|  |  |
| --- | --- |
| 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 | Ratio |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

ANS) If a coin is tossed the outcomes are Head , Tail

Let, Head = H

Tail = T

Three Coins are tossed then outcomes are

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

Total number of outcomes ,n(S)= 8

Let,’A’ be the Event of getting 2 heads and 1 tail

Number of favourable outcomes , n(A)= 3 , {THH,HTH,HHT}

P(A) = n(A)/n(S)

= 3/8

Probability of getting 2 heads and 1 tail = 3/8.Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4

c) Sum is divisible by 2 and 3

**ANS)** Given that Two Dice are rolled,

Outcomes 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)

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

Let,Total number of outcomes = n(S)

n(S) = 36

1. Let ‘A’ be the event of getting sum is equal to 1,

Total number of Outcomes = n(S) = 36

Number of favourable outcomes = n(A) = 0

Probability of getting sum is equal to , P(A) = n(A)/n(S)

=0/36

= 0

1. Let ‘B’ be the event of getting Less than or equal to 4,

Total number of outcomes = n(S) = 36

Number of favourable outcomes = n(B) = {(1,1),(1,2),(1,3)

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

Probability of getting less than or equal to 4 = P(B)= n(B)/n(S)

= 6/36

= 1/6

1. Let ‘C’ be the event of getting Sum is divisible by 2 and 3,

Total number of Outcomes = n(S) = 36

Number of favourable outcomes = n(C) = {(1,6),(2,3),(3,2),(6,1)}

n(C) = 4

Probability of getting a sum is divisible by 2 and 3 = P(C) = n(C)/n(S)

=4/36

= 1/9

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) Given , A bag contains 2 red balls, 3 green balls , 2 blue balls

Total no 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

n(S) = 7C2

= (7×6)/(2×1)

= 21

Let , ‘E’ be the event of getting none of the balls drawn is blue.

Therefore, n(E) = Number of ways of drawing 2 balls out of (2+3) balls

n(E) = 5C2

n(E) = (5×4)/(2×1)

= 10

Probability of getting that none of the balls drawn is blue =

P(E) = n(E)/n(S)

= 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:

To calculate the expected number of candies for a randomly selected child, you need to multiply each child's candy count by their respective probabilities and then sum up these products. The formula for calculating the expected value (mean) is:

Expected Value (E) = Σ (x \* P(x))

Expected number of candies 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

Expected number of candies for a randomly selected child = 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** | **Weigh** |
| mean | 3.59656 | 3.217 | 17.85 |
| median | 3.695 | 3.325 | 17.71 |
| mode | 3.92 | 3.44 | 17.02 |
| variance | 0.28588 | 0.957 | 3.193 |
| standard deviation | 0.53468 | 0.978 | 1.787 |
| MAX | 4.93 | 5.424 | 22.9 |
| MIN | 2.76 | 1.513 | 14.5 |
| Range | 2.17 | 3.911 | 8.4 |

**Points:**

The mean Points is approximately 3.60.

The median is also around 3.69, indicating a relatively symmetric distribution.

The mode is approximately 3.07.

The variance is relatively low, suggesting that the data points are closely clustered around the mean.

The standard deviation is relatively small, indicating low variability.

The range of Points is about 2.17.

**Score:**

The mean Score is approximately 3.22.

The median is approximately 3.44, suggesting a slightly right-skewed distribution.

The mode is approximately 3.44.

The variance is moderate, indicating some spread in the data.

The standard deviation is relatively small, suggesting moderate variability.

The range of Score is about 3.81.

**Weigh:**

The mean Weigh is approximately 17.85.

The median is approximately 17.82, indicating a relatively symmetric distribution.

The mode is approximately 17.02.

The variance is relatively low, suggesting that the data points are closely clustered around the mean.

The standard deviation is relatively small, indicating low variability.

The range of Weigh is about 8.43.

Q8) Calculate Expected Value for the problem below

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 = Sum (X \* Probability of X)

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

= 148

Therefore, Expected value = 148

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

**Cars speed and distance**

**Use Q9\_a.csv**

import pandas as pd

data=pd.read\_csv("C:/Users\Desktop\Excelr\_Assignments\Basic1\Q9\_a.csv")

data

data["speed"].skew().round(2) = -0.12

data["speed"].kurt().round(2) = -0.51

data["dist"].skew().round(2) = 0.81

data["dist"].kurt().round(2) = 0.41

**Inference :**

For Speed skewness is negatively skewed (-0.12) and it is Platykurtic Curve (-0.51)

For Distance skewness is Positively skewed (0.81) and it is Leptokurtic Curve (0.41)

**SP and Weight(WT)**

**Use Q9\_b.csv**

import pandas as pd

data=pd.read\_csv("C:/Users\Desktop\Excelr\_Assignments\Basic1\Q9\_b.csv")

data

data["SP"].skew().round(2) = 1.61

data["SP"].kurt().round(2) = 2.98

data["WT"].skew().round(2) = -0.61

data["WT"].kurt().round(2) = 0.95

**Inference :**

For SP skewness is Positively skewed (1.61) and it is Leptokurtic Curve (2.98)

For WT skewness is negatively skewed (-0.61) and it is Leptokurtic Curve (0.95)

**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

**BOXPLOT**:

- The data is right skewed.

- There are outliers at upper side

The most of the data points are concentrated in the range 50-100 with frequency 200. And least range of weight is 400 somewhere around 0-10.So the expected value the above distribution is 75.Skewness-we can notice a long tail towards right so it is heavily right skewed or positively skewed.

