9.9 STATISTICS, BASIC DEFINITIONS AND NOTATIONS

Statistics is the science of collecting, analysing, presenting, and interpreting data. A population includes all the elements from a set of data.

A sample consists of one observation drawn from the population (a small part of population).

Studying a population, we are interested in collecting information about different characteristics of the subject (like their length, or weight, or age) in a sample. Those characteristics are called variables.

There are two main branches of statistics: descriptive and inferential. Descriptive statistics is used to say something about a set of information that has been gathered from a small part. Inferential statistics is used to make predictions or comparisons about a larger group (a population) using information gathered about a small part of that population.

Other distinctions are made between data types. Discrete data consist of integer numbers and usually describes a number of objects. For instance, one study might describe how many children different families own.

Measured data, in contrast to discrete data, are continuous, and thus may take on any real value. For example, the amount of time a group of children spent playing computer games would be measured data, since they could spend any number of hours playing.

For a given random sample of data:

$$x_1, x_2, \dots, x_n$$

we define a few numbers that characterise a given sample.

Definition: Sample mean

The *sample mean* is defined by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Definition: Sample variance

The *sample variance* is defined by

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Co-funded by the
Erasmus+ Programme
of the European Union

Remark:

The sample variance is a measure of how items in a sample are dispersed about their mean.

Definition: Sample standard deviation

The sample *standard deviation* is a square root of the variance

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

Example 9.30

Sample:

$$3,\ 3,\ 3,\ 3,\ 3$$

mean:

$$\bar{x} = \frac{1}{5}(3 + 3 + 3 + 3 + 3) = 3$$

variance:

$$s^2 = \frac{1}{5}[(3-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2] = 0$$

standard deviation:

$$s = \sqrt{0^2} = 0$$

There is no dispersion of data around the mean.

Example 9.31

Sample:

$$3,\ 2,\ 4,\ 3,\ 3$$

mean:

$$\bar{x} = \frac{1}{5}(3 + 2 + 4 + 3 + 3) = 3$$

variance:

$$s^2 = \frac{1}{5}[(3-3)^2 + (2-3)^2 + (4-3)^2 + (3-3)^2 + (3-3)^2] = \frac{2}{5} = 0.4$$

standard deviation:

$$s = \sqrt{\frac{2}{5}} = 0.63$$

There is some dispersion of the data around the mean.

Co-funded by the
Erasmus+ Programme
of the European Union

Example 9.32

sample:

$$5,\ 2,\ 4,\ 2,\ 2$$

mean:

$$\bar{x} = \frac{1}{5}(5 + 2 + 4 + 2 + 2) = 3$$

variance:

$$s^2 = \frac{1}{5}[(5-3)^2 + (2-3)^2 + (4-3)^2 + (2-3)^2 + (2-3)^2] = \frac{8}{5} = 1.6$$

Standard deviation

$$s = \sqrt{\frac{8}{5}} = 1.26$$

<u>There is a significant dispersion of the data around the mean.</u>

Remark

We will define one more characteristic of the sample.

Standard deviation and variance are positive unless all elements in the sample are the same. In this case are always equal to zero.

Definition: Estimator of variance

The *estimator of variance* is given by

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The *standard deviation* of the estimator of variance is given by

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

Example 9.33

Sample:

$$5,\ 2,\ 4,\ 2,\ 2$$

mean:

$$\bar{x} = \frac{1}{5}(5 + 2 + 4 + 2 + 2) = 3$$

estimator of variance:

$$\hat{s}^2 = \frac{1}{4}[(5-3)^2 + (2-3)^2 + (4-3)^2 + (2-3)^2 + (2-3)^2] = \frac{8}{4} = 2$$

standard deviation of the estimator of variance:

$\hat{s} = \sqrt{2} = 1.41$

Co-funded by the
Erasmus+ Programme
of the European Union