Descriptive statistics

Business Mathematics

VU UNIVERSITY AMSTERDAM
Faculty of Economics and Business Administration

CONTENTS

Data reduction

Measures of location

Measures of dispersion

Measure of association

Example on housing market

Old exam question

DATA REDUCTION

If we have n measurements on the same variable x, we can organize this in a data set $x_i, i = 1, \ldots, n$

- or in a data vector $\mathbf{x}$ of size n

This data set can be summarized by means of descriptive statistics (data reduction):

- measures of location (mean, median, etc.)
- measures of disperion (variance, standard deviation, etc.)

MEASURES OF LOCATION (CENTRAL TENDENCY)

Mean:

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

- alternative names: average, arithmetic mean




MEASURES OF DISPERSION (VARIABILITY)

Measure of dispersion (variability)

Variance:

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

- also denoted by ${s_x}^2$ to emphasize that it is s^2 of x

Why $\frac{1}{n-1}$ and not $\frac{1}{n}$?
Not clear for this moment, but we will return to this in the statistics course

MEASURES OF DISPERSION (VARIABILITY)

Two more measures of dispersion

In original units:
Standard deviation:

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

- or s_x if there are more data vectors

Relative (dimensionless) measure of dispersion:
Coefficient of variation:

$$\boxed{CV = \frac{s}{\bar{x}}}$$

- or CV_x if there are more data vectors

MEASURES OF ASSOCIATION

Measures of association between two paired data vectors

- so, considering $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, condensed in two data vectors $\mathbf{x}$ and $\mathbf{y}$ of equal size

Covariance:

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

What happens if you calculate the covariance of a data vector and the same data vector?

- $s_{x,x} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = {s_x}^2$

so the covariance between x and x is the variance of x

MEASURES OF ASSOCIATION

Alternative (relative, dimensionless) measure of association

Correlation coefficient:

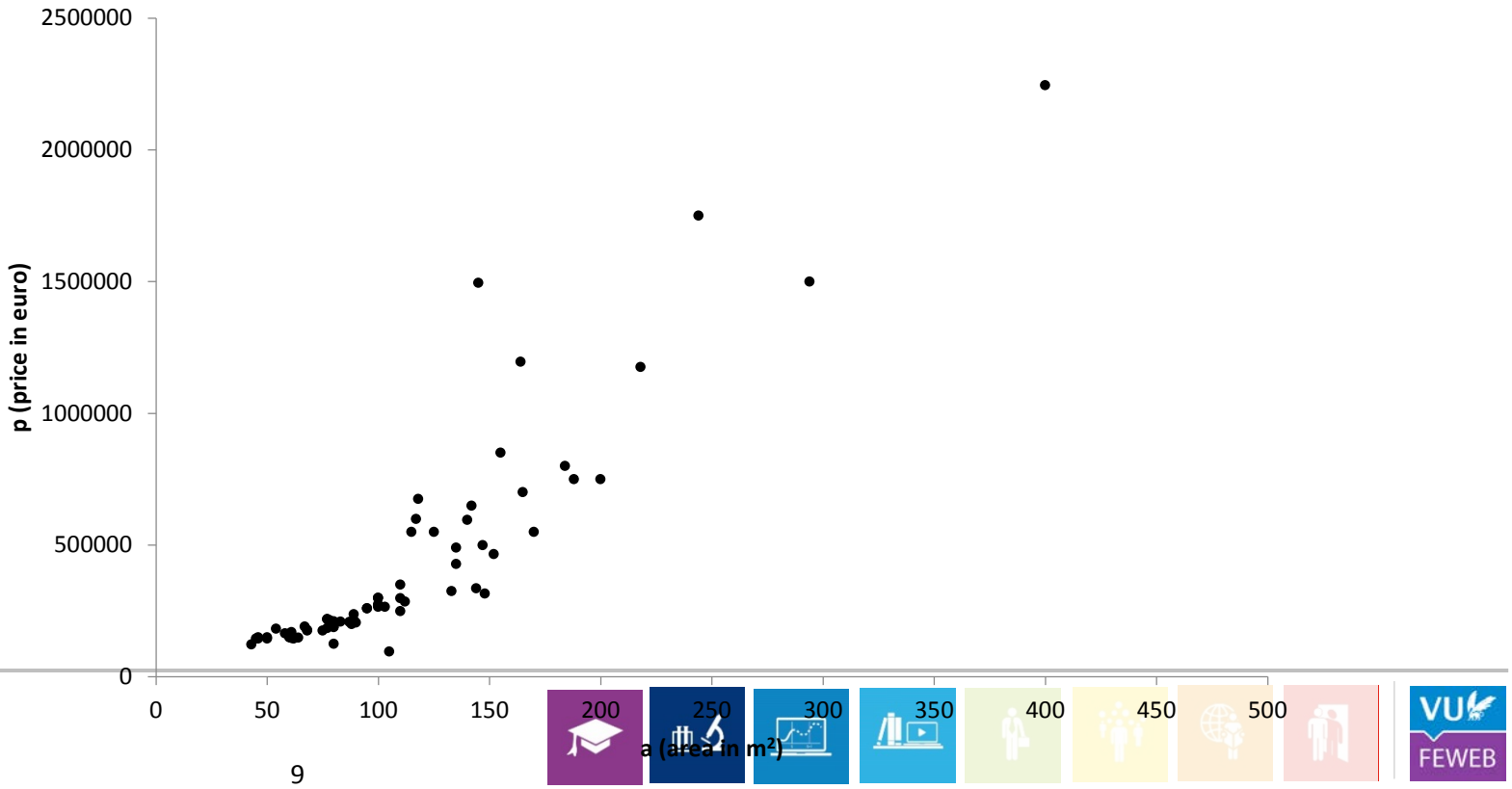
$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

