جامعة التكنولوجية
قسم الهندسة الكيميائية
المرحلة الثانية
الاحصاء والقياسات
م. ابتسام حسين
References

1. Calculus and analytic geometry
   By Thomas / Finney
   Sixth Edition

2. Advanced engineering mathematics
   By C. Ray Wylie
   Louis C. Barrett
   Fifth Edition

3. Mathematical methods for science students
   By G. Stephenson
   Second Edition

4. Mathematical methods in chemical engineering
   By V. G. Jenson and
   G. V. Jeffreys
   Second Edition
Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

- **Population**: set of all possible measurements.

- **Finite**: all bots produced in a factory catalyst pellets.

- **Infinite**: all possible outcomes (heads, tails) in successive tosses of a coin.

- **Sample**: a set of measurements taken to represent an infinite or large finite population, which is selected randomly.

- **Random Sample**: is selected so that all elements of the population have an equal chance of being measured.
- Sample array: is the set of measurements of sample elements.

- Inductive or statistical inference: if a sample is representative of a population, important conclusions about the population can often be inferred from the analysis of the sample. The phase of statistics dealing with the conditions under which such inference is valid is called inductive statistics.

- Deductive or descriptive statistics: the phase of statistics which seeks only to describe and analyze a given group without drawing any conclusions or inferences about a large group.

- Variable: is a symbol, such as X, Y, which can assume any of a prescribed set of values, called the domain of the variable.

- If the variable can assume only one value, it is called a constant.

- A variable which can theoretically assume any value between two given values is called a continuous variable, otherwise it is
Called a discrete variable:

- The no. N of children in a family, which can assume any of the values 0, 1, 2, 3,... but cannot be 2.5 or 3.842, is a discrete variable.

- The age A of an individual, which can be 62 years, 63.8, etc., depending on accuracy of measurement, is a continuous variable.

- Size of data: number of measurements

- Range: highest - lowest measurement
Chapter 2

Frequency Distribution

- When elements of pop. are unequal in a certain parameter and/or measurement error is involved a statistical estimation is needed. This involves:
  1. Data sampling for repeated measurements
  2. Classification of data (frequency distribution)
  3. Presentation of classified data
  4. Estimation of statistical parameters
  5. Analysis of statistical parameters and hypotheses.

- Frequency distribution into classes: The sample range is subdivided into a number of classes. Usually:

  \[
  \begin{align*}
  \text{for size} & > 50 & \text{10-20 classes} \\
  \text{for size} & \leq 50 & \text{5-10 classes are used.}
  \end{align*}
  \]
Example: The life of electric bulbs in hours was sampled:

<table>
<thead>
<tr>
<th>690</th>
<th>701</th>
<th>722</th>
<th>684</th>
<th>680</th>
</tr>
</thead>
<tbody>
<tr>
<td>728</td>
<td>705</td>
<td>693</td>
<td>691</td>
<td>688</td>
</tr>
<tr>
<td>740</td>
<td>663</td>
<td>676</td>
<td>738</td>
<td>714</td>
</tr>
<tr>
<td>698</td>
<td>687</td>
<td>703</td>
<td>726</td>
<td>699</td>
</tr>
<tr>
<td>694</td>
<td>705</td>
<td>717</td>
<td>682</td>
<td>717</td>
</tr>
<tr>
<td>712</td>
<td>733</td>
<td>705</td>
<td>673</td>
<td>694</td>
</tr>
<tr>
<td>679</td>
<td>680</td>
<td>664</td>
<td>691</td>
<td>669</td>
</tr>
<tr>
<td>689</td>
<td>702</td>
<td>710</td>
<td>696</td>
<td>697</td>
</tr>
<tr>
<td>685</td>
<td>724</td>
<td>726</td>
<td>698</td>
<td>688</td>
</tr>
<tr>
<td>702</td>
<td>696</td>
<td>708</td>
<td>696</td>
<td>710</td>
</tr>
</tbody>
</table>

