Chapter 4

Measures of Variation/Dispersion

Alemakef Wagnew(Bsc and MPH in EPi-BIO)

Institute of Public Health
Departement of Epidemiology and Biostatistics

April 3, 2019

Schedule

Introduction

Introduction

Example

Consider the following two sets of scores:Set 1: 40, 50, 60, 60, 40, 50Set 2: 0, 100, 25, 75, 80, 20

Alert block

- Both these sets have the same mean (50),
- but the second set is a lot more widely dispersed ("scattered") than the first....

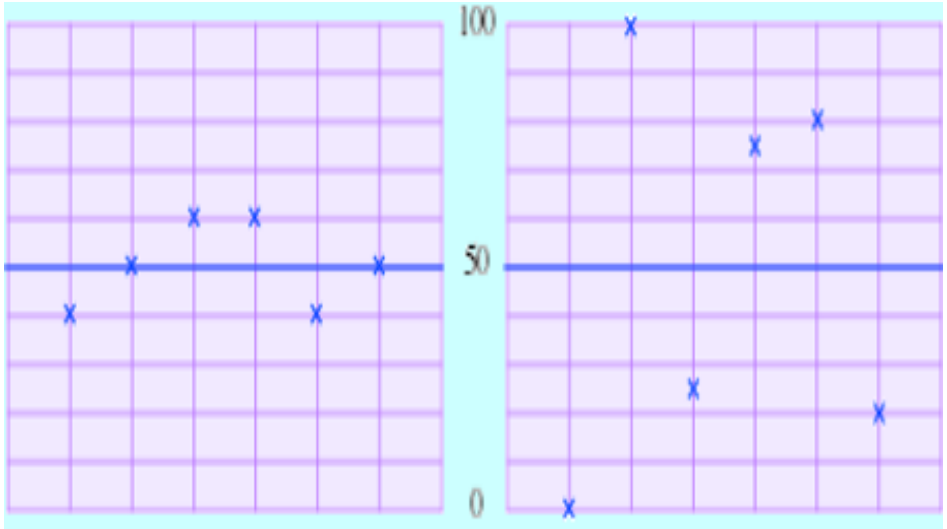


Figure: Example images from the data set

Measure of variation/dispersion

- The scatter or spread of items of a distribution is known as dispersion or variation.
- In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data.

- Measures of dispersion are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.

Objective of Measuring Variation

Objective of measuring variation

- To determine the reliability of an average by pointing out as how far an average is representative of the entire data.
- To determine the nature and cause of variation in order to control the variation itself.
- Enable comparison of two or more distribution with regard to their variability.
- Measuring variability is of great importance to other statistical analysis. E.g., it is the basis of statistical quality control

A good measure of variation

- It should be easy to compute and understand.
- It should be based on all observations.
- It should be Uniquely defined
- It should be capable of further algebraic treatment.
- It should be as little as affected by extreme values

Types of measure of variation

Types of measure of variation

Absolute measure:

- Range
- Quartile deviation
- Mean deviation
- Variance
- Standard deviation

Relative measures:

- Relative range
- Coefficient of quartile deviation
- Coefficient of mean deviation
- Coefficient of variation
- Standard scores

The range

Several measures of dispersion are available. We will discuss the common ones below.
The Range:

- The difference between the largest (maximum) and smallest (minimum) values.

$$Range = Maximum - Minimum \tag{1}$$

- **For frequency distributed data, the range is:**
- The difference between the upper class boundary of the last class and the lower class boundary of the first class.

