|  |  |
| --- | --- |
| 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 |

**Name: - UPENDRA DAMA**

**Batch: - 25/01/2020 (Weekend)**

**Module: - 2**

Q1) Identify the Data type for the Following:

**Explanation: -** Discrete Data is counted whereas continuous data is measured. In other words, Discrete Random variables can take on distinct or separate values. Whereas continuous random variables that can take on any value in a range.

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) | Interval |
| 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 | Interval |
| Years of Education | Ratio |

**Explanation: -**

*1st Level of Measurement – Nominal Scale: -* Also called the categorical variable scale, is defined as a scale used for labeling variables into distinct classifications and doesn’t involve a quantitative value or order.

*2nd Level of Measurement – Ordinal Scale: -* Is defined as a variable measurement scale used to simply describe the order of variables and not the difference between each of the variables.

*3rd Level of Measurement – Interval Scale: -* Is defined as a numerical scale where the order of the variables is known as well as the difference between these variables.

*4th Level of Measurement – Ratio Scale: -* Is defined as a variable measurement scale that not only produces the order of variables but also makes the difference between variables known along with information on the value of true zero.

In Simple words: -

Nominal – Named Variables

Ordinal – Named + Ordered Variables

Interval – Named + Ordered + Proportionate Interval between variables

Ratio – Named + Ordered + Proportionate Interval between variables + Can accommodate absolute Zero.

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

**Explanation: -** As we know for each coin there are two possibilities of getting head or tail. So, for three coins the total number of possibilities are 8 and they are (HHH, HHT, HTH, HTT, THH, THT, TTH, TTT).

Hence the P (Two heads & one tail) = # of interested events/Total # of events = **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

**Explanation: -** First of all I assume the two dice are fair and independent. As we know the total # of possibilities for a single die is 6. So, the total number of possibilities for two dice is 6 \* 6 = 36.

1. P (Sum = 1) = 0 /36 = **0**, because the possibilities start with (1,1) so there is no possibility to have sum is 1. The minimum available sum is 2.
2. P (sum <= 4) = 6/36 = **1/6**, because the available possibilities are {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)}
3. P (sum is divisible by 2 & 3) = 6/36 = **1/6**, because the possibilities are starting from (1,1) till (6,6). So, the sum is between 2 to 12. Between these two numbers only 6 and 12 are divisible by 2 & 3. Hence the total possibilities for sum divisible by 2 & 3 are {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}.

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?

**Explanation: -** A bag contains 2 red, 3 green and 2 blue balls. So, the total possibility of selecting two balls out of total 7 balls are 7C2 = 7! / {(7-2)! \* 2!} = 21.

But the selected balls should not be blue. So, the total possible ways of selecting non-blue balls (2 red + 3 green) are 5C2 = 5! / {(5-2)! \* 2!} = 10

The required probability = P (none of the balls drawn are blue) = **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

**Explanation: -** Formula for expected value = **Σ(Entity \* Probability) = ΣX\*P(X).**

The expected value of candies for a randomly selected child is = 1\*0.015 + 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.



**Explanation: -**

*Mean: - The mean is the arithmetic average of a set of numbers, or distribution. It is the most commonly used measure of central tendency of a set of numbers. But since it is* ***largely influenced by outliers****, we will go for Median. The mean is used for* ***normal distributions****.*

Points = 115.09/32 = 3.5965

Score = 102.952/32 = 3.2173

Weigh = 571.16/32 = 17.8488

*Median: - The median is described as the numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half.* *The median is generally used for* ***skewed distributions****.*

Points = (3.69 + 3.7)/2 = 3.695

Here median is greater than the mean so the distribution is negatively skewed.

Score = (3.215 + 3.435)/2 = 3.325

Here median is greater than the mean so the distribution is negatively skewed.

Weigh = (17.6 + 17.82)/2 = 17.71

Here median is less than the mean so the distribution is positively skewed.

*Mode: - The mode of a set of data is the number with the highest frequency. In this case below are the modes for Points, Score & Weigh, since below numbers occurs frequently compared to rest of the outcomes. If we see for Points & Weigh, they have two modes. This is called* ***bimodal.***

Points = 3.07 & 3.92

Score = 3.44

Weigh = 17.02 & 18.9

*Note: - Above three are used for 1st moment business decision.*

*Variance: - Variance measures how far a set of random numbers are spread out from their mean. It is always positive.*

Points = 8.862/31 = 0.2858814

Score = 29.678748/31 = 0.957379

Weigh = 98.98815/31 = 3.193166

*Standard Deviation: - Since the variance is a squared quantity, it cannot be directly compared to the data values or the mean value of a data set. So, we always consider the square root of the variance which is known as the standard deviation.*

