CSC 1311: STATISTICS FOR PHYSICAL SCIENCE AND ENGINEERING

1) Measures of location, partition and dispersion by H. A. Kakudi
2) Elements of Probability
3) Probability distribution: binomial Poisson, geometric, hypergeometric, negative-binomial, normal Poisson
4) Estimation (Point and interval) and tests of hypotheses concerning population means, proportions and variances
5) Regression and correlation
6) Non-parametric tests
7) Contingency table analysis
8) Introduction to design of experiments
9) Analysis of variance

1) Measures of location, partition and dispersion by H. A. Kakudi

After studying this lesson, you will be able to : ·

- explain the meaning of dispersion through examples;
- define various measures of dispersion - range, mean deviation, variance and standard deviation;
- calculate mean deviation from the mean of raw and grouped data;
- calculate variance and standard deviation for raw and grouped data; and illustrate the properties of variance and standard deviation.

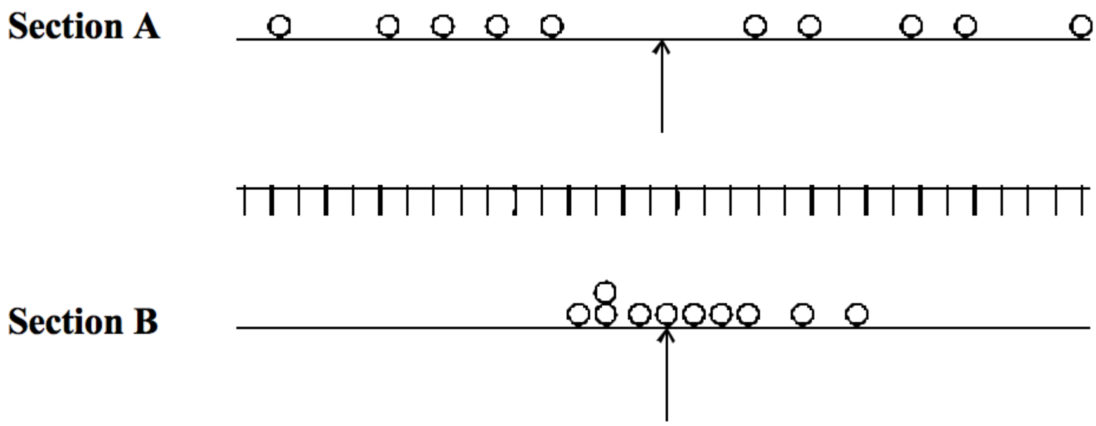
- Dispersion (a.k.a., variability, scatter, or spread)) characterizes how stretched or squeezed of the data.
- A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.
- Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.
- There are many types of dispersion measures:
  - Range
  - Mean Absolute Deviation
  - Variance/Standard Deviation

To explain the meaning of dispersion, let us consider an example.

Two sections of 10 students each in class X in a certain school were given a common test in Mathematics (40 maximum marks). The scores of the students are given below :

Section A :	6	9	11	13	15	21	23	28	29	35
Section B:	15	16	16	17	18	19	20	21	23	25

The average score in section A is 19.
The average score in section B is 19.
Let us construct a dot diagram, on the same scale for section A and section B
The position of mean is marked by an arrow in the dot diagram.



Clearly, the extent of spread or dispersion of the data is different in section A from that of B.
The measurement of the scatter of the given data about the average is said to be a measure of dispersion or scatter.

Range

- $$\text{Range} = \max - \min$$

In the above cited example, we observe that

(i) the scores of all the students in section A are ranging from 6 to 35;
ii) the scores of the students in section B are ranging from 15 to 25.
iii) The difference between the largest and the smallest scores in section A is 29 (35-6).
iv) The difference between the largest and smallest scores in section B is 10 (25-15).
v) Thus, the difference between the largest and the smallest value of a data, is termed as the range of the distribution.

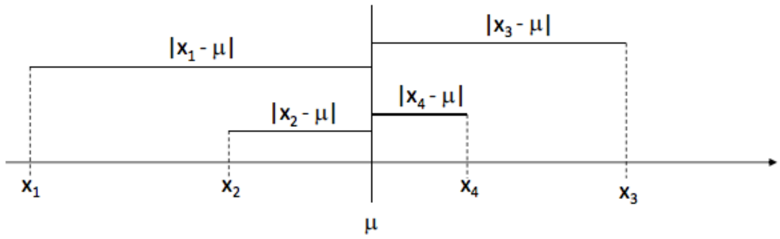
Properties of Range

- Only two values are used in its calculation.
- It is influenced by an extreme value (non-robust).
- It is easy to compute and understand.

