Statistical Methods for Economists

Lecture 5

Correlation Analysis

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistical Methods for Economists
INM/BASTE

Outline of the lecture

- Revision: Scalar product, Expected value, Covariance
- Pearson's Correlation Coefficient
- Regression Coefficient
- Multiple Correlation Coefficient
- Coefficient of Partial Correlation
- Hypothesis Testing
- Non-parametric and robust methods:

  Spearman's Rank Correlation Coefficient

Simple Linear Regression: Summary

We are given an underlying probability space $(\Omega, \mathcal{F}, P)$ and n independent random variables

$$Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$$

such that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta x_i, \sigma^2) \qquad \text{for} \quad i = 1, 2, \dots, n$$

We then perform n random experiments and obtain the outcomes $\omega_1, \omega_2, \dots, \omega_n \in \Omega$ as well as the n numerical outcomes $y_1, y_2, \dots, y_n$ of the random experiments $(y_i = Y_i(\omega_i)$ for $i = 1, 2, \dots, n)$.

Simple Linear Correlation: Motivation

We are given the underlying probability space $(\Omega, \mathcal{F}, P)$ and

<u>two</u> random variables

$$Y: \Omega \to \mathbb{R} \qquad \text{and} \qquad X: \Omega \to \mathbb{R}$$

We then perform n random experiments and obtain the outcomes

$\omega_1, \omega_2, \dots, \omega_n \in \Omega$ as well as the corresponding numerical outcomes

$$y_i = Y(\omega_i) \qquad \text{and} \qquad x_i = X(\omega_i) \qquad \text{for} \quad i = 1, 2, \dots, n$$

The purpose is to decide whether there is (linear) correlation between

the values of the random variable X and the values of the random variable Y.

Multiple Linear Regression: Summary

We are given the underlying probability space $(\Omega, \mathcal{F}, P)$ and n <u>independent</u> random variables

$$Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$$

such that

$$Y_i \sim \mathcal{N}(\boldsymbol{x}_i \boldsymbol{\beta}, \sigma^2) \qquad \text{for} \quad i = 1, 2, \dots, n$$

We then perform n random experiments and obtain the outcomes $\omega_1, \omega_2, \dots, \omega_n \in \Omega$ as well as the n numerical outcomes $y_1, y_2, \dots, y_n$ of the random experiments $(y_i = Y_i(\omega_i)$ for $i = 1, 2, \dots, n)$.

Multiple Linear Correlation: Motivation

We are given the underlying probability space $(\Omega, \mathcal{F}, P)$ and

$k+2$ random variables

$$Y: \Omega \to \mathbb{R} \qquad \text{and} \qquad X_0, X_1, \dots, X_k: \Omega \to \mathbb{R}$$

We then perform n random experiments and obtain the outcomes

$\omega_1, \omega_2, \dots, \omega_n \in \Omega$ as well as the corresponding numerical outcomes

$$y_i = Y_i(\omega_i) \qquad \text{and} \qquad x_{i0} = X_0(\omega_i),\ x_{i1} = X_1(\omega_i),\ \dots,\ x_{ik} = X_k(\omega_i) \qquad \text{for } i = 1, 2, \dots, n$$

The purpose is to decide whether there is (linear) correlation between

the values of the group of the random variables $X_0, X_1, \dots, X_k$ and

Revision: Scalar Product

- Scalar Product & the Length of a vector
- Geometrical interpretation
- Useful conclusions

Revision: Scalar Product & the Length of a vector

The **scalar product** of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is

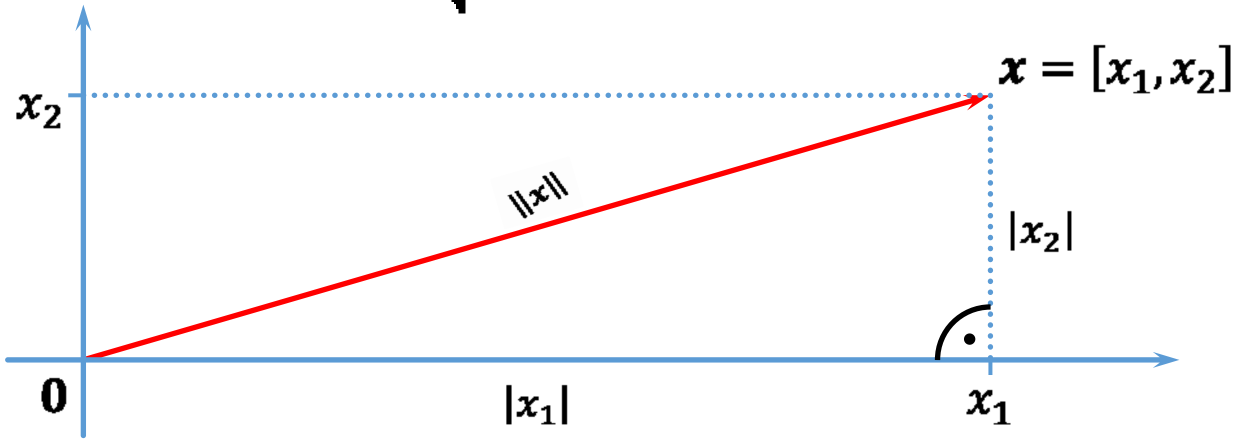
$$(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

The (Euclidean) **length** of the vector $\boldsymbol{x} \in \mathbb{R}^n$ is

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

By the Pythagoras Theorem:

$$\|\boldsymbol{x}\|^2 = |x_1|^2 + |x_2|^2$$



Revision: Scalar Product & the Length of a vector

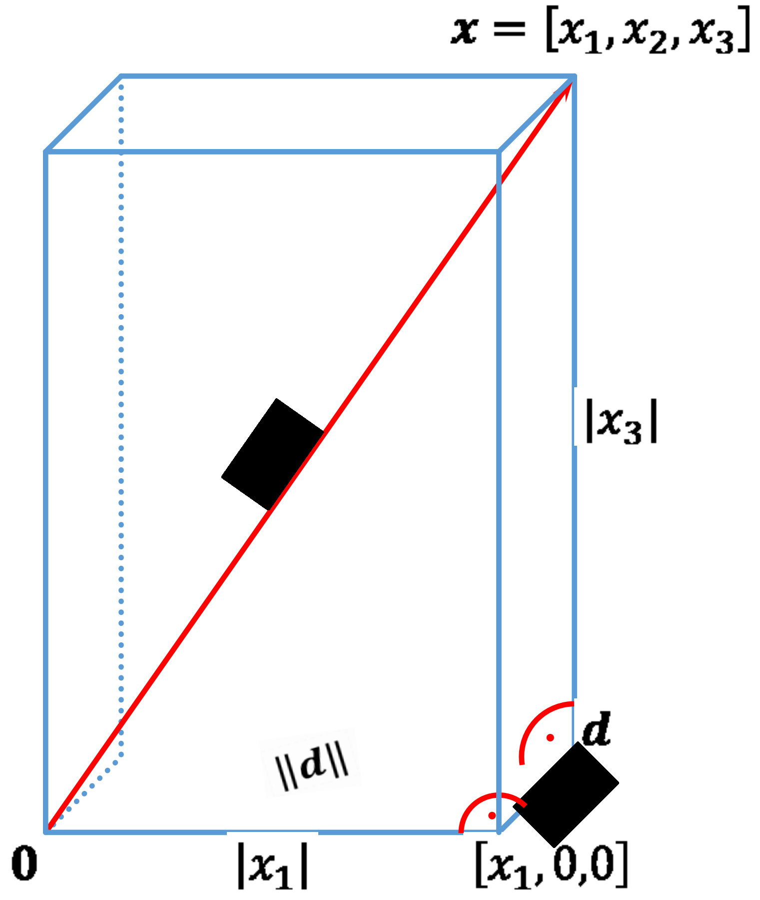
The (Euclidean) length of the vector $\boldsymbol{x} \in \mathbb{R}^n$ is

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

By the Pythagoras Theorem:

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{d}\|^2 + |x_3|^2 =$$

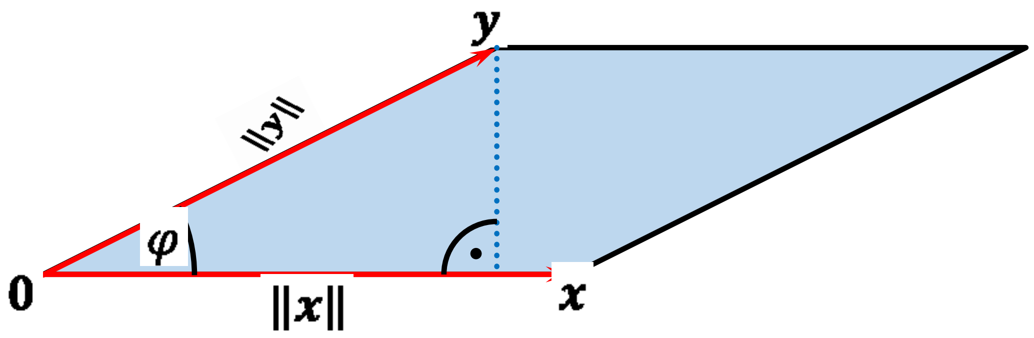
$$= |x_1|^2 + |x_2|^2 + |x_3|^2$$



Revision: Scalar Product: Geometrical interpretation

Given two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have:

$$(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y} = \|\boldsymbol{x}\| \times \|\boldsymbol{y}\| \times \cos \varphi$$

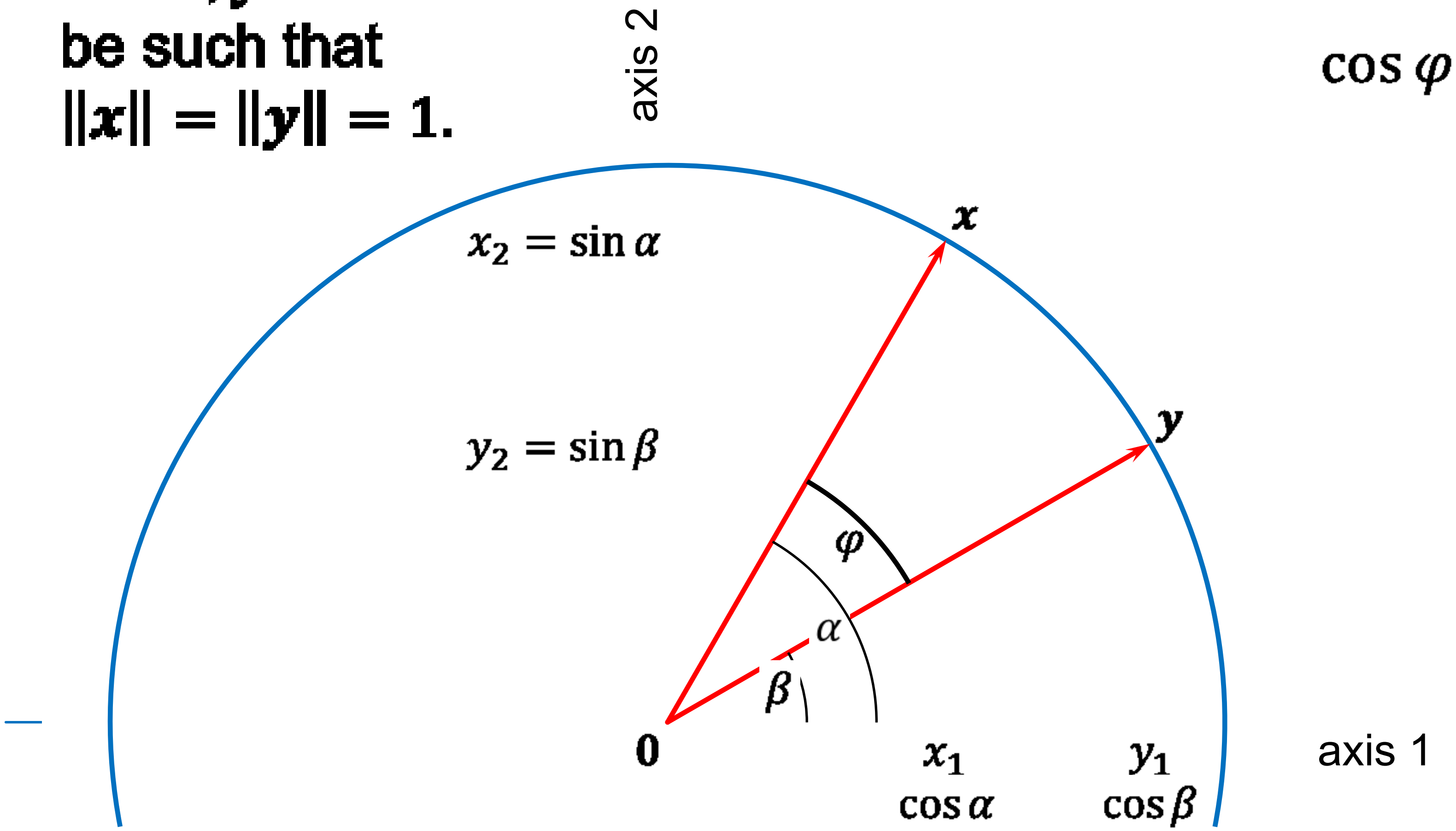


Remark:

The (absolute value of the) scalar product $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y} = \|\boldsymbol{x}\| \times \|\boldsymbol{y}\| \times \cos \varphi$

Revision: Scalar Product: Why $\cos \varphi$?

Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$
be such that
$\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1.$



$$\cos \varphi = \cos(\alpha - \beta) =$$

$$= \cos \alpha \cos \beta + \sin \alpha \sin \beta =$$

$$= x_1 y_1 + x_2 y_2$$

Revision: Scalar Product

Let $x, y \in \mathbb{R}^n$ be non-zero vectors $(x \neq 0 \neq y)$. Since $(x, y) = \|x\| \times \|y\| \times \cos\varphi$, it follows

$$\frac{(x, y)}{\|x\|\,\|y\|} = \cos\varphi$$

Therefore, it always holds:

$$-1 \leq \frac{(x, y)}{\|x\|\,\|y\|} \leq +1$$

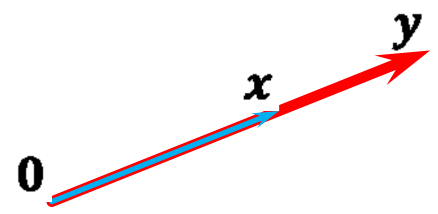
Recall:

$\cos\varphi = +1$	if and only if	$\varphi = 0°$
$\cos\varphi = -1$	if and only if	$\varphi = 180°$

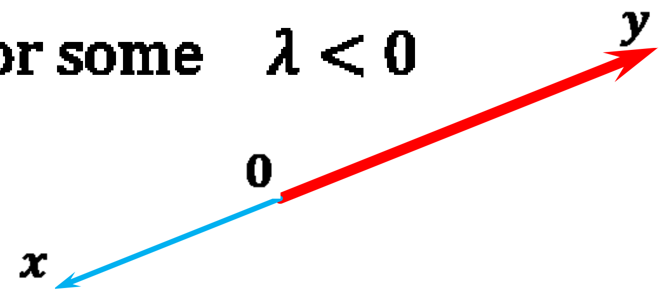
Revision: Scalar Product

Let $x, y \in \mathbb{R}^n$ be non-zero vectors $(x \neq 0 \neq y)$.

It then holds:



$$\frac{(x,y)}{\|x\|\,\|y\|} = +1 \qquad \text{if and only if} \qquad y = \lambda x \qquad \text{for some} \quad \lambda > 0$$



$$\frac{(x,y)}{\|x\|\,\|y\|} = -1 \qquad \text{if and only if} \qquad y = \lambda x \qquad \text{for some} \quad \lambda < 0$$

otherwise

$$-1 < \frac{(x,y)}{\|x\|\,\|y\|} < -1$$

Revision: Expected value, Covariance

- Expected value
- Covariance
- Variance
- Standard deviation
- Geometrical interpretation
- Uncorrelated random variables

Revision: Expected value

Let an underlying probability space $(\Omega, \mathcal{F}, P)$

and a random variable $X: \Omega \to \mathbb{R}$ be given.

- If the sample space Ω is finite $(\Omega = \{1, 2, \dots, N\})$ or countable $(\Omega = \{1, 2, 3, \dots\})$ and $p: \Omega \to \mathbb{R}$ is the probability mass function of the probability measure P, then

$$\mu_X = \mathrm{E}[X] = \sum_{\omega \in \Omega} p(\omega) X(\omega)$$

Revision: Expected value

Let an underlying probability space $(\Omega, \mathcal{F}, P)$

and a random variable $X: \Omega \to \mathbb{R}$ be given.

- If $\Omega = \mathbb{R}$ and $f: \Omega \to \mathbb{R}$ is the probability density function of the probability measure P, then

$$\mu_X = \mathrm{E}[X] = \int_{\omega \in \Omega} f(\omega) X(\omega)\, \mathrm{d}\omega = \int_{-\infty}^{+\infty} f(x) X(x)\, \mathrm{d}x$$

- If $X(x) = x$, then

$$\mu_X = \mathrm{E}[X] = \int_{-\infty}^{+\infty} x f(x)\, \mathrm{d}x$$

Revision: Covariance & Variance

Let an underlying probability space $(\Omega, \mathcal{F}, P)$

and two random variables $X, Y: \Omega \to \mathbb{R}$ be given.

The **covariance** of the random variables X and Y is

$$\mathrm{cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] =$$

$$= \mathrm{E}\big[XY - X\mathrm{E}[Y] - \mathrm{E}[X]Y + \mathrm{E}[X]\mathrm{E}[Y]\big] =$$

$$= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[Y] + \mathrm{E}[X]\mathrm{E}[Y] =$$

$$= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$$

Variance: Geometrical interpretation

Assume for simplicity that the sample space is finite $(\Omega = \{1, 2, \dots, N\})$ and that the probability mass function is uniform $(p(\omega) = 1/N$ for every $\omega \in \Omega)$.

Then, given a random variable $X: \Omega \to \mathbb{R}$, we have:

$$\mu_X = \mathrm{E}[X] = \sum_{\omega \in \Omega} p(\omega) X(\omega) = \frac{1}{N} \sum_{i=1}^{N} X_i$$

and

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)^2] = \sum_{\omega \in \Omega} p(\omega) [X(\omega) - \mu_X]^2 = \frac{1}{N} \sum_{i=1}^{N} [X_i - \mu_X]^2$$

Variance: Geometrical interpretation

We have:

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2] = \sum_{\omega\in\Omega} p(\omega)[X(\omega)-\mu_X]^2 = \frac{1}{N}\sum_{i=1}^{N}[X_i-\mu_X]^2$$

The random variable X can be seen as a vector $X \in \mathbb{R}^N$.

Let

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{be the vector of } N \text{ ones, so} \quad \mathbf{1}\mu_X = \begin{pmatrix} \mu_X \\ \mu_X \\ \vdots \\ \mu_X \end{pmatrix} \Bigg\} N$$

Variance: Geometrical interpretation

We then have:

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2] = \sum_{\omega\in\Omega} p(\omega)[X(\omega)-\mu_X]^2 = \frac{1}{N}\sum_{i=1}^{N}[X_i-\mu_X]^2 =$$

$$= \frac{1}{N}(\boldsymbol{X}-\mathbf{1}\mu_X)^{\mathrm{T}}(\boldsymbol{X}-\mathbf{1}\mu_X) = \big((\boldsymbol{X}-\mathbf{1}\mu_X),(\boldsymbol{X}-\mathbf{1}\mu_X)\big)$$

← scalar product

The standard deviation:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\big((\boldsymbol{X}-\mathbf{1}\mu_X),(\boldsymbol{X}-\mathbf{1}\mu_X)\big)}$$

← the **length** of the vector $\vec{\boldsymbol{x}} = \boldsymbol{X}-\mathbf{1}\mu_X$

Standard deviation: Geometrical interpretation

The standard deviation

$$\sigma_X = \sqrt{\left((X - \mathbf{1}\mu_X), (X - \mathbf{1}\mu_X)\right)} = \sqrt{\frac{1}{N}(X - \mathbf{1}\mu_X)^{\mathrm{T}}(X - \mathbf{1}\mu_X)} =$$

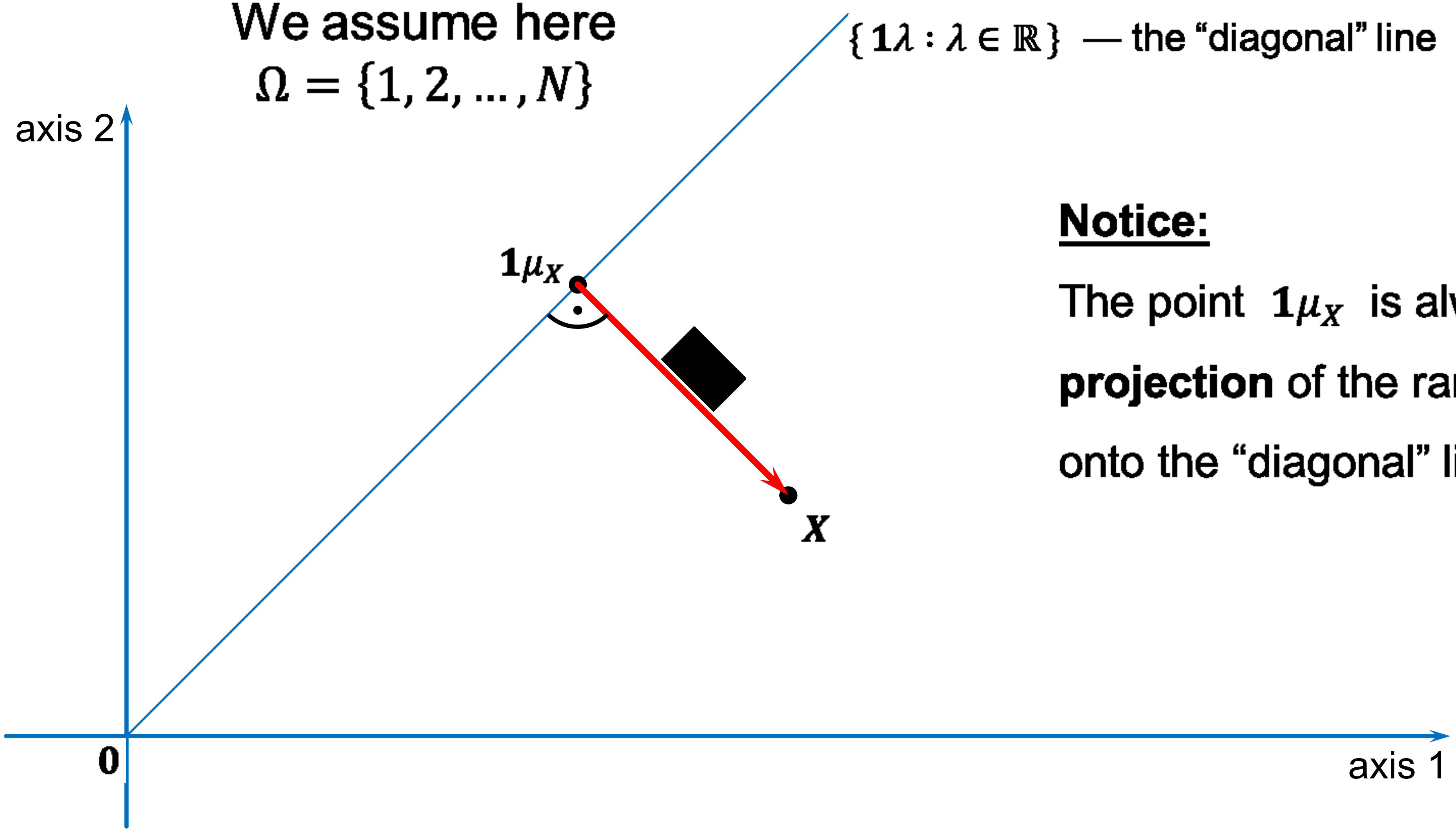
$$= \sqrt{(X - \mathbf{1}\mu_X)^{\mathrm{T}}(X - \mathbf{1}\mu_X)} \Big/ \sqrt{N}$$

is **the length of the vector** $\vec{x} = X - \mathbf{1}\mu_X$

that is the Euclidean length of the vector divided by $\sqrt{N}$.

Standard deviation: Geometrical interpretation

We assume here
$\Omega = \{1, 2, \dots, N\}$



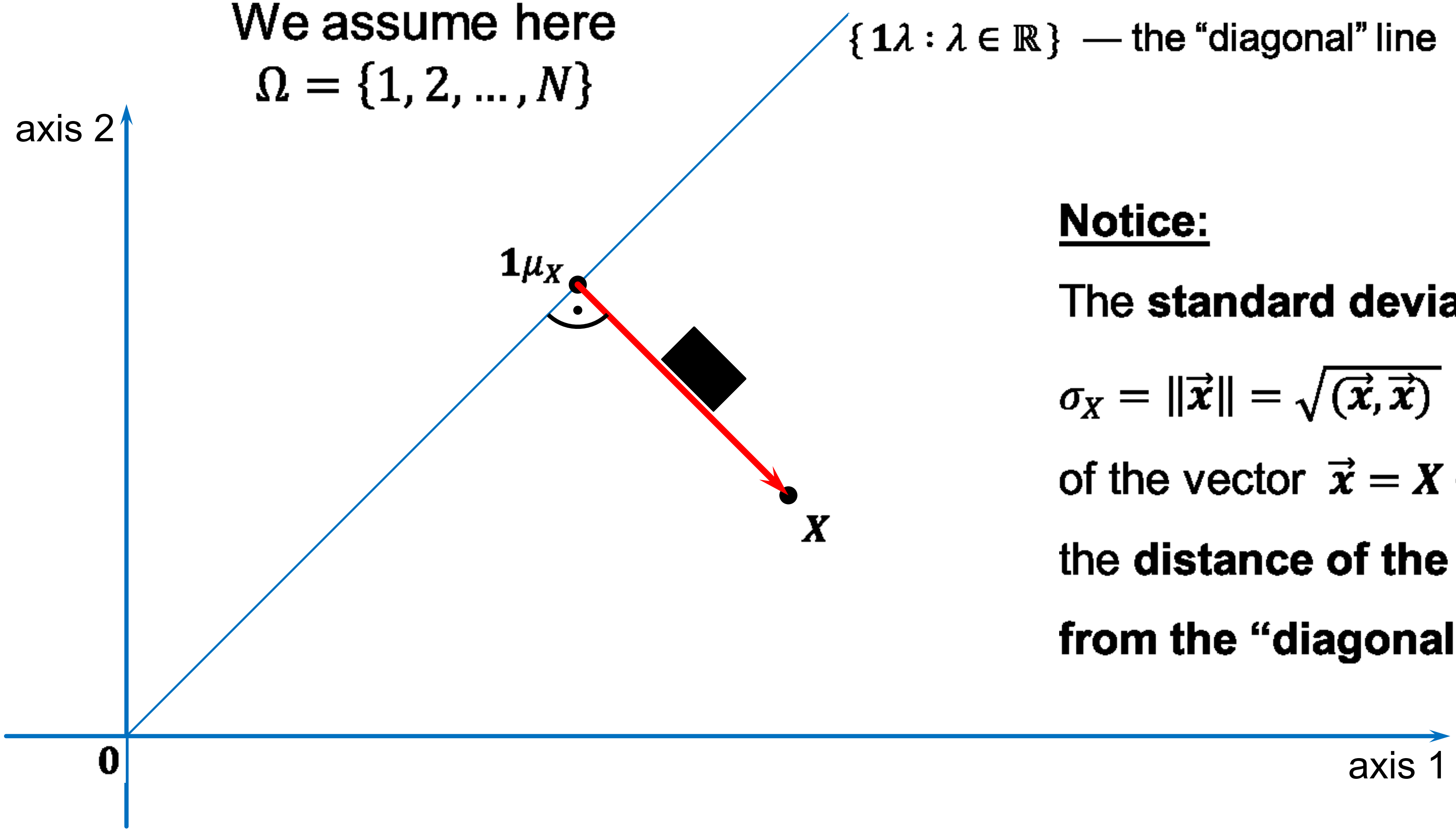
Notice:

The point $\mathbf{1}\mu_X$ is always the **orthogonal projection** of the random variable X onto the "diagonal" line $\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$.