Points = 0.5346787

Score = 0.9784574

Weigh = 1.786943

*Note: - Above two are used for 2nd moment business decision.*

|  |  |
| --- | --- |
| **R Code** | **Python Code** |
| > library(readxl)  > assignment <- read\_excel("~/Desktop/Digi 360/Module 2/assignment.xlsx")  > View(assignment)  > attach(assignment)  The following objects are masked from assignment (pos = 3):  Points, Score, Weigh  #Mean Calculation  > mean(Points)  [1] 3.596563  > mean(Score)  [1] 3.21725  > mean(Weigh)  [1] 17.84875  #Median Calculation  > median(Points)  [1] 3.695  > median(Score)  [1] 3.325  > median(Weigh)  [1] 17.71  #Mode Calculation  > x=table(Points)  > names(x)[which(x==max(x))]  [1] "3.07" "3.92"  > y=table(Score)  > names(y)[which(y==max(y))]  [1] "3.44"  > z=table(Weigh)  > names(z)[which(z==max(z))]  [1] "17.02" "18.9"  #Variance Calculation  > var(Points)  [1] 0.2858814  > var(Score)  [1] 0.957379  > var(Weigh)  [1] 3.193166  #Standard Deviation Calculation  > sd(Points)  [1] 0.5346787  > sd(Score)  [1] 0.9784574  > sd(Weigh)  [1] 1.786943 | import pandas as pd  assignment\_df = pd.read\_excel ("~/desktop/Digi 360/Module 2/assignment.xlsx")  #Mean Calculation.  print("Points Mean:", assignment\_df.Points.mean())  print("Score Mean:", assignment\_df.Score.mean())  print("Weigh Mean:", assignment\_df.Weigh.mean())  #Median Calculation.  print("Points Median:", assignment\_df.Points.median())  print("Score Median:", assignment\_df.Score.median())  print("Weigh Median:", assignment\_df.Weigh.median())  #Mode Calculation.  print("Points Mode:", assignment\_df.Points.mode())  print("Score Mode:", assignment\_df.Score.mode())  print("Weigh Mode:", assignment\_df.Weigh.mode())  #Variance Calculation  print("Points Variance:", assignment\_df.Points.var())  print("Score Variance:",assignment\_df.Score.var())  print("Weigh Variance:", assignment\_df.Weigh.var())  #Standard Deviation Calculation  print("Points Std Deviation :", assignment\_df.Points.std())  print("Score Std Deviation :", assignment\_df.Score.std())  print("Weigh Std Deviation :", assignment\_df.Weigh.std())  **Output: -**  Points Mean: 3.5965625000000006  Score Mean: 3.2172499999999995  Weigh Mean: 17.848750000000003  Points Median: 3.6950000000000003  Score Median: 3.325  Weigh Median: 17.71  Points Mode: 0 3.07  1 3.92  dtype: float64  Score Mode: 0 3.44  dtype: float64  Weigh Mode: 0 17.02  1 18.90  dtype: float64  Points Variance: 0.28588135080645166  Score Variance: 0.9573789677419356  Weigh Variance: 3.193166129032258  Points Std Deviation : 0.5346787360709716  Score Std Deviation : 0.9784574429896967  Weigh Std Deviation : 1.7869432360968431 |

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?

**Explanation: -** The probability of choosing one patient randomly is 1/9. So, the expected value E(X) = ΣXP(X) = 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) = 145.333 which is equal to mean.

**i.e. The expected value for the discrete random variable is mean.**

Q9) Look at the data given below. Plot the data, find the outliers and find out

|  |  |
| --- | --- |
| **Name of company** | **Measure X** |
| Allied Signal | 24.23% |
| Bankers Trust | 25.53% |
| General Mills | 25.41% |
| ITT Industries | 24.14% |
| J.P.Morgan & Co. | 29.62% |
| Lehman Brothers | 28.25% |
| Marriott | 25.81% |
| MCI | 24.39% |
| Merrill Lynch | 40.26% |
| Microsoft | 32.95% |
| Morgan Stanley | 91.36% |
| Sun Microsystems | 25.99% |
| Travelers | 39.42% |
| US Airways | 26.71% |
| Warner-Lambert | 35.00% |

**Explanation: -** *A boxplot, is a way to show the spread and*[*centers*](https://www.statisticshowto.datasciencecentral.com/center-of-a-distribution/)*of a data set. Measures of spread include the interquartile range and the mean of the data set. Measures of center include the mean and median.*

*Below five-piece information from boxplot gives us how the data is spread out.*

*The minimum - Q1 -1.5\*IQR*

*First Quartile Q1 (25th Percentile) - The middle number between the smallest number (not the “minimum”) and the median of the dataset.*

*The median – Q2 (50th Percentile) – The middle of the dataset.*

*Third Quartile Q3 (75th Percentile) - The middle value between the median and the highest value (not the “maximum”) of the dataset.*

*The Maximum - Q3 + 1.5\*IQR*

***Calculations: -***

***Q1*** *= Book2['Measure X'].quantile(0.25)*

*Print(Q1) =* ***0.2547***

***Q3*** *= Book2['Measure X'].quantile(0.75)*

*print(Q3) =* ***0.33975***

*Inter Quartile Region =* ***IQR*** *= 25th Percentile to 75th Percentile = Q3 – Q1 =* ***0.08505***

***Q2*** *= Median =* ***0.2671***

***The minimum*** *= Q1-1.5\*IQR = 0.2547 - 1.5(0.08505) = 0.1271*

***Th maximum*** *= Q3+1.5\*IQR = 0.33975 + 1.5(0.08505) = 0.4673*

***Outliers =*** *Book2[(Book2['Measure X'] < Q1-1.5\*IQR ) | (Book2['Measure X'] > Q3+1.5\*IQR)]['Measure X']*

*print("Outliers:", Outlr) = 10* ***0.9136***

|  |  |
| --- | --- |
| **R Code** | **Python Code** |
| [Workspace loaded from ~/Desktop/Digi 360/R/Upendra/.RData]  > library(readxl)  > assignment <- read\_excel("~/Desktop/Digi 360/Module 2/assignment.xlsx")  New names:  \* Mean -> Mean...2  \* `X-Mu` -> `X-Mu...3`  \* `sqrt(X-Mu)` -> `sqrt(X-Mu)...4`  \* Mean -> Mean...6  \* `X-Mu` -> `X-Mu...7`  \* … and 2 more problems  > View(assignment)  > attach(assignment)  >  > attach(assignment)  The following objects are masked from assignment (pos = 3):  ...10, Mean...2, Mean...6, Measure X, Name of company, Points, Score,  sqrt(X-Mu)...4, sqrt(X-Mu)...8, Weigh, X-Mu...3, X-Mu...7  > boxplot(`Measure X`)  > boxplot(`Measure X`,col=c("#FF000099"),  + medcol=c("#FFDB00FF"),  + whiskcol=c("#49FF00FF"),  + staplecol=c("#00FF92FF"),  + boxcol=c("#0092FFFF"),  + outcol=c("#4900FFFF"),  + outbg=c("#FF00DB66"),  + outcex=3, outpch=21)    > **mean**(`Measure X`)  [1] 0.3327133  > **var**(`Measure X`)  [1] 0.02871466  > **sd**(`Measure X`)  [1] 0.169454  **#Finding outliers**  > y <- boxplot(`Measure X`)  > y$out  [1] 0.9136 | import pandas as pd  import matplotlib.pyplot as plt  import numpy as np  import seaborn as sns  Book2 = pd.read\_excel ("~/desktop/Digi 360/Module 2/Data Sets/Book2.xlsx")  sns.boxplot(y=Book2['Measure X'])  plt.show()  print("Mean:", Book2['Measure X'].mean())  print("Variance:", Book2['Measure X'].var())  print("Std Dvtn:", Book2['Measure X'].std())  Q1 = Book2['Measure X'].quantile(0.25)  Q3 = Book2['Measure X'].quantile(0.75)  IQR = Q3 - Q1  Outlr = Book2[(Book2['Measure X'] < Q1-1.5\*IQR ) | (Book2['Measure X'] > Q3+1.5\*IQR)]['Measure X']  print("Outliers:", Outlr)  **Output:-**  /var/folders/kv/w79zffc14fd2hj518gqdhnmc0000gn/T/com.microsoft.Word/Content.MSO/4D7EB8E1.tmp  **Mean**: 0.3327133333333333  **Variance**: 0.028714661238095233  **Std Dvtn**: 0.16945400921222029  **Outliers**: 10 0.9136  Name: Measure X, dtype: float64 |

Q10) AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that “could happen.” Suppose that one in 200 long-distance telephone calls is misdirected. What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

**Explanation: -**

**The probability of one wrong number is 1/200.**

**Probability of not a wrong number is 1-(1/200) = 199/200**

**Probability of at least one out of five is a wrong number = 1 - Probability that all five calls are not wrong numbers = 1 – (199/200 \* 199/200 \* 199/200 \* 199/200 \* 199/200) = 1 – (199/200) ^5 = 1 – 0.9752 = 0.02475 ~ 0.025 = 25%**

Q11) Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution

|  |  |
| --- | --- |
| x | P(x) |
| -2,000 | 0.1 |
| -1,000 | 0.1 |
| 0 | 0.2 |
| 1000 | 0.2 |
| 2000 | 0.3 |
| 3000 | 0.1 |

1. What is the most likely monetary outcome of the business venture?

**Explanation: - Maximum possibility is 0.3. So, the corresponding outcome is 2000 which is the most likely monetary outcome.**

1. Is the venture likely to be successful? Explain

**Explanation: - Total probability for X > 0 is 0.6 (02 + 0.3 + 0.1). That means there is 60% chance that venture would give profits or more than the expected returns.**

**Total probability for X < 0 is 0.2 (0.1 + 0.1). That means there is 20% chance that venture would give losses.**

**So, over all venture likely to be successful.**

1. What is the long-term average earning of business ventures of this kind? Explain

**Explanation: - Long term average earning is nothing but expected earnings which is ΣXP(X) = -2000\*0.1 + -1000\*0.1 + 0\*0.2 + 1000\*0.2 + 2000\*0.3 + 3000 \* 0.1 = 800**

**This means the average expected earnings over a long period of time would be 800 including all losses and gains.**

1. What is the good measure of the risk involved in a venture of this kind? Compute this measure

**Explanation: - Total probability for X < 0 is 0.2 (0.1 + 0.1). That means the risk associated with this venture is 20% which is pretty good to start a new venture.**