Measure of dispersion

Measure of variation-dispersionn

- Find the Range of 54.5, 55.0, 55.7, 51.8, 54.2, 52.4 Solution:
- range(R) = 55.7- 51.8 = 3.9cm

Given the following frequency distribution. Find the range

Class	frequency
52.5-63.5	6
63.5-74.5	12
74.5-85.5	25
85.5-96.5	18
96.5-107.5	14
107.5-118.5	5

Solution: Range =UCB$_l$ $-$ LCB_f $=$ 118.5 $-$ 52.5 $=$ 66

Measure of variation

The relative range:

$$RR = relative\ range = \frac{l - s}{l + s} \tag{2}$$

we adjust similarly for Grouped data
Quartile deviation (QD):

The range expresses the extreme variability of observations of a variable. is half of the inter quartile range.

$$InterquartileRange = Q_3 - Q_1$$

$$QD = \frac{Interquartilerange}{2}$$

$$QD = \frac{Q3 - Q1}{2}$$

Measure of variation

- **Coefficient of quartile deviation (CQD):**
- It gives the average amount by which the two quartiles differ from the median

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{3}$$

- **Mean Deviation(M.D):**
  - The average deviation measures the scatter of the individual observations around a central value usually the mean or the median of a distribution.
  - The mean deviation is defined as the arithmetic mean of positive deviations of each observation from either the mean or the median of a distribution.
  - If the deviations are taken from the mean then it is called mean deviation about the mean. On the other hand, if the deviations are taken from the median we call it mean deviation about the median.

mean deviation about the mean

- The mean Deviation (M.D) is the arithmetic mean of the absolute deviations of the values from the mean.
- It is the "average absolute deviation of the values from the mean".

$$Mean\ deviation = \frac{\sum_{i=1}^{n} \mid X_i - \bar{x} \mid}{n} \tag{4}$$

- **Note that:**while dealing with population values, it is adjusted accordingly
- **Mean Deviations for Grouped data (discrete or continuous)**

$$Mean\ deviation = \frac{\sum_{i=1}^{n} f_i \mid X_i - \bar{x} \mid}{n} \tag{5}$$

- Where m = number of classes and x_i = *class mark of the* i^{th} *class*, n = *number of observation*

mean deviation

Mean deviation about the median (MD())

ungrouped data:

$$MD(\hat{X}) = \frac{\sum_{i=1}^{n} \mid X_i - \hat{x} \mid}{n} \tag{6}$$

grouped Frequency Distribution:

$$MD(\hat{X}) = \frac{\sum_{i=1}^{k} f_i \mid X_i - \hat{x} \mid}{n} \tag{7}$$

Example

- The weights of a sample of six students from a class (in kilograms) is measured as: 53, 56, 57, 59, 63 and 66. Find the mean deviation about the mean and the mean deviation from the median.
- **solution:**First find the mean and the median. The mean is 59 kg and the median is 58 kg. Then take the deviations of each observation from these averages as shown below

weight X_i	Ad from mean$\mid x_i - \bar{x} \mid$	AD from median$\mid x_i - \hat{x} \mid$
53	6	5
56	3	2
57	2	1
59	0	1
63	4	5
66	2	8
Total	22	22

example continued

mean deviation a bout the mean:

$$MD(\bar{X}) = \frac{\sum_{i=1}^{n} f_i \mid X_i - \bar{x} \mid}{n}$$

$$= MD(\bar{x}) = \frac{22}{6}$$

$$= 3.67$$

mean deviation about median:

$$MD(\hat{X}) = \frac{\sum_{i=1}^{k} f_i \mid X_i - \hat{x} \mid}{n}$$

$$MD(\hat{x}) = \frac{22}{6} = 3.67$$

Example 4.4: Calculate the mean deviation from the mean and median for the following data.