- Sample size = 50 measurements.
- Sample range = 740 - 663 = 77 hr.
- Class limits = highest and lowest measurements in the class.
- Class interval = upper limit - lower limit.
- Class boundaries = limits ± \(\frac{1}{2}\) unit in LSD
- Class width = upper - lower boundaries
- Class mark = mid-point of class.
- e.g. if the class limits are 670 → 678:
  * class interval = 678 - 670 = 8
  * class boundaries are:
    * lower bound = 670 - 0.5 = 669.5
    * upper bound = 678 + 0.5 = 678.5
  * class width = 678.5 - 669.5 = 9
  * class mark = \( \frac{670 + 678}{2} = 674 \)

  OR:
  * class mark = \( \frac{669.5 + 678.5}{2} = 674 \)

- e.g. if class limits are 5.87 → 6.32:
  * class interval = 6.32 - 5.87 = 0.45
  * class boundaries are:
    * lower bound = 5.87 - 0.005 = 5.865
    * upper bound = 6.32 + 0.005 = 6.325
  * class width = 6.325 - 5.865 = 0.46
  * class mark = \( \frac{5.87 + 6.32}{2} = 6.095 \)

  OR:
  * class mark = \( \frac{5.865 + 6.325}{2} = 6.095 \)
Determination of classes:

1. Determine the range:
2. Determine the total width:
   \[ \text{total width} = \text{range} + \text{one unit in LSD} \]
3. Divide the total width into a convenient no. of classes:
   \[ \text{class width} = \frac{\text{total width}}{\text{no. of Classes}} \]

Note: (Adjust the total width by adding one or two units in LSD if necessary, to select a suitable no. of classes, so that the class width is of a similar accuracy to the measurements.)

4. Determine class interval:
   \[ \text{class interval} = \text{class width} - \text{one unit} \]
5. Starting at lowest measurement, calculate the limits of successive classes.
Solution of example (electric bulb example)

1. range = $740 - 663 = 77$ hr

2. one unit in LSD = 1

   $\therefore$ total width $= 77 + 1 = 78$

3. Select no. of classes (for example take 5 classes) so the class width $= \frac{78}{5} = 15.6$
   
   which is not the same accuracy as the data.

   So take 6 classes:

   
   Class width $= \frac{78}{6} = 13$. (which is the same accuracy as the data).

4. Class interval $= 13 - 1 = 12$

Freq. dis. table

<table>
<thead>
<tr>
<th>Class limits</th>
<th>Class bound</th>
<th>Class mark</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 663-675</td>
<td>662.5-675.5</td>
<td>669</td>
<td>4</td>
</tr>
<tr>
<td>2. 676-688</td>
<td>675.5-688.5</td>
<td>682</td>
<td>10</td>
</tr>
<tr>
<td>3. 689-701</td>
<td>688.5-701.5</td>
<td>695</td>
<td>15</td>
</tr>
<tr>
<td>4. 702-714</td>
<td>701.5-714.5</td>
<td>708</td>
<td>11</td>
</tr>
<tr>
<td>5. 715-727</td>
<td>714.5-727.5</td>
<td>721</td>
<td>6</td>
</tr>
<tr>
<td>6. 728-740</td>
<td>727.5-740.5</td>
<td>734</td>
<td>4</td>
</tr>
</tbody>
</table>

$N = 50$
Types of Frequency:

1. Numeric Frequency: \( f \rightarrow \sum f_i = N \)
2. Relative Frequency: \( f_r = \frac{f}{N} \rightarrow \sum f_r = 1 \)
3. Percent Frequency: \( f_p = f_r \times 100 \rightarrow \sum f_p = 100 \)
4. Cumulative Frequency: The freq. is also expressed cumulatively of: \( f, f_r, f_p \).

- Cumulative freq. of class \( K \) is the sum of frequencies of all classes up to \( K \).

\[
\begin{align*}
 f_{C_K} &= \sum_{i=1}^{K} f_i, \\
 f_{CyK} &= \sum_{i=1}^{K} f_{ri} = 1 \\
 f_{CP_K} &= \sum_{i=1}^{K} f_{pi} = 100
\end{align*}
\]

<table>
<thead>
<tr>
<th>( f )</th>
<th>( f_r )</th>
<th>( f_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.08</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>2.0</td>
</tr>
<tr>
<td>15</td>
<td>0.3</td>
<td>3.0</td>
</tr>
<tr>
<td>11</td>
<td>0.22</td>
<td>2.2</td>
</tr>
<tr>
<td>6</td>
<td>0.12</td>
<td>1.2</td>
</tr>
<tr>
<td>4</td>
<td>0.08</td>
<td>8</td>
</tr>
</tbody>
</table>

\( \sum f = 50 \) \quad \sum f_r = 1 \quad \sum f_p = 100 \)
Graphical presentation of freq. dist.:  

Classified data may be presented as graphical plot with freq. as vertical axis versus measurement as horizontal axis.

