Smt. Indira Gandhi College of Engineering
Computer Engineering Department
Ghansoli – Navi Mumbai
Academic Year 2023-24 (Even Sem)

Student Name: Khyati Garude **Roll No.:** 13 **Class:** BE **Sem:** VIII

Course Name: Applied Data Science Lab

Course Code: CSL8023

Experiment No. 02

Experiment Title: Descriptive Statistics Univariate Analysis:

A) Computation Of Central Tendency Using Mean , Median , Mode. (For Non Grouped And Grouped Data)

B) Computation Of Dispersion Using Range , Variance And Standard Deviation (For Non Grouped And Grouped Data)

C) Figure Out The Skewness Of Dataset

Date of Performance	Date of Submission	Marks (10)					Sign / Remark
		A	B	C	D	E	
		2	3	2	2	1	
18/1/24	25/1/24	2	3	2	2	0	20/3/24
		Total Marks					
		09					

A: Prerequisite Knowledge
B: Implementation
C: Oral
D: Content
E: Punctuality & Discipline

DATE	EXPERIMENT-2	SIGN
25/1/24	Descriptive statistics Univariate Analysis: A) Computation of central Tendency using Mean, Median, Mode (for non-grouped and grouped data.) B) Computation of dispersion using Range, variance and standard deviation (for non-grouped and grouped data) c) Figure out the skewness of the Dataset.	20/3/24

AIM: Descriptive statistics Univariate Analysis –

A) Computation of central Tendency using Mean, Median, Mode (for non-grouped and grouped data).

B) Computation of dispersion using Range, variance and standard deviation (for non-grouped and grouped data)

c) Figure out the skewness of the dataset.

Date:_____

THEORY:

Descriptive statistics refers to the study of the aggregate quantities of a dataset.

Examples – average annual income, median home price in a neighbourhood, range of credit scores of a population, etc.

Univariate data exploration denotes analysis of one attribute at a time.

A) Measure of Central Tendency:

The objective of finding the central location of an attribute is to quantify the dataset with one central or most common number.

(i) Mean:

The mean is the arithmetic average of all observations in the dataset. It is calculated by summing all the data points and dividing by the number of data points.

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{(for ungrouped data)}$$

$$\text{Mean} = \frac{\sum_{i=1}^{n} fx}{\sum f} \quad \text{(for grouped data)}$$

Date:_____

(ii) Median:

The median is the value of the central point in the distribution. The median is calculated by sorting all the observations from small to large and selecting the mid-point observation in the sorted list. If the number of data points is even, then the average of the middle two data points is used as the median.

For grouped data: $L + \left[\frac{N/2 - C.f}{f}\right] \times h$

(iii) Mode:

The mode is the most frequently occuring observation. In the dataset, data points may be repetitive, and the most repetitive data point is the mode of the dataset. For grouped data, we consider the frequency of each data.

$$\text{Mode} = L + \left[\frac{f_1 + f_0}{2f_1 - f_0 - f_2}\right] \times h$$

B) Computation of Dispersion:

(i) Range:

Range is the difference between the maximum and the minimum value in a dataset.

(ii) Variance:

Variance measures the sum of the squared deviations of all data points divided by the number of

Date:_____

data points.

$$\sigma^2 = \frac{\Sigma f(x-\bar{x})^2}{n-1} \quad \text{(for grouped data)}$$

$$\sigma^2 = \frac{\Sigma (x-\bar{x})^2}{n-1} \quad \text{(for ungrouped data)}$$

(iii) Standard deviation:

It is the square root of the variance. High standard deviation means the data points are spread widely around the central point. Low standard deviation means data points are closer to the central point.

$$\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{n-1}} \quad \text{(for grouped data)}$$

$$\sigma = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}} \quad \text{(for ungrouped data)}$$

c) Skewness of the dataset:

Skewness measures the asymmetry of the data distribution, that deviates from the symmetrical normal distribution (bellcurve) in a given set of data. In positively skewed, the mean of the data is greater than the median (a large number of data is pushed on the right-hand side).

Date:_____

In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side)

ALGORITHM:

(i) Import libraries: pandas for data handling, variance and stdev functions from statistic model.
(ii) Read data from csv file into a pandas dataframe.
(iii) Calculate mean, median and median:
- for ungrouped data, use inbuilt .mean(), .median(), .mode() functions.
- for grouped data, use formulas
(iv) Calculate range, variance and standard deviation:-
- for ungrouped data, use (max-min) for range, variance() for variance and stdev for standard deviation.
- for grouped data use formulas.
(v) Calculate skewness using the .skew() function
(vi) Print the results.