Standard deviation: Geometrical interpretation

We assume here

$\Omega = \{1, 2, \dots, N\}$



<u>Notice:</u>

The **standard deviation**

$\sigma_X = \|\vec{\boldsymbol{x}}\| = \sqrt{(\vec{\boldsymbol{x}}, \vec{\boldsymbol{x}})}$ is the length

of the vector $\vec{\boldsymbol{x}} = \boldsymbol{X} - \mathbf{1}\mu_X$, that is

the **distance of the random variable** X

from the “diagonal” line $\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$.

Covariance: Geometrical interpretation

Assume for simplicity that the sample space is finite ($\Omega = \{1, 2, \ldots, N\}$) and that the probability mass function is uniform ($p(\omega) = 1/N$ for every $\omega \in \Omega$). Then, given two random variables $X, Y: \Omega \to \mathbb{R}$, we have:

$$\mu_X = \mathrm{E}[X] = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \text{and} \qquad \mu_Y = \mathrm{E}[Y] = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

and also

$$\sigma_{XY} = \mathrm{cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$$

Covariance: Geometrical interpretation

We then have:

$$\sigma_{XY} = \mathrm{cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y) =$$

$$= \frac{1}{N}(X-\mathbf{1}\mu_X)^{\mathrm{T}}(Y-\mathbf{1}\mu_Y) = \left((X-\mathbf{1}\mu_X), (Y-\mathbf{1}\mu_Y)\right) =$$

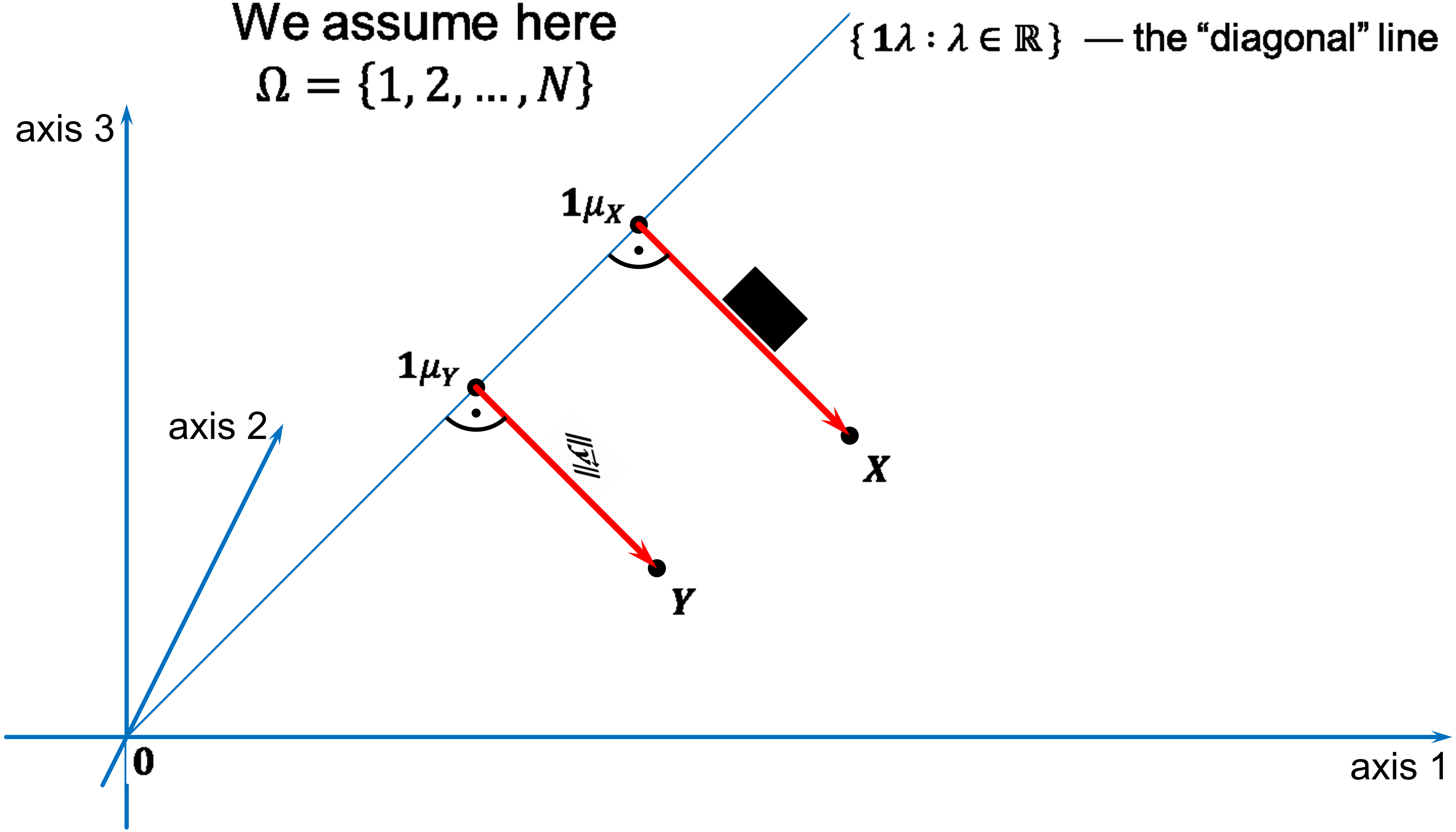
$$= \quad \|\vec{x}\|\,\|\vec{y}\|\cos\varphi \quad = \quad \sigma_X\sigma_Y\cos\varphi$$

The covariance is the scalar product of the vectors $\vec{x} = X - \mathbf{1}\mu_X$ and $\vec{y} = Y - \mathbf{1}\mu_Y$.

Covariance: Geometrical interpretation

We assume here

$$\Omega = \{1, 2, \dots, N\}$$



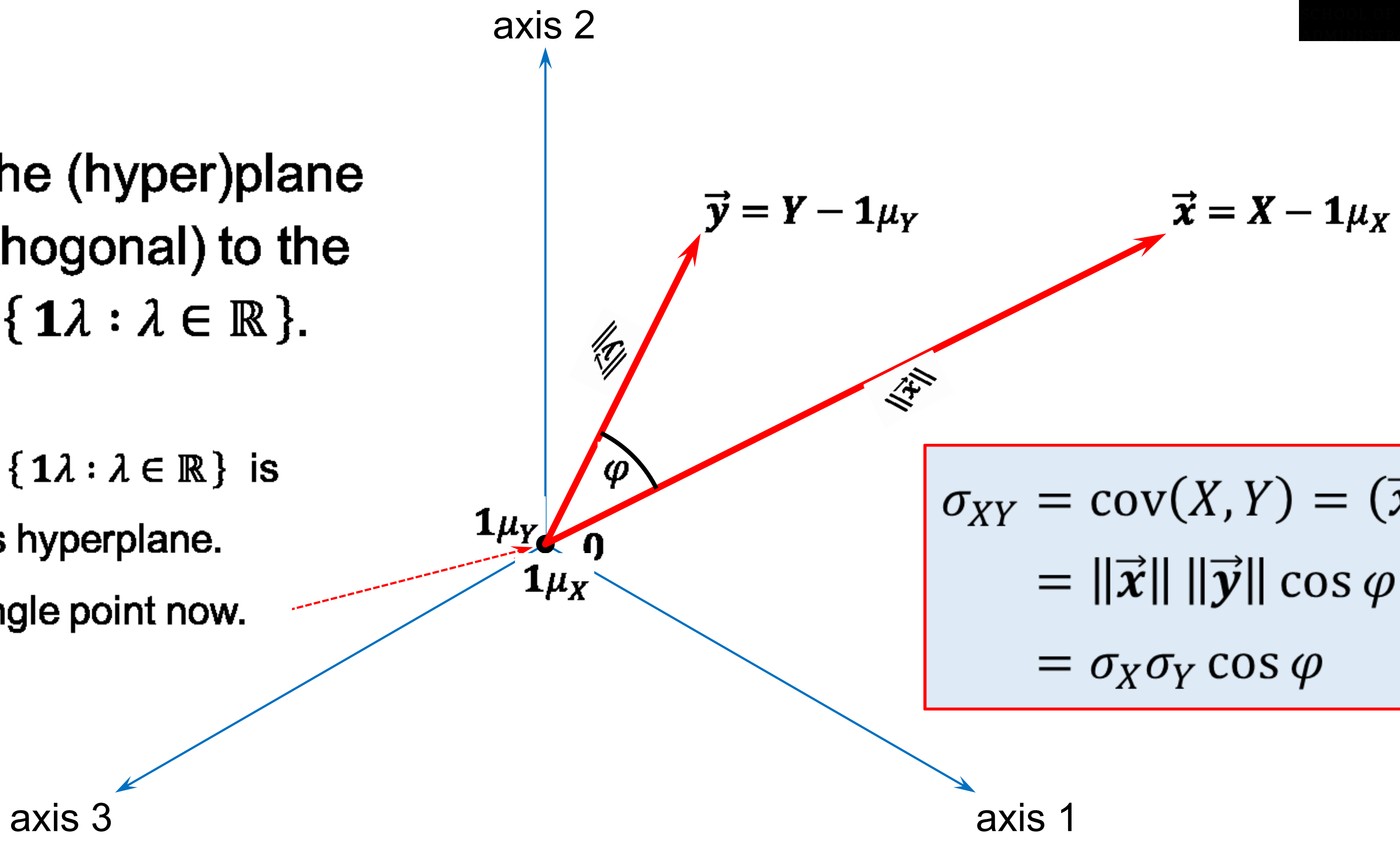
Covariance: Geometrical interpretation

We assume here

$\Omega = \{1, 2, \dots, N\}$

This view is onto the (hyper)plane perpendicular (orthogonal) to the "diagonal" line $\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$.

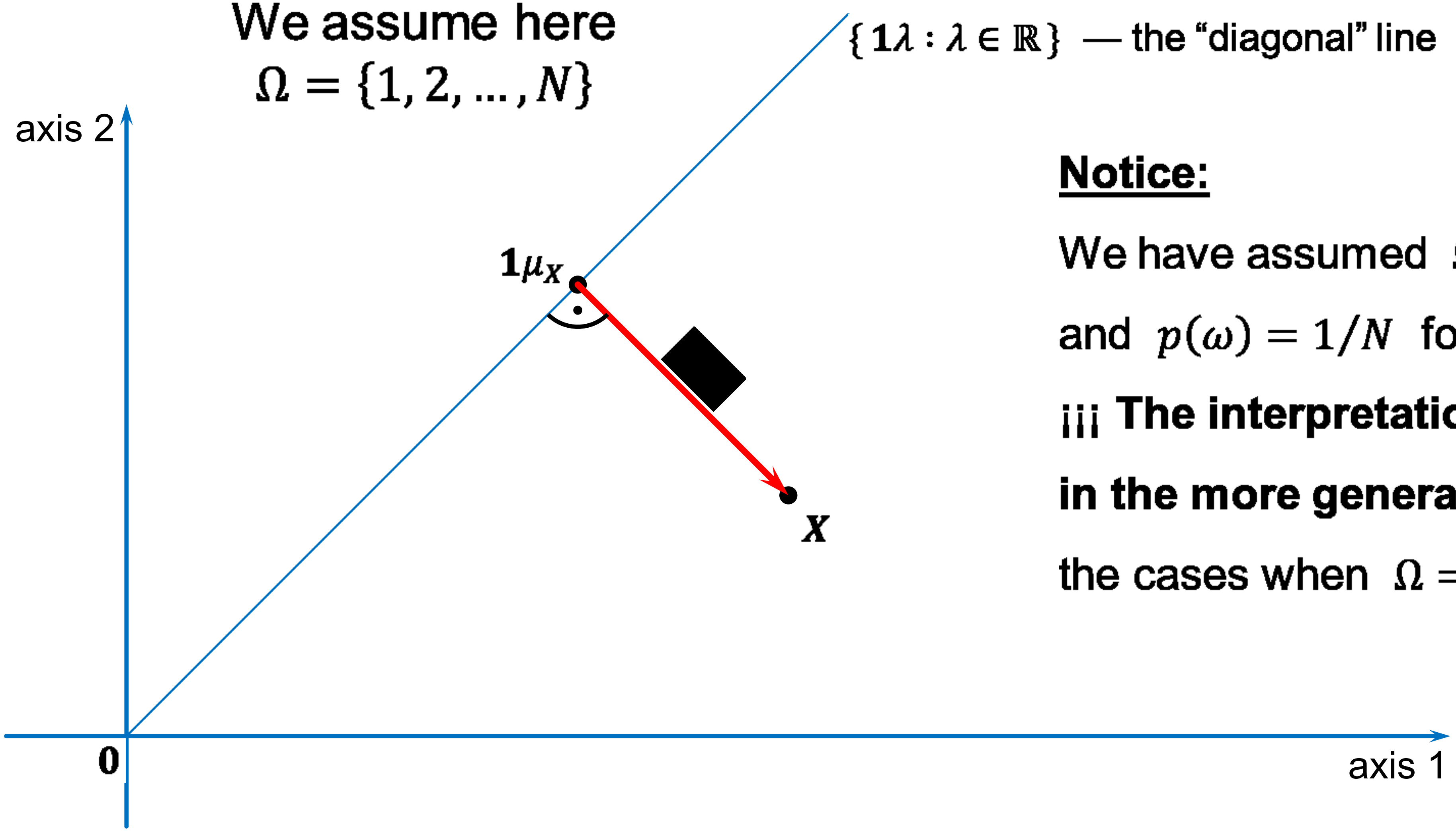
The "diagonal" line $\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$ is orthogonal to this hyperplane. It is seen as a single point now.



$$\sigma_{XY} = \operatorname{cov}(X, Y) = (\vec{x}, \vec{y}) = \\ = \|\vec{x}\|\,\|\vec{y}\| \cos\varphi = \\ = \sigma_X \sigma_Y \cos\varphi$$

¡¡¡ Notice !!!