1. **Histogram** is a bar chart, in which each class is represented by a rectangle, whose base extends between the class boundaries and the area proportional to frequency.

2. **Frequency Polygon** consists of lines joining the class mark with freq. It may be obtained from the histogram by joining the mid-points of the bar tops.

3. **Frequency Curve** is a smoothed frequency polygon into a continuous curve.

4. **Cumulative freq. curve (Ogive)** is a smoothed cumulative freq. polygon. It is usually S-shape.
Graphical presentation of frequency distribution.

1. Histogram:

2. Frequency Polygon:

3. Frequency Curve:
4. cumulative freq.

Ascending cum. freq.  

<table>
<thead>
<tr>
<th>Upper Boundaries</th>
<th>Cum. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 662.5</td>
<td>0</td>
</tr>
<tr>
<td>5 ≤ 675.5</td>
<td>4</td>
</tr>
<tr>
<td>6 ≤ 688.5</td>
<td>14</td>
</tr>
<tr>
<td>6 ≤ 701.5</td>
<td>29</td>
</tr>
<tr>
<td>6 ≤ 714.5</td>
<td>40</td>
</tr>
<tr>
<td>6 ≤ 727.5</td>
<td>46</td>
</tr>
<tr>
<td>6 ≤ 740.5</td>
<td>50</td>
</tr>
</tbody>
</table>

Descending cum. freq.  

<table>
<thead>
<tr>
<th>Upper Boundaries</th>
<th>Cum. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than 662.5</td>
<td>50</td>
</tr>
<tr>
<td>5 ≤ 675.5</td>
<td>46</td>
</tr>
<tr>
<td>5 ≤ 688.5</td>
<td>36</td>
</tr>
<tr>
<td>5 ≤ 701.5</td>
<td>21</td>
</tr>
<tr>
<td>5 ≤ 714.5</td>
<td>10</td>
</tr>
<tr>
<td>5 ≤ 727.5</td>
<td>4</td>
</tr>
<tr>
<td>5 ≤ 740.5</td>
<td>0</td>
</tr>
</tbody>
</table>

Cumulation = Freq. - Cum.  

(0give)
For the following data groups obtain:

1. Frequency distribution table.
2. FR, FP, FC.
3. Histogram, freq. polygon, and ogive.

Data 1:

6.3 7.0 7.5 9.0 7.7 7.8 7.1 8.1
6.6 7.2 8.3 8.5 6.9 7.7 8.0 7.3
8.6 7.1 8.7 6.4 7.7 7.4 8.0 7.6
7.5 7.2 7.5 8.8 7.8 7.9 7.3 7.0
6.8 8.1 8.4 6.7 7.1 8.2 8.1 7.7

Data 2:

5.4 4.1 5.2 2.8 4.9 5.6 4.0 4.1 4.3
3.9 4.5 6.1 3.7 2.3 4.5 4.9 5.6 4.3
4.2 3.2 5.0 4.8 3.7 4.6 5.5 1.8 5.1
5.1 6.3 3.3 5.8 4.4 4.8 3.0 4.3 4.7
| 12.16 | 12.38 | 12.21 | 12.55 | 12.22 | 12.40 | 12.43 | 12.35 |
| 12.31 | 12.07 | 12.31 | 12.33 | 12.56 | 12.41 | 12.42 | 12.44 |
| 12.35 | 12.20 | 12.39 | 12.54 | 12.59 | 12.29 | 12.16 | 12.09 |
Ch. 3

Measures of Location

When raw data is classified into a frequency distribution table and presented graphically, the major features of the sample become apparent. However, to make quantitative decisions, further condensation into a number of statistical parameters is needed. Measures of location are statistical parameters giving an estimate of the data centres, being typical of all measurement.