Date:_____

WORKING:

→ Calculating mean, median, mode, range, standard deviation, variance and skewness for ungrouped data:

Age	Sleep efficiency (x)	$(x-\bar{x})$	$(x-\bar{x})^2$
11	0.55	-0.188	0.035344
27	0.54	-0.198	0.039204
36	0.9	0.162	0.026244
40	0.51	-0.228	0.051984
40	0.89	0.152	0.023104
41	0.79	0.052	0.002704
53	0.9	0.162	0.026244
57	0.76	0.022	0.000484
65	0.88	0.142	0.020164
69	0.66	-0.078	0.006084
	$\Sigma x = 7.38$		$0.23156 = \Sigma$

(i) Mean $(\bar{x}) = \frac{\Sigma x}{n} = \frac{7.38}{10} = 0.738$

(ii) Median $= \left[\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}\right]$ term $= \left[\frac{\left(\frac{10}{2}\right)^{th} + \left(\frac{10}{2}+1\right)^{th}}{2}\right]$ term

$= \frac{[5^{th} + 6^{th}]}{2}$ term $= \frac{0.89+0.79}{2} = 0.84$

(iii) Mode = 0.9

Date:_____

(iv) Range = Highest value − Lowest value

= 0.9 − 0.51

= 0.39

(v) Variance = $\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{0.23156}{9} = 0.0257$

(vi) Standard deviation = $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

$= \sqrt{0.0257}$

$= 0.16$

(vii) Skewness = $\frac{\Sigma_{i=1}^{n}(x_i-\bar{x})^3/n}{\sigma^3}$

$= \frac{-0.01313868/10}{(0.16)^3}$

$= -0.3207$

Date:_____

→ Calculating mean, median, mode, range, standard deviation, variance and skewness for grouped data:

Sleep efficiency (Class Interval)	Frequency (f)	Class Mark (x)	fx	cf	$(x-\bar{x})^2 \times f$
0.5 - 0.55	38	0.525	19.95	38	2.7476
0.55 - 0.6	22	0.575	12.65	60	1.0541
0.6 - 0.65	29	0.625	18.125	89	0.8272
0.65 - 0.7	29	0.675	19.575	118	0.4099
0.7 - 0.75	36	0.725	26.1	154	0.1708
0.75 - 0.8	46	0.775	35.65	200	0.0164
median class → 0.8 - 0.85	54	0.825	44.55	254	0.0522
0.85 - 0.9	68	0.875	59.5	322	0.4472
mode class → 0.9 - 0.95	98	0.925	90.65	420	1.6843
0.95 - 1	37	0.975	36.075	457	1.2134
Σ =	457		362.825		8.6231

Date:_____

(i) Mean $(\bar{x}) = \frac{\Sigma fx}{\Sigma f} = \frac{362.825}{457} = \underline{0.7939}$

(ii) Median $= L + \left[\frac{(N/2) - Cf.}{f}\right] h$

$= 0.8 + \left[\frac{228.5 - 200}{54}\right] 0.05$

$= \underline{0.82638}$

(iii) Mode $= L + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] h$

$= 0.9 + \left[\frac{98 - 68}{2(98) - 68 - 37}\right] 0.05$

$= 0.9 + \left[\frac{30}{91}\right] 0.05 = \underline{0.9164}$

(iv) Range $= UCB_H - LCB_L = 1 - 0.5 = \underline{0.5}$

(V) Variance $= \sigma^2 = \frac{\Sigma f(x - \bar{x})^2}{n-1} = \frac{8.6231}{456} = \underline{0.0189}$

(vi) Standard deviation $= \sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{n-1}} = \sqrt{0.0189}$

$= \underline{0.1373}$

Date:_____

(vii) Skewness $= \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$

$= \frac{3\,(0.7939 - 0.82638)}{0.1373}$

$= -0.709$

CONCLUSION:

In this analysis, we computed central tendency using mean, median, and mode for both non-grouped and grouped data, shedding light on the dataset's typical values. Dispersion measures like range, variance, and standard deviation provided insights into data spread. Additionally, skewness analysis illuminated the distribution's asymmetry. These computations offer a comprehensive overview of the dataset's characteristics, crucial for understanding its nature and making informed decisions.