Mean Absolute Deviation of Raw Data

- The Mean Absolute Deviation of a set of n numbers

$$\text{MAD} = \frac{|x_1 - \mu| + \ldots + |x_n - \mu|}{n}$$



- **Example:** A sample of four executives received the following bonuses last year ($000): 14.0 15.0 17.0 16.0
- **Problem:** Determine the MAD.
- **Solution:**

$$\begin{aligned}\bar{x} &= \frac{14 + 15 + 17 + 16}{4} = \frac{62}{4} = 15.5.\\ \text{MAD} &= \frac{|14 - 15.5| + |15 - 15.5| + |17 - 15.5| + |16 - 15.5|}{4}\\ &= \frac{4}{4} = 1.\end{aligned}$$

Properties of MAD

- All values are used in the calculation.
- It is not unduly influenced by large or small values (robust)
- The absolute values are difficult to manipulate.

Variance

- The variance of a set of n numbers as population:

$$\begin{aligned}\text{Var} := \sigma^2 &= \frac{(x_1-\mu)^2+\ldots+(x_n-\mu)^2}{n} \quad \text{(conceptual formula)}\\ &= \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n} \quad \text{(computational formula).}\end{aligned}$$

- The variance of a set of n numbers as sample:

$$\begin{aligned}S^2 &= \frac{(x_1-\mu)^2+\ldots+(x_n-\mu)^2}{n-1} \quad \text{(conceptual formula)}\\ &= \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} \quad \text{(computational formula).}\end{aligned}$$

Standard Deviation of Raw Data

- The Standard Deviation is the square root of variance
- Other notations: σ for population and S for sample.

- **Example**: The hourly wages earned by three students are: \$10, \$11, \$13.
- **Problem:** Find the mean, variance, and Standard Deviation.
- **Solution**: Mean and variance

$$
\begin{aligned}
\mu &= \frac{10+11+13}{3} = \frac{34}{3} \approx 11.33333333333 \\
\sigma^2 &\approx \frac{(10-11.33)^2+(11-11.33)^2+(13-11.33)^2}{3} \\
&= \frac{1.769+0.1089+2.7889}{3} = \frac{4.6668}{3} = 1.5556.
\end{aligned}
$$

- Standard Deviation

$$\sigma \approx 1.247237.$$

- **Example**: The hourly wages earned by three students are: \$10, \$11, \$13.
- **Problem:** Find the variance, and Standard Deviation.
- **Solution**:
  - Variance

$$
\begin{aligned}
\sigma^2 &= \frac{(10^2+11^2+13^2)-\frac{(10+11+13)^2}{3}}{3} \\
&= \frac{390-\frac{1156}{3}}{3} = \frac{390-385.33}{3} = \frac{4.67}{3} = 1.555667.
\end{aligned}
$$

  - Standard Deviation

$$\sigma \approx 1.247665.$$

- If the above is sample, then $\sigma^2 \approx 2.33335$ and $\sigma \approx 1.527531$.

- Conceptual formula may have accumulated rounding error.
- Computational formula only has rounding error towards the end!

Properties of Variance/Standard deviation

- All values are used in the calculation.
- It is not extremely influenced by outliers (non-robust).
- The units of variance are awkward: the square of the original units. Therefore standard deviation is more natural since it recovers he original units.

Range of Grouped Data

- The range of a sample of data organized in a frequency distribution is computed by the following formula:

  Range = upper limit of the last class - lower limit of the first class

Mean Absolute Deviation for Grouped Data

$$\text{Mean deviation from mean of grouped data} = \frac{\sum_{i=1}^{n}[f_i|x_i - \overline{x}|]}{N}$$

where $N = \sum_{i=1}^{n} f_i$, $\overline{x} = \frac{1}{N}\sum_{i=1}^{n}(f_i x_i)$

Example 29.1 Find the mean deviation from the mean of the following data :

Size of items x_i	4	6	8	10	12	14	16
Frequency f_i	2	5	5	3	2	1	4

Mean is 10

Solution :

x_i	f_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
4	2	−5.7	5.7	11.4
6	4	−3.7	3.7	14.8
8	5	−1.7	1.7	8.5
10	3	0.3	0.3	0.9
12	2	2.3	2.3	4.6
14	1	4.3	4.3	4.3
16	4	6.3	6.3	25.2
	21			69.7

