Review: Arithmetic mean & variances

Econometrics

herbert.stocker@uibk.ac.at
www.hsto.info/econometrics2

Means: Arithmetic mean

- Mean: Umbrella term for different measures of location.
- Mean: often used for Arithmetic mean

Definition Arithmetic mean

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

- only useful for metrically scaled variables!
- sensitive to extreme values.

Arithmetic mean

Arithmetic mean: If some of the n observations have the same numerical value, they can be summarized

$$\bar{x} = \frac{1}{n}\left(\underbrace{x_1 + \cdots + x_1}_{n_1\text{-times}} + \underbrace{x_2 + \cdots + x_2}_{n_2\text{-times}} + \cdots + \underbrace{x_k + \cdots + x_k}_{n_k\text{-times}}\right)$$

with frequencies $n_1, n_2, \ldots, n_k$

$$\bar{x} = \frac{1}{n}(x_1 n_1 + \cdots + x_k n_k) = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \sum_{j=1}^{k} x_j \frac{n_j}{n}$$

and with $\sum_{j=1}^{k} n_j = n$

Arithmetic mean

Arithmetic mean:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \sum_{j=1}^{k} x_j \frac{n_j}{n} = \quad \text{with } j = 1, \ldots, k$$

respectively with $f_j := \frac{n_j}{n}$ (*relative* frequencies)

$$\boxed{\bar{x} = \sum_{j=1}^{k} x_j f_j}$$

$\Rightarrow$ **weighted arithmetic mean**

Arithmetic mean

Arithmetic mean: 4 Properties

1. **Center of gravity property:** The sum of the deviations of the individual values from the arithmetic mean $\bar{x}$ is zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_i x_i - \sum_i \bar{x} = n\bar{x} - n\bar{x} = 0$$

because from $\bar{x} = \frac{1}{n}\sum_i x_i$ follows $\sum_i x_i = n\bar{x}$, and $\sum_i \bar{x} = n\bar{x}$

$\rightarrow$ Center of gravity of a distribution.

Arithmetic mean

Arithmetic mean: 4 Properties

2. The sum of the squared deviations from $\bar{x}$ is smaller than from any other fixed value z

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - z)^2 \quad \text{für } \bar{x} \neq z$$

- Why?

$$\begin{aligned}
\sum_i (x_i - z)^2 &= \sum_i (x_i - \bar{x} + \bar{x} - z)^2 = \sum_i \left[(x_i - \bar{x}) + (\bar{x} - z)\right]^2 \\
&= \sum_i (x_i - \bar{x})^2 + 2(\bar{x} - z)\underbrace{\sum_i (x_i - \bar{x})}_{=0} + \sum_i (\bar{x} - z)^2 \\
&= \sum_i (x_i - \bar{x})^2 + \sum_i (\bar{x} - z)^2 \quad \text{with } \sum_i (\bar{x} - z)^2 > 0
\end{aligned}$$

- $\Rightarrow$ arithm. mean $\bar{x}$ generates the smallest possible dispersion (variance)!

Arithmetic mean

Arithmetic mean: 4 Properties

3. Any linear transformation of individual values $x_i^* = b_1 + b_2 x_i$ results in a equivalent transformation of the arithmetic mean

$$\bar{x}^* = b_1 + b_2\bar{x}$$

- Why?

$$\begin{aligned}
\bar{x}^* &= \frac{1}{n}\sum_i (b_1 + b_2 x_i) \\
&= \frac{1}{n}\left(nb_1 + b_2\sum_i x_i\right) \\
&= b_1 + b_2\frac{1}{n}\sum_i x_i = b_1 + b_2\bar{x}
\end{aligned}$$

Arithmetic mean

Arithmetic mean: 4 Properties

4. **Weighted arithm. mean:** Two (or more) subpopulations whose size and arithm. means are known ($n_1, \bar{x}_1, n_2, \bar{x}_2$ with $n_1 + n_2 = n$)

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\left(\frac{x_1}{x_1}\sum_{i=1}^{n_1} x_{1i} + \frac{n_2}{n_2}\sum_{i=1}^{n_2} x_{2i}\right) \\
&= \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}\right) + \frac{n_2}{n}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}\right) = \frac{n_1}{n}\,\bar{x}_1 + \frac{n_2}{n}\,\bar{x}_2
\end{aligned}$$

- Why?
  From the definition of the arithm. mean follows

$$\sum_{i=1}^{n_1} x_{1i} = n_1\bar{x}_1 \text{ and } \sum_{i=1}^{n_2} x_{2i} = n_2\bar{x}_2$$

Variance

Definition Variance

Variance s^2: Mean square deviation from the arithmetic mean $\bar{x}$.

$$\begin{aligned} \operatorname{var}(x) := s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 := \overline{x^2} - \bar{x}^2 \end{aligned}$$

Variance

Why?

