|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Continuous |
| 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 | Continuous |
| Number of tickets in Indian railways | Continuous |
| Number of times married | Continuous |
| 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 | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ (Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| 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?

Combinations will be 2 \* 2 \* 2 = 8

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

P (Head = 2, Tail = 1) = 3 /8

Ans 3: Probability is **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.

Combination will be 2C6 = 36

X = {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}

Ans:

P (Sum = 1) = 0 / 36 = **0**

P (Sum <\_ 4) = 6/36 = **1/6**

P (Sum div 2, 3) = 24 / 36 = **2 / 3**

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?

No of balls = 2 + 3 + 2 = 7

Balls {R1, R2, G1, G2, G3, B1, B2}

X = {(R1, R2), (R1, G1), (R1, G2), (R1, G3), (R1, B1), (R1, B2)

(R2, G1), (R2, G2), (R2, G3), (R2, B1), (R2, B2), (R2, R1)

(G1, G2), (G1, G3), (G1, B1), (G1, B2), (G1, R1), (G1, R2)

(G2, G3), (G2, B1), (G2, B2), (G2, R1), (G2, R2), (G2, G1)

(G3, B1), (G3, B2), (G3, R1), (G3, R2), (G3, G1), (G3, G2)

(B1, B2), (B1, R1), (B1, R2), (B1, G1), (B1, G2), (B1, G3)

(B2, R1), (B2, R2), (B2, G1), (B2, G2), (B2, G3), (B2, B1)}

P (No ball Not equal to Blue) = 20/42 = **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

Expected value / mean = (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**

import pandas as pd

import NumPy as np

from Scipy import stats

df=pd.read\_csv("Q7.csv")

1. **Mean**

**df. mean ()**

Points 3.596563

Score 3.217250

Weigh 17.848750

1. **Median**

**df. median ()**

Points 3.695

Score 3.325

Weigh 17.710

1. **Mode**

**df.mode().iloc[0:1, 1:4]**

|  | Points | Score | Weigh |
| --- | --- | --- | --- |
| **0** | 3.07 | 3.44 | 17.02 |

1. **Variance**

**np.var(df)**

Points 0.276948

Score 0.927461

Weigh 3.093380

1. **Standard Deviation**

**np.std(df)**

Points 0.526258

Score 0.963048

Weigh 1.758801

1. **Range**

**max(df["Points"])-min(df["Points"])** = 2.17

**max(df["Score"])-min(df["Score"])** =3.910

**max(df["Weigh"])-min(df["Weigh"])** = 8.399

Since mode, median and mean are almost equal , we could conclude that the data are **symmetrica**l in nature.

Compared to the variance, standard deviation, range of Points and score , weight has higher value . Hence the **measure of dispersion is high** for weigh data.

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?

108+110+123+134+135+145+167+187+199 = 1308

|  |  |
| --- | --- |
| Weight (X) | P(X) |
| 108 | 108/ 1308 = 0.082 |
| 110 | 110 / 1308 = 0.084 |
| 123 | 123/1308 = 0.094 |
| 134 | 134 / 1308 = 0.102 |
| 135 | 135 / 1308 = 0. 103 |
| 145 | 145 / 1308 =0.111 |
| 167 | 167 / 1308 = 0.128 |
| 187 | 187 / 1308 = 0.143 |
| 199 | 199 / 1308 = 0.152 |

E(X) = 108 \* 0.082 + 110 \* 0.084 + 123 \* 0.094 + 134 \* 0.102 + 135 \* 0.103 + 145

\* 0.111 + 167 \* 0.128 + 187 \* 0.143 + 199 \* 0.152

= 8.856 + 9.24 + 11.562 + 13.668 + 13.905+16.905+21.376+26.741+30.248

= **152.501**

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

**Cars speed and distance**

**Use Q9\_a.csv**

import pandas as pd

import NumPy as np

from Scipy import stats

data = pd.read\_csv("Q9\_a.csv")

data["speed"].skew() = **-0.11750**

data["dist"].skew() = **0.80689**

data["speed"].kurt() = **-0.50899**

data["dist"].kurt() = **0.40505**

speed implies a negative skewness which indicates a left skewed distribution.

dist implies a positive skewness which is a indicates a right skewed distribution.

speed implies a negative kurtosis which indicates a flatter distribution.

Dist implies a positive kurtosis which indicates a peaked distribution.

**SP and Weight(WT)**

**Use Q9\_b.csv**

dataset = pd.read\_csv("Q9\_b.csv")

dataset

dataset["SP"].skew() = **1.61145**

dataset["WT"].skew() = **-0.61475**

dataset["SP"].kurt() = **2.97733**

dataset["WT"].kurt() = **0.95029**

SP implies a positive skewness which indicates a right skewed distribution.

WT implies a negative skewness which is a indicates left skewed distribution.

SP implies a positive kurtosis which indicates a peaked distribution.

WT implies a positive kurtosis which indicates a peaked distribution.

