Chapter 1
Introduction to Statistics and Data Analysis

1.1 Overview

Objectives:

1. To define some important concepts concerning statistics.

Definition: **Statistics** is the science that is interested in collecting, organizing, summarizing data, then analyzing it and getting results to reach decisions.
Definition: **Population** is the collection of all elements under study.
Definition: **Sample** is a random subset of the population.

1.4 Measures of Location

Objectives:

1. To define and calculate two important measures of location.
2. To identify two advantages of the sample median.

Definition: The **sample mean** is the arithmetic average and denoted by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Ex.1: Find the mean of the following observations,

56 , 65 , 60 , 58 , 57

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{56, 65, 60, 58, 57}{5} = \frac{296}{5} = 59.2$$

Note: The **population** mean, denoted by μ is given by,

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$, N: Population size

Definition: The **sample median** is the observation that lies on the middle denoted by $\tilde{x}$, given that the observations are sorted in ascending order and is given by, $\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{, n is odd} \\ \dfrac{x_{(n/2)} + x_{\left(n/2+1\right)}}{2} & \text{, n is even} \end{cases}$

Ex.2: (1/20) b. Arrange the data in ascending order as follows,

110 , 111 , 116 , 117 , 118 , 122 , 123 , 125 , 126 , 175

Advantages of sample median:

1. Easier to calculate since it depends only on one or two observations only.
2. Unlike the mean, it is not affected by outliers (extreme values).

Definition: The **trimmed mean** is a mean computed by trimming out a certain percent of both largest and smallest set of values.

Ex.3: Refer to (1/20) and find the trimmed mean by eliminating 10% of the data. Answer is 119.75 (closer to the mean).

1.5 Measures of Variability

Objectives:

1. To define and calculate two important measures of variability.
2. To identify two advantages of the sample variance.

They express the degree to which individual observations differ from each other.

Definition: The sample **range** is given by; $R = X_{max} - X_{min}$.

Definition: The sample **variance**, denoted by s^2, is given by;

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ or } s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Definition: The sample **standard deviation**, denoted by *s*, is given by;

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

where n-1 are called the degrees of freedom.

Notes:

1. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, the sum of the deviations of the observations about their arithmetic mean is always 0.
2. If all the observations are the same ($x_1 = x_2 = ... = x_n$), then $s^2 = s = 0$.

Ex.4: (2/20) The data are 572 572 573 568 569 575 565 570.

a. The mean = 570.5 mm.

The sorted data are 565 568 569 570 572 572 573 575.

b. The range = 10 mm.

The sample variance = $\frac{2603832 - 8(570.5)^2}{7} = \frac{70}{7} = 10 \text{ mm}^2$.

The sample standard deviation = $\sqrt{10} = 3.16228$ mm.

Notes:

1. The units for s^2 and *s* are *unit*^{2} and *unit* respectively.
2. The **population variance** and **standard deviation** are given by,

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} X^2{}_i - n\mu^2\right) \quad \text{and} \quad \sigma = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N} X^2{}_i - n\mu^2\right)}$$

respectively.

1.6 Discrete and Continuous Data

Objectives:

1. To define both discrete and continuous data.

Definition: The **discrete** data is any *countable* set of observations.

Definition: The **continuous** data is any *uncountable* set of observations (intervals).

1.8 Graphical Methods and Data Description.

Objectives:

1. To plot some graphical presentations of the data.
2. To construct a frequency distribution.

I. Stem-and-Leaf plot

It is a way of arranging the raw data by separating the values into a stem digit and a leaf digit.

Ex.5: Consider table 1.1 for the battery life in p.16. Construct a stem-and-leaf plot for the data.

An ordered stem-and- leaf plot

Stem	Leaf	Frequency
1	6 9	2
2	2 5 6 6 9	5
3	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4 5 5 6 7 7 7 8 8 9 9	25
4	1 1 2 3 4 5 7 7	8

Since the plot does not provide an adequate picture of the distribution, so the stem value can be written twice as follows

1*	
1●	6 9
2*	2
2●	5 6 6 9
3*	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4
3●	5 5 6 7 7 7 8 8 9 9
4*	1 1 2 3 4
4●	5 7 7

Note: Choosing the stem value depends on the data set such that the smallest **significant** digit is assigned to the leaf and the other greater digits are assigned to the stems.

II. Frequency Distribution

It is a **table** made by arranging the raw data into intervals or classes.

Definition: The class **frequency** is the number of observations falling in the class.

Definition: The class **relative frequency** is the frequency divided by the frequency total.

Definition: The class **midpoint** for a class $[a\ ,\ b)$ is $\frac{a+b}{2}$.

