Biostatistics

Lecture 4

Measures of Dispersion (Variation)

The mean, mode and median do a nice job in telling where the center of the data set is, but often we are interested in more.

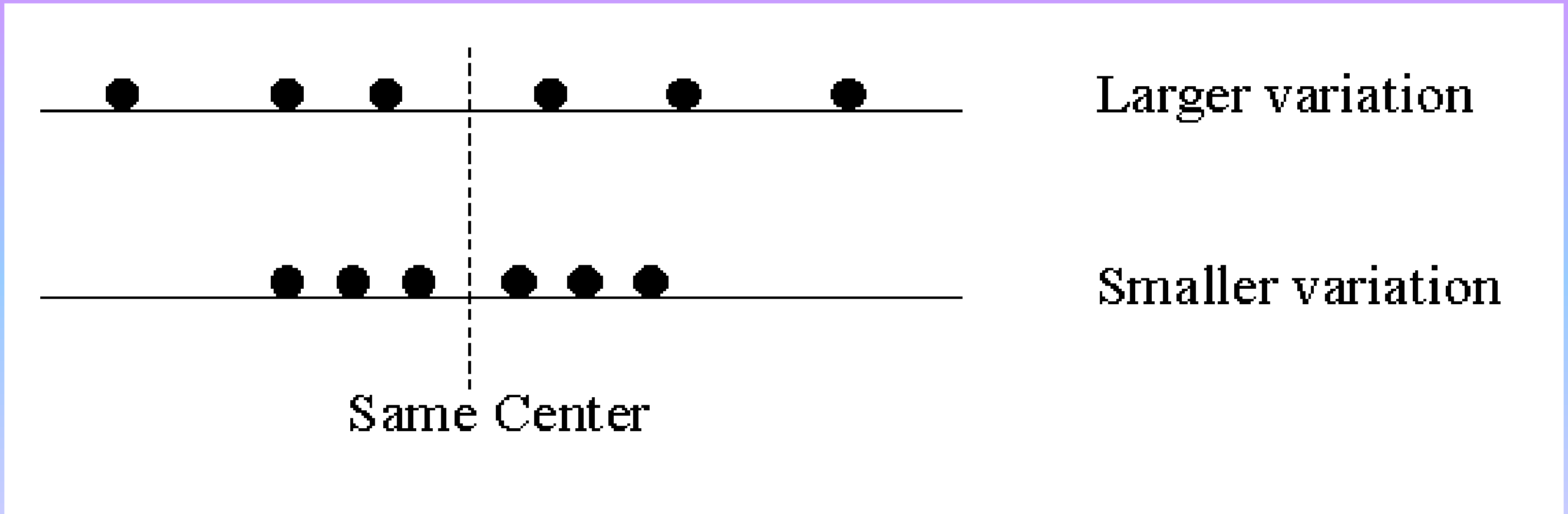
For example: Pharmaceutical engineer develops a new drug that regulator iron in the blood. Suppose he finds out that the average sugar content after taking the medication is optimal level.

This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half has dangerously high content. Instead of the drug being an effective regulator, it is a deadly poison.

What the pharmacist needs, is a measure of how far the data is spread apart. This is what the variance and standard deviation do.

*** The variation or dispersion in a set of values refers to how spread out the values are from each other.**

- **The variation is small when the values are close together.**
- **There is no variation if the values are the same.**



Some measures of dispersion:

Range, Variance, Standard deviation, Coefficient of variation

Range:

The range is the difference between the largest (XL) and the smallest (XS) values in a set of observations.

$$R = XL - XS$$

Example:

Find the range for the sample values:

26, 25, 35, 27, 29, 29.

Solution:

Range = 35 – 25 = 10 (unit)

Note: The range is poor measure of dispersion? because it only takes into account two of the values.

Variance:

The variance is the most commonly used to measure of spread in biological statistics.

For a population is defined as the sum of squares of the deviation from the mean (SS), dividing by the total number of the deviations, and by one less than the total number of the deviation (df) for a sample.