$$\begin{aligned} \operatorname{var}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{n}\left(\sum_i x_i^2 - 2\bar{x}\sum_i x_i + \sum_i \bar{x}^2\right) \\ &= \frac{1}{n}\left(\sum_i x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 := \overline{x^2} - \bar{x}^2 \end{aligned}$$

because $\sum_i x_i = n\bar{x}$ und $\sum_i \bar{x}^2 = n\bar{x}^2$

Varianz

Linear transformed data:

Let $x_i^* = b_1 + b_2 x_i$ for $i = 1, \dots, n$

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

as have previously shown $\bar{x}^* = b_1 + b_2\bar{x}$.

$$\begin{aligned} \operatorname{var}(x^*) &= \frac{1}{n}\sum_i (b_1 + b_2 x_i - b_1 - b_2\bar{x})^2 \\ &= \frac{1}{n}\sum_i (b_2[x_i - \bar{x}])^2 \\ &= b_2^2\frac{1}{n}\sum_i (x_i-\bar{x})^2 = b_2^2 \operatorname{var}(x_i) \end{aligned}$$

$\Rightarrow$ Addition or subtraction of a constant has no effect on the variance!

Varianz

Two types of variances

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

- Most programs calculate the variance according to the 2nd formula, i.e. they use the pre-factor $1/(n-1)$
- the use of the pre-factor $1/(n-1)$ instead of $1/n$ is appropriate when the variance is calculated from a sample and serves as an *estimate* for the variance of the population.
- The reason for this lies in the concept of *unbiasedness* discussed later.

Standard deviation

- The variance is sometimes difficult to interpret, e.g. if x is measured in euros, the variance has the dimension Euro^2.
- The *standard deviation* has the advantage over the variance that it is measured in the same unit as the observed values.

Definition Standard deviation

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

Covariance

a measure of the joint variability of two metrically scaled variables

Definition Covariance

Covariance is a (non-standardized) measure of the relationship between two metrically scaled statistical variables x and y.

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

with $\bar{x} := \frac{1}{n}\sum_i x_i$ und $\bar{y} := \frac{1}{n}\sum_i y_i$

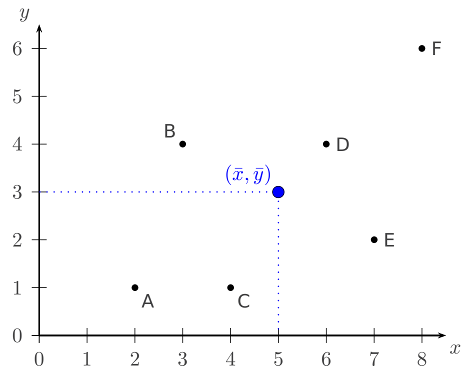
Covariance

Example:

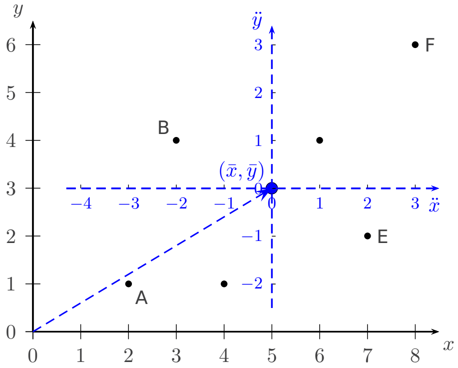
	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
	2	1	-3	-2	6
	3	4	-2	1	-2
	4	1	-1	-2	2
	6	4	1	1	1
	7	2	2	-1	-2
	8	6	3	3	9
$\sum$	30	18	0	0	14

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{14}{6} = 2.33$$

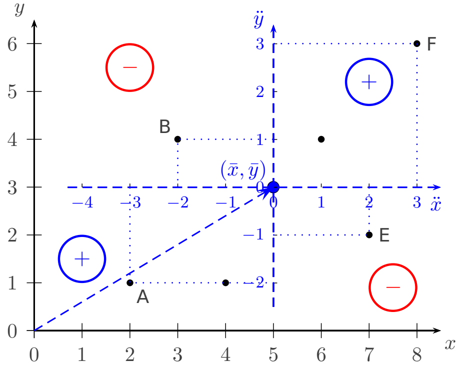
Covariance: mean transformation



Covariance: mean transformation



Sign of the covariance



Kovarianz

Covariance:

- The covariance is positive if x and y tend to have a (linear) relationship in the same direction, i.e. high values of x are associated with high values of y and low values of x with low values of y.
- The covariance is negative if x and y show an opposite linear relationship.
- If the covariance is zero, there is no *linear relationship* (but there may still be a non-linear relationship, e.g. U-shaped).