Class interval	1-5	6-10	11-15	16-20
Frequency	4	1	2	3

solution:

$$Mean = \frac{3 * 4 + 8 * 1 + 13 * 2 + 18 * 3}{10} \tag{8}$$

CL	X_i	f_i	$f_i \mid x_i - 10 \mid$
1-5	3	4	28
6-10	8	1	2
11-15	13	2	6
16-20	18	3	24
Total		10	60

Therefore MD around the mean :

$$MD = \frac{60}{10}$$
$$= 6$$

Exercise: Mean Deviation about the Median

Coefficients of Mean Deviation(C.M.D)

- C.M.D = M.D/Average about which deviations are taken
- Coefficient of mean deviation about the

$$CMD(\bar{X}) = \frac{MD\bar{X}}{\bar{X}}$$

- Coefficient of mean deviation about the median=

$$CMD(\hat{X}) = \frac{MD\hat{X}}{\hat{X}}$$

Example

Example 4.5

- Find the coefficient of mean deviation about the mean and mean deviation about the median for the weights of six students in example above.
- **Solution:**
- Coefficient of mean deviation about the mean

$$\begin{aligned} CMD(\bar{X}) &= \frac{MD\bar{X}}{\bar{X}} \\ &= \frac{3.67kg}{59kg} \\ &= 0.0622 \end{aligned}$$

- Coefficient of mean deviation about the median

$$\begin{aligned} CMD(\hat{X}) &= \frac{MD\hat{X}}{\hat{X}} \\ &= \frac{3.67kg}{58kg} \\ &= 0.0632 \end{aligned}$$

Variance and Standard Deviation

Variance and Standard Deviation

- The variance and standard deviation are the most superior and widely used measures of dispersion
- Both measures the average dispersion of the observations around the mean.
- The variance is defined as the average of the squared deviation from the mean.

population and sample variance

$$population\ variance = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)}{N} i = 1, 2, 3.....N$$

for the case of frequency distribution

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \mu)}{N} \quad (9)$$

Where i=1,2,3..K and x_i *areclassmarks*
Sample variance:the sample variance is dented by S^2 *andisgivenby*

$$sample\ variance = S^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{X})}{n-1}, i = 1, 2, 3...n$$

$$for\ the\ case\ of\ frequency\ distribution = S^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{X})}{n-1}$$

$$n = \sum f_i$$

$$where\ i = 1, 2, 3..K\ and\ x_i\ are\ class\ marks$$

Standard Deviation

The positive square root of the variance is called standard deviation. Therefore

$$\textit{population standard deviation} \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \tag{10}$$

$$\textit{sample standard deviation} = \sqrt{S^2} = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$$

Example

24, 25, 29,29,30,31 Find variance and standard deviation ?

Solution:

Value	value-mean($x_i - \bar{X}$)	difference	Difference square($X_i - \bar{X}$)^2
24	24-28	-4	16
25	25-28	-3	9
29	29-28	1	1
29	29-28	1	1
30	30-28	2	4
31	31-28	3	9
total		0	40

$$S^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{X})}{n-1}, i = 1, 2, 3...n$$

$$= \frac{40}{6-1}, \qquad = \frac{40}{5}$$

$S^2 = 8$, *standard deviation*, $S = \sqrt{8} = 2.83$

Exercise

Find the variance and standard deviation of the following sample data

i 5, 17, 12, 10,8

ii The data is given in the form of frequency distribution.

class	Frequency
40-44	7
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Special properties of standard deviation /variance

- The main drawback of variance => unit is squared and this is difficult to interpret.
- Variance gives weight to extreme values than those near to the mean value. This is because the difference is squared.
- Variance will be zero for distributions with equal magnitude. The greater the difference in the values, the greater the variance and vise versa.
- $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} < \sqrt{\frac{\sum_{i=1}^{n}(X_i - A)^2}{n-1}}, \bar{X} \neq A$

- If the standard deviation of $X_1, X_2, \ldots X_n$ *is* S, **Then the standard deviation 0f:**
  - $X_1 + K, X_2 + k \ldots\ldots X_n + k$ *will alsoS*
- $KX_1, KX_2, \ldots\ldots KX_n$ *wouldbe* $|k|S$
- a+$KX_1, a + KX_2, \ldots\ldots a + KX_n$ *would be* $|k|S$

special properties of standard deviation/variance

- If a sample of n_1 *observations has a variance* S_1^2
- and a sample of n_2 *observations have a variance of* S_2^2
- then the combined variance called the pooled variance (S_p^2) *is given by* :