Variance for a population

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X-\mu)^2}{N}$$

OR

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} Xi^2 - \frac{\left(\sum_{i=1}^{N} Xi\right)^2}{N}\right]$$

Variance for a sample

$$S^2 = \frac{\sum_{i=1}^{n} (Xi - \bar{X})^2}{n-1}$$

OR

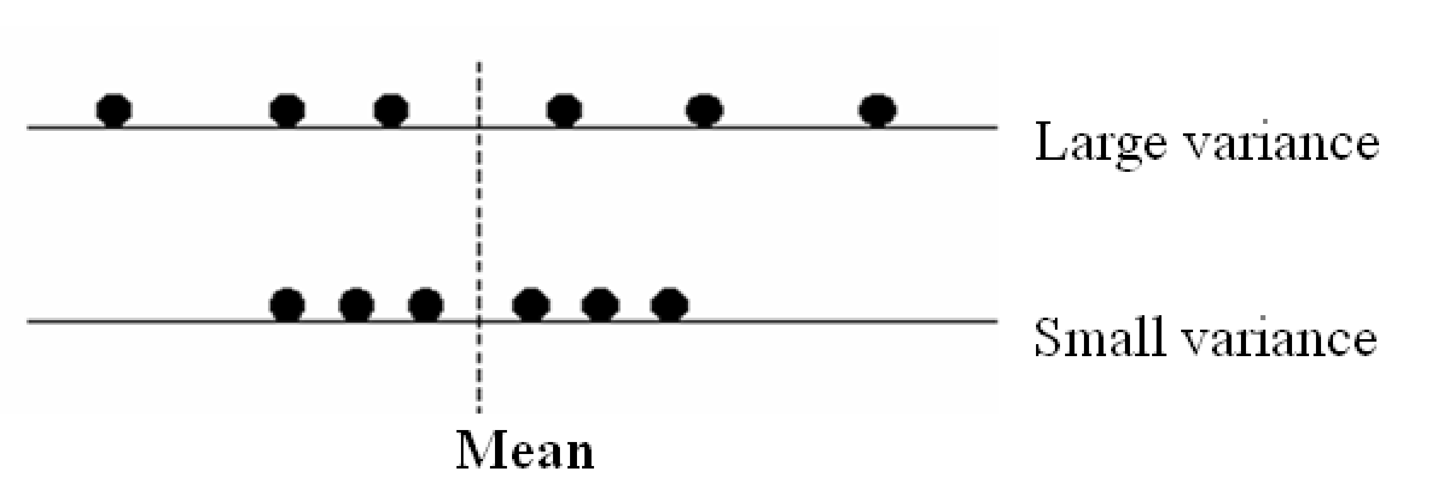
$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Xi^2 - \frac{\left(\sum_{i=1}^{n} Xi\right)^2}{n}\right]$$

Notes:

1) The variance is a measure that uses the mean as a point of reference.

2) The variance is small when all values are close to the mean.

3) The variance is large when all values are spread away from the mean.



Why the separate formula for sample?

The formula for sample divided by n-1 to:

1. Correct for probability that most extreme cases will be excluded from a smaller sample.

2. Makes the sample more representative of the population for every small sample.

3. Reduces the denominator to a larger extent (If n=5 they we have a 20% reduction in the denominator). But in large samples the n-1 correction does not have as large effect.

<u>Example:</u> **We want to compute the sample variance of the following sample values:**

10, 21, 33, 53, 54

* First method

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{10+21+33+53+54}{5} = \frac{171}{5} = 34.2 \quad (\text{unit})$$

$$\therefore S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} = \frac{\sum_{i=1}^{5} \left(x_i - 34.2\right)^2}{5-1}$$

$$S^2 = \frac{(10-34.2)^2 + (21-34.2)^2 + (33-34.2)^2 + (53-34.2)^2 + (54-34.2)^2}{4}$$

$$= \frac{1506.8}{4} = 376.7 \quad (\text{unit})^2$$

* Second method:

x_i	$(x_i - \bar{x}) = (x_i - 34.2)$	$(x_i - \bar{x})^2 = (x_i - 34.2)^2$	$\bar{x} = \dfrac{\sum_{i=1}^{5} x_i}{5}$
10 21 33 53 54	-24.2 -13.2 -1.2 18.8 19.8	585.64 174.24 1.44 353.44 392.04	$= \dfrac{171}{5} = 34.2$ $S^2 = \dfrac{1506.8}{4}$
$\sum_{i=1}^{5} x_i = 171$	$\sum_{i=1}^{5} (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 1506.8$	$= 376.7$