Mode is the measurement that occurs with the greatest frequency.
- For sample: 14, 19, 16, 21, 19, 21, 18, 19
  mode = 19
- For sample: 6, 7, 7, 3, 8, 3, 9, 5
  mode = 3, 7 (bimodal).

For grouped data, the mode corresponds to the top of the frequency curve.

\[ \text{mode} = L_m + \frac{\Delta L}{\Delta L + \Delta h} \times C_m \]

where \( L_m \) is the lower boundary of modal class.
\( \Delta L = f_m - f \) lower class,
\( \Delta h = f_m - f \) higher class,
\( C_m \) = width of modal class.

E.g. For electric bulbs sample:

mode = 688.5 + \frac{15-10}{(15-10) + (15-11)} \times (15) = 695.7
Median: is the middle measurement of an ordered array (odd), or the arithmetic mean of the two middle values (even).

\[ \text{For Sample: 3, 4, 4, 5, 6, 8, 8, 10, 11} \]
\[ \text{median} = 6 \]
\[ 5, 5, 7, 9, 11, 12, 15, 18 \]
\[ \text{median} = 10 \]

*For grouped data, the median line is the value that divides the area under the frequency curve.

\[ \text{median} = Lm + \frac{N}{2} \cdot \frac{f_{cl}}{Fm} - Cm \]

Where
- \( Lm \) is lower boundary of median class
- \( N \) is sample size
- \( f_{cl} \) is cumulative frequency of lower class
- \( Fm \) is frequency of median class
- \( Cm \) is width of median class

\[ \text{Eg. for electric bulbs sample:} \]

3rd class is median class, since \( f_c = 29 > \frac{N}{2} \)

\[ \text{median} = 688.5 + \frac{50 - 14}{15} (13) = 698.0 \]

Arithmetic Mean: is the sum of measurements divided by sample size

\[ \bar{x} = \frac{\sum f_i x_i}{\sum f_i} \]

For grouped data:

\[ \bar{x} = \frac{\sum f_i x_i}{\sum f_i} \]
e.g. For electric bulbs sample:
\[ \bar{x} = \frac{(4)(669) + (10)(682) + (15)(675) + (11)(708) + (6)(721) + (4)(734)}{50} \]
\[ \bar{x} = 699.4. \]

Relation between mode/median/mean

For symmetrical distributions, the three measures coincide. E.g., the mean is further removed from mode than is the median. For moderately skewed unimodal distributions:
\[ \text{mean} - \text{mode} \approx 3 \times (\text{mean} - \text{median}). \]

Other Mean Measures:

* Geometric Mean \( G = \left( \prod x_i \right)^{\frac{1}{N}} \)

* Harmonic Mean \( H = \frac{N}{\sum \frac{1}{x_i}} \)

* Root Mean Square \( \text{RMS} = \sqrt{\frac{\sum x_i^2}{N}} \)

\( \log G = \frac{\sum \log x_i}{N} \)
For a sample of positive measurements,

\[ H \leq G \leq \bar{x} \leq RMS \]

e.g. for electric bulbs sample:

\( \bar{x} = 699.4 \)
\( G = 699.2 \)
\( H = 699.0 \)
\( RMS = 699.6 \)

Properties of the Arithmetic Mean:

1. The sum of deviations of the data from their arithmetic mean is zero:

   \[ \sum (x_i - \bar{x}) = 0 \quad \text{(prove)} \]

2. For several samples, the combined mean is given by:

   \[ \bar{x} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2 + \ldots}{N_1 + N_2 + \ldots} \]

3. If the deviations (\( d_i \)) from any value (A) are available, then:

   \[ \bar{x} = A + \frac{\sum d_i}{N} \quad \text{where} \ d_i = x_i - A \quad \text{(prove)} \]

   or \[ \bar{x} = A + \frac{\sum f_i d_i}{N} \quad \text{(grouped data)} \]
Summation Notation:

1. \[ \sum_{i=1}^{6} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \]

2. \[ \sum_{i=1}^{4} (y_i - 3)^2 = (y_1 - 3)^2 + (y_2 - 3)^2 + (y_3 - 3)^2 + (y_4 - 3)^2 \]

