DCS404-Artificial Intelligence and Machine Learning
March-24 Batch;Presidential Graduate School
WESTCLIFF UNIVERSITY

Problem Set-02
A Review on Descriptive Statistics

March 7 2024

Contents

0.1 Graphical Methods for Describing Data.

For this section you can either:

- Use pen and paper to complete the task.
- Use Excel/Google Sheet or any spreadsheet or graphical calculator/programming language to finish the task.

1. Classify each of the following attributes as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.
   (a) Brand of computer purchased by a customer.
   (b) Province of Birth for someone born in Nepal.
   (c) Price of textbook.
   (d) Concentration of a contaminant(micrograms per cubic centimeter)in a water sample.
   (e) Zip Code.
   (f) Actual weight of coffee in a 1-lb can.
   (g) The length of 1 year old cat.
   (h) Brand of a motorcycle purchased

---

2. The scores (out of 100) obtained by 33 students in a mathematics test are as follows: 69, 48, 84, 58, 48, 73, 83, 48, 66, 58, 84, 66, 64, 71, 64, 66, 69, 66, 83, 66, 69, 71, 81, 71, 73, 69, 66, 66, 64, 58, 64, 69, 69. Represent this data in the form of a frequency distribution.

---

3. Suppose we have the following test scores (out of 100) for a class of 40 students:

$$72, 88, 62, 95, 78, 90, 85, 68, 77, 82, 93, 75, 88, 72, 98, 65, 88, 72, 85, 92, 80, 78, 88, 69, 77, 84$$
$$75, 90, 68, 82, 76, 88, 73, 81, 89, 70, 78, 87, 94, 79$$

   Create a grouped frequency table with class intervals and frequencies.
   Review some Definition:
   To create a grouped frequency table, we need to decide on the class intervals. Determining an appropriate class interval involves considering the range of the data and the number of data points. The goal is to create intervals that capture the variability of the data while maintaining a reasonable number of classes for clarity in presentation. One of the common methods for determining class intervals:

   - **Square Root Rule:**

$$\textit{Class Width} = \frac{\textit{Range of Data}}{\textit{Number of Classes}}$$

   - Here:
     - Number of classes: Square root of total number of data points.
     - Range: The difference between the maximum and minimum values (data points).

4. Following is recorded weight of apples:

$$[106, 107, 123, 70, 139, 186, 111, 115, 107, 104107, 115, 125, 126, 119, 84, 141, 98, 81, 110$$
$$76, 82, 109, 93, 187, 95, 111, 92, 86, 68, 130129, 115, 128, 100, 99, 113, 204, 90, 123, 136$$
$$110, 131, 80, 78, 75, 118, 90, 84, 82]$$

   (a) Construct a Simple Frequency Table.

   (b) Construct a Grouped Frequency Table.

   (c) Construct a Relative Frequency Table.

   (d) Construct a Cumulative Frequency Table.

---

5. You conducted a survey to collect the ages of 50 participants in a fitness program. The ages were recorded as follows:

$$32, 40, 38, 45, 50, 34, 42, 36, 48, 55,$$
$$39, 41, 37, 44, 49, 43, 35, 52, 46, 40,$$
$$33, 47, 39, 51, 38, 44, 42, 36, 53, 41,$$
$$37, 45, 39, 54, 40, 38, 36, 43, 50, 42,$$
$$38, 55, 39, 37, 49, 46, 44, 40, 35, 52$$

   Construct a Histogram.

   Steps in Creating a Histogram:

   (a) Determine the number of classes using the Square Root Rule.

   (b) Calculate the class width.

   (c) Create a frequency distribution table with equal intervals.

   (d) Draw a histogram representing the age distribution using the calculated class width and intervals.

   (e) Discuss any patterns or insights you observe from the histogram.