* Third method:

x_i	10	21	33	53	54	$\sum x_i = 171$
x_i^2	100	441	1089	2809	2916	$\sum x_i^2 = 7355$

$$S^2 = \frac{7355 - (5)(34.2)^2}{5-1} = \frac{1506.8}{4} = 376.7$$

Standard Deviation (s) or (sd)

Is defined as a positive square root of variance

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

Coefficient of Variation (C.V.):

The S^2 and the sd are useful as measures of variation of the values of a single variable for a single population or sample.

If we want to compare the variation of two variables we can not use the S^2 or sd because:

1. The variables might have different units.
2. The variables might have same means.

Coefficient of Variation is defined as the ratio of the standard deviation to the mean. ***It is independent of the units employed (unit less).***

$$C.V = \frac{\sigma}{\mu}100\% \quad \text{................ (Population)}$$

$$C.V = \frac{s}{\overline{X}}100\% \quad \text{.................. (Sample)}$$

Note:

In biological experiments if coefficient of variation (C.V):

1) Is 5% or less that is refer to homogenous data with less variance.
2) Is 10% to 15% the variance between a data is acceptable.
3) Is 25% or more indicating very considerable variance.

1. Ungrouped Data

Example: The data below present the technician from two different laboratories, all making the same specific blood chemistry determination.

Laboratories	Technician	Mean	Standard deviation
1	5, 7, 6, 6	6	0.82
2	6, 4, 9, 5	6	2.16

$$C.V = \frac{0.82}{6} \times 100 = 13.6\%$$

$$C.V = \frac{2.16}{6} \times 100 = 36\%$$

Lab. 1 gives most accurate result.

Example: A set of data (4, 6, 3, 4, 5 and 2) compute: R, S^2, sd and C.V?

$R = X_L - X_S$ …………………………… $R = 6-2=4$

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Xi^2 - \frac{\left(\sum_{i=1}^{n} Xi\right)^2}{n}\right]$$

$$S^2 = \frac{1}{6-1}\left[(4^2+6^2+3^2+4^2+5^2+2^2) - \frac{(4+6+3+4+5+2)^2}{6}\right] = 2$$

$$s=\sqrt{s^2} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\; s=\sqrt{2} = 1.414$$

$$\bar{X}=\frac{\sum_{i=1}^{n} Xi}{N} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\; \bar{X}=\frac{4+6+3+4+5+2}{6}=4$$

$$C.V=\frac{s}{\bar{X}}\times 100 \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\; C.V=\frac{1.414}{4}\times 100=35.35\%$$

2. Grouped Data(for view only)

Example: **The following shows the hemoglobin values (g/100ml) of 30 children receiving treatment for hemolytic anemia, compute: S^2, sd and C.V?**

Hemoglobin	Midpoint (mi)	Frequency (fi)
6.5 – 7.5	7	1
7.5 – 8.5	8	5
8.5 – 9.5	9	11
9.5 – 10.5	10	9
10.5 – 11.5	11	3
11.5 – 12.5	12	1
		$\sum fi = 30$

Variance for grouped data .... $$S^2 = \frac{1}{\sum_{i=1}^{k} fi - 1}\left[\sum_{i=1}^{k} mi^2 fi - \frac{\left(\sum_{i=1}^{k} mifi\right)^2}{\sum_{i=1}^{k} fi}\right]$$

$$S^2 = \frac{1}{30-1}\left[(7^2\times 1 + 8^2\times 5 + \ldots + 12^2\times 1) - \frac{(7\times 1 + 8\times 5 + \ldots + 12\times 1)^2}{30}\right] = 1.199$$

$s = \sqrt{s^2}$ .... $s = \sqrt{1.199} = 1.095$

$\bar{X} = \frac{\sum_{i=1}^{k} mifi}{\sum_{i=1}^{k} fi}$ .... $\frac{(7\times 1 + 8\times 5 + 9\times 11 + 10\times 9 + 11\times 3 + 12\times 1)}{30} = 9.367$

$$C.V = \frac{1.095}{9.367}\times 100 = 11.69\%$$