3. \[ \sum_{i=1}^{n} a = a + a + \ldots + a = a \cdot n \]

4. \[ \sum_{k=1}^{5} p_k x_k = p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4 + p_5 x_5 \]

5. \[ \sum_{i=1}^{3} (x_i - a) = (x_1 - a) + (x_2 - a) + (x_3 - a) = x_1 + x_2 + x_3 - 3a \]

Example: Express each of the following using the summation notation:

1. \[ x_1^2 + x_2^2 + \ldots + x_{10}^2 = \sum_{i=1}^{10} x_i^2 \]

2. \[ (x_1 + x_1) + (x_2 + x_2) + \ldots + (x_8 + x_8) = \sum_{i=1}^{8} (x_i + x_i) \]

3. \[ p_1 x_1^3 + p_2 x_2^3 + \ldots + p_{20} x_{20}^3 = \sum_{i=1}^{20} p_i x_i^3 \]

4. \[ a_1 b_1 + a_2 b_2 + \ldots + a_\nu b_\nu = \sum_{i=1}^{\nu} a_i b_i \]

5. \[ p_1 x_1 y_1 + p_2 x_2 y_2 + p_3 x_3 y_3 = \sum_{i=1}^{3} p_i x_i y_i \]
Examples: Prove that:
\[ \frac{1}{n} \sum_{i=1}^{n} (aX_i + bY_i - cZ_i) = a \frac{1}{n} \sum_{i=1}^{n} X_i + b \frac{1}{n} \sum_{i=1}^{n} Y_i - c \frac{1}{n} \sum_{i=1}^{n} Z_i \]

2. If \( Z_1 = X_1 + Y_1 \), \( Z_2 = X_2 + Y_2 \), ..., \( Z_n = X_n + Y_n \)
Prove that \( Z = X + Y \)

3. If \( N \) nos. \( X_1, X_2, \ldots, X_n \) have deviations from any no. \( A \) given respectively by \( d_1 = X_1 - A \), \( d_2 = X_2 - A \), ..., \( d_n = X_n - A \), prove that:
\[ \bar{X} = A + \frac{1}{n} \sum_{i=1}^{n} d_i = A + \frac{1}{n} \sum_{i=1}^{n} d_i \]

b) In case \( X_1, X_2, \ldots, X_k \) have respective frequencies \( f_1, f_2, \ldots, f_k \) and \( d_1 = X_1 - A \), ..., \( d_k = X_k - A \), show that the result in c) is replaced by:
\[ \bar{X} = A + \frac{\sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i} = A + \frac{\sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i} \text{ where } \sum_{i=1}^{k} f_i = k' = n' \]
Solution of Data

1. range = 6.3 - 1.8 = 4.5
2. total width = 4.5 + 0.1 = 4.6
   total width = 4.6 + 2(0.1) = 4.8
3. class width = \( \frac{4.8}{6} = 0.8 \)
4. class interval = 0.8 - \( \frac{6}{1} = 0.7 \)

<table>
<thead>
<tr>
<th>Class limit</th>
<th>Class bound.</th>
<th>Class mark</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8 - 2.5</td>
<td>1.75 - 2.55</td>
<td>2.15</td>
<td>2</td>
</tr>
<tr>
<td>2.6 - 3.3</td>
<td>2.55 - 3.35</td>
<td>2.95</td>
<td>4</td>
</tr>
<tr>
<td>3.4 - 4.1</td>
<td>3.35 - 4.15</td>
<td>3.75</td>
<td>6</td>
</tr>
<tr>
<td>4.2 - 4.9</td>
<td>4.15 - 4.95</td>
<td>4.55</td>
<td>13</td>
</tr>
<tr>
<td>5.0 - 5.7</td>
<td>4.95 - 5.75</td>
<td>5.35</td>
<td>8</td>
</tr>
<tr>
<td>5.8 - [6.5]</td>
<td>5.75 - 6.55</td>
<td>6.15</td>
<td>3</td>
</tr>
</tbody>
</table>