---

6. Consider the data given in the table below: Note:

Weight(gms.)	Frequency
$0 \leq w < 40$	5
$40 \leq w < 50$	6
$50 \leq w < 60$	8
$60 \leq w < 70$	4
$70 \leq w < 100$	2

   In order to keep the histogram fair, the area of the bars rather than the height, must be proportional to the frequency. So on the vertical scale we plot frequency density instead of frequency, where

$$\mathbf{FrequencyDensity} = \frac{\text{Frequency}}{\text{Class Width}}$$

7. The value of π up to 50 decimal places is given below:

$$3.14159265358979323846264338327950288419716939937510$$

(a) Make a frequency distribution of the digits from 0 to 9 after the decimal point.

(b) What are the most and the least frequently occurring digits?

0.2 Measurement of Central Tendency

1. The lengths of time (in minutes) that ten patients at a doctor's clinic wait to see their doctor are as follows: 5, 17, 8, 2, 55, 9, 22, 11, 16, 5.
   (a) What are the mean, median and mode for these datasets?
   (b) What measure of Central Tendency would you use here? Explain and Understand.

   Notes:
   - **Mean:**To calculate the mean $(\bar{X})$, use the formula:
   $$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$
   where n is the number of observations.
   - **Median:**The formula for calculating the median of a dataset is:
   $$\text{Median} = \begin{cases} \frac{n+1}{2}\text{-th observation}, & \text{if } n \text{ is odd}, \\ \frac{(\frac{n}{2})\text{-th observation}+(\frac{n}{2}+1)\text{-th observation}}{2}, & \text{if } n \text{ is even}, \end{cases}$$
   where n is the number of observations.
   - **Mode:**The mode is the value(s) that occur most frequently.

---

2. Given a provided collection D with a mean of 14, find the missing value of y.[**23**].
$$D = [12, 14, 10, 15, 10, 10, y, 14, 13, 11, 10, 10]$$

---

3. Find the mean of 30 given numbers, when it is given that the mean of 10 of them is 12 and the mean of the remaining 20 is 9.[**10**]

---

4. A teacher wants to calculate the overall performance of a student who scored 70 in the midterm (weighted 40%) and 90 in the final exam (weighted 60%). Calculate the weighted mean of the student's scores.[**82**]

---

5. Find out the median from the following data:

Daily wages(Rs)	5	7	8	10	11
No. of Workers	20	15	12	15	18

---

6. For a moderately skewed distribution, the mean and median are respectively 26.8 and 27.9. What can be the mode of distribution? Discuss.(Show your reasoning)

---

7. Calculate the mean and median marks of students from the following distribution.

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Students	7	10	10	20	20	X	8

{Hint: First find the value of X}

---

8. In a class of 100 students, the average amount of pocket money is Rs 35 per student, if the average is Rs 25 for girls and Rs 50 for boys, then the number of girls in the class will be?[**60**].

---

9. Noah scored 20 points in a game. Mai's score was 30 points. The mean score for Noah, Mai, and Clare was 40 points. Explain.[**70**].

0.3 Measurement of Dispersion/Variance

1. Calculate the range and coefficient of range of the following dataset:
   (a) 2,3,5,7,12,15,8,20 [**Range:18; CR:0.82**]
   (b) 2,3,5,5,5,7,14,18,6,25[**Range:23; CR:0.85**]
   (c) Frequency Table: [**Range:20; CR:0.4**]

X	15-20	20-25	25-30	30-35
f	8	21	15	4

   Notes:
   The range (R) is calculated as the difference between the maximum (X_{max}) and minimum (X_{min}) values in a dataset:

$$R = X_{\text{max}} - X_{\text{min}}$$

   The coefficient of range (CR) is a measure of relative variability and is calculated using the formula:

$$CR = \frac{X_{\text{max}} - X_{\text{min}}}{X_{\text{max}} + X_{\text{min}}}$$

$$\text{Range} = \text{Upper limit of highest class interval} - \text{Lower limit of lowest class interval}$$

