University of BRISTOL

Applied Statistics

Lectures 13

David Barton & Sabine Hauert

Department of Engineering Mathematics

Outline

- Sample covariance
- Correlation and sample correlation
- Hypothesis testing for correlations

OpenIntro Statistics

Chapter 7

Dependent random variables

random variables

$X, Y \sim N\left(\begin{matrix}\mu_x \\ \mu_y\end{matrix}, \Sigma\right)$

The covariance matrix Σ is defined as

covariance term
when 0 → no link

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}$$

where

$$\sigma_X^2 = E\left[(X - \mu_X)^2\right]$$
$$\sigma_Y^2 = E\left[(Y - \mu_Y)^2\right]$$
$$\sigma_{X,Y} = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

population level quantities
→ tells us the link between our data

It tells us about the links (at a linear level) between two (or more) random variables

Dependent random variables

Often this is written in vector form

vector

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \qquad \begin{bmatrix} X_1 \\ \vdots \\ X_{10} \end{bmatrix}$$

→ many variables

with

vector of all random variable

row vector

$$\Sigma = E\left[(Z - \mu_Z)(Z - \mu_Z)^T\right]$$

covariance matrix

column vector

vector of the means

which simplifies down to the previous expression ($\mu_Z = E[Z]$).

The covariance matrix tells us the link between the random variables at a *linear level* — more on this later.

covariance of 0 means there is no linear link

The covariance matrix has some nice properties; it is

- symmetric: $\Sigma = \Sigma^T$, and it is ⇒ eigenvalues are real and ≥ 0
- positive semi-definite: all it's eigenvalues $\lambda \geqslant 0$.

Because of the symmetry, it's eigenvalues are also real.

not examined on this bit cool

Sampling — covariance and correlation

The mean and covariance is used a lot even when the distribution is non-normal — generating a full joint distribution requires a lot of data!

The sample variances s_X^2 and s_Y^2 are estimated in the standard way from n samples

sample variance

bessel's correction → wikipedia

mean

Need to build a model based on data

$$s_X^2 = \boxed{\frac{1}{n-1}} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

value

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{y})^2$$

estimate of population values

Following this pattern, the *sample covariance* $q_{X,Y}$ is

sample covariance

$$q_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})(Y_i - \bar{y})$$

Sampling — covariance and correlation

Hence the *sample covariance matrix* from n samples is given by

$$Q = \frac{1}{n-1}\sum_{i=1}^{n}\begin{bmatrix}(X_i-\bar{x})^2 & (X_i-\bar{x})(Y_i-\bar{y})\\(X_i-\bar{x})(Y_i-\bar{y}) & (Y_i-\bar{y})^2\end{bmatrix}$$

More generally, the sample covariance matrix is given by

$$Q = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{u}_i-\bar{\mathbf{u}})(\mathbf{u}_i-\bar{\mathbf{u}})^T$$

single line in matlab!

where $\mathbf{u}_i \in \mathbb{R}^m$ is a column vector with m elements (one element for each of the m measured random variables). (Above $\mathbf{u}_i = [X_i, Y_i]^T$.)