\( \sum f = 36 \)
\[
\begin{array}{|c|c|c|c|c|}
\hline
X_i & P_i & x_i f_i & \frac{P_i}{x_i} & \frac{P_i}{x_i^2} & \log P_i \\
\hline
2.15 & 2 & 4.3 & 0.93 & 9.25 & 0.66 \\
2.95 & 4 & 11.8 & 1.36 & 34.81 & 1.88 \\
3.75 & 6 & 22.5 & 1.60 & 84.38 & 3.44 \\
4.55 & 13 & 59.15 & 2.86 & 267.17 & 8.55 \\
5.35 & 8 & 42.8 & 1.49 & 228.96 & 6.83 \\
6.15 & 3 & 18.45 & 0.49 & 113.47 & 2.37 \\
\hline
\end{array}
\]

\[\bar{X} = \frac{\sum x_i f_i}{\sum f_i} = \frac{159}{36} = 4.412\]

\[\bar{y} = \frac{\sum y_i f_i}{\sum f_i} = \frac{8.73}{36} = 0.242\]

\[\text{RMS} = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i}} = \sqrt{\frac{740.02}{36}} = 4.53\]

\[\log G = \frac{\sum f_i \log x_i}{\sum f_i} = \frac{22.73}{36} = 0.63\]

\[G = 4.28\]

\[\text{RMS} > \bar{X} > G > \bar{y}\]
mode = L_1 + \frac{D_1}{D_1 + D_2} \times C_m

= 4.15 + \frac{7}{7+5} \times 8 = 4.62

median = L_1 + \left( \frac{N/2 - F_<}{f_{med}} \right) \times C_m

= 4.15 + \frac{36/2 - 12}{13} \times 8 = 4.52

Empirical Relation between mean, median and mode

For unimodal freq. curves which are moderately skewed (asymmetrical), we have the empirical relation

Mean - Mode = 3(Mean - Median).

In Figs. below are shown the relative positions of the mean, median and mode for freq.-curves which are skewed to the right and left resp. For symmetrical curves the mean, mode, and median all coincide.
Symmetrical frequency curve
Measures of Dispersion

Dispersion is the degree of data spread about an average. Several measures are used including:
- range
- mean absolute deviation
- standard deviation
- variance
- coefficient of variation

Mean Absolute Deviation

is the arithmetic mean of the absolute deviations.

\[ \text{M.A.D.} = \frac{\sum |x_i - \bar{x}|}{N} \text{ for raw data.} \]

\[ = \frac{\sum f_i |x_i - \bar{x}|}{N} \text{ for grouped data} \]

Means other than \( \bar{x} \) may be used to obtain M.A.D from the respective mean.

Standard Deviation

is the root mean square of the deviations.

\[ s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \text{ for raw material} \]

\[ = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}} \text{ for grouped data} \]
Means other than \( \bar{x} \) may be used to obtain the standard deviation from the respective mean.

Standard deviation of a sample (\( S \)) is related to the standard deviation of the population (\( \sigma \)) by:

\[
S = \sqrt{\frac{N}{N-1}} \cdot \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}}
\]

Variance is the square of the standard deviation i.e. \( S^2 \) for sample, \( \sigma^2 \) for population.

Coefficient of Variation is a relative dispersion measure (dimensionless).

Relative Dispersion = \( \frac{\text{absolute dispersion}}{\text{average}} \)

Coefficient of Variation = \( \frac{S}{\bar{x}} \)
Properties of Standard Deviation

- Of all standard deviations, the min. is that from the arithmetic mean.
- For ideal normal distributions:
  - within $x \pm 5$ 68.27% of data
  - within $x \pm 25$ 95.45% of data
  - within $x \pm 35$ 99.73% of data
- For several samples, the combined $S$ is given by:
  $$S^2 = \frac{N_1S_1^2 + N_2S_2^2 + \cdots}{N_1 + N_2 + \cdots}$$

Standard Variable

The dimensional measurement's $x_i$ may be expressed as dimensionless standardized variables $Z_i$:

$$Z_i = \frac{x_i - \bar{x}}{S}$$

i.e. When $Z = 1$, the measurement is removed by one standard deviation from the mean.

Properties of Standard scores:
standardized variable, standard scores

① The variable \( z = \frac{X - \mu}{\sigma} \)

which measures the deviation from the mean in units of the standard deviation and it is a dimensionless quantity.

