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

Name: Chinmai Shivaji Dandekar

Batch: 18 December (Thane)

**ASSIGNMENT 1**

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

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

**Solution**:

Given: Three coins are tossed so the sample space is,

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

Number of sample,

n(S) = 8

Let A be the probability of getting two heads and one tail,

A = {HHT, THH, HTH}

n(A) = 3

Answer: P(A) = n(A)/n(S) = 3/8 = 0.375

**The probability of getting two heads and one tail 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

**Solution:**

Given: Two Dice are rolled so the sample space is,

S = {(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)}

n(S) = 36

1. Let A be the sample but we cannot have sum as 1 because here both the dices are starting from sample as (1, 1), so the smallest sum will be 2.

P(A) = 0

**Probability of sum equal to 1 is 0**

1. Let B be the sample having sum less than equal to 4,

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

n(B) = 6

P(B) = n(A)/n(S) = 6/36 = 1/6 = 0.1667

**Probability of sum less than equal to 4 is 1/6 or 0.1667**

1. Let C be the sample having sum divisible by 2 and 3.

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

n(C) = 6

P(C) = n(B)/ n(S) = 6/36 = 1/6 = 0.1667

**Probability of sum divisible by 2 and 3 is 1/6 or 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:**

Given: There are total 7 balls 2 red, 3 green and 2 blue,

So the total number of chances to draw 2 balls at a random from 7 colored balls is,

n(S) = 7C2 = 7\*6/2 = 7\*3 = 21

Now, Let A be the event none of the ball drawn is blue,

Number of chances of drawing two balls other than blue is,

n(A) = 2C2 + 3C2 + 2C1\* 3C1  = 1 + 3 + 2\*3 = 1 + 3 + 6 = 10

**Answer: Probability of two balls drawn at random other than blue is**

**P(A) = n(A)/n(S) = 10/21 = 0.476**

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:**

Expected Value = ∑ (probability \* Value)

 ∑ P(x).E(x)

|  |  |
| --- | --- |
| Candies count | Probability |
| E(x) | P(x) |
| 1 | 0.015 |
| 4 | 0.20 |
| 3 | 0.65 |
| 5 | 0.005 |
| 6 | 0.01 |
| 2 | 0.120 |

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

**Answer: 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.

**Solution:** [**Q7 Solution.xlsx**](Q7%20Solution.xlsx)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Solution | | | | | | |
|
|  | Mean | Median | Mode | Variance | Standard Deviation | Range |
| Points | 3.5965625 | 3.695 | 3.92 | 0.285881351 | 0.534678736 | 2.17 |
| Score | 3.21725 | 3.325 | 3.44 | 0.957378968 | 0.978457443 | 3.911 |
| Weigh | 17.84875 | 17.71 | 17.02 | 3.193166129 | 1.786943236 | 8.4 |

**Inferences:**

Here in this dataset Mean and Median values of all columns are approximately close to each other so we can conclude that the dataset has symmetrical distribution.

For mode 3.92, 3.44 and 17.02 is the most frequently occurring data.

We can say that there is not much variance (difference between the values of dataset) in this dataset.

As we can see that the values of standard deviation for the given dataset are low so we can conclude that the values are clustered close to the mean.

Ranges are in quiet in the center so the values are less than or close to mean so there is less number of chances of having an outlier in the dataset.

So this data set may have normal distribution.

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:**

Expected Value = ∑ (Probability \* Value)

∑ P(x).E(x)

There are 9 patients Probability of selecting each patient = 1/9

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| E(x) | 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 = (1/9) (108) + (1/9)110 + (1/9)123 + (1/9)134 + (1/9)135 +

Value (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

SP and Weight (WT)

Use Q9\_b.csv

**Solution:**

For cars speed and distance

[Q9\_a Solution.xlsx](Q9_a%20Solution.xlsx)

|  |  |  |
| --- | --- | --- |
| Solution | | |
|
|  | Skewness | Kurtosis |
| Speed | -0.11751 | -0.509 |
| Distance | 0.806895 | 0.40505 |

**Inference:**

**Skewness**

Here skewness of Speed is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer. If skewness = 0, the data are perfectly symmetrical. So dataset contains outliers at the left tail as majority of data is towards right.

Here skewness of Distance is positive; the data are positively skewed or skewed right, meaning that the right tail is longer. So there are outliers in the right tail as majority of data is towards left side.

**Kurtosis:**

Kurtosis of Speed is negative which indicates that the distribution has lighter tails than the normal distribution. For example, data that follow a beta distribution with first and second shape parameters equal to 2 have a negative kurtosis value.

Kurtosis of Distance is positive which indicate that distribution is peaked and possesses thick tails.

For SP and Weight (WT)

[Q9\_b Solution.xlsx](Q9_b%20Solution.xlsx)

|  |  |  |
| --- | --- | --- |
| Solution | | |
|
|  | Skewness | Kurtosis |
| SP | 1.611450196 | 2.977328944 |
| WT | -0.614753326 | 0.950291491 |

**Inference:**

**Skewness**

Here skewness of SP is positive; the data are positively skewed or skewed right, meaning that the right tail is longer. So there are outliers in the right tail as majority of data is towards left side

Here skewness of WT is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer. If skewness = 0, the data are perfectly symmetrical. So dataset contains outliers at the left tail as majority of data is towards right.

**Kurtosis:**

Kurtosis of SP and WT are positive which indicate that distribution is peaked and possesses thick tails.

Q10) Draw inferences about the following boxplot & histogram



**Solution:**

**Inference**

**Histogram**

This histogram represents the data of ChickWeights and their frequency. This histogram is not normally distributed it contains some outliers at the right tail. So, if majority of data is in left and the tail is in right then it is a positively skewed histogram. We can square the data to make the graph distribution normal. Or we can treat the outliers to overcome skewness.