not the population covariance matrix
→ generated from data

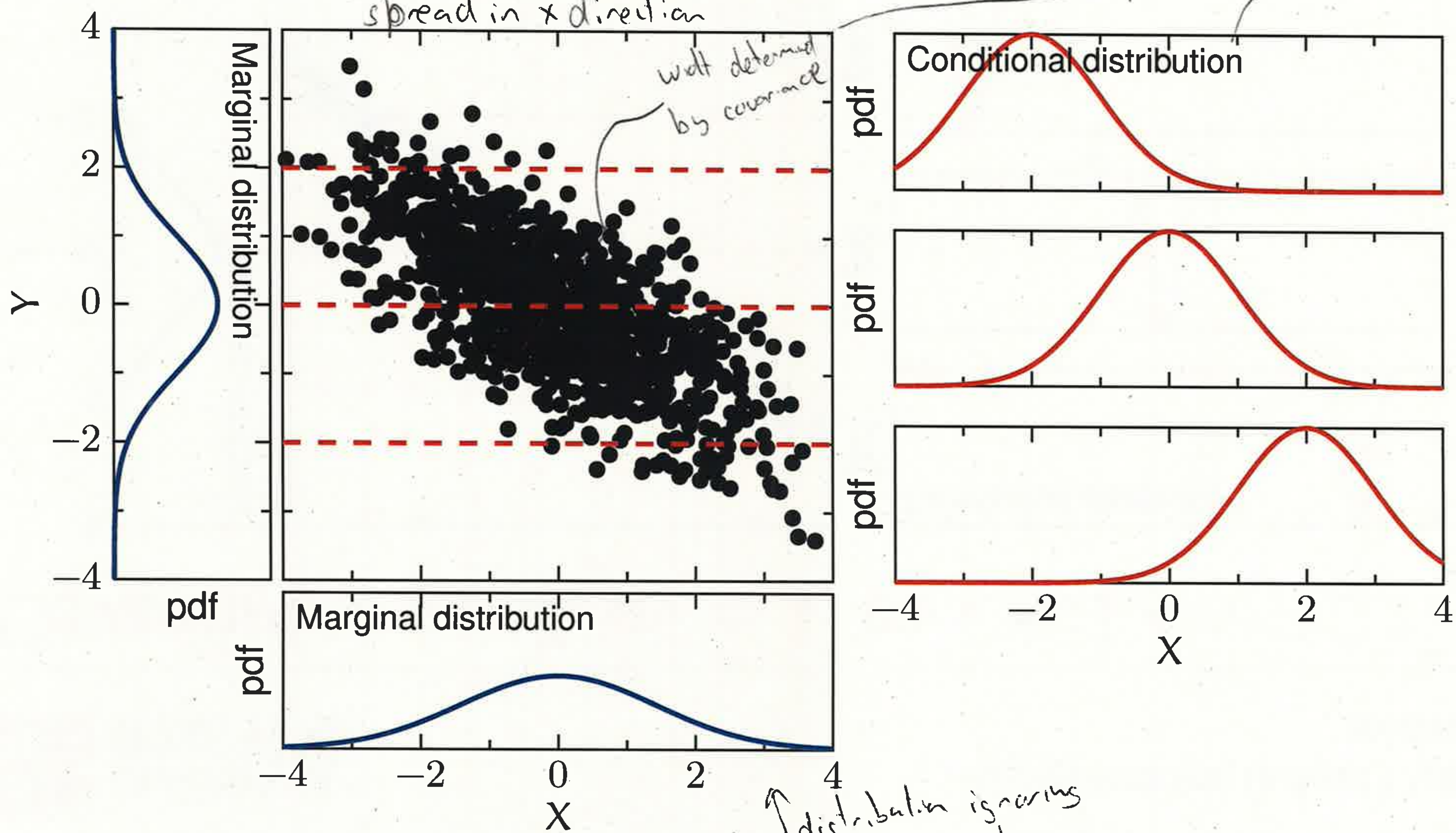
† The sample covariance matrix is a random variable $Q \sim W_p(\Sigma, n-1)$ where W_p is the Wishart distribution with $n-1$ degrees of freedom. (A generalisation of the χ^2 distribution — no details are needed.)