Properties of \( z \):

① The arithmetic mean for the standard scores equal to \( z = 0 \).

\[ z = \bar{z} = 0 \]

② The standard deviation for the standard scores equal to \( s = 1 \).

\[ s^2 = \sqrt{\frac{\sum (z - \bar{z})^2}{N}} = 1 \]

\[ s = \sqrt{\frac{\sum (z - \bar{z})^2}{N}} = 1 \]
Q. 1 / Prove the following:

A. \[ S = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2} \]

B. \[ S = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2} \]

where \( d_i = X_i - A \) where \( A \) is constant.

Q. 2 / For the following data:

<table>
<thead>
<tr>
<th>Class Limits</th>
<th>( f )</th>
<th>( ? )</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 - 62</td>
<td>5</td>
<td>( \sqrt{S} )</td>
</tr>
<tr>
<td>63 - 65</td>
<td>18</td>
<td>( \sqrt{S} )</td>
</tr>
<tr>
<td>66 - 68</td>
<td>42</td>
<td>( \sqrt{S} )</td>
</tr>
<tr>
<td>69 - 71</td>
<td>27</td>
<td>( \sqrt{S} )</td>
</tr>
<tr>
<td>72 - 74</td>
<td>8</td>
<td>( \sqrt{S} )</td>
</tr>
</tbody>
</table>

obtain:

A. \( S, \sqrt{S} \)

B. \( Z \)

C. \( Z, \sqrt{S} \)
Chapter 7

Curve Fitting and Method of Least Squares

* Relationship between variables: Very often in practice a relationship is found to exist between two (or more) variables. For example, circumferences of circles depend on their radii; and the pressure of a given mass of gas depends on its temperature and volume. It is frequently desirable to express this relationship in mathematical form by determining an equation connecting the variables.

Curve fitting procedure:

1. Plot set of data points \((x, y)\)
2. Suggest a form of relation defining \(y = f(x)\) from:
   a. Theoretical considerations
   b. Observation of the trend of data points
3. Evaluate constants in the suggested function, so that the deviations of data points from the function are minimized.
4. Calculate statistical measures of the degree of fit.

5. Other functions may be proposed, and procedure is repeated.

Method of least squares:

The simplest situation is a linear or straight-line relation between a single input and the response: \[ E(Y) = \alpha + \beta x \]

where \( \alpha \) and \( \beta \) are constant parameters that we want to estimate, (regression coefficients). For a sample of \( n \) pairs of data \((x_i, y_i)\) we calculate \( \alpha \), for \( \alpha \) and \( \beta \) to fit. If at \( x = x_i \), \( \hat{y}_i \) is the estimated value of \( E(Y) \), we have the fitted regression line:

\[ \hat{y}_i = \alpha + \beta x_i \]

Let \( e_i = y_i - \hat{y}_i \) be the deviation in the \( y \)-direction of any data point from the fitted regression line. Then the estimates \( \alpha \) and \( \beta \) are chosen so that the sum of the squares of deviations of all the \( e_i \) is minimized.
is smaller than for any other choice of \(a\) and \(b\). So that:

\[
\sum e_i^2 = \sum (y_i - \hat{y}_i)^2
\]

has a min. value. This is called the method of Least Squares and the resulting eqn. is called the regression line of \(y\) on \(x\), where \(y\) is the response (dependent) and \(x\) is the input (independent variable).

If the estimated eqn. \(\hat{y} = a+bx\) then \(e_i = y_i - (a+bx)\) These deviation called residuals

\[
e_i^2 = [y_i - (a+bx)]^2
\]

\[
\sum e_i^2 = \sum [y_i - (a+bx)]^2
\]

This sum of the squares of the deviations or error or residuals for all \(n\) pts is abbreviated as SSE. So the principle of L.S.M. is to minimize SSE:

\[
\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (a+bx)]^2
\]
To minimize a quantity we take the derivative with respect to the independent variable and set it equal to zero.

\[
\frac{\partial}{\partial a} (SS\bar{E}) = \frac{\partial}{\partial a} \sum (y_i - (a + bxi))^2
\]

\[
= -2 \sum y_i - na - b \sum xi = 0
\]

\[
\frac{\partial}{\partial b} (SS\bar{E}) = \frac{\partial}{\partial b} \sum (y_i - (a + bxi))^2
\]

\[
= -2 \sum xi (y_i - na - b \sum xi) = 0
\]

Eqs. 1 and 2 are called the least squares eqns. (or normal equations).

