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

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

**Nominal:** the data can only be categorized.

**Ordinal:** the data can be categorized and ranked.

**Interval:** the data can be categorized and ranked, and evenly spaced.

**Ratio**: the data can be categorized, ranked, evenly spaced and has a natural zero.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | nominal |
| High School Class Ranking | ordinal |
| Celsius Temperature | ratio |
| Weight | ordinal |
| Hair Color | nominal |
| Socioeconomic Status | interval |
| Fahrenheit Temperature | ratio |
| Height | ordinal |
| Type of living accommodation | Interval |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | ordinal |
| Sales Figures | ratio |
| Blood Group | nominal |
| Time Of Day | nominal |
| Time on a Clock with Hands | ordinal |
| Number of Children | ordinal |
| Religious Preference | nominal |
| Barometer Pressure | ordinal |
| SAT Scores | ordinal |
| Years of Education | nominal |

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

Answer= total possible combinations is = 8

Therefore

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 ] = 3

Probability = 3/8

= 3 = 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**

The total possible combinations are

(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),

Total = 36 combinations

The probability of sum is equal to 1 is = 0

The Probability of less than or equal to 4 is (1,1) , (1,2) , (1,3) ,(2,1) , (2,2) (3,1)

= 6/36

= 0.1666

Sum is divisible by 2and 3

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

9/36

= 0.25

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?

1/3 =0.33

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| child | candies count | Probability |  |  |
|  |  |  |  |  |
| A | 1 | 0.015 |  | 0.015 |
| B | 4 | 0.2 |  | 0.8 |
| C | 3 | 0.65 |  | 1.95 |
| D | 5 | 0.005 |  | 0.025 |
| E | 6 | 0.01 |  | 0.06 |
| F | 2 | 0.12 |  | 0.24 |
|  |  |  |  |  |
|  |  |  |  | 3.09 |
|  |  |  |  |  |

The Expected value is 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**

Points Score

Length:32 Min. :2.760 Min. :1.513

Class :character 1st Qu.:3.080 1st Qu.:2.581

Mode :character Median :3.695 Median :3.325

Mean :3.597 Mean :3.217

3rd Qu.:3.920 3rd Qu.:3.610

Max. :4.930 Max. :5.424

Weigh

Min. :14.50

1st Qu.:16.89

Median :17.71

Mean :17.85

3rd Qu.:18.90

Max. :22.90

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| skewness(Points)  [1] 0.2788734  > skewness(Weigh)  [1] 0.3870456  > skewness(Score)  [1] 0.4437855  > kurtosis((Points))  [1] 2.435116  > kurtosis(Weigh)  [1] 3.553753  > kurtosis(Score)  [1] 3.172471   |  | | --- | | var(Score)  [1] 0.957379  > var(Weigh)  [1] 3.193166  > var(Points)  [1] 0.2858814  > sd(Score)  [1] 0.9784574  > sd(Weigh)  [1] 1.786943  > sd(Points)  [1] 0.5346787 | |  | | |  | | --- | | > the range of every column is given in above table  Skewness for Points, Weigh and Score is 0.2788734,0.3870456 and 0.4437855 respectively hence the data distribution is **approximately symmetric** | | |
|  |
| |  | | --- | | > | |

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?

Expected Value  =  ∑ ( probability  \* Value )

 ∑ P(x).E(x)

Total 9 patients are there

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**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 108 | 110 | 123 | 134 | 135 | 145 | 167 | 187 | 199 |  | 1308 |
|  |  |  |  |  |  |  |  |  |  | 145.3333 |

Expected value of Weight of Patient is 145.3 from Excel

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

**Cars speed and distance**

**Use Q9\_a.csv**

skewness(dist)

[1] 0.7824835

> skewness(Index)

[1] 0

> skewness(speed)

[1] -0.1139548

> kurtosis(dist)

[1] 3.248019

> kurtosis(speed)

[1] 2.422853

> kurtosis(Index)

[1] 1.79904

> summary(df)

Index speed dist

Min. : 1.00 Min. : 4.0 Min. : 2.00

1st Qu.:13.25 1st Qu.:12.0 1st Qu.: 26.00

Median :25.50 Median :15.0 Median : 36.00

Mean :25.50 Mean :15.4 Mean : 42.98

3rd Qu.:37.75 3rd Qu.:19.0 3rd Qu.: 56.00

Max. :50.00 Max. :25.0 Max. :120.00

For speed : Skewness for dist is 0.7824835 which is greator than 0.5 hence the distance data is **moderately skewed** ,mean > median, skewness is positive

For dita : Skewness of speed is -0.1139548 which is less than -1 hence speed is **highly skewed if mean is greator than median the +ve skewness is theire**

**SP and Weight(WT)**

**Use Q9\_b.csv**

summary(df)

X SP WT

Min. : 1 Min. : 99.56 Min. :15.71

1st Qu.:21 1st Qu.:113.83 1st Qu.:29.59

Median :41 Median :118.21 Median :32.73

Mean :41 Mean :121.54 Mean :32.41

3rd Qu.:61 3rd Qu.:126.40 3rd Qu.:37.39

Max. :81 Max. :169.60 Max. :53.00

> library(moments)

> skewness(SP)

[1] 1.581454

> skewness(WT)

[1] -0.6033099

> kurtosis(SP)

[1] 5.723521

> kurtosis(WT)