Calculation rules for empirical covariances

1) Symmetry:

$$\boxed{\mathrm{cov}(x, y) = \mathrm{cov}(y, x)}$$

Why?

$$\begin{aligned} \mathrm{cov}(x, y) &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \mathrm{cov}(y, x) \end{aligned}$$

Calculation rules for empirical covariances

2) Constant factors can be factored out: for $x, y \in \mathbb{R}^n$ and constants $a, b \in \mathbb{R}$

$$\boxed{\operatorname{cov}(ax, by) = ab \operatorname{cov}(x, y)}$$

Why?

$$\begin{aligned}
\operatorname{cov}(ax, by) &= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})(by_i - b\bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n} a(x_i - \bar{x})b(y_i - \bar{y}) \\
&= ab\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= ab \operatorname{cov}(x, y)
\end{aligned}$$

Calculation rules for empirical covariances

3) Additivity: for $x, y, z \in \mathbb{R}^n$

$$\boxed{\operatorname{cov}[x, (y+z)] = \operatorname{cov}(x, y) + \operatorname{cov}(x, z)}$$

Why?

$$\begin{aligned}
\operatorname{cov}[x, (y+z)] &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})[(y_i + z_i) - (\overline{y+z})] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})[(y_i + z_i) - (\bar{y} + \bar{z})] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})[(y_i - \bar{y}) + (z_i - \bar{z})] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z}) \\
&= \operatorname{cov}(x, y) + \operatorname{cov}(x, z)
\end{aligned}$$

Calculation rules for empirical covariances

4) Connection with empirical variance: for $x \in \mathbb{R}^n$

$$\boxed{\operatorname{cov}(x, x) = \operatorname{var}(x)}$$

Why?

$$\begin{aligned}
\operatorname{cov}(x, x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \operatorname{var}(x)
\end{aligned}$$

Calculation rules for empirical covariances

5) Variance of a sum: for $x, y \in \mathbb{R}^n$

$$\boxed{\operatorname{var}(x+y) = \operatorname{var}(x) + \operatorname{var}(y) + 2\operatorname{cov}(x, y)}$$

Why?

$$\begin{aligned}
\operatorname{var}(x+y) &= \frac{1}{n}\sum_{i=1}^{n}[(x_i + y_i) - (\bar{x} + \bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x}) + (y_i - \bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \\
&\quad +2\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \operatorname{var}(x) + \operatorname{var}(y) + 2\operatorname{cov}(x, y)
\end{aligned}$$

Covariance

Two types of covariance

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

versus

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

- most programs calculate the covariance according to the 2nd formula, i.e. they use the prefactor $1/(n-1)$
- the use of the prefactor $1/(n-1)$ instead of $1/n$ is appropriate if the covariance is calculated from a sample and serves as an *estimate* for the covariance of the population ($\rightarrow$ unbiasedness!).

Correlation

Correlations:

- *Covariances* depend on units of measurement! To make comparable covariance must be normalized $\Rightarrow$ correlation coefficients
- Correlations are special statistical indicators to measure the association between two variables.
- Do variables move in the same direction? How strongly are they related?

Correlation coefficient according to Bravais-Pearson

Definition *Correlation coefficient according to Bravais-Pearson*

Correlation coefficient r: a dimensionless measure of the degree of linear association between two *at least interval-scaled* characteristics.

$$\operatorname{corr}(x,y) := r_{xy} = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)}\sqrt{\operatorname{var}(y)}}$$

Korrelationskoeffizient nach Bravais-Pearson

Eigenschaften des Korrelationskoeffizient nach Bravais-Pearson: for vectors $x, y \in \mathbb{R}^n$ and constants $a, b, c, d \in \mathbb{R}$

1. $r_{x,y}$ can only take values between -1 and $+1$

$$-1 \leq \operatorname{corr}(x,y) \leq +1$$

2. $r_{x,y}$ does not change with a linear transformation

$$\operatorname{corr}(ax+b, cy+d) = \operatorname{corr}(x,y)$$

3. If the $\operatorname{corr}(x,y) = 0$, the two features are linearly independent (but they can still be non-linearly dependent); if $|\operatorname{corr}(x,y)| = 1$, the features are exactly linearly dependent.
   - $\operatorname{corr}(x,y) = +1$ if $y = a + bx$
   - $\operatorname{corr}(x,y) = -1$ if $y = a - bx$

Correlation coefficient according to Bravais-Pearson

Example: mit $\ddot{x} := x - \bar{x}$, $\ddot{y} := y - \bar{y}$

	x	y	$\ddot{x}$	$\ddot{x}^2$	$\ddot{y}$	$\ddot{y}^2$	$\ddot{x}\ddot{y}$
	2	1	-3	9	-2	4	6
	3	4	-2	4	1	1	-2
	4	1	-1	1	-2	4	2
	6	4	1	1	1	1	1
	7	2	2	4	-1	1	-2
	8	6	3	9	3	9	9
$\sum$	30	18	0	28	0	20	14

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{14}{\sqrt{28 \cdot 20}} = 0.591608$$

Correlation coefficient according to Bravais-Pearson

Exercise: Show that the correlation coefficient

$$\begin{aligned} r &= \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} \\ &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \end{aligned}$$

can alternatively be calculated as

$$r = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_i x_i^2 - n\bar{x})(\sum_i y_i^2 - n\bar{y}^2)}}$$