Summarization of data

(Measures of variation)

Professor Narjis A-H Ajeel
Dept. of Community Medicine, College of Medicine,
University of Basrah

Learning objectives

At the end of the lecture you should be able to:

1. Define and Calculate four measures of variation for a set of data:
   - Range
   - Variance
   - Standard deviation
   - Coefficient of Variation
2. Choose the appropriate measure of variability

Measures of variation

- Measures of central location fail to answer the following question:

  (How much spread out are the measurements about the average value (the mean)?)

Measures of Variation

- Measures of variation are measures that describe how spread out or scattered a set of data is.
- They are also known as measures of dispersion or measures of spread.

There are 4 common measures of variation:

1. Range
2. Variance
3. Standard Deviation
4. Coefficient of Variation

1.The range

The range is the difference between the highest and the lowest values.

Range = Highest value – Lowest value

Example: Range of

{7, 8, 9, 10, 11, 12} is 12-7=5.

{7, 8, 9, 10, 11, 30} is 30-7=23.

Properties of the Range

1. The range is the simplest measure of variation.
2. It is easy to calculate.
3. The major disadvantages of the range are:

a. The range takes only two values into account (the highest and the lowest)
b. It is largely affected by extreme values.
c. It does not provide information on the spread of the values between the lowest and highest values.

2.The Variance

The Variance is the average squared deviation (distance) of the observations from the Mean.

The variance of the sample is denoted by (S^2)
The variance of the population is denoted by σ^2

For ungrouped data

$$S^2 = \frac{\sum (X - \bar{x})^2}{n-1}$$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

Example: Calculate the variance for the following set of data

{82, 68, 74, 86, 90, 88, 62, 75, 80, 55}

1. The first formula: $s^2 = \Sigma(x - \bar{x})^2/n-1$

x	$x - \bar{x}$	$(x - \bar{x})^2$
82	6	36
68	-8	64
74	-2	4
86	10	100
90	14	196
88	12	144
62	-14	196
75	-1	1
80	4	16
55	-21	441

$\sum x - \bar{x} = 0$ $\sum (x - \bar{x})^2 = 1198$

$$S^2 = \frac{\sum (X - \bar{x})^2}{n - 1}$$

$S^2 = 1198/9 = 133.11 \text{ year}^2$

Note: The sum of deviations from the mean for any set of data is zero

The sum of deviations= {(82-76) + (68-76) + (74-76) + (86-76) + (90-76) + (88-76) + (62-76) + (75-76) + (80-76) + (55-76)} =

(6 – 8 – 2 +10 + 14 + 12 – 14 – 1 + 4 – 21)= 0

2. The second formula

x	x^2
82	6724
68	4724
74	5476
86	7396
90	8100
88	7744
62	3844
75	5625
80	6400
55	3025
$\sum x=760$	$\sum x^2=58958$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

s^2 = (58958)-{$(760)^2$/10}/ 10-1

= (58958-57760)/ 9

= 133.11 year^2

Variance for grouped data

$$S^2 = \frac{\sum fX^2 - \frac{(\sum fX)^2}{n}}{n-1}$$

Example

Age (yrs)	F	Midpoint (X)	fX	fX^2
1 –	18	2	36	18x4=72
3 –	20	4	80	20x16=320
5 –	39	6	234	39x36=1404
7 –	17	8	136	17x64=1088
9 – 11	6	10	60	6x100=600
Total	∑f=100	∑fx=546		$\sum fX^2$ =3484

$$S^2 = \frac{3484 - \frac{546^2}{100}}{100-1} = 5.079 \text{ year}^2$$

3.Standard Deviation

- It is the **average deviation** of the observations **from the Mean**
- It is the square root of variance

-The standard deviation of a **sample** is denoted by the letter **(s) or (sd)**.
-The standard deviation of the **population** is denoted by **(σ)**.

Standard deviation for ungrouped data

$$\boxed{Sd = \sqrt{\frac{(X_i - \overline{X})^2}{n-1}}}$$

$$Sd = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$$

Standard Deviation for grouped data

$$Sd = \sqrt{\frac{\sum fX^2 - \frac{(\sum fX)^2}{n}}{n-1}}$$

Properties of the standard deviation

1. It is the most commonly used measure of variation.
2. All values are taken into account in its calculation.
3. It has the same units as the original data

4. The coefficient of variation

- It is the ratio of the standard deviation to the mean
- It is a measure of relative variation rather than absolute variation and expressed as a percent

:

$$CV = \frac{sd}{\bar{x}} \times 100$$

The coefficient of variation is used to:-

1. Compare the variation of two or more distributions if they have different units (CV has no unit this makes comparing data sets with different units possible).
2. Compare the variation of two or more distributions having widely different means.

Example: Compare the variability in heights and weights of a sample of 40 men?

	Sample mean	Sample sd
Height	68.34 in	3.02 inches
Weight	172.55 lb	26.33 lb

$$CV(height) = \frac{s}{\bar{x}} * 100 = \frac{3.02}{68.34} x100 = 4.42\%$$

$$CV(weight) = \frac{s}{\bar{x}} * 100 = \frac{26.33}{172.55} x100 = 15.26\%$$

Conclusion:

Heights (with CV=4.42%) have less variation than weights (with CV=15.26%)

Terminology

	SYMBOL/ NOTATION	
Quantity	Population Parameter	Sample Statistics
Mean	**μ**	$\bar{X}$
Standard Deviation	**σ**	**Sd**
Variance	σ^2	s^2

Questions

1. What is the relationship between the variance and the standard deviation?

a. Variance is the square root of the standard deviation

b. Variance is the square of the standard deviation

c. No constant relationship between the variance and the standard deviation

d. None of the above

2. Which of the following is a disadvantage of the range as a measure of dispersion?

a. Can be distorted by a large mean

b. Not in the same units as the original data

c. Based on only two observations

d. Has no disadvantage

3. In general, which of the following statements is FALSE?

a. The sample mean is affected by extreme values.

b. The sample range is more affected by extreme values than the standard deviation.

c. The sample standard deviation is a measure of spread around the sample mean.

d. The sample standard deviation is a measure of variation around the median.

Thank You