20/3/24

```
# Ungrouped data
import csv
import statistics
from scipy.stats import skew,mode

# Update the file path to '/content/Sleep_Efficiency.csv'
file_path = '/content/Sleep_Efficiency.csv'

def read_sleep_data(file_path):
    sleep_efficiency_data = []

    try:
        with open(file_path, 'r') as file:
            reader = csv.reader(file)
            next(reader)  # Skip header if exists

            for row in reader:
                sleep_efficiency_data.append(float(row[6]))  # Assuming sleep

        return sleep_efficiency_data
    except FileNotFoundError:
        print(f"File not found at {file_path}. Please check the file path.")
        return None

def calculate_additional_statistics(data):
    data_range = max(data) - min(data)
    variance_value = statistics.variance(data)
    standard_deviation_value = statistics.stdev(data)
    skewness_value = skew(data)

    return data_range, variance_value, standard_deviation_value, skewness_valu

sleep_efficiency_data = read_sleep_data(file_path)

if sleep_efficiency_data is not None:
    # Calculate sleep statistics
    mean_value = statistics.mean(sleep_efficiency_data)
    median_value = statistics.median(sleep_efficiency_data)
    # Handle multiple modes
    mode_result = mode(sleep_efficiency_data)
    mode_values = mode_result.mode
    mode_counts = mode_result.count
    if isinstance(mode_values, float):
        mode_values = [mode_values]
    mode_value = mode_values[0] if len(mode_values) == 1 else "Multiple modes"

    # Display the results
```

```
    print(f"Sleep Efficiency -")
    print(f"Mean: {mean_value}")
    print(f"Median: {median_value}")
    print(f"Mode: {mode_value} (Counts: {mode_counts})")

    # Calculate additional statistics
    data_range, variance_value, standard_deviation_value, skewness_value = cal

    # Display additional results
    print(f"Range: {data_range}")
    print(f"Variance: {variance_value}")
    print(f"Standard Deviation: {standard_deviation_value}")
    print(f"Skewness: {skewness_value}")
```

```
Sleep Efficiency -
Mean: 0.7889159292035398
Median: 0.82
Mode: 0.9 (Counts: 29)
Range: 0.49
Variance: 0.01828906608716127
Standard Deviation: 0.13523707364166554
Skewness: -0.6481135894464444
```

```
Grouped data
mport pandas as pd
mport numpy as np
rom scipy.stats import skew

 Load the dataset
ile_path = '/content/Sleep_grouped.csv'
ata = pd.read_csv(file_path)
 Convert the column to numeric values
ata['Sleep efficiency'] = data['Sleep efficiency'].apply(lambda x: np.mean(lis
 Drop rows with NaN values
ata = data.dropna()

 Check if there is at least one valid row
f data.shape[0] > 0:
    # Extract Sleep Efficiency and Frequency columns
    sleep_efficiency = data['Sleep efficiency']
    frequency = data['Frequency']

    # Mean
    mean = (sleep_efficiency * frequency).sum() / frequency.sum()

    # Median
    # Assuming the data is already grouped, you can directly calculate the medi
```

```
# using the midpoint of the cumulative frequency
cumulative_frequency = frequency.cumsum()
median_index = cumulative_frequency[cumulative_frequency >= frequency.sum()
median = sleep_efficiency.iloc[median_index]

# Mode
mode_index = frequency.idxmax()
mode = sleep_efficiency.iloc[mode_index]

# Range
data_range = sleep_efficiency.max() - sleep_efficiency.min()

# Standard Deviation and Variance
weighted_sum_squares = ((sleep_efficiency - mean) ** 2 * frequency).sum()
variance = weighted_sum_squares / frequency.sum()
std_deviation = variance ** 0.5

# Skewness
deviation_from_mean = sleep_efficiency - mean
cubed_deviation = deviation_from_mean ** 3
weighted_cubed_deviation = cubed_deviation * frequency
skewness = (weighted_cubed_deviation.sum() / frequency.sum()) / (std_deviat

# Output the results
print(f"Sleep Efficiency -")
print(f"Mean: {mean}")
print(f"Median: {median}")
print(f"Mode: {mode}")
print(f"Range: {data_range}")
print(f"Standard Deviation: {std_deviation}")
print(f"Variance: {variance}")
print(f"Skewness: {skewness}")
lse:
print("No valid data to calculate statistics.")
```

```
Sleep Efficiency -
Mean: 0.7939277899343544
Median: 0.825
Mode: 0.925
Range: 0.44999999999999996
Standard Deviation: 0.13736958897834126
Variance: 0.018870403976078413
Skewness: -0.5844884403176561
```

20/3/24