We assume here
$\Omega = \{1, 2, \ldots, N\}$



Notice:

We have assumed $\Omega = \{1, 2, \ldots, N\}$

and $p(\omega) = 1/N$ for simplicity here.

¡¡¡ **The interpretation is analogous in the more general cases**, including the cases when $\Omega = \{1, 2, 3, \ldots\}$ and $\Omega = \mathbb{R}$!!!

The geometrical interpretations: ¡¡¡ Notice !!!

For simplicity, we have assumed $\Omega = \{1, 2, \dots, N\}$ and $p(\omega) = 1/N$, so the expected value has been

$$\mu_X = \mathrm{E}[X] = \frac{1}{N} \sum_{\omega \in \Omega} X(\omega)$$

In the more general cases, when $\Omega = \{1, 2, \dots, N\}$ or $\Omega = \{1, 2, 3, \dots\}$ and the expected value is

$$\mu_X = \mathrm{E}[X] = \sum_{\omega \in \Omega} p(\omega) X(\omega)$$

or $\Omega = \mathbb{R}$ and

$$\mu_X = \mathrm{E}[X] = \int_{\Omega} f(x) X(x) \, \mathrm{d}x$$

Pearson's Correlation Coefficient

- Covariance
- Independent random variables
- Uncorrelated random variables
- Pearson's Correlation Coefficient

Revision: Covariance

Recall that, **if** the random variables X and Y are **independent**, that is

$$P\left(\begin{matrix}\{\omega \in \Omega : a < X(\omega) < b\} \cap \\ \cap \{\omega \in \Omega : c < Y(\omega) < d\}\end{matrix}\right) = \begin{matrix} P(\{\omega \in \Omega : a < X(\omega) < b\}) \times \\ \times P(\{\omega \in \Omega : c < Y(\omega) < d\})\end{matrix}$$

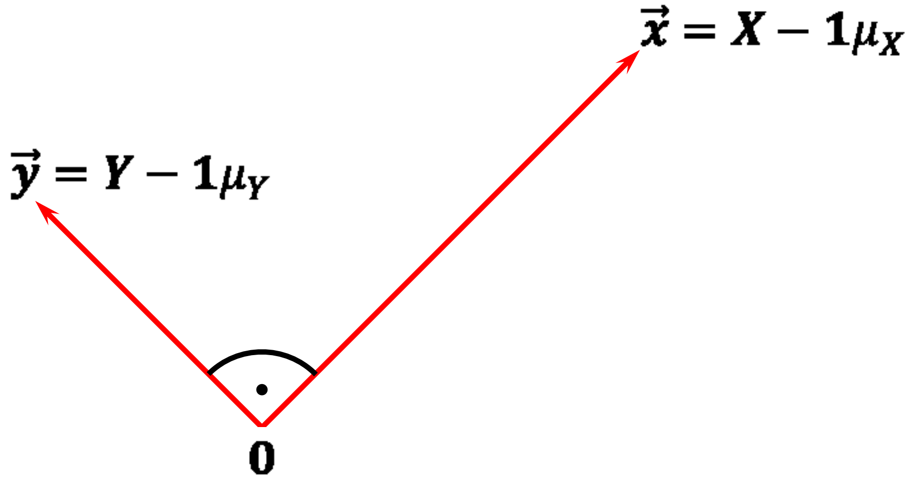
for every $a, b, c, d \in \mathbb{R} \cup \{\pm\infty\}$ such that $a < b$ and $c < d$

then

$$\text{cov}(X, Y) = 0$$

that is,

the random variables X and Y are **uncorrelated**:



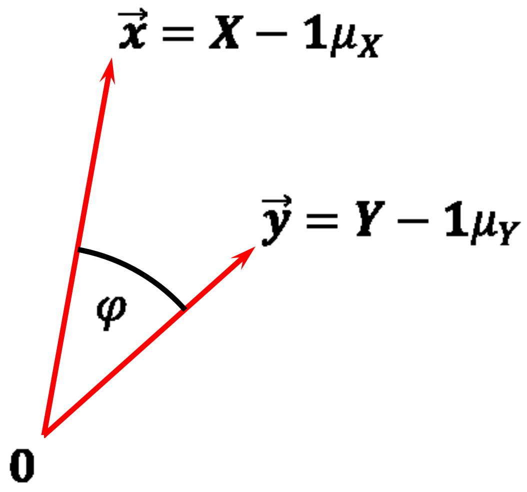
Pearson's Correlation Coefficient

Let the underlying probability space $(\Omega, \mathcal{F}, P)$

and two random variables $X, Y: \Omega \to \mathbb{R}$ be given.

Pearson's Correlation Coefficient between the two random variables X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} \qquad \text{if } \operatorname{Var}(X) \neq 0 \neq \operatorname{Var}(Y)$$



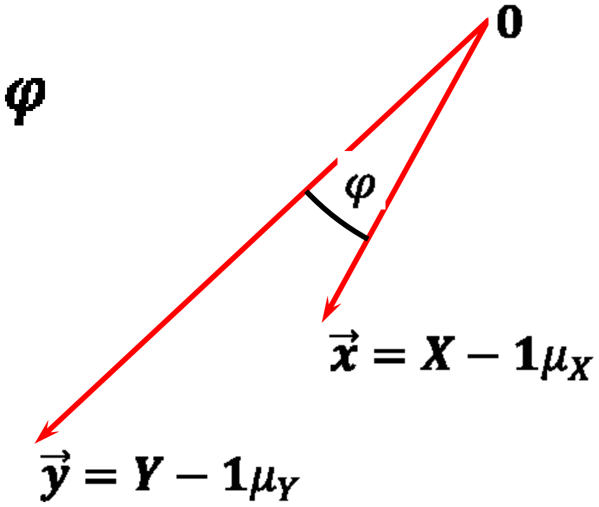
<u>Actually:</u>

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} = \frac{(\vec{x}, \vec{y})}{\|\vec{x}\|\,\|\vec{y}\|} = \cos\varphi$$

Pearson's Correlation Coefficient

We have:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} = \frac{(\vec{x},\vec{y})}{\|\vec{x}\|\,\|\vec{y}\|} = \cos\varphi$$



Notice:

- It holds $$\rho_{XY} = \rho_{YX}$$
- It holds $$-1 \le \rho_{XY} \le +1$$
- It holds $$\rho_{a+bX,\,c+dY} = \operatorname{sgn}(bd) \times \rho_{XY}$$ if $b \neq 0 \neq d$

for every $a, b, c, d \in \mathbb{R}$

Pearson’s Correlation Coefficient

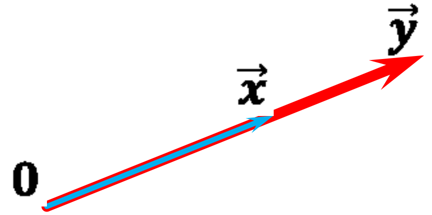
We have:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} = \frac{(\vec{x},\vec{y})}{\|\vec{x}\|\,\|\vec{y}\|} = \cos\varphi$$

It holds

$\rho_{XY} = +1$ if and only if $\vec{y} = b\vec{x}$ for some $b > 0$

that is



$$Y - \mathbf{1}\mu_Y = b(X - \mathbf{1}\mu_X)$$

$$Y - \mu_Y = b(X - \mu_X)$$

$$Y = bX + (\mu_Y - \mu_X)$$

Pearson's Correlation Coefficient

We have:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} = \frac{(\vec{x},\vec{y})}{\|\vec{x}\|\,\|\vec{y}\|} = \cos\varphi$$

It holds

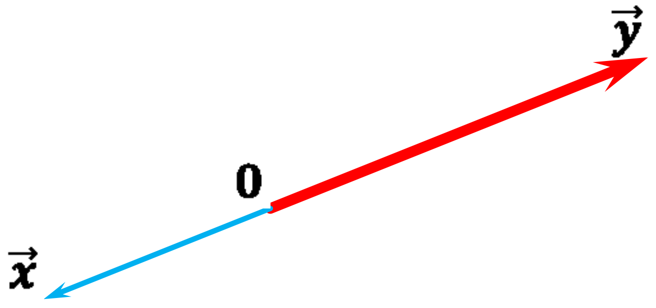
$$\rho_{XY} = -1 \qquad \text{if and only if} \qquad \vec{y} = b\vec{x} \qquad \text{for some} \quad b < 0$$

that is

$$\boldsymbol{Y} - \boldsymbol{1}\mu_Y = b(\boldsymbol{X} - \boldsymbol{1}\mu_X)$$

$$Y - \mu_Y = b(X - \mu_X)$$

$$Y = bX + (\mu_Y - \mu_X)$$



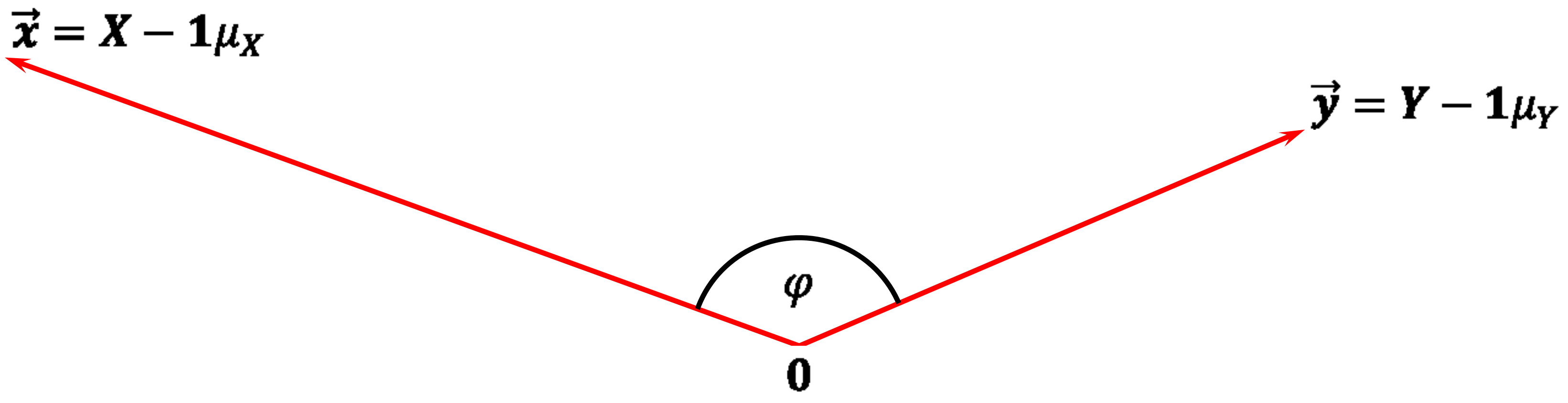
Pearson’s Correlation Coefficient

We have:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} = \frac{(\vec{x},\vec{y})}{\|\vec{x}\|\,\|\vec{y}\|} = \cos\varphi$$

It holds

$$-1 < \rho_{XY} < +1 \qquad \text{otherwise}$$



Regression Coefficient

- Regression Coefficient
- Regression Lines
- Coefficients of Regression

Regression Coefficient

Let the underlying probability space $(\Omega, \mathcal{F}, P)$ and two random variables

$$X : \Omega \to \mathbb{R} \qquad \text{and} \qquad Y : \Omega \to \mathbb{R}$$

be given.