**Box plot**

Box plot are graphical plots used to check if my dataset contains any outliers. So here in this graph we can clearly see that there are 7 dots or 7 data entries which are outliers. Looking carefully, we can see that this box plot is derived from the same data used to plot histogram. Since majority of the data is towards lower 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?

**Solution:**

We are given the standard deviation for the sample, which is why the **t-distribution** is used to solve this question.

The information given is:

Sample **mean**of  = 200.

Sample **standard deviation** of  *s = 30*.

Sample **size**of *n = 2000*.

The interval is:

In which, **t** is the critical value for the two-tailed confidence interval.

Considering a **94%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 1.8916**, hence:

The **94%** confidence interval is **(198.73, 201.27).**

Considering a **96%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.0673**, hence:

The **96%** confidence interval is **(198.61, 201.39).**

Considering a **98%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.3452**, hence:

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:**

1. Solution for mean, median, mode, variance and standard deviation is as follows:

|  |  |  |
| --- | --- | --- |
| *x* |  |  |
| 34 | -7 | 49 |
| 36 | -5 | 25 |
| 36 | -5 | 25 |
| 38 | -3 | 9 |
| 38 | -3 | 9 |
| 39 | -2 | 4 |
| 39 | -2 | 4 |
| 40 | -1 | 1 |
| 40 | -1 | 1 |
| 41 | 0 | 0 |
| 41 | 0 | 0 |
| 41 | 0 | 0 |
| 41 | 0 | 0 |
| 42 | 1 | 1 |
| 42 | 1 | 1 |
| 45 | 4 | 16 |
| 49 | 8 | 64 |
| 56 | 15 | 225 |
| = 738 |  | = 434 |

Mean =

=

=

= **41**

Median is the middle value of the series ordered in ascending order or descending order if N value is odd. If N is even like in our case so we will use formula given below:

Median = = =

= = = = **40.5**

Mode is the frequent number of time the data is occurred,

Mode = **41**

Variance = S2 = = = **25.5294**

Standard Deviation = S = = = = **5.0526**

**Answer:**

**Mean = 41**

**Median = 40.5**

**Mode = 41**

**Variance = 25.5294**

**Standard Deviation = 5.0526**

1. Student’s marks are in a range 30 to 60 and it is arranged in ascending order so we can say that his scores are increasing in every tests.

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

Answer: The skewness of such data set will be 0 and it will be a symmetric distribution.

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

Answer: If the mean is greater than the median, the distribution is positively skewed.

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

Answer: If the mean is less than the median, the distribution is negatively skewed.

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: Negative excess values of kurtosis indicate that a distribution is flat and has thin tails.

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



1. What can we say about the distribution of the data?
2. What is nature of skewness of the data?
3. What will be the IQR of the data (approximately)?

**Solution:**

1. Majority of data is on the right side.

The data is not equally distributed across the plane.

The median of the data is approximately 15.2

25 percent of the data lies between 0-10

50 percent of the data lies between 10-18

25 percent of the data lies after 18-20

1. The nature of the skewness is negative as it is skewed towards left.
2. The IQR approximately is 8, i.e. Q1 = 10, Q3 = 18 so IQR = Q3 – Q1

IQR = 18 – 10 = 8(approximately form the diagram).

Q19) Comment on the below Boxplot visualizations?



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

**Solution:**

The lines coming out from each box extend from the maximum to the minimum values of each set. Together with the box, the whiskers show how big a range there is between those two extremes. Larger ranges indicate wider distribution, that is, more scattered data. The same thing can be said about the boxes. Short boxes mean their data points consistently hover around the center values (Boxplot 1). Taller boxes imply more variable data (Boxplot 2).

We can see that the median of both the Boxplot 1 and 2 lies approximately at the same point on the scale.

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

**Solution:**

[Cars Solution.xlsx](Cars%20Solution.xlsx)

|  |  |  |
| --- | --- | --- |
| **Solution** | | |
|
|  | Mean | Standard Deviation |
| MPG | 34.4221 | 9.131444732 |

* 1. P(MPG>38)

= mean(MPG)

= 34.42208

=sd(MPG)

=9.131445

=1- pnorm(38, mean(MPG), sd(MPG))

= 0.330

= 33%

* 1. P(MPG<40)

=pnorm(40, mean(MPG), sd(MPG))

=0.7293499

=72.3%

* 1. P (20<MPG<50)

=pnorm(50,mean(MPG),sd(MPG))-pnorm(20,mean(MPG),sd(MPG))

=0.955 -0.057

=0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Solution:**

After analyzing the graph, we can see that majority of the data lies near the line so we can state that our data is normally distributed.

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

Dataset: wc-at.csv

**Solution:**

[WC-AT](wc-at%20Solution.xlsx)

After looking at the QQ-plot for both data, we can say that the points are not near the line. Data contains some outliers. Hence the data may have skewness (positive or negative).

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

**Solution:**

1. 90%

* 95+2.5
* 97.5
* qnorm(0.975) =1.96

1. 94%

* 94+4
* 97
* qnorm(0.97) =1.88

1. 60%

* 60 + 20
* 80
* qnorm(0.80) = 0.841

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

**Solution:**

T-Score Calculation

T((1, alpha),(n-1))

Here n = 25, n-1 = 24

1. 95%

* qt(0.975,24)
* 2.063899

1. 96%

* qt(0.98,24)
* 2.171545

1. 99%

* qt(0.995,24)
* 2.7969

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:**

Sample size = n = 18

Sample mean = x = 260 days

Sample standard deviation = s = 90 days

* 260 – 270/90/SQRT(18)
* -10/9.487
* -1.054