spreading direction

Covariant: $\sigma_X^2 = 2,\ \sigma_Y^2 = 1,\ \sigma_{X,Y} = -1$

spread in x direction

distribution if Y is fixed

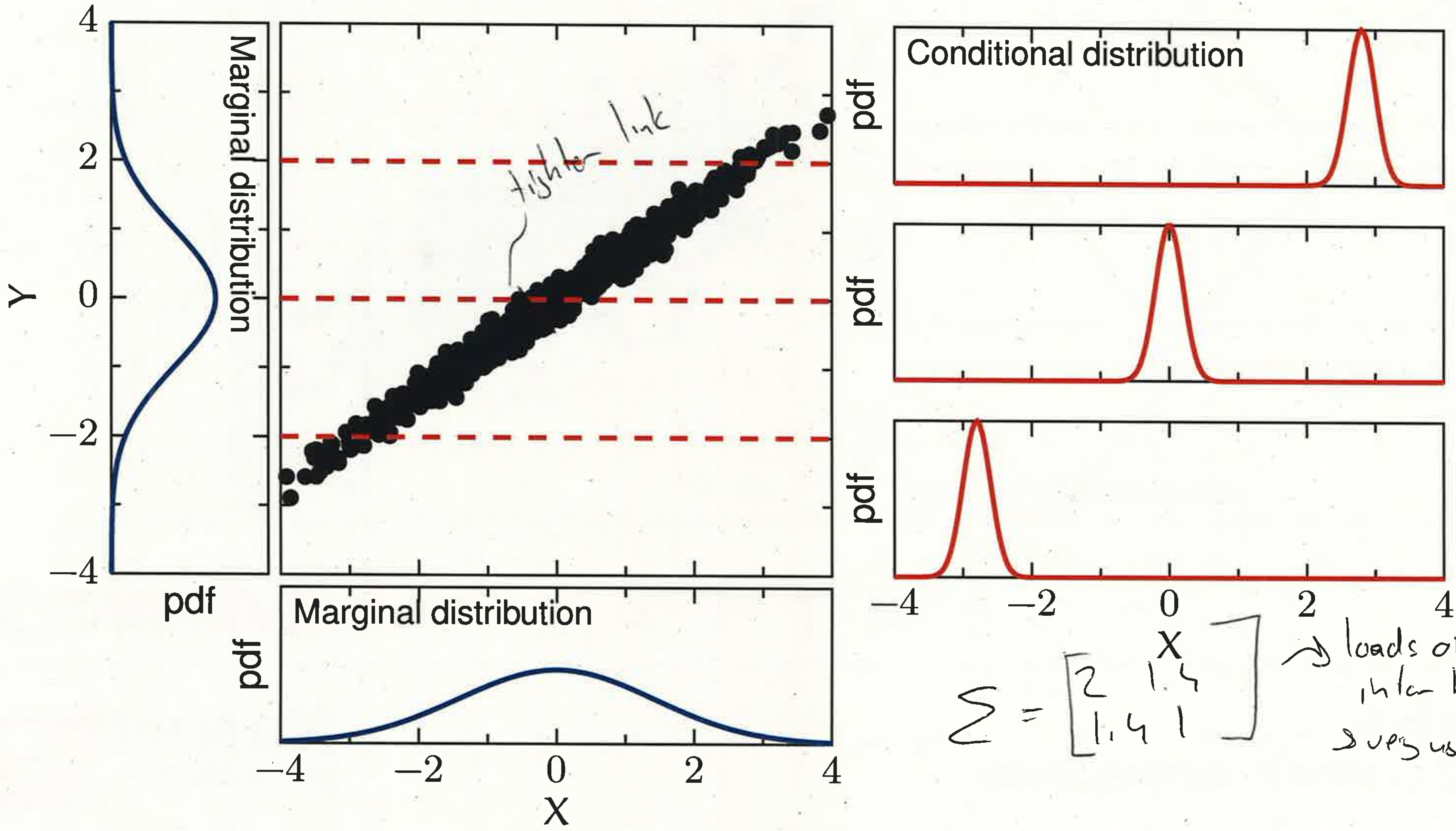


width determined by covariance

distribution ignoring Y values

Covariant: $\sigma_X = 2$, $\sigma_Y = 1$, $\sigma_{X,Y} = 1.4$

~ is there a maximum?