Let us find the best (linear) approximation of the random variable Y

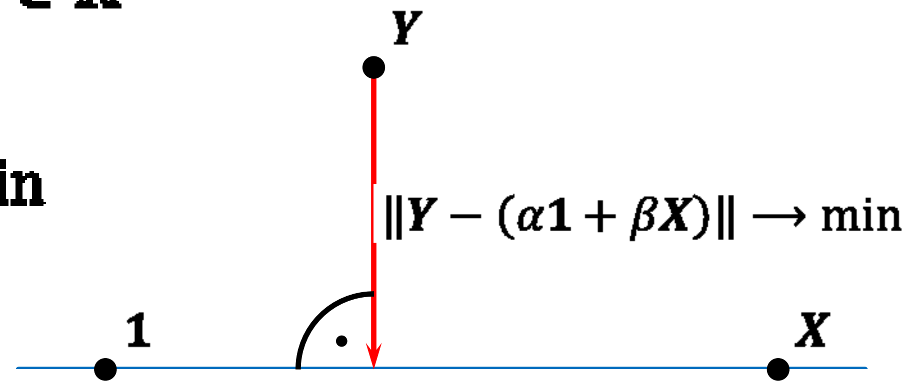
by the random variable X, that is

$$Y \approx \alpha + \beta X \qquad \text{for some} \quad \alpha, \beta \in \mathbb{R}$$

in such a way that

$$\mathrm{E}\left[\left(Y - (\alpha + \beta X)\right)^2\right] \longrightarrow \min$$

that is:



the linear subspace $\{\alpha\mathbf{1} + \beta X : \alpha, \beta \in \mathbb{R}\}$

Regression Coefficient

Denote and calculate:

$$f(\alpha,\beta) = \mathrm{E}\left[\left(Y - (\alpha + \beta X)\right)^2\right] = \mathrm{E}[Y^2 + \alpha^2 + \beta^2 X^2 - 2\alpha Y - 2\beta XY + 2\alpha\beta X] =$$

$$= \mathrm{E}[Y^2] + \alpha^2 + \beta^2 \mathrm{E}[X^2] - 2\alpha \mathrm{E}[Y] - 2\beta \mathrm{E}[XY] + 2\alpha\beta \mathrm{E}[X]$$

To find the minimum, calculate:

$$\frac{\partial f}{\partial \alpha} = 2\alpha - 2\mathrm{E}[Y] + 2\beta \mathrm{E}[X]$$

$$\frac{\partial f}{\partial \beta} = 2\beta \mathrm{E}[X^2] - 2\mathrm{E}[XY] + 2\alpha \mathrm{E}[X]$$

Regression Coefficient

We thus have:

$$2\alpha - 2\mathrm{E}[Y] + 2\beta\mathrm{E}[X] = 0$$

$$2\beta\mathrm{E}[X^2] - 2\mathrm{E}[XY] + 2\alpha\mathrm{E}[X] = 0$$

Hence

$$\alpha = \mathrm{E}[Y] - \beta\mathrm{E}[X]$$

and

$$\beta\mathrm{E}[X^2] - \mathrm{E}[XY] + (\mathrm{E}[Y] - \beta\mathrm{E}[X])\mathrm{E}[X] = 0$$

$$\beta(\mathrm{E}[X^2] - \mathrm{E}^2[X]) = \mathrm{E}[XY] - \mathrm{E}[Y]\mathrm{E}[X]$$

$$\beta\mathrm{Var}(X) = \mathrm{cov}(X, Y)$$

$$\beta_{YX} = \frac{\mathrm{cov}(X, Y)}{\mathrm{Var}(X)} \quad \text{if} \quad \mathrm{Var}(X) \neq 0$$

the regression coefficient of the random variable Y on X

Regression Coefficient

Our purpose has been to approximate the random variable Y by using the random variable X linearly ($Y \approx \alpha + \beta X$ for some $\alpha, \beta \in \mathbb{R}$ to be found).

We have found the **coefficient of regression** of Y on X as follows:

$$\beta_{YX} = \frac{\operatorname{cov}(X,Y)}{\operatorname{Var}(X)} \qquad \text{if} \quad \operatorname{Var}(X) \neq 0$$

Regression Coefficient

Similarly, we can define the coefficient of regression of X on Y:

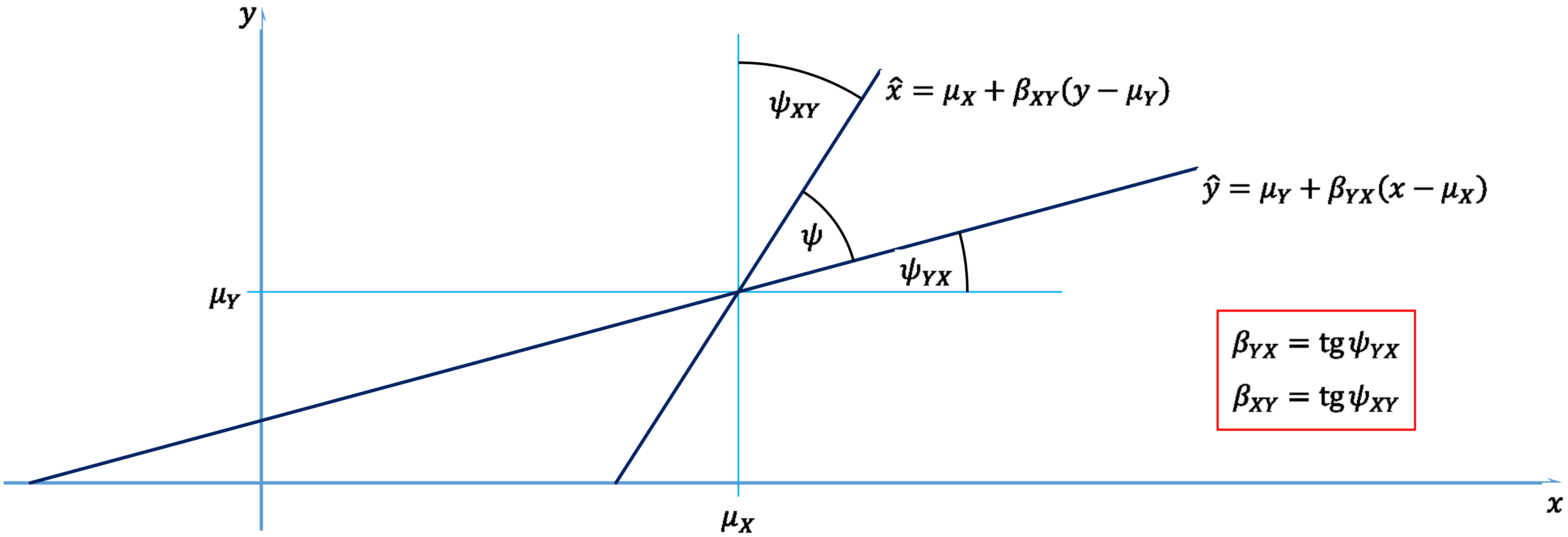
$$\beta_{XY} = \frac{\operatorname{cov}(Y,X)}{\operatorname{Var}(Y)} \qquad \text{if} \quad \operatorname{Var}(Y) \neq 0$$

We observe that

$$\beta_{YX} \times \beta_{XY} = \frac{\operatorname{cov}(X,Y)}{\operatorname{Var}(X)} \times \frac{\operatorname{cov}(Y,X)}{\operatorname{Var}(Y)} = \left(\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}\right)^2 = \rho_{XY}^2$$

The Regression Lines

$$\operatorname{tg}\psi = \operatorname{tg}\left(\frac{\pi}{2} - \psi_{XY} - \psi_{YX}\right) = \operatorname{cotg}(\psi_{XY} + \psi_{YX}) = \frac{1}{\operatorname{tg}(\psi_{XY} + \psi_{YX})} = \frac{1 - \operatorname{tg}\psi_{XY}\operatorname{tg}\psi_{YX}}{\operatorname{tg}\psi_{XY} + \operatorname{tg}\psi_{YX}} = \frac{1 - \rho_{XY}^2}{\beta_{XY} + \beta_{YX}}$$



$$\beta_{YX} = \operatorname{tg}\psi_{YX}$$

$$\beta_{XY} = \operatorname{tg}\psi_{XY}$$

Coefficients of Regression

More generally, let $n+1$ random variables

$$X_1, X_2, \dots, X_n : \Omega \to \mathbb{R} \qquad \text{and} \qquad Y : \Omega \to \mathbb{R}$$

be given.

Let us find the best (linear) approximation of the random variable Y

by the random variables $X_1, X_2, \dots, X_n,$ that is

$$Y \approx \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \qquad \text{for some} \quad \alpha, \beta_1, \beta_2, \dots, \beta_n \in \mathbb{R}$$

in such a way that

$$\mathrm{E}\left[\left(Y - (\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)\right)^2\right] \longrightarrow \min$$

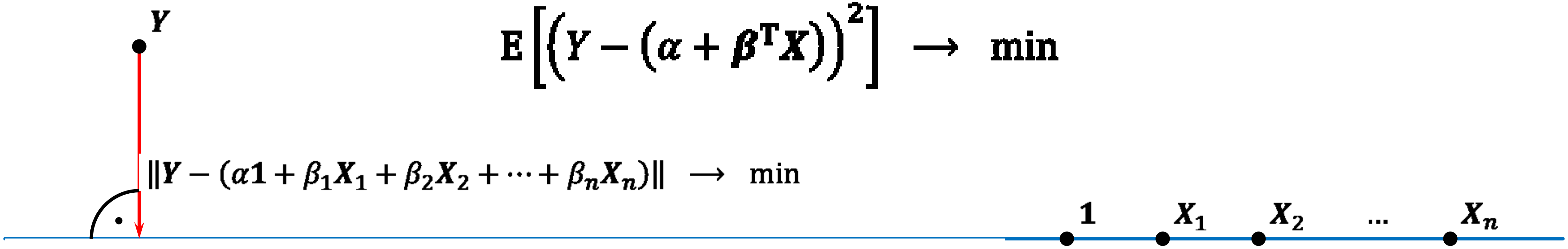
Coefficients of Regression

We stack the random variables $X_1, X_2, \ldots, X_n$ into a random vector:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

And we rewrite the problem: find $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^n$ so that

$$\mathrm{E}\left[\left(Y - (\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X})\right)^2\right] \longrightarrow \min$$



the linear subspace $\{\alpha \mathbf{1} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n : \alpha, \beta_1, \beta_2, \ldots, \beta_n \in \mathbb{R}\}$

Coefficients of Regression

Denoting

$$f(\alpha, \beta_1, \beta_2, \dots, \beta_n) = \mathrm{E}\left[\left(Y - (\alpha + \boldsymbol{\beta}\boldsymbol{X})\right)^2\right]$$

and letting

$$\frac{\partial f}{\partial \alpha} = 0 \qquad \frac{\partial f}{\partial \beta_1} = 0 \qquad \frac{\partial f}{\partial \beta_2} = 0 \qquad \dots \qquad \frac{\partial f}{\partial \beta_n} = 0$$

we obtain

$$\alpha = \mathrm{E}[Y] - \boldsymbol{\beta}_{YX}^{\mathrm{T}}\mathrm{E}[\boldsymbol{X}]$$

and

$$\boldsymbol{\beta}_{YX} = \boldsymbol{\beta}_{Y(X_1 X_2 \dots X_n)} = \left(\mathrm{Var}(\boldsymbol{X})\right)^{-1}\mathrm{cov}(\boldsymbol{X}, Y)$$

Multiple Correlation Coefficient & Coefficient of Partial Correlation

- Multiple Correlation Coefficient
- Coefficient of Partial Correlation

Multiple Correlation Coefficient

Let the underlying probability space $(\Omega, \mathcal{F}, P)$ and $n+1$ random variables

$$X_1, X_2, \dots, X_n : \Omega \to \mathbb{R} \qquad \text{and} \qquad Y : \Omega \to \mathbb{R}$$

be given, and stack the random variables $X_1, X_2, \dots, X_n$ into the random vector $\boldsymbol{X}$.

Assume that the variance-covariance matrix $\mathrm{Var}(\boldsymbol{X})$ is non-singular and calculate the regression coefficients

$$\boldsymbol{\beta}_{Y\boldsymbol{X}} = \beta_{Y(X_1 X_2 \dots X_n)} = \left(\mathrm{Var}(\boldsymbol{X})\right)^{-1} \mathrm{cov}(\boldsymbol{X}, Y) \qquad \text{and} \qquad \alpha = \mathrm{E}[Y] - \boldsymbol{\beta}_{Y\boldsymbol{X}}^{\mathrm{T}} \mathrm{E}[\boldsymbol{X}]$$

The **multiple correlation coefficient** is

$$\rho_{Y(\boldsymbol{X})} = \rho_{Y(X_1 X_2 \dots X_n)} = \rho_{Y, \alpha + \boldsymbol{\beta}_{Y\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{X}}$$

Multiple Correlation Coefficient

In other words, the **multiple correlation coefficient**

$$\rho_{Y(X)} = \rho_{Y(X_1 X_2 \dots X_n)} = \rho_{Y,\alpha+\boldsymbol{\beta}_{YX}^{\mathrm{T}}\boldsymbol{X}}$$

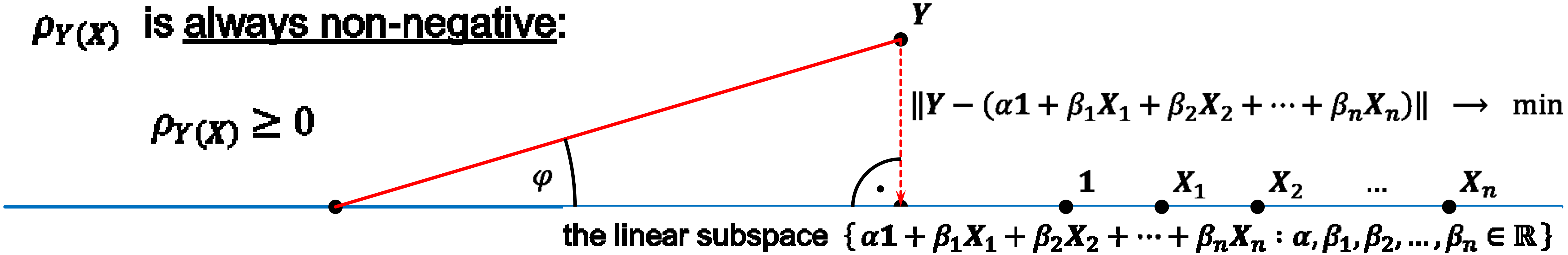
is **Pearson's Correlation Coefficient**

of the random variable Y and its best linear approximation $\alpha + \boldsymbol{\beta}_{YX}^{\mathrm{T}}\boldsymbol{X}$.