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{11}$$

Coefficient of variation (CV)

Coefficient of variation (CV)

- In situations where either two series have different units of measurements, or their means differ sufficiently in size, the CV should be used as a measure of dispersion.

$$coefficient\ of\ variation(CV) = \frac{standard\ deviation}{mean} * 100\%$$

$$CV = \frac{S}{\bar{X}} * 100\% for\ sample\ and$$

$$CV = \frac{\sigma}{\mu} * 100\% for\ population$$

- In spite of the fact that the C.V. is broadly applied, its disadvantage is that it's not useful when the mean is negative or zero or very close to zero.

- **Interpretation of the coefficient of variation:** the distribution having less CV is said to be less variable or more consistent

EXAMPLE

Example 4.7: For the garment length data mean = 53.6 and standard deviation = 1.46cm, so that the coefficient of deviation is

Solution:

$$\text{CV}= \frac{S}{\bar{X}} * 100\%$$

$$\text{CV}= \frac{1.46}{53.6} * 100\%,$$

CV=2.72%

Example 4.8 Suppose that the mean weight of a group of students is 165 pounds with a S.D of 8 pounds. If the height of the same group of students has a mean of 60 inches with a S.D of 3 inches, compare the variability in weight and height measurements.

Solution:

$$\text{for weight CV}= \frac{8lb}{165lb} * 100\% = 4.85\%$$

and for height:

$$\text{CV}= \frac{3\ in}{60\ in} * 100\% = 5\%$$

=> The height data is more variable/less consistent than the weight data.

Standard Scores (Z-Scores)

- Are not measures of relative dispersion, but one of the applications of standard deviation.
- We define the standard score as:

$$Z = \frac{X - \bar{X}}{S} \text{ or } Z = \frac{X - \mu}{\sigma} \tag{12}$$

- Tells us how many standard deviations a value lies above (if positive) or below (if negative) the mean.
- Standard score gives the deviations from the mean in units of standard deviation
- It is used to compare two observations coming from different groups.

Standard z-score

Questions: Two third year Medical laboratory sections were given introduction to bio-statistics examinations. The following information was given.

Value	Section_1	section_2
mean	78	90
Standard deviation	6	5

Student A from section1 scored 90 and student B from section2 scored 95.
Relatively speaking who performed better ?

$$Z_1 = \frac{X_1 - \bar{X}_1}{S_1} = \frac{90 - 78}{6} = 2$$

$$Z_2 = \frac{X_2 - \bar{X}_2}{S_2} = \frac{95 - 90}{5} = 1$$

Student A performed better relative to his section because the score of student A is two standard deviation above the mean score of his section while, the score of student B is only one standard deviation above the mean score of his section.

Moments

The rth *moment about the mean* (*the rth central moment*) *defined as* :

$$M_r = \frac{\sum(X_i - \bar{X})^r}{n}, r = 0, 1, 2.. \tag{13}$$

for continuous grouped data it is given by:

$$M_r = \frac{\sum f_i(X_i - \bar{X})^r}{n}, \textit{where } X_i \textit{ are class marks} \tag{14}$$

Example: Find the first three central moments of the numbers 2, 3 and 7
Solution first find the mean:

$$\text{mean} == \frac{2+3+7}{3} = 4$$

$$\Rightarrow m_1 = \frac{\sum (X_i - \bar{X})^1}{n} = \frac{(2-4)+(3-4)+(7-4)}{3} = 0$$

$$m_2 = \frac{\sum (X_i - \bar{X})^2}{n} = \frac{(2-4)^2+(3-4)^2+(7-4)^2}{3} = 4.67$$

$$m_3 = \frac{\sum (X_i - \bar{X})^3}{n} = \frac{(2-4)^3+(3-4)^3+(7-4)^3}{3} = 6$$