Eqs. 1 and 2 can be solved simultaneously.

The results are:
\[ b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \]

\[ a = \frac{\sum y_i - b \sum x_i}{n} = \overline{y} - b \overline{x} \]

Then we have:

The sum of squares for \( x \):

\[ S_{xx} = \sum x_i^2 - \frac{1}{n} \left[ \sum x_i \right]^2 \]

\[ S_{xx} = \sum x_i^2 - \frac{1}{n} \left[ \sum x_i \right]^2 \]

\[ S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{1}{n} \left[ \sum y_i \right]^2 \]

\[ S_{x,y} = \sum (x_i - \overline{x})(y_i - \overline{y}) \]

\[ = \sum x_i y_i - \frac{1}{n} \left[ \sum x_i \right] \left[ \sum y_i \right] \]
Eqn. 3 and 4 can be written compactly as:
\[ b = \frac{\sum xy}{\sum xx} \quad \text{and} \quad a = \bar{y} - b \bar{x} \]

If we substitute the \( y = a + bx_i \), we get
\[ (\hat{y}_i, -\bar{y}) = b (x_i - \bar{x}) \]

This indicates that the best-fit line passes through the pt. \((\bar{x}, \bar{y})\), which is called the Centroidal Pt. and is the Centre of Mass of the data pts.
Example 0: Data for Simple Linear Regression

\[ X: 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \]
\[ Y: 3.85 \quad 0.03 \quad 3.50 \quad 6.13 \quad 4.07 \quad 7.07 \quad 8.66 \quad 11.65 \]
\[ X: 8 \quad 9 \quad 10 \quad 11 \quad 12 \]
\[ Y: 15.23 \quad 12.29 \quad 14.74 \quad 16.02 \quad 16.86 \]

So

\[ N = 13, \quad \Sigma X_i = 78, \quad \Sigma Y_i = 120.1 \]
\[ \Sigma X_i^2 = 650, \quad \Sigma Y_i^2 = 1483.0828 \]
\[ \Sigma X_i Y_i = 968.95 \]

The Centroidal Pt. \((\bar{X}, \bar{Y}) = (6, 9.23846)\)

\[ S_{XX} = \Sigma X_i^2 - \frac{1}{N} \left( \Sigma X_i \right)^2 = 650 - \frac{1}{13} (78)^2 \]
\[ S_{XX} = 182 \]
\[ S_{YY} = \Sigma Y_i^2 - \frac{1}{N} \left( \Sigma Y_i \right)^2 = 1483.08 - \frac{1}{13} (120.1)^2 \]
\[ S_{YY} = 373.5436 \]
\[ S_{XY} = \Sigma X_i Y_i - \frac{1}{N} \left( \Sigma X_i \right) \left( \Sigma Y_i \right) = 968.95 - \frac{1}{13} (78) (120.1) \]
\[ S_{XY} = 248.35 \]
\[ b = \frac{S_{xy}}{S_{xx}} = \frac{248.35}{182} = 1.36456 \]
\[ a = \bar{y} - b \bar{x} = 9.23846 - (1.36456)(6) \]
\[ a = 1.0511 \]

The best-fit regression equation is
\[ y = 1.0511 + 1.36456x \]

**Variance of Experimental Points Around the Line:**
This must be found from the residuals,
\[ e_i = y_i - \hat{y} = y_i - (a + bx_i) = y_i - a - bx_i \]
\[ SSE = \sum (y_i - a - bx_i)^2 \]
since \( \bar{y} = a + bx \) \( \rightarrow a = \bar{y} - b\bar{x} \)
\[ SSE = \sum [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \]
\[ = \sum (y_i - \bar{y})^2 - 2b \sum (x_i - \bar{x})(y_i - \bar{y}) + b^2 \sum (x_i - \bar{x})^2 \]
\[ SSE = S_{yy} - 2bS_{xy} + b^2 S_{xx} \]