Notice that the multiple correlation coefficient

$\rho_{Y(X)}$ is <u>always non-negative</u>:

$$\rho_{Y(X)} \geq 0$$



Multiple Correlation Coefficient

In other words, the **multiple correlation coefficient**

$$\rho_{Y(X)} = \rho_{Y(X_1 X_2 \dots X_n)} = \rho_{Y,\alpha+\boldsymbol{\beta}_{YX}^{\mathrm{T}}\boldsymbol{X}}$$

is **Pearson's Correlation Coefficient**

of the random variable Y and its best linear approximation $\alpha + \boldsymbol{\beta}_{YX}^{\mathrm{T}}\boldsymbol{X}$.

Substituting and calculating, we obtain:

$$\rho_{Y(X)}^2 = \frac{\boldsymbol{\beta}_{YX}^{\mathrm{T}}(\mathrm{Var}(\boldsymbol{X}))\boldsymbol{\beta}_{YX}}{\mathrm{Var}(Y)} = \frac{\mathrm{cov}(Y,\boldsymbol{X})(\mathrm{Var}(\boldsymbol{X}))^{-1}\mathrm{cov}(\boldsymbol{X},Y)}{\mathrm{Var}(Y)}$$

Coefficient of Partial Correlation

<u>Motivation:</u>

Consider two random variables X and Y.

It happens sometimes that the random variables X and Y are highly correlated (that is ρ_{XY} is close to ±1), but there is no statistical dependence between them actually. For example:

- $X =$ the birth-rate (i.e. natality) in some region in Germany
- $Y =$ the size of the population of stork in the region

The correlation may be caused by the effect of some other factors Z behind.

Our purpose is to eliminate the effect of the factors Z (the controlling variables).

Coefficient of Partial Correlation

Let the underlying probability space $(\Omega, \mathcal{F}, P)$, the two random variables

$$X: \Omega \to \mathbb{R} \quad \text{and} \quad Y: \Omega \to \mathbb{R}$$

and a random vector

$$\boldsymbol{Z}: \Omega \to \mathbb{R}^n$$

be given.

Assuming that the variance-covariance matrix $\mathrm{Var}(\boldsymbol{Z})$ is non-singular, find the best linear approximations of X and Y based on $\boldsymbol{Z}$. That is, calculate

$$\alpha_{X\boldsymbol{Z}} = \mathrm{E}[X] - \boldsymbol{\beta}_{X\boldsymbol{Z}}^{\mathrm{T}}\mathrm{E}[\boldsymbol{Z}] \quad \text{and} \quad \boldsymbol{\beta}_{X\boldsymbol{Z}}^{\mathrm{T}} = \left(\mathrm{Var}(\boldsymbol{Z})\right)^{-1}\mathrm{cov}(\boldsymbol{Z}, X)$$

and

Coefficient of Partial Correlation

Then

$$\alpha_{XZ} + \boldsymbol{\beta}_{XZ}^{\mathrm{T}}\boldsymbol{Z} \qquad \text{and} \qquad \alpha_{YZ} + \boldsymbol{\beta}_{YZ}^{\mathrm{T}}\boldsymbol{Z}$$

is the best linear approximation of X and Y based on $\boldsymbol{Z}$, respectively.

The **Coefficient of Partial Correlation** between the random variables X and Y with the effect of the controlling random variables $\boldsymbol{Z}$ removed is

$$\rho_{XY\cdot\boldsymbol{Z}} = \rho_{X-\alpha_{XZ}-\boldsymbol{\beta}_{XZ}^{\mathrm{T}}\boldsymbol{Z},\, Y-\alpha_{YZ}-\boldsymbol{\beta}_{YZ}^{\mathrm{T}}\boldsymbol{Z}}$$

In words, it is Pearson's Correlation Coefficient between the residuals

$$X - \left(\alpha_{XZ} + \boldsymbol{\beta}_{XZ}^{\mathrm{T}}\boldsymbol{Z}\right) \qquad \text{and} \qquad Y - \left(\alpha_{YZ} + \boldsymbol{\beta}_{YZ}^{\mathrm{T}}\boldsymbol{Z}\right)$$

Coefficient of Partial Correlation

If $n = 1$, that is $Z = Z_1 = Z$, then the Coefficient of Partial Correlation between the random variables X and Y subject to a fixed value of Z takes the form

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}$$

Hypothesis Testing

- Motivation
- Pearson’s sample correlation coefficient
- Sample multiple correlation coefficient
- Sample coefficient of partial correlation

Hypothesis Testing: Motivation

Until now, we have presented the theoretical correlation coefficients:

- Pearson's correlation coefficient ρ_{XY}
- Multiple correlation coefficient $\rho_{Y(X)}$
- Partial correlation coefficient $\rho_{XY \cdot Z}$

Notice that these coefficients are defined by the intrinsic properties of the random variables X and Y (or X or Z) themselves.

That is, **no random experiments were necessary to define their values.**

Hypothesis Testing: Motivation

The true values of the coefficients ρ_{XY}, $\rho_{Y(X)}$, $\rho_{XY\cdot Z}$ do exist in theory, but are often unknown to us in practice. This is because we do not know the true functions (random variables) $X, Y: \Omega \to \mathbb{R}$, perhaps not even the underlying probability space $(\Omega, \mathcal{F}, P)$, very well.

This is the reason why we explore the properties of the random variables by the means of doing random experiments.

We wish to test the null hypotheses that

$$H_0: \quad \rho_{XY} = 0 \qquad \text{or} \qquad H_0: \quad \rho_{Y(X)} = 0 \qquad \text{or} \qquad H_0: \quad \rho_{XY\cdot Z} = 0$$

respectively.

Pearson's sample correlation coefficient

Consider the underlying probability space $(\Omega, \mathcal{F}, P)$ and the two random variables

$$X: \Omega \to \mathbb{R} \qquad \text{and} \qquad Y: \Omega \to \mathbb{R}$$

Perform the underlying random experiment n-times, where $n \geq 3$.

Let $\omega_1, \omega_2, \dots, \omega_n \in \Omega$ be the outcomes of the trials.

(It is assumed that each trial is independent of the others.)

Then, let

$$\begin{matrix} x_1 = X(\omega_1) & x_2 = X(\omega_2) & \dots & x_n = X(\omega_n) \\ y_1 = Y(\omega_1) & y_2 = Y(\omega_2) & \dots & y_n = Y(\omega_n) \end{matrix}$$

be the numerical outcomes of the trials; that is, we have n pairs

Pearson's sample correlation coefficient

Having the sample (x_1, y_1), (x_2, y_2), …, (x_n, y_n) of the observations of the random variables X, Y, we define *Pearson's sample correlation coefficient* like *Pearson's correlation coefficient*, but the sample variance and the sample covariance is used instead of the variance and the covariance, respectively.

Pearson's sample correlation coefficient is:

$$r_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

Pearson’s sample correlation coefficient

Equivalently, **Pearson’s sample correlation coefficient** is:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Pearson's sample correlation coefficient: Theorem

Let the random vector $\binom{X}{Y}$ follow a bivariate normal (Gaussian) distribution, that is

$$\binom{X}{Y} \sim \mathcal{N}\left(\binom{\mu_X}{\mu_Y}, \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right) \quad \text{with } \sigma_X^2 > 0 \text{ and } \sigma_Y^2 > 0$$

and let (x_1, y_1), (x_2, y_2), …, (x_n, y_n) be a sample of n observations of the random vector. If the null hypothesis

$$H_0: \quad \rho_{XY} = 0$$

holds true, then

$$\frac{r_{XY}}{\sqrt{1 - r_{XY}^2}}\sqrt{n-2} \sim t_{n-2}$$

Pearson's sample correlation coefficient: Hyp. Test

The null hypothesis is $H_0\colon \rho_{XY} = 0$ and the alternative hypothesis is $H_1\colon \rho_{XY} \neq 0$.

Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Calculate the statistic

$$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}}\sqrt{n-2}$$

The critical value is

$$c = t_{n-2}\left(1 - \frac{\alpha}{2}\right)$$

where $t_{n-2}(q)$ is the quantile function of Student's *t*-distribution with $n-2$ d.f.

If $T \in (-\infty, -c] \cup [+c, +\infty)$, the **critical region**, then **reject** the null hypothesis.

If $T \in (-c, +c)$, then **do not reject** (or fail to reject) the null hypothesis.

Sample Multiple Correlation Coefficient

Consider the underlying probability space $(\Omega, \mathcal{F}, P)$, the random vector

$$X : \Omega \to \mathbb{R}^k$$

and the random variable

$$Y : \Omega \to \mathbb{R}$$

where $n \geq k + 2$

Perform the underlying random experiment n-times. Let $\omega_1, \omega_2, \dots, \omega_n \in \Omega$ be the outcomes of the trials. Assume that each trial is independent of the others. Then, let

$$x_1 = X(\omega_1) \qquad x_2 = X(\omega_2) \qquad \dots \qquad x_n = X(\omega_n)$$

$$y_1 = Y(\omega_1) \qquad y_2 = Y(\omega_2) \qquad \dots \qquad y_n = Y(\omega_n)$$

be the numerical outcomes of the trials; that is, we have n pairs

Sample Multiple Correlation Coefficient

Having the sample $(x_1, y_1),\ (x_2, y_2),\ \dots,\ (x_n, y_n)$ of the observations of the random vector $X: \Omega \to \mathbb{R}^k$ and of the random variable $Y: \Omega \to \mathbb{R}$, calculate the **sample correlation vectors**

$$r_{YX} = \begin{pmatrix} r_{YX_1} & r_{YX_2} & \cdots & r_{YX_k} \end{pmatrix} \quad \text{and} \quad r_{XY} = \begin{pmatrix} r_{X_1Y} \\ r_{X_2Y} \\ \vdots \\ r_{X_kY} \end{pmatrix}$$

where

$$r_{YX_\kappa} = r_{X_\kappa Y} = \frac{\sum_{i=1}^{n}(x_{\kappa i} - \bar{x}_\kappa)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{\kappa i} - \bar{x}_\kappa)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{for} \quad \kappa = 1, 2, \dots, k$$

is Pearson’s sample correlation coefficient

Sample Multiple Correlation Coefficient

Having the sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the observations of the random vector $\boldsymbol{X}: \Omega \to \mathbb{R}^k$, calculate also the sample correlation matrix

$$\boldsymbol{R}_{\boldsymbol{XX}} = \begin{pmatrix} r_{X_1X_1} & r_{X_1X_2} & \cdots & r_{X_1X_k} \\ r_{X_2X_1} & r_{X_2X_2} & \cdots & r_{X_2X_k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_kX_1} & r_{X_kX_2} & \cdots & r_{X_kX_k} \end{pmatrix}$$

where

$$r_{X_pX_q} = \frac{\sum_{i=1}^{n}(x_{pi} - \bar{x}_p)(x_{qi} - \bar{x}_q)}{\sqrt{\sum_{i=1}^{n}(x_{pi} - \bar{x}_p)^2}\sqrt{\sum_{i=1}^{n}(x_{qi} - \bar{x}_q)^2}} \qquad \text{for} \quad p, q = 1, 2, \ldots, k$$

is Pearson's sample correlation coefficient

Sample Multiple Correlation Coefficient

If the sample correlation matrix $\boldsymbol{R_{XX}}$ is non-singular,

the **sample multiple correlation coefficient** squared is

$$r^2_{Y(X)} = \boldsymbol{r_{YX}} \times \boldsymbol{R^{-1}_{XX}} \times \boldsymbol{r_{XY}}$$

Remark: We know that the multiple correlation coefficient is always non-negative.

So we can define the **sample multiple correlation coefficient** as

$$r_{Y(X)} = \sqrt{r^2_{Y(X)}}$$

Sample Multiple Correlation Coefficient: Theorem

Let the random vector $\binom{X}{Y}$ follow a $(k+1)$-dimensional normal (Gaussian) distribution, that is

$$\binom{X}{Y} \sim \mathcal{N}\left(\binom{\mu_X}{\mu_Y}, \begin{pmatrix} \mathrm{Var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{Var}(Y) \end{pmatrix}\right) \quad \text{with} \quad \begin{pmatrix} \mathrm{Var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{Var}(Y) \end{pmatrix}$$

non-singular.

Let (x_1, y_1), (x_2, y_2), …, (x_n, y_n) be a sample of n observations of the random vector. If the null hypothesis

$$H_0: \quad \rho_{X(Y)} = 0$$

holds true, then

$$\frac{r^2_{Y(X)}}{1 - r^2_{Y(X)}} \Big/ \frac{k}{n-k-1} \sim F_{k,n-k-1}$$

Sample Multiple Correlation Coefficient: Hyp. Test

The null hypothesis is $H_0\colon \rho_{X(Y)} = 0$ and the alternative hypothesis is $H_1\colon \rho_{X(Y)} \neq 0$. Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$. Calculate the statistic

$$F = \frac{r^2_{Y(X)}}{1 - r^2_{Y(X)}} \Big/ \frac{k}{n-k-1}$$

The critical value is

$$c = F_{k,n-k-1}(1-\alpha)$$

with k and $n-k-1$ d.f.

where $F_{k,n-k-1}(q)$ is the quantile function of Fisher's *F*-distribution

If $F \in (-\infty, -c] \cup [+c, +\infty)$, the **critical region**, then **reject** the null hypothesis.