$$\Sigma = \begin{bmatrix} 2 & 1.4 \\ 1.4 & 1 \end{bmatrix}$$

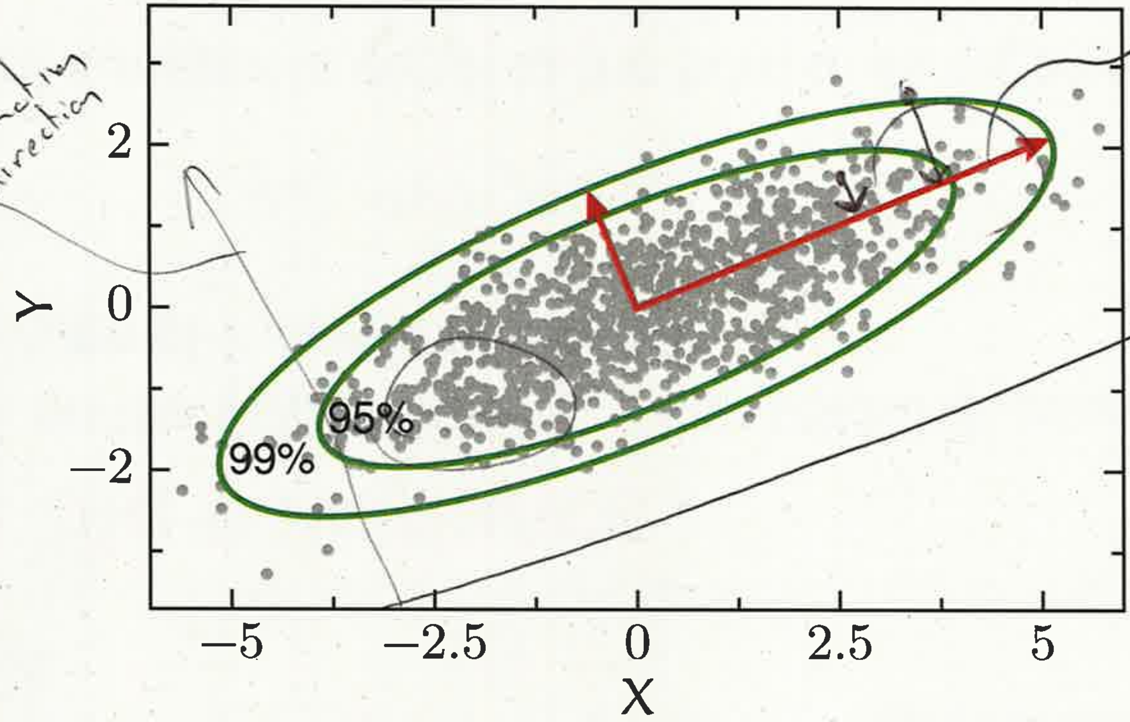
→ loads of info here → very useful

Eigenvalues and eigenvectors — PCA† not examinable

The eigenvalues and eigenvectors of the covariance matrix have the useful property that they provide a coordinate system that *linearly decouples* the random variables.

eigenvectors of covariance matrix!

less information in this direction



most of the information is contained in this direction

Goal: reduce the number of variables

These are the *principle components* or *principle directions*.

Values for the covariance

There are limits on the values that $\sigma_{X,Y}$ can take that come from the fact that the eigenvalues of Σ are non-negative —

symmetric and ≥ 0

$$-\sigma_X \cdot \sigma_Y \leqslant \sigma_{X,Y} \leqslant \sigma_X \cdot \sigma_Y$$

- The trace of a matrix is equal to the sum of its eigenvalues, so

sum of diagonal sum of eigenvalues

$$\text{trace}\left(\begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}\right) = \sigma_X^2 + \sigma_Y^2 = \lambda_1 + \lambda_2 \geqslant 0$$

$\neq$

Note that $\lambda_1 \neq \sigma_X^2$ and $\lambda_2 \neq \sigma_Y^2$!

- The determinant of a matrix is equal to the product of its eigenvalues

$$\det\left(\begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}\right) = \sigma_X^2\sigma_Y^2 - \sigma_{X,Y}^2 = \lambda_1\lambda_2 \geqslant 0$$

Positive semi-definite covariance matrix†

†For interested students only. Why are the eigenvalues non-negative? (That is, why is the matrix positive semi-definite?)

Positive semi-definite implies that for any $\mathbf{u}$ (column vector) we have

$$\mathbf{u}^{\mathsf{T}}\Sigma\mathbf{u} \geqslant 0$$

To show this consider that

$$\Sigma = \mathsf{E}\left[(Z-\mu_Z)(Z-\mu_Z)^{\mathsf{T}}\right]$$

and so, by linearity of expectations we have

$$\mathbf{u}^{\mathsf{T}}\Sigma\mathbf{u} = \mathsf{E}\left[\mathbf{u}^{\mathsf{T}}(Z-\mu_Z)(Z-\mu_Z)^{\mathsf{T}}\mathbf{u}\right]$$

The product $\mathbf{u}^{\mathsf{T}}(Z-\mu_Z)$ is a scalar (row vector · column vector) and so

$$\mathbf{u}^{\mathsf{T}}\Sigma\mathbf{u} = \mathsf{E}\left[\nu^2\right]$$

where ν is a scalar and so $\mathsf{E}\left[\nu^2\right] \geqslant 0$.