[1] 3.819466

For SP : skewness is 1.581454 hence Data is highly skewed

Foe WT : skewness(-0.6033099) is less than 0 hence data is moderately skewed

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

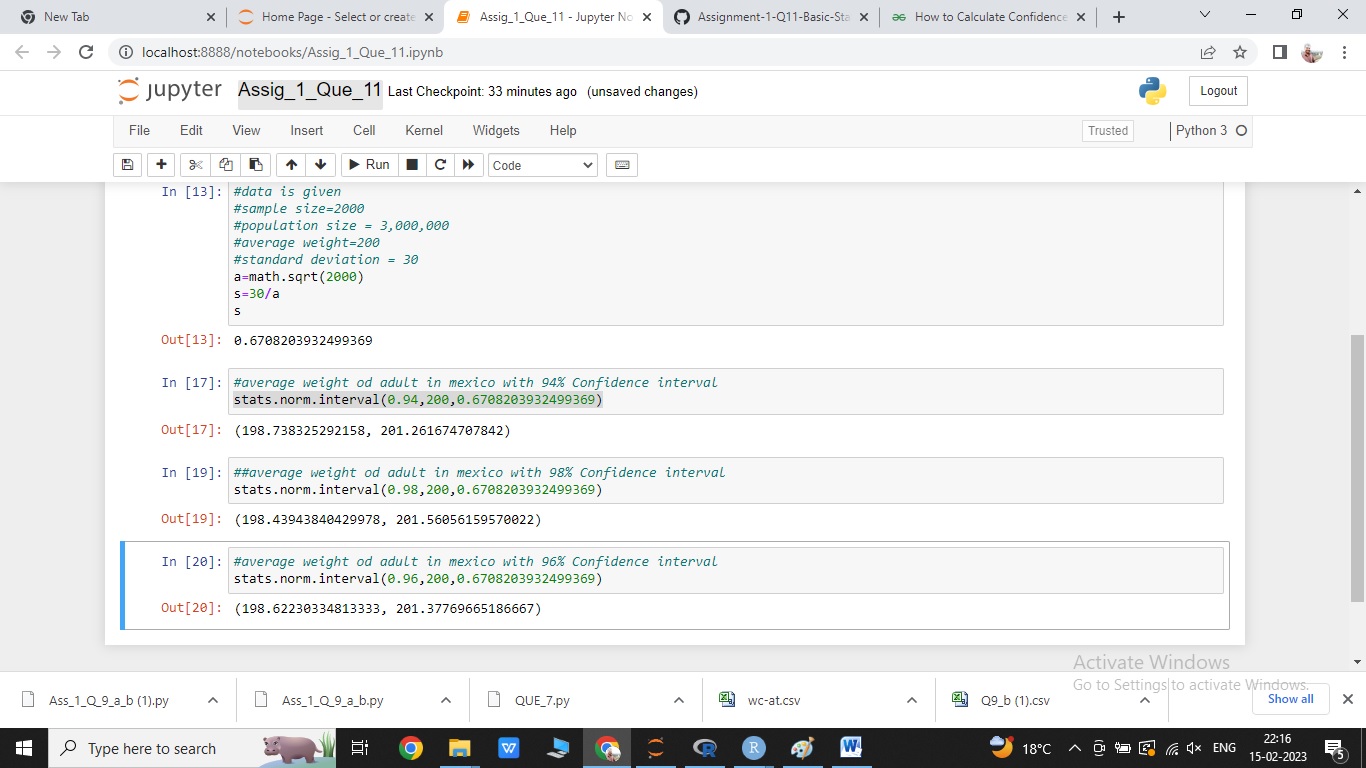


1. Positive skewness is present in above Histogram
2. Outliers is present in upper range of dataset
3. The max weight of chick is AT LOWER RANGE OF WEIGHT



1. The outliers present in data sets
2. The outliers is at the uppear range of the dataset
3. The value of upper whiskers is high
4. The box plot is uneven in size

**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: (198.62230334813333, 201.37769665186667)**

**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?

|  |  |
| --- | --- |
| mean | 41 |
|  |  |
| median | 40.5 |
|  |  |
| std deviation | 4.886957 |
|  |  |
| variance | 25.52941 |
|  |  |

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

Answer: the skewness is zero.

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

Answer : the skewness is positive

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

Answer: the skewness is negative

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

Answer: Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.

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

Answer: a negative value means that there are light tails

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



What can we say about the distribution of the data?

Distribution of data is uneven

What is nature of skewness of the data?

Positive , skewness is left skewed

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

10 to 18 are IQR

Q19) Comment on the below Boxplot visualizations?



Answer: In the Box plot 1 the range of data is short but it follows normal distributions

In box plot 2 the range is high as compared the box plot 1 but it also follows normal distributions.

The interqurtile range of box plot 2 is high , the median of both the plot is same

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

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)
  2. P(MPG<40)

c. P (20<MPG<50)

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

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

[1]

2s

import pandas as pd  
from scipy import stats  
from scipy.stats import norm

[10]

0s

#95%  
zcritical = stats.norm.ppf(1-0.025)  
zcritical

1.959963984540054

[13]

0s

#90%

zcritical = stats.norm.ppf(1-0.05)

zcritical



1.6448536269514722

[14]

0s

#94%

zcritical = stats.norm.ppf(1-0.03)

zcritical



1.8807936081512509

[15]

0s

#60%

zcritical = stats.norm.ppf(1-0.20)

zcritical



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