If $F \in (-c, +c)$, then **do not reject** (or fail to reject) the null hypothesis.

Sample Coefficient of Partial Correlation

Consider the underlying probability space $(\Omega, \mathcal{F}, P)$, the two random variables

$$X: \Omega \to \mathbb{R} \quad \text{and} \quad Y: \Omega \to \mathbb{R}$$

and the random vector

$$Z: \Omega \to \mathbb{R}^k$$

where $n \geq k + 3$

Perform the underlying random experiment n-times. Let $\omega_1, \omega_2, \dots, \omega_n \in \Omega$ be the outcomes of the trials. Assume that each trial is independent of the others. Then, let

$$\begin{array}{cccc} x_1 = X(\omega_1) & x_2 = X(\omega_2) & \dots & x_n = X(\omega_n) \\ y_1 = Y(\omega_1) & y_2 = Y(\omega_2) & \dots & y_n = Y(\omega_n) \\ z_1 = Z(\omega_1) & z_2 = Z(\omega_2) & \dots & z_n = Z(\omega_n) \end{array}$$

Sample Coefficient of Partial Correlation

That is, we have n triples of the observations of the random variables and vector:

$$(x_1, y_1, z_1) \qquad (x_2, y_2, z_2) \qquad \dots \qquad (x_n, y_n, z_n)$$

Then, calculate the **sample correlation vectors**

$$r_{XZ} = \begin{pmatrix} r_{XZ_1} & r_{XZ_2} & \cdots & r_{XZ_k} \end{pmatrix} \qquad \text{and} \qquad r_{ZX} = \begin{pmatrix} r_{Z_1X} \\ r_{Z_2X} \\ \vdots \\ r_{Z_kX} \end{pmatrix}$$

where

$$r_{XZ_\kappa} = r_{Z_\kappa X} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_{\kappa i} - \bar{z}_\kappa)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(z_{\kappa i} - \bar{z}_\kappa)^2}} \qquad \text{for} \quad \kappa = 1, 2, \dots, k$$

is Pearson's sample correlation coefficient

Sample Coefficient of Partial Correlation

Having the n triples of the observations of the random variables and vector

$$(x_1, y_1, z_1) \qquad (x_2, y_2, z_2) \qquad \dots \qquad (x_n, y_n, z_n)$$

calculate also the **sample correlation vectors**

$$r_{YZ} = \begin{pmatrix} r_{YZ_1} & r_{YZ_2} & \dots & r_{YZ_k} \end{pmatrix} \quad \text{and} \quad r_{ZY} = \begin{pmatrix} r_{Z_1Y} \\ r_{Z_2Y} \\ \vdots \\ r_{Z_kY} \end{pmatrix}$$

where

$$r_{YZ_\kappa} = r_{Z_\kappa Y} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(z_{\kappa i} - \bar{z}_\kappa)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(z_{\kappa i} - \bar{z}_\kappa)^2}} \qquad \text{for} \quad \kappa = 1, 2, \dots, k$$

is Pearson's sample correlation coefficient

Sample Coefficient of Partial Correlation

And, having the sample $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ of the observations of the random vector $\boldsymbol{Z}: \Omega \to \mathbb{R}^k$, calculate the **sample correlation matrix**

$$\boldsymbol{R}_{ZZ} = \begin{pmatrix} r_{Z_1 Z_1} & r_{Z_1 Z_2} & \cdots & r_{Z_1 Z_k} \\ r_{Z_2 Z_1} & r_{Z_2 Z_2} & \cdots & r_{Z_2 Z_k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{Z_k Z_1} & r_{Z_k Z_2} & \cdots & r_{Z_k Z_k} \end{pmatrix}$$

where

$$r_{z_p z_q} = \frac{\sum_{i=1}^{n} (z_{pi} - \bar{z}_p)(z_{qi} - \bar{z}_q)}{\sqrt{\sum_{i=1}^{n} (z_{pi} - \bar{z}_p)^2} \sqrt{\sum_{i=1}^{n} (z_{qi} - \bar{z}_q)^2}} \qquad \text{for} \quad p, q = 1, 2, \ldots, k$$

is Pearson's sample correlation coefficient

Sample Coefficient of Partial Correlation

If the sample correlation matrix $\boldsymbol{R}_{ZZ}$ is non-singular,

the sample coefficient of partial correlation is

$$r_{XY\cdot Z} = \frac{r_{XY} - \boldsymbol{r}_{XZ}\boldsymbol{R}_{ZZ}^{-1}\boldsymbol{r}_{ZY}}{\sqrt{1 - \boldsymbol{r}_{XZ}\boldsymbol{R}_{ZZ}^{-1}\boldsymbol{r}_{ZX}}\sqrt{1 - \boldsymbol{r}_{YZ}\boldsymbol{R}_{ZZ}^{-1}\boldsymbol{r}_{ZY}}}$$

Sample Coefficient of Partial Correlation

If $k = 1$, that is $Z = Z_1 = Z$, then the **sample coefficient of partial correlation** takes the form

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}}$$

Sample Coefficient of Partial Correlation: Theorem…

Let the random vector $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ follow a $(k+2)$-dimensional normal (Gaussian) distribution, that is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \mathrm{Var}(X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z) \\ \mathrm{cov}(Y,X) & \mathrm{Var}(Y) & \mathrm{cov}(Y,Z) \\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{Var}(Z) \end{pmatrix}\right)$$

with

$$\begin{pmatrix} \mathrm{Var}(X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z) \\ \mathrm{cov}(Y,X) & \mathrm{Var}(Y) & \mathrm{cov}(Y,Z) \\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{Var}(Z) \end{pmatrix} \quad \text{being non–singular}$$

Sample Coefficient of Partial Correlation: …Theorem

Let the random vector $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ follow a $(k+2)$-dimensional normal (Gaussian) distribution with a non-singular variance-covariance matrix.

Let (x_1, y_1, z_1), (x_2, y_2, z_2), …, (x_n, y_n, z_n) be a sample of n observations of the random vector. If the null hypothesis

$$H_0: \quad \rho_{XY \cdot Z} = 0$$

holds true, then

$$\frac{r_{XY \cdot Z}}{\sqrt{1 - r_{XY \cdot Z}^2}} \sqrt{n-k-2} \quad \sim \quad t_{n-k-2}$$

Sample Coefficient of Partial Correlation: Hyp. Test

The null hypothesis is $H_0\colon \rho_{XY \cdot Z} = 0$ and the alt. hypothesis is $H_1\colon \rho_{XY \cdot Z} \neq 0$.

Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Calculate the statistic

$$T = \frac{r_{XY \cdot Z}}{\sqrt{1 - r_{XY \cdot Z}^2}} \sqrt{n - k - 2}$$

The critical value is

$$c = t_{n-k-2}\left(1 - \frac{\alpha}{2}\right)$$

where $t_{n-k-2}(q)$ is the quantile function of Student's t-distrib. with $n - k - 2$ d.f.

If $T \in (-\infty, -c] \cup [+c, +\infty)$, the **critical region**, then **reject** the null hypothesis.

Non-parametric and robust methods: Spearman’s Rank Correlation Coefficient

- Introduction
- Definition
- Simplification
- Theorems & Hypothesis Testing
- Remarks

Spearman's rank correlation coefficient: Introduction

Given the underlying probability space $(\Omega, \mathcal{F}, P)$ and two random variables

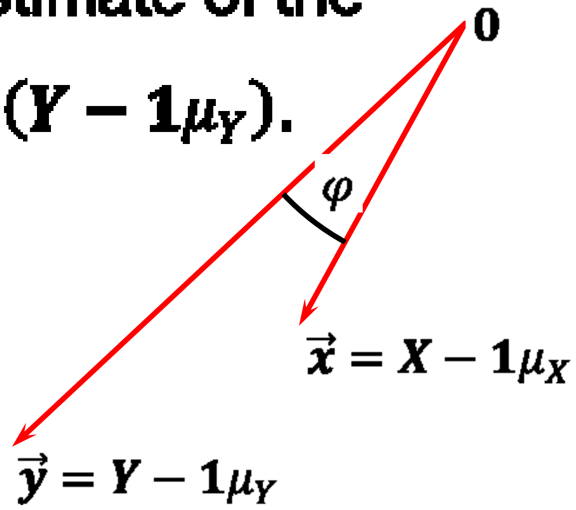
$$X: \Omega \to \mathbb{R} \qquad \text{and} \qquad Y: \Omega \to \mathbb{R}$$

and asking whether there is a (linear) correlation between the variables X and Y:

- We can use Pearson's sample correlation coefficient r_{XY} under the assumption that the vector $\binom{X}{Y}$ follows a bivariate normal (Gaussian) distribution with $\sigma_X^2 > 0$ and $\sigma_Y^2 > 0$.
- If we are not sure whether the distribution of the random vector $\binom{X}{Y}$ is normal, then the use of Pearson's sample correlation coefficient is questionable,

Spearman's rank correlation coefficient: Introduction

Recall also that Pearson's sample correlation coefficient is an estimate of the cosine of the angle between the vectors $\vec{x} = (X - \mathbf{1}\mu_X)$ and $\vec{y} = (Y - \mathbf{1}\mu_Y)$.



Therefore, Pearson's (sample) correlation coefficient can detect only the linear dependence between the variables X and Y.

The idea behind Spearman's rank correlation coefficient is different.

Spearman's rank correlation coefficient: Introduction

- Spearman's rank correlation coefficient detects whether there is any monotonic dependence between the variables X and Y, which means:
  if one variable increases, then the other variable increases too (linearly or not);
  if one variable decreases, then the other variable decreases too (linearly or not).

- That is, Spearman's rank correlation coefficient is more general than Pearson's sample correlation coefficient (in the above sense).

Spearman's rank correlation coefficient

Let the underlying probability space $(\Omega, \mathcal{F}, P)$ and random variables

$$X_1, X_2, \dots, X_n : \Omega \to \mathbb{R} \qquad \text{and} \qquad Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$$

be given. We test the following **null hypothesis** $\boldsymbol{H_0}$:

- the random variables $X_1, X_2, \dots, X_n$ are mutually independent and their cumulative distribution functions are all the same $F: \mathbb{R} \to \mathbb{R}$, and
- the random variables $Y_1, Y_2, \dots, Y_n$ are mutually independent and their cumulative distribution functions are all the same $G: \mathbb{R} \to \mathbb{R}$, and
- the random vectors $(X_1, X_2, \dots, X_n)^{\mathrm{T}}$ and $(Y_1, Y_2, \dots, Y_n)^{\mathrm{T}}$ are mutually independent.

Spearman's rank correlation coefficient

Given the random variables $X_1, X_2, \dots, X_n : \Omega \to \mathbb{R}$, define new random variables

$$R_1, R_2, \dots, R_n : \Omega \to \mathbb{N}$$

as follows:

$$R_i(\omega) = \left|\left\{ j \in \{1, 2, \dots, n\} : X_j(\omega) \leq X_i(\omega) \right\}\right| \qquad \text{for} \quad i = 1, 2, \dots, n$$

and $\omega \in \Omega$

The variable R_i is the **rank** of the variable X_i.

It is the number of the variables X_j that are less than or equal to X_i (at $\omega \in \Omega$).

Spearman's rank correlation coefficient

Given the random variables $Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$, define new random variables

$$Q_1, Q_2, \dots, Q_n : \Omega \to \mathbb{N}$$

as follows:

$$Q_i(\omega) = \left|\{ j \in \{1, 2, \dots, n\} : Y_j(\omega) \le Y_i(\omega) \}\right| \qquad \text{for} \quad i = 1, 2, \dots, n \quad \text{and} \quad \omega \in \Omega$$

The variable Q_i is the **rank** of the variable Y_i.

It is the number of the variables Y_j that are less than or equal to Y_i (at $\omega \in \Omega$).

Spearman's rank correlation coefficient

Given the underlying probability space $(\Omega, \mathcal{F}, P)$, the random variables

$$X_1, X_2, \dots, X_n\colon \Omega \to \mathbb{R} \qquad \text{and} \qquad Y_1, Y_2, \dots, Y_n\colon \Omega \to \mathbb{R}$$

with their ranks

$$R_1, R_2, \dots, R_n\colon \Omega \to \mathbb{R} \qquad \text{and} \qquad Q_1, Q_2, \dots, Q_n\colon \Omega \to \mathbb{R}$$

perform the underlying random experiment.

Let $\omega \in \Omega$ be the outcome of the random experiment.

Then the ranks $R_1, R_2, \dots, R_n$ and $Q_1, Q_2, \dots, Q_n$ can be seen as random variables on the set $\Omega' = \{1, 2, \dots, n\}$:

$$R(\omega)\colon \{1, 2, \dots, n\} \longrightarrow \mathbb{R} \qquad \text{and} \qquad Q(\omega)\colon \{1, 2, \dots, n\} \longrightarrow \mathbb{R}$$

Spearman's rank correlation coefficient

Then **Spearman's rank correlation coefficient** (at $\omega \in \Omega$) is simply

Pearson's correlation coefficient of the new random variables

$$R(\omega)\colon \{1,2,\dots,n\} \longrightarrow \mathbb{R} \qquad \text{and} \qquad Q(\omega)\colon \{1,2,\dots,n\} \longrightarrow \mathbb{R}$$

Then **Spearman's rank correlation coefficient** (at $\omega \in \Omega$) is

$$\rho(\omega) = \rho_{R(\omega),Q(\omega)} = \frac{\operatorname{cov}\big(R(\omega),Q(\omega)\big)}{\sqrt{\operatorname{Var}\big(R(\omega)\big)}\sqrt{\operatorname{Var}\big(Q(\omega)\big)}}$$

Spearman's rank correlation coefficient

In other words, **Spearman's rank correlation coefficient** (at $\omega \in \Omega$) is

$$\rho(\omega) = \rho_{R(\omega),Q(\omega)} = \frac{\operatorname{cov}\big(R(\omega), Q(\omega)\big)}{\sqrt{\operatorname{Var}\big(R(\omega)\big)}\sqrt{\operatorname{Var}\big(Q(\omega)\big)}} =$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(R_i(\omega) - \mathrm{E}[R(\omega)])(Q_i(\omega) - \mathrm{E}[Q(\omega)])}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_i(\omega) - \mathrm{E}[R(\omega)])^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i(\omega) - \mathrm{E}[Q(\omega)])^2}}$$

Spearman's rank correlation coefficient: Simplification

From now on, **assume for simplicity** that **the random variables**

$X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_n$ **are continuous** — that is,

their cumulative distribution functions $F: \mathbb{R} \to \mathbb{R}$ and $G: \mathbb{R} \to \mathbb{R}$ are continuous.