Correlation

The limits on $\sigma_{X,Y}$ suggest a normalisation —

population level quantity
– deterministic / true value

$$\rho := \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \quad \text{such that} \quad -1 \leqslant \rho \leqslant 1$$

This is called *Pearson's correlation coefficient*. It tells us the *linear* correlation between two random variables.

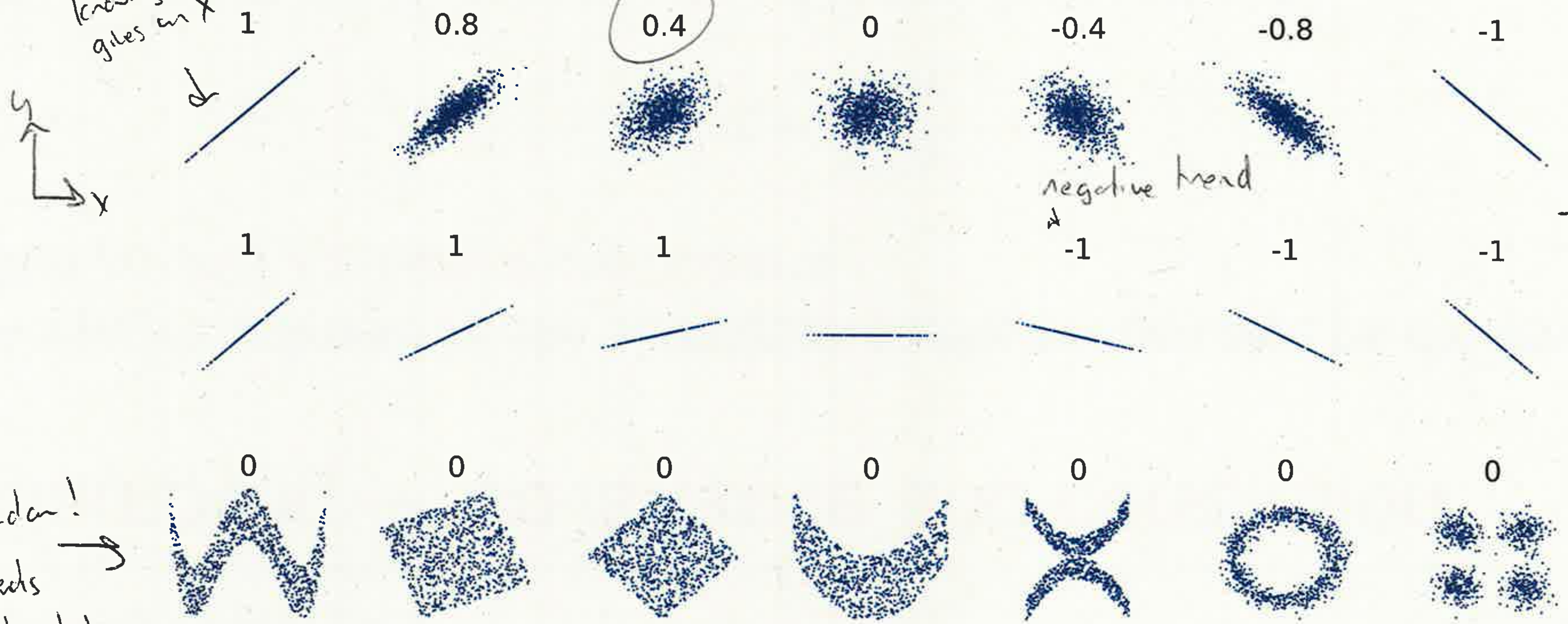
This is defined at the level of the population, not individual samples (see later), and so it is a deterministic quantity.

Correlation examples

pearson

A few examples of correlation coefficients [from Wikipedia]

ρ for A level marks and 1st year exam



→ value of correlation doesn't give the slope of the data

not random! clear trends not revealed by correlation coefficient → non linear relations between X & Y

simple example X X^2 → definite relationship but 0 correlation!

Sampling — covariance and correlation

A sample estimate for the correlation coefficient can also be determined. The *sample correlation coefficient* is

use estimated quantities

$$r = \frac{q_{X,Y}}{s_X s_Y}$$

sample covariance

sample std