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



|  |  |
| --- | --- |
| ChickWeight $ Weight | Frequency |
| 0 to less than 50 | 80 |
| 50 to less than 100 | 190 |
| 100 to less than 150 | 130 |
| 150 to less than 200 | 90 |
| 200 to less than 250 | 40 |
| 250 to less than 300 | 30 |
| 300 to less than 350 | 20 |
| 350 to less than 400 | 10 |

Histogram indicates that the distribution is **right skewed distribution**.



The box plot indicates that the right whisker/hinge is extended and long hence the data is **Right-Skewed**. The data points outside the whisker indicates the presence of **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?

n = 2000

mean = 200

Standard deviation = 30

import pandas as pd

import numpy as np

from scipy import stats

ConfidenceInterval94 = stats.norm.interval(0.94,200,30/np.sqrt(2000))

ConfidenceInterval96 = stats.norm.interval(0.96,200,30/np.sqrt(2000))

ConfidenceInterval98 = stats.norm.interval(0.98,200,30/np.sqrt(2000))

print('Confidence Interval in 94% =',np.round(ConfidenceInterval94,4))

print('Confidence Interval in 96% =',np.round(ConfidenceInterval96,4))

print('Confidence Interval in 98% =',np.round(ConfidenceInterval98,4))

**Answer**

Confidence Interval in 94% = [198.7383 , 201.2617]

Confidence Interval in 96% = [198.6223 , 201.3777]

Confidence Interval in 98% = [198.4394 , 201.5606]

**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?

import statistics as s

import Numpy as np

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

print (" mean =" ,s.mean(scores))

print (" median =", s.median(scores))

print (" variance=", np.var(scores))

print (" standard deviation=", np.std(scores))

**Answer:**

mean = 41

median = 40.5

variance= 24.11111111111111

standard deviation= 4.910306620885412

**On an average the students scored 41 marks in the test.**

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

Skewness will be 0. The normal distribution will be symmetrical.

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

The normal distribution will be asymmetrical. The value of skewness will be positive. The tail distribution will be longer towards the right side of the curve.

It is Right-Skewed Distribution

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

The normal distribution will be asymmetrical. The value of skewness will be Negative. The tail distribution will be longer towards the left side of the curve. It is a left-skewed Distribution.

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

If kurtosis has positive value, then the distribution is peaked.

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

If kurtosis has negative value, then the distribution is flatter than normal distribution.

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



What can we say about the distribution of the data?

The median of the data is distributed to the right. Mean and median is not equal hence the distribution is not symmetric.

What is nature of skewness of the data?

Median > mean and the left whiskers are extended, hence it is a left skewed distribution.

What will be the IQR of the data (approximately)?   
  
IQR = Upper quartile – Lower quartile

= 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.

ANSWER

1. Both the boxplot 1 and boxplot 2 have same median
2. Both the boxplot 1 and boxplot 2 are symmetrically distributed
3. Boxplot 1 has less variance, whereas boxplot 2 has more variance.
4. IQR of boxplot 1 is different from IQR of boxplot2
5. No suspected outliers in boxplot 1 and 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)

Refer BasicStatistics\_Assigment1.ipynb

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Refer BasicStatistics\_Assigment1.ipynb

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

Dataset: wc-at.csv

Refer BasicStatistics\_Assigment1.ipynb

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

|  |  |  |  |
| --- | --- | --- | --- |
| Confidence Interval | 90% | 94% | 60% |
| Significance level (alpha) | 1. 0.90=0.1 | 1. 0.94=0.06 | 1-0.60=0.40 |
| Significance level (alpha / 2) | 0.1/2 = 0.05 | 0.06/2 = 0.03 | 0.40/2 = 0.20 |
| Value to look in z table | 1-0.05 = 0.95 | 1. 0.03= 0.97 | 1-0.20 = 0.80 |
| Z score | 1.65 | 1.89 | 0.85 |
|  |  |  |  |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Confidence Level | 95% | 99% | 96% | notes |
| 1. Degree of Freedom, 25-1 | 24 | 24 | 24 | sample size, n = 25 |
| 2. Significance Level αα | 1-0.95=0.05 | 1-0.99=0.01 | 1-0.96=0.04 | αα = (1 - confidence level) |
| 3. Significance Level α/2α/2 | 0.05/2 = 0.025 | 0.01/2 = 0.005 | 0.04/2 = 0.02 | two-tailed |
| 4. value to lookup in t-table | 24,0.025 | 24,0.005 | 24,0.02 | use t-value calculator |
| 5. critical value | 2.064 | 2.797 | 2.492 | t-critical |
|  |  |  |  |  |

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

Answer : Refer BasicStatistics\_Assigment1.ipynb

x – bar = 260

mu = 270

n = 18

sample std ( s) = 90

t – score = 260 – 270 / 90 / (sqrt (18))

= -0.47140

pvalue= stats.t.cdf(tscore,17)

pvalue

= 0.3217