Then the probability that the values of the random variables $X_1, X_2, \dots, X_n$ and

those of the random variables $Y_1, Y_2, \dots, Y_n$ are pairwise distinct is equal to one:

$$P(\{\omega \in \Omega : X_i(\omega) \neq X_j(\omega) \text{ if } i \neq j \quad \text{for } i, j = 1, 2, \dots, n\}) = 1$$

$$P(\{\omega \in \Omega : Y_i(\omega) \neq Y_j(\omega) \text{ if } i \neq j \quad \text{for } i, j = 1, 2, \dots, n\}) = 1$$

Spearman's rank correlation coefficient: Simplification

That is, we may assume (further) for simplicity that $\omega \in \Omega$ is such that

- the values $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ are pairwise distinct and
- the values $Y_1(\omega), Y_2(\omega), \ldots, Y_n(\omega)$ are pairwise distinct.

Recall that $R_i(\omega)$ or $Q_i(\omega)$ is the rank of the value $X_i(\omega)$ or $Y_i(\omega)$, respectively, that is the number of values $X_j(\omega)$ or $Y_j(\omega)$ that are less than $X_i(\omega)$ or $Y_i(\omega)$, respectively.

Observe then that $R_1(\omega), R_2(\omega), \ldots, R_n(\omega)$ as well as $Q_1(\omega), Q_2(\omega), \ldots, Q_n(\omega)$ are **permutations** of $1, 2, \ldots, n$.

Spearman's rank correlation coefficient: Simplification

Then, if $R_1(\omega), R_2(\omega), \ldots, R_n(\omega)$ as well as $Q_1(\omega), Q_2(\omega), \ldots, Q_n(\omega)$ are permutations of $1, 2, \ldots, n,$ we can simplify the formula for Spearman's rank correlation coefficient:

$$\rho(\omega) = \rho_{R(\omega),Q(\omega)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_i(\omega) - \mathrm{E}[R(\omega)])(Q_i(\omega) - \mathrm{E}[Q(\omega)])}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_i(\omega) - \mathrm{E}[R(\omega)])^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i(\omega) - \mathrm{E}[Q(\omega)])^2}}$$

Firstly,

$$\mathrm{E}[R(\omega)] = \frac{1}{n}\sum_{k=1}^{n} k = \frac{1}{n}\frac{n^2+n}{2} = \frac{n+1}{2} \qquad \text{and analogously} \quad \mathrm{E}[Q(\omega)] = \frac{n+1}{2}$$

Spearman's rank correlation coefficient: Simplification

Secondly,

$$\mathrm{Var}(R(\omega)) = \frac{1}{n}\sum_{k=1}^{n}(k - \mathrm{E}[R(\omega)])^2 = \frac{1}{n}\left(\sum_{k=1}^{n}k^2 - 2\sum_{k=1}^{n}k\frac{n+1}{2} + \sum_{k=1}^{n}\left(\frac{n+1}{2}\right)^2\right) =$$

$$= \frac{1}{n}\left(\frac{2n^3+3n^2+n}{6} - 2\frac{n^2+n}{2}\frac{n+1}{2} + n\frac{(n+1)^2}{4}\right) =$$

$$= \frac{1}{n}\left(\frac{2n^3+3n^2+n}{6} - n\frac{(n+1)^2}{4}\right) =$$

$$= \frac{1}{n}\frac{4n^3+6n^2+2n-3n^3-6n^2-3n}{12} = \frac{n^2-1}{12}$$

Spearman's rank correlation coefficient: Simplification

Finally,

$$\rho(\omega) = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_i(\omega) - \mathrm{E}[R(\omega)])(Q_i(\omega) - \mathrm{E}[Q(\omega)])}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_i(\omega) - \mathrm{E}[R(\omega)])^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i(\omega) - \mathrm{E}[Q(\omega)])^2}} =$$

$$= \frac{1}{n}\sqrt{\frac{12}{n^2-1}}\sqrt{\frac{12}{n^2-1}}\sum_{i=1}^{n}\left(R_i(\omega) - \frac{n+1}{2}\right)\left(Q_i(\omega) - \frac{n+1}{2}\right) =$$

$$= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \sum_{k=1}^{n} k\frac{n+1}{2} - \sum_{k=1}^{n}\frac{n+1}{2}k + \sum_{k=1}^{n}\left(\frac{n+1}{2}\right)^2\right) =$$

Spearman’s rank correlation coefficient: Simplification

Finally, $\rho(\omega) =$

$$= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \sum_{k=1}^{n} k\frac{n+1}{2} - \sum_{k=1}^{n}\frac{n+1}{2}k + \sum_{k=1}^{n}\left(\frac{n+1}{2}\right)^2\right) =$$

$$= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - 2\frac{n^2+n}{2}\frac{n+1}{2} + n\frac{(n+1)^2}{4}\right) =$$

$$= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \frac{n^3+2n^2+n}{4}\right) =$$

Spearman's rank correlation coefficient: Simplification

Finally, $\rho(\omega) =$

$$= \frac{12}{n(n^2-1)}\left(\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \frac{n^3+2n^2+n}{4}\right) =$$

$$= \frac{6}{n^3-n}\left(2\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \frac{3n^3+6n^2+3n}{6}\right) =$$

$$= \frac{6}{n^3-n}\left(-\frac{2n^3+3n^2+n}{6} + 2\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \frac{2n^3+3n^2+n}{6} + \frac{n^3+n}{6}\right) =$$

Spearman's rank correlation coefficient: Simplification

Finally, $\rho(\omega) =$

$$= \frac{6}{n^3 - n}\left(-\frac{2n^3 + 3n^2 + n}{6} + 2\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \frac{2n^3 + 3n^2 + n}{6} + \frac{n^3 + n}{6}\right) =$$

$$= \frac{6}{n^3 - n}\left(-\sum_{k=1}^{n} k^2 + 2\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \sum_{k=1}^{n} k^2 + \frac{n^3 + n}{6}\right) =$$

$$= \frac{6}{n^3 - n}\left(-\sum_{i=1}^{n} \left(R_i(\omega)\right)^2 + 2\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \sum_{i=1}^{n} \left(Q_i(\omega)\right)^2 + \frac{n^3 + n}{6}\right) =$$

Spearman's rank correlation coefficient: Simplification

Finally, $\rho(\omega) =$

$$= \frac{6}{n^3 - n}\left(-\sum_{i=1}^{n}\left(R_i(\omega)\right)^2 + 2\sum_{i=1}^{n} R_i(\omega)Q_i(\omega) - \sum_{i=1}^{n}\left(Q_i(\omega)\right)^2 + \frac{n^3 + n}{6}\right) =$$

$$= 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(R_i(\omega) - Q_i(\omega)\right)^2$$

Spearman's rank correlation coefficient

In practice, we often do not know the underlying probability space $(\Omega, \mathcal{F}, P)$.

This is the reason why we shall omit the symbol "ω" $(\omega \in \Omega)$ from now on.

In practice, we only have the numerical outcomes

$$x_1 = X_1(\omega) \qquad x_2 = X_2(\omega) \qquad \dots \qquad x_n = X_n(\omega)$$

$$y_1 = Y_1(\omega) \qquad y_2 = Y_2(\omega) \qquad \dots \qquad y_n = Y_n(\omega)$$

of the random experiment.

Moreover, we do assume that the values are pairwise distinct:

$$x_i \neq x_j \quad \text{if} \quad i \neq j \quad \text{for} \quad i, j = 1, 2, \dots, n$$

$$y_i \neq y_j \quad \text{if} \quad i \neq j \quad \text{for} \quad i, j = 1, 2, \dots, n$$

Spearman's rank correlation coefficient

We then calculate the ranks:

$$R_i = \left|\{j \in \{1, 2, \ldots, n\} : x_j < x_i\}\right| \qquad \text{for} \quad i = 1, 2, \ldots, n$$

and

$$Q_i = \left|\{j \in \{1, 2, \ldots, n\} : y_j < y_i\}\right| \qquad \text{for} \quad i = 1, 2, \ldots, n$$

And calculate **Spearman's rank correlation coefficient:**

$$r_s = \rho = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_i - Q_i)^2 = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where

Spearman's rank correlation coefficient: Hyp. testing

Assume that the null hypothesis H_0 (the random variables $X_1, X_2, \dots, X_n$ have the same (continuous) cumulative distributive function, the random variables $Y_1, Y_2, \dots, Y_n$ also have the same (continuous) cumulative distributive function, and the variables $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_n$ are mutually independent) **holds true.**

¿ What is the distribution of Spearman's rank correlation coefficient $r_s = \rho$?

→ If H_0 holds true, then every permutation $Q_1, Q_2, \dots, Q_n$ of the numbers $1, 2, \dots, n$ is equally probable.

Spearman's rank correlation coefficient: Hyp. testing

We may assume without loss of generality that $R_1 = 1,\ R_2 = 2,\ \ldots,\ R_n = n.$

Then, assuming that every permutation $Q_1, Q_2, \ldots, Q_n$ of the numbers $1, 2, \ldots, n$ is equally probable (which is true if H_0 holds), we shall evaluate the expression

$$\rho = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (i - Q_i)^2$$

over all the permutations $Q_1, Q_2, \ldots, Q_n$ of the numbers $1, 2, \ldots, n.$

We get the values ρ and their probabilities. If H_0 holds true, then large values of $|\rho|$ are improbable. That is, if $|\rho| \geq c,$ the critical value, we reject the null hyp.

Spearman's rank correlation coefficient: Hyp. testing

The above procedure (to evaluate $\rho = 1 - 6\sum_{i=1}^{n}(i - Q_i)^2/(n^3 - n)$ over all the permutations) is practically hardly feasible.

Special statistical tables of the critical values (for $n \leq 30$) exists.

Or, we can use approximation:

<u>Theorem:</u> If the null hypothesis H_0 holds true and n is large, then

$$Z = r_s\sqrt{n-1} \sim \mathcal{N}(0,1) \qquad \textit{approximately}$$

Spearman's rank correlation coefficient: Theorems

Theorem: If the null hypothesis H_0 holds true and n is large, then

$$Z = r_s\sqrt{n-1} \sim \mathcal{N}(0,1) \qquad \textit{approximately}$$

Another Theorem: If the null hypothesis H_0 holds true and n is large, then

$$T = \frac{r_s}{\sqrt{1-r_s^2}}\sqrt{n-2} \sim t_{n-2} \qquad \textit{approximately}$$

Spearman's rank correlation coefficient: Hyp. test

- The null hypothesis is H_0 that the values in the pairs (x_1,y_1), (x_2,y_2), …, (x_n,y_n) of the sample are monotonically independent.
- The alternative hypothesis H_1 is $\neg H_0$ (two-sided).
  (One-sided alternative hypotheses can also be considered.)
- Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.
- Calculate the ranks

$$R_i = \left|\{j \in \{1,2,\dots,n\} : x_j < x_i\}\right| \qquad \text{for} \quad i = 1,2,\dots,n$$

and

$$Q_i = \left|\{j \in \{1,2,\dots,n\} : y_j < y_i\}\right| \qquad \text{for} \quad i = 1,2,\dots,n$$

Spearman's rank correlation coefficient: Hyp. test

- Calculate Spearman's rank correlation coefficient:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n-1)} = 1 - \frac{6}{n(n-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

- Calculate the statistic

$$T = \frac{r_s}{\sqrt{1-r_s^2}}\sqrt{n-2} \quad \text{or} \quad Z = r_s\sqrt{n-1}$$

Spearman's rank correlation coefficient: Hyp. test

- The critical value is

$$c = t_{n-2}\left(1-\frac{\alpha}{2}\right) \quad \text{or} \quad c = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$

where $t_{n-2}(q)$ or $\Phi^{-1}(q)$ is the quantile function of Student's *t*-distribution with $n-2$ degrees of freedom or the quantile function of the normalized normal distribution, respectively.

- If $T \text{ or } Z \in (-\infty, -c] \cup [+c, +\infty)$,

  the **critical region**, then **reject** the null hypothesis.

- If $T \text{ or } Z \in (-c, +c)$, then **do not reject** (or fail to reject) the null hypothesis.

Spearman's rank correlation coefficient: Remarks

The statistical test by using Spearman's rank correlation coefficient is suitable whenever

- we cannot assume the normal distribution of the observed random variables X and Y $(X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n)$;
- the sample is small (the number n is small)
  — special statistical tables are necessary then;
- the random variables X and Y are not numerical (quantitative), but take on qualitative values from linearly ordered scales S_X and S_Y (possibly $S_X = S_Y$), that is $X_1, X_2, \ldots, X_n : \Omega \longrightarrow S_X$ and $Y_1, Y_2, \ldots, Y_n : \Omega \longrightarrow S_Y$