**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) import scipy.stats as stats

import numpy as np

# Sample data

sample\_mean = 200 # Average weight of the sample

sample\_std\_dev = 30 # Standard deviation of the sample

sample\_size = 2000 # Sample size

degrees\_of\_freedom = sample\_size - 1

from scipy import stats

from scipy.stats import t

# for 94% confidence interval

ci\_min,ci\_max = stats.t.interval(0.94,df=degrees\_of\_freedom, loc = sample\_mean, scale = sample\_std\_dev/np.sqrt(sample\_size))

print(ci\_min,ci\_max)

for 94% confidence interval is(198.73 , 201.26)

# for 98% confidence interval

ci\_min,ci\_max = stats.t.interval(0.98,df=degrees\_of\_freedom, loc = sample\_mean, scale = sample\_std\_dev/np.sqrt(sample\_size))

print(ci\_min,ci\_max)

for 98% confidence interval (198.43 , 201.56)

# for 96% confidence interval

ci\_min,ci\_max = stats.t.interval(0.96,df=degrees\_of\_freedom, loc = sample\_mean, scale = sample\_std\_dev/np.sqrt(sample\_size))

print(ci\_min,ci\_max)

for 96% confidence interval (198.62,201.37)

**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**

Find mean, median, variance, standard deviation.

**ANS)**

import numpy as np

import pandas as pd

score = np.array([[34],[36],[36],[38],[38],[39],[39],[40],[40],[41],[41],[41],[41],[42],

[42],[45],[49],[56]])

score = pd.DataFrame(score)

score

score.mean()

score.median()

score.var()

score.std()

**Mean**: The mean is the average of all the student marks. In this case, it is approximately 41.0. This indicates that, on average, the student marks are close to 41.0.

**Median**: The median is the middle value of the data when it is sorted in ascending order. In this case, the median is 40.5. This implies that half of the student marks are below 40.5 and other half are above 40.5.

**Variance**: variance measures the spread of data points from the mean. A Higher variance suggests that the data points are more spread out. In this case, the variance is approximately 24.11. The higher variance indicates that the student marks are somewhat spread out from the mean, suggesting some variability in scores.

**Standard Deviation**: The standard deviation is another measure of the spread of the data points from the mean .It is the square root of the variance. In this case , the standard deviation is approximately 4.91. The larger standard deviation indicates that there is a noticeable amount of variability in the student marks from the meanWhat can we say about the student marks?

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

Ans:

When the mean and median of the dataset are equal, it indicates that the data is symmetrically distributed around the center. In this case , the skewness of the data is zero. Skewness is the measure of the asymmetry of the probability distribution of a real valued random variables. It tells us whether the data is skewed to the left(negative skewness ) or to the right (positive skewness) relative to the mean.

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

Ans:

The nature of skewness will be “Positively skewed” in nature

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

Ans:

The nature of skewness will be “Negatively skewed” in nature

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

Ans: Positive Kurtosis value indicates that the dataset has heavier tails or more/highly peaked or leptokurtic distribution .Positive kurtosis value always have the kurtosis value when calculated as (>0).

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

Ans: Negative Kurtosis value indicates that the dataset has lighter tails or less/flatter peaked or platykurtic distribution .Positive kurtosis value always have the kurtosis value when calculated as (<0).

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



What can we say about the distribution of the data?

Ans:

* The data is distributed across a range from around 2 to above 18.
* The upper whisker length extending beyond 18 suggests that the possibility of outliers or data points that are more spread out.
* The position of the median line between 14 and 16 suggests that the central value of the data is shifted slightly towards the higher end of the range (the line inside the boxplot indicates the median line)
* The point that the quartile range 1 being starts exactly at 10 shows that the minimum value of the data set starts from 10 and the lower 25% of data is clustered around a value around the 10.
* In the same way, the slight exceeding of Q3 beyond the 18 shows that the upper 25% of the data extends beyond 18.

What is nature of skewness of the data?

Ans: The data is a skewed towards left. The whisker range of minimum value is greater than maximum . so negatively skewed in nature.

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

The Inter Quantile Range = Q3 Upper quartile – Q1 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.

**ANS)**  First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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

P(MPG>38)

P(MPG<40)

P (20<MPG<50)

**ANS)**

import pandas as pd

cars = pd.read\_csv("C:/Users\Desktop\Excelr\_Assignments\Basic1\Cars.csv")

cars

MPG = cars["MPG"]

MPG

Avg=MPG.mean()

Avg

std = MPG.std()

std

from scipy.stats import norm

z=norm(Avg,std) #mean =34.42,sd = 9.13

z

#P(MPG>38)

1-z.cdf(38)

#P(MPG<40)

z.cdf(40)

#P(20<MPG<50)

z.cdf(50)-(1-z.cdf(20))

P(MPG>38) = 0.347

P(MPG<40) = 0.729

P(20<MPG<50) = 0.013

Q 21) Check whether the data follows normal distribution

Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANS)** MPG of cars follows Normal Distribution

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

Dataset: wc-at.csv

**ANS)** Adipose Tissue (AT) and Waist Circumference(Waist) does not follows Normal Distribution

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

**ANS**)

*Z score of* 90%​ =1.645

*Z score of* 94%=1.8808

*Z score of* 60%=0.8416

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

**ANS)**

*t score of* 95%​ =2.06

*t score of* 96%=2.17

*t score of* 99%=2.79

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

**ANS)**

µ=270, =260, SD=90, n=18, df=17

t-score = =-0.4714  
Degree of freedom = 17   
P(t) = 0.3216725

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32%