Mean deviation from mean $= \dfrac{\sum\left[f_i \lvert x_i - \bar{x}\rvert \right]}{21}$

$$= \frac{69.7}{21} = 3.319$$

Example 29.2 Calculate the mean deviation from mean of the following distribution :

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of Students	5	8	15	16	6

Mean is 27 marks

Solution :

Marks	Class Marks x_i	f_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
0 – 10	5	5	−22	22	110
10 – 20	15	8	−12	12	96
20 – 30	25	15	−2	2	30
30 – 40	35	16	8	8	128
40 – 50	45	6	18	18	108
Total		50			472

$$\text{Mean deviation from Mean} = \frac{\sum [f_i |x_i - \bar{x}|]}{N}$$

$$= \frac{472}{50} \text{ Marks} = 9.44 \text{ Marks}$$

Variance/Standard Deviation for Grouped Data-Method I

We are given k classes and their corresponding frequencies. We will denote the variance and the standard deviation of grouped data by σ_g^2 and σ_g respectively. The formulae are given below :

$$\sigma_g^2 = \frac{\sum_{i=1}^{K} [f_i (x_i - \bar{x})^2]}{N}, \qquad N = \sum_{i=1}^{K} f_i$$

and $\sigma_g = +\sqrt{\sigma_g^2}$

Example 29.6 In a study to test the effectiveness of a new variety of wheat, an experiment was performed with 50 experimental fields and the following results were obtained :

Yield per Hectare (in quintals)	Number of Fields
31 – 35	2
36 – 40	3
41 – 45	8
46 – 50	12
51 – 55	16
56 – 60	5
61 – 65	2
66 – 70	2

The mean yield per hectare is 50 quintals. Determine the variance and the standard deviation of the above distribution.

Solution :

Yield per Hectare (in quintal)	No. of Fields	Class Marks	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
31 – 35	2	33	– 17	289	578
36 – 40	3	38	– 12	144	432
41 – 45	8	43	– 7	49	392
46 – 50	12	48	– 2	4	48
51 – 55	16	53	+3	9	144
56 – 60	5	58	+8	64	320
61 – 65	2	63	+13	169	338
66 – 70	2	68	+18	324	648
Total	50				2900

Thus $\sigma_g^2 = \dfrac{\sum_{i=1}^{n}\left[f_i(x_i - \bar{x})^2\right]}{N} = \dfrac{2900}{50} = 58$ and $\sigma_g = +\sqrt{58} = 7.61$ (approx)

Variance/Standard Deviation for Grouped Data-Method II

If $\bar{x}$ is not given or if $\bar{x}$ is in decimals in which case the calculations become rather tedious, we employ the alternative formula for the calculation of SD as given below:

- The variance of a sample of data organized in a frequency distribution is computed by the following formula:

$$S^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \dfrac{\left(\sum_{i1}^{k} f_i x_i\right)^2}{n}}{n-1}$$

- where f_i is the class frequency and x_i is the class midpoint for Class $i = 1, \ldots, k$.

- **Example:** Consider the guessed weights (lbm) collected in our first class on Sept. 5, 2013 from 62 students (the e-version of this data will be available online on my website).

140 135 140 160 175 150 152 155 155 165 145 150 154 160 143
160 170 155 140 160 160 175 140 145 150 150 152 159 160 165
145 155 150 150 165 148 152 155 155 160 172 180 141 147 155
165 170 160 140 150 150 152 155 130 155 163 170 139 165 180
180 190

class	freq. (f_i)	mid point (x_i)	$f_i x_i$	$f_i x_i^2$
$[130, 140)$	3	135	405	54675
$[140, 150)$	12	145	1740	252300
$[150, 160)$	23	155	3565	552575
$[160, 170)$	14	165	2310	381150
$[170, 180)$	6	175	1050	183750
$[180, 190]$	4	185	740	136900
	62		9,810	1,561,350

- **Solution:** The Variance/Standard Deviation are:

$$S^2 = \frac{1,561,350 - \frac{9,810^2}{62}}{62-1} \approx 150.0793.$$
$$S \approx 12.25069$$

- The real sample variance/SD for the raw data is 146.3228/12.0964.

Range for grouped data

- The range of a sample of data organized in a frequency distribution is computed by the following formula:

Range = upper limit of the last class - lower limit of the first class