Skewness

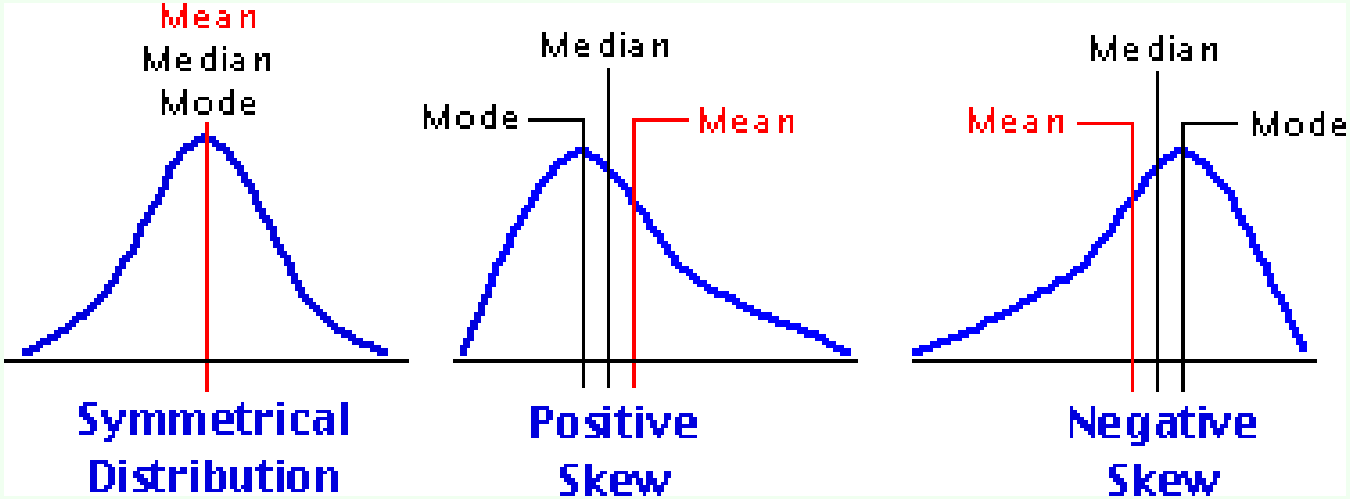
- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- A skewed frequency distribution is one that is not symmetrical.
- Skewness is concerned with the shape of the curve not size
- If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

$$Mean - Mode = 3(Mean - Median) \tag{15}$$

Skewness

Remarks:

- In a positively skewed distribution, smaller observations are more frequent than larger observations. i.e. the majority of the observations have a value below an average.
- In a negatively skewed distribution, smaller observations are less frequent than larger observations. i.e. the majority of the observations have a value above an average.



Skewness

Example. Suppose the mean, the mode, and the standard deviation of a certain distribution are 32, 30.5 and 10 respectively. What is the shape of the curve representing the distribution?
Solution:

$$S_k = \frac{mean - mode}{standard\ deviation}$$
$$= \frac{32 - 30.5}{10} = 0.15$$

The distribution is positively skewed

Measures of Skewness

The Karl Pearson's Coefficient of Skewness (SK):

$$S_k = \frac{mean - mode}{standard\ deviation}, S_k = \frac{3(mean - median)}{standard\ deviation}$$

skewness

- If SK = 0, then the distribution is symmetrical.
- If SK > 0, then the distribution is positively skewed.
- If SK < 0, then the distribution is negatively skewed

Kurtosis

- Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution.

- When the curve of a distribution is relatively:

  - flatter than normal it is known as **platykurtic** and
  - the distribution is more peaked than normal, it is called **leptokuric.**
  - The normal distribution which is not very high peaked or flat topped is called **mesokurtic.**

kurtosis

- The moment coefficient of skewness (β_2)

$$\beta_2 = \frac{m_4}{(m_2)^2} \tag{16}$$

- If B2 =3, then the distribution is **mesokurtic.**
- If B2 > 3, then the distribution is **leptokurtic.**
- If B2 < 3, then the distribution is **platykurtic.**

Thank You