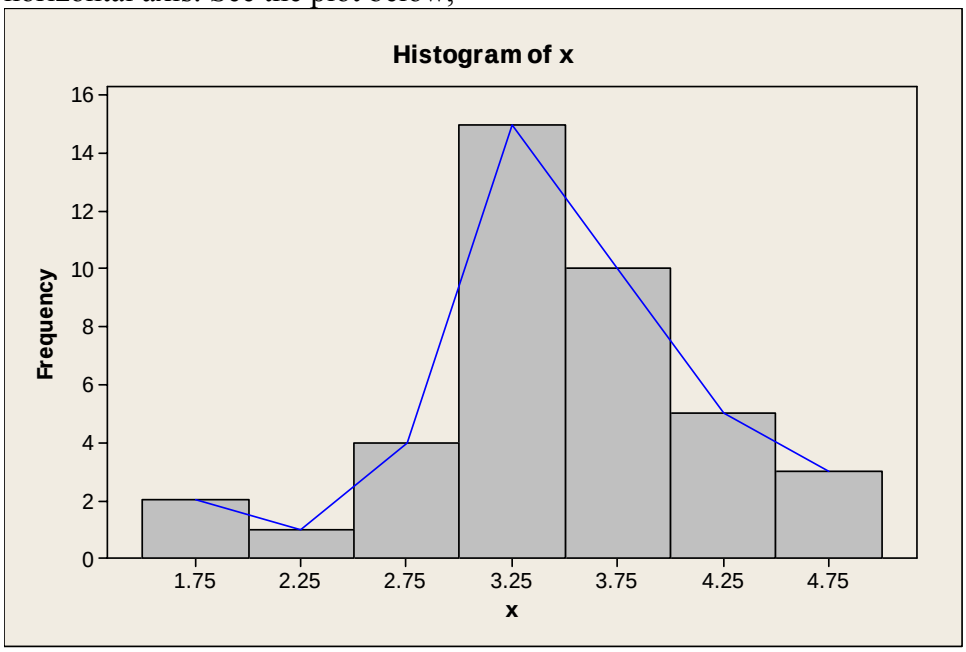
Ex.6: Construct a frequency table (distribution) for the data in table 1.1.

Class interval	Frequency	Midpoint	Relative frequency
[1.5 , 2.0)	2	1.75	2/40=0.05
[2.0 , 2.5)	1	2.25	1/40=0.025
[2.5 , 3.0)	4	2.75	4/40=0.100
[3.0 , 3.5)	15	3.25	15/40=0.375
[3.5 , 4.0)	10	3.75	10/40=0.25
[4.0 , 4.5)	5	4.25	5/40=0.125

[4.5 , 5.0)	3	4.75	3/40=0.075
Total	**40**		**1.000**

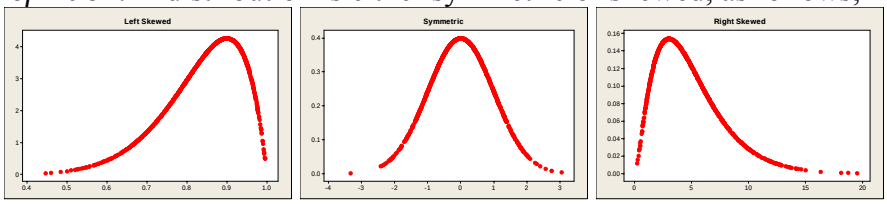
III. Frequency (Relative Frequency) Histogram

To plot a histogram, or a relative frequency histogram, we assign the frequencies, or the r.f., to the vertical axis and the class limits to the horizontal axis. See the plot below,



If the top centers of the rectangles are connected by line segments then the plot is called a frequency, or a r.f., **polygon**, and if the polygon is smoothed by hand it is called a **curve**.

Definition: A distribution is either symmetric or skewed, as follows;



and the skewed one is either a right-skewed (positive-skewed) or a left-skewed (negative-skewed).

The Empirical Rule (ER) and Grouped Data

Objectives:

1. **To verify the empirical rule for a set of data.**
2. **To define and calculate coefficients of variation & skewness.**
3. **To define and calculate the approximate mean and variance of the grouped data.**

Definition: If the relative frequency distribution of the data is approximately bell shaped, i.e. unimodal and symmetric, then

a. $P[\mu-\sigma < X < \mu+\sigma] = P[\bar{x}-s < X < \bar{x}+s] \approx 68\%$

b. $P[\mu-\frac{3}{2}\sigma < X < \mu+\frac{3}{2}\sigma] = P[\bar{x}-\frac{3}{2}s < X < \bar{x}+\frac{3}{2}s] \approx 86\%$