- alternative name: Pearson product-moment correlation coefficient
- always: $-1 \leq r \leq 1$

EXAMPLE ON HOUSING MARKET

Data on Amstelveen 2014 ($n = 71$):

- house prices in €: p: $(649000,125000,599000,145000, \cdots)$
- floor areas in m²: a: $(142,80,117,46, \cdots)$



EXAMPLE ON HOUSING MARKET

Mean: $\bar{p} = 414014$ €; $\bar{a} = 110.4$ m^2

Variance: ${s_p}^2 = 1.69 \times 10^{11}$ €^{2}; ${s_a}^2 = 3673$ m^4

Standard deviation: $s_p = 411864$ €; $s_a = 60.6$ m^2

Coefficient of variation: $CV_p = 0.99$; $CV_a = 0.55$

Covariance: $s_{p,a} = 2.26 \times 10^7$ €×m^2

Correlation coefficient: $r_{p,a} = 0.90$

OLD EXAM QUESTION

22 October 2014, Q2c

Price data for products 1 and 2 are available, by observations on 251 days. The average prices are $\overline{p_1} = 23.6$ euro and $\overline{p_2} = 71.2$ euro, with variances $s_{p_1}^2 = 5.4$ euro^2 and $s_{p_2}^2 = 8.3$ euro^2. The correlation coefficient between the price vectors is $r_{p_1,p_2} = 0.71$. Compute the coefficient of variation for the price data of product 1 (CV_{p_1}) as well as the covariance between the two price vectors s_{p_1,p_2}. (6 points)

OLD EXAM QUESTION

27 March 2015, Q1j

Someone claims that a data vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$ has $\bar{x} = -23.6$ and $\sum_{i=1}^{n} x_i = 498.3$. Which statements can you deduce? (choose one or more)

(A) $n > 10$

(B) all data elements x_i are ≥ 0

(C) there must be a mistake in $\bar{x}$ and/or $\sum_{i=1}^{n} x_i$

(D) no data element x_i is > 500

(E) None of the above is correct.