 0.05, so we do not

reject that and conclude that the average life of bulbs

are 260 days.