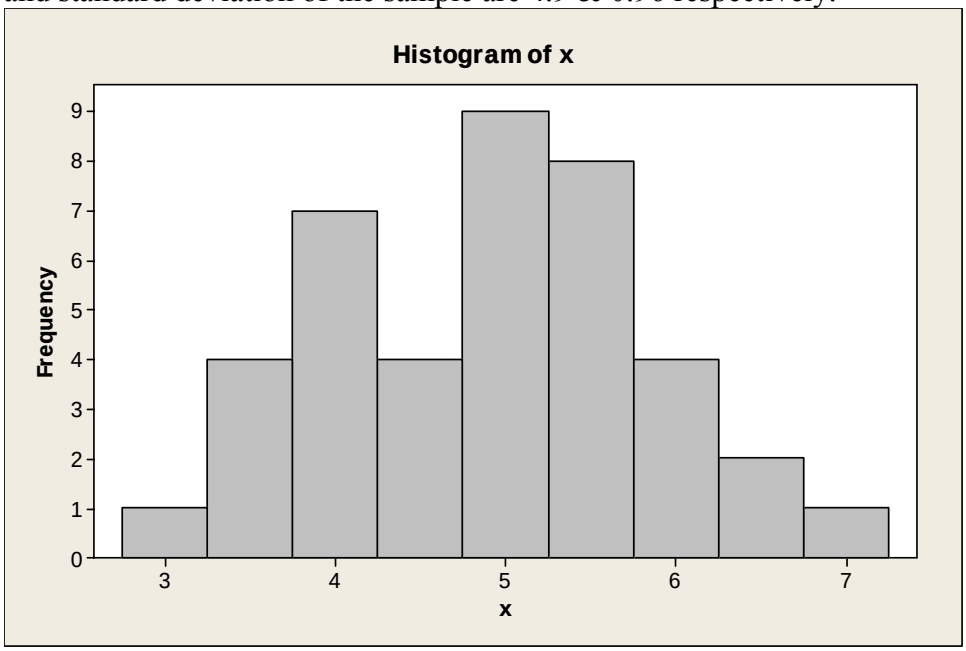
c. $P[\mu-2\sigma < X < \mu+2\sigma] = P[\bar{x}-2s < X < \bar{x}+2s] \approx 95\%$

d. $P[\mu-3\sigma < X < \mu+3\sigma] = P[\bar{x}-3s < X < \bar{x}+3s] \approx 100\%$

A population, or sample, satisfying these properties is said to satisfy ER.

Ex.7: Verify that the following sample satisfies ER

2.8 3.3 3.3 3.5 3.7 3.8 4.0 4.1 4.1 4.1 4.2 4.2 4.3 4.4 4.4
4.5 4.8 4.8 4.8 4.9 5.0 5.1 5.1 5.2 5.2 5.3 5.4 5.4 5.4 5.5
5.5 5.5 5.7 5.9 6.0 6.0 6.2 6.3 6.5 7.0, given that the mean and standard deviation of the sample are 4.9 & 0.96 respectively.



Definition: The sample **coefficient of variation** is the proportion of the sample standard deviation to the sample mean in a percentage form as follows; $CV = \frac{s}{\bar{x}} \times 100\%$.

Definition: The sample **coefficient of skewness** is given by $CS = \frac{\bar{x} - \tilde{x}}{s/3}$ which indicates the direction of the relative frequency distribution.

Ex.8: Referring to example 5, calculate the coefficient of variation and the coefficient of skewness for the battery lives.

Mean and Variance of Grouped Data

If the sample data are grouped in a frequency table with k groups, or classes, then the approximate mean and variance are given by;

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j} = \frac{x_1 f_1 + x_2 f_2 + ... + x_k f_k}{f_1 + f_2 + ... + f_k} \text{ and}$$

$$s^2 = \frac{\sum_{j=1}^{k} \left(x_j - \bar{x}\right)^2 f_j}{\sum_{j=1}^{k} f_j - 1} = \frac{\sum_{j=1}^{k} X_j^2 f_j - n\bar{x}^2}{\sum_{j=1}^{k} f_j - 1} \text{ respectively.}$$

Ex.9: Referring to example 6, calculate the approximate mean and variance for the frequency table of the battery lives sample.

Consider the following table;

$\bar{x}_{raw} = 3.4125$ and $s^2_{raw} = 0.702810$ where $\bar{x} = 3.46250$ and $s^2 = 0.485737$

Class interval	**Frequency (*f*)**	**Midpoint (*x*)**	***xf***	***x*²*f***
[1.5 , 2.0)	2	1.75	3.50	6.125
[2.0 , 2.5)	1	2.25	2.25	5.063
[2.5 , 3.0)	4	2.75	11	30.250
[3.0 , 3.5)	15	3.25	48.75	158.438
[3.5 , 4.0)	10	3.75	37.5	140.625
[4.0 , 4.5)	5	4.25	21.25	90.313
[4.5 , 5.0)	3	4.75	14.25	67.688
Total	**40**		**138.5**	**498.5**