---

2. For the dataset: [10, 15, 20, 25, 30, 35, 40, 45]. Find the InterQuartile Range (IQR).

---

3. Find the inter-quartile range for the following data: [56, 14, 84, 21, 85, 2, 35, 74, 66, 52,45].[**42**].

---

4. Find any outliers for the following set of data: [1, 3, 4, 6, 13, 20, 25, 26, 28, 62, 95].

---

5. For the dataset:

$$\text{D} = [27, 41, 23, 56, 76, 54, 53, 49, 50, 92, 47, 23, 56, 65, 71, 73, 76, 77]$$

   Calculate following:
   - (Population) Variance [$\approx$ **215.30**] and Standard Deviation [$\approx$ **14.67**].
   - (Sample) Variance [$\approx$ **227.97**] and Standard Deviation[$\approx$ **15.10**].

   Formula Review:

$$\textbf{Population Variance}(\sigma^2) = \frac{\sum(X_i - \mu)^2}{\textbf{N}} \qquad \textbf{Standard Deviation}(\sigma) = \sqrt{\textbf{Variance}(\sigma)^2}$$

$$\textbf{Sample Variance}(S^2) = \frac{\sum(X_i - \bar{X})^2}{\textbf{n-1}} \qquad \textbf{Standard Deviation}(S) = \sqrt{\textbf{Variance}(S)^2}$$

---

6. The following distribution is related to time spent by visitors in the newly opened E-sports center in Kathmandu.

Time spent(mins.)	10	20	30	40	60
Visitors	2	7	15	10	20

Find Variance[**188.08**] and standard deviation [**13.71**].
Formula Review:

$$\textbf{Variance} = \frac{\sum (X - \bar{X})^2 \cdot f_i}{\sum f_i} \qquad \textbf{Standard Deviation} = \sqrt{\text{Variance}}$$

---

7. Following is the marks obtained by 20 students:

Marks	5	9	11	15	18
No. of Students	2	4	8	5	1

Calculate Variance[≈ 10.73] and Standard Deviation [≈ 3.28].

---

8. Find the Standard deviation for the following distribution: [≈ 14.3]

X	10-20	20-30	30-40	40-50	50-60	60-70	70 -80
f	5	12	15	20	10	4	2

---

9. Following table shows the distribution of Nepalese women in economic activities:

Economic activity	1981	1991
Labour Force	34.6	40.4
Agricultural Work	36.4	45.0
Non-agricultural Work	16.6	20.3
Manufacturing	14.9	22.9
Various	46.2	45.8

Find Coefficient of Variance and interpret the result.[$\mathbf{CV_{1981} = 37.20\%; CV_{1991} = 25.20\%}$]

---

10. Suppose you are analyzing the annual returns of two investment portfolios, Portfolio A and Portfolio B, over the past 5 year. The returns are as follows (in percentage):

$$\mathbf{PortfolioA} : [\mathbf{10, 8, 12, 15, 11}]$$

$$\mathbf{PortfolioB} : [\mathbf{5, 9, 11, 14, 10}]$$

Find the Coefficient of Variance and Interpret the result.[$\mathbf{CV_A = 13.30\%}$, $\mathbf{CV_B = 15.41\%}$].

---

11. Two plants C and D of a factory show the following results about the number of workers and the wages paid to them.

Plant	C	D
No. of Workers	5000	6000
Average Monthly Wages ($)	2500	2500
Standard Deviation	9	10

Using coefficients of variation formulas, find in which plant, C or D is there greater variability in individual wages. Interpret the result.[$\mathbf{CV_C \approx 0.36\%}$, $\mathbf{CV_D \approx 0.40\%}$].

---

12. If the coefficient of variation of two distributions are 60 and 70, and their standard deviations are 25 and 16, respectively, find their arithmetic means.[$\mathbf{am_1 = 41.77; am_2 = 22.86}$]

0.4 Bi-Variate Statistics: Co-relation and Co-variance.

1. You are investigating the relationship between the number of hours students spend studying and their exam scores.You collect data from 10 students and want to explore if there is a correlation between study hours and exam scores.

$$\mathbf{StudyHours(X)} : [\mathbf{3, 5, 2, 7, 4, 6, 1, 8, 5, 6}]$$

$$\mathbf{ExamScores(Y)} : [\mathbf{70, 75, 60, 85, 72, 80, 55, 90, 78, 82}]$$

Find the correlation coefficients and interpret the result.[**0.9874**].
Formula Review

$$\textbf{Correlation Coefiicient}(\mathbf{r_{XY}}) = \frac{\sum(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{Y} - \bar{\mathbf{Y}})}{\sqrt{\sum(\mathbf{X} - \bar{\mathbf{X}})^2 \times \sum(\mathbf{Y} - \bar{\mathbf{Y}})^2}}$$

---

2. Suppose you are studying the relationship between the time spent on homework,the number of extracurricular activities and students' overall grades. You collect data from 10 students; and want to explore the correlation among variables.{Hint: Explore the correlation between each pair of variables. i.e $\mathbf{r_{XY}}$; $\mathbf{r_{XZ}}$; $\mathbf{r_{YZ}}$}

$X : [2, 3, 1, 4, 2, 3, 1, 5, 3, 4]$ (hours spent on homework)
$Y : [3, 2, 1, 4, 3, 2, 1, 5, 4, 3]$ (number of extracurricular activities)
$Z : [75, 80, 70, 85, 78, 82, 68, 90, 85, 88]$ (overall grades)

---

3. Look at the following bi-variate data table. It represents the age and average height of a group of babies and kids. Find the correlation coefficient and interpret the result:

Age (mnths.)	Height(cms.)
3	58.5
6	64
9	68.5
12	74
24	81.2
36	89.1
48	95
60	102.5

4. Let's say you have to study the relationship between the age and the systolic blood pressure in a company. You have a sample of 10 workers aged thirty to fifty-five years. The results are presented in the following bi-variate data table:

S.No	Age	Systolic Blood Pressure
1	37	130
2	38	140
3	40	132
4	42	149
5	45	144
6	48	157
7	50	161
8	52	145
9	53	165
10	55	162

Construct a Covariance Matrix.
Notes:
The covariance matrix is a matrix that summarizes the co-variances between multiple variables. If you have two variables, the covariance matrix will be a 2x2 matrix. The general formula for the covariance between two variables X and Y is given by:

$$\mathbf{Cov}(\mathbf{X},\mathbf{Y}) = \frac{\sum(\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{Y_i} - \bar{\mathbf{Y}})}{\mathbf{n-1}}$$

The covariance matrix for two variables X and Y is given by:

$$\mathbf{CovarianceMatrix} = \begin{bmatrix} \mathbf{Cov}(\mathbf{X},\mathbf{X}) & \mathbf{Cov}(\mathbf{X},\mathbf{X}) \\ \mathbf{Cov}(\mathbf{Y},\mathbf{X}) & \mathbf{Cov}(\mathbf{Y},\mathbf{Y}) \end{bmatrix} = \begin{bmatrix} \mathbf{Var}(\mathbf{X}) & \mathbf{Cov}(\mathbf{X},\mathbf{X}) \\ \mathbf{Cov}(\mathbf{Y},\mathbf{X}) & \mathbf{Var}(\mathbf{Y}) \end{bmatrix}$$

---

5. Construct a Covariance Matrix for Question:[2].
Hint:

$$\textbf{Covaraince Matrix}(\Sigma) = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) & \mathrm{Cov}(X,Z) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) & \mathrm{Cov}(Y,Z) \\ \mathrm{Cov}(Z,X) & \mathrm{Cov}(Z,Y) & \mathrm{Var}(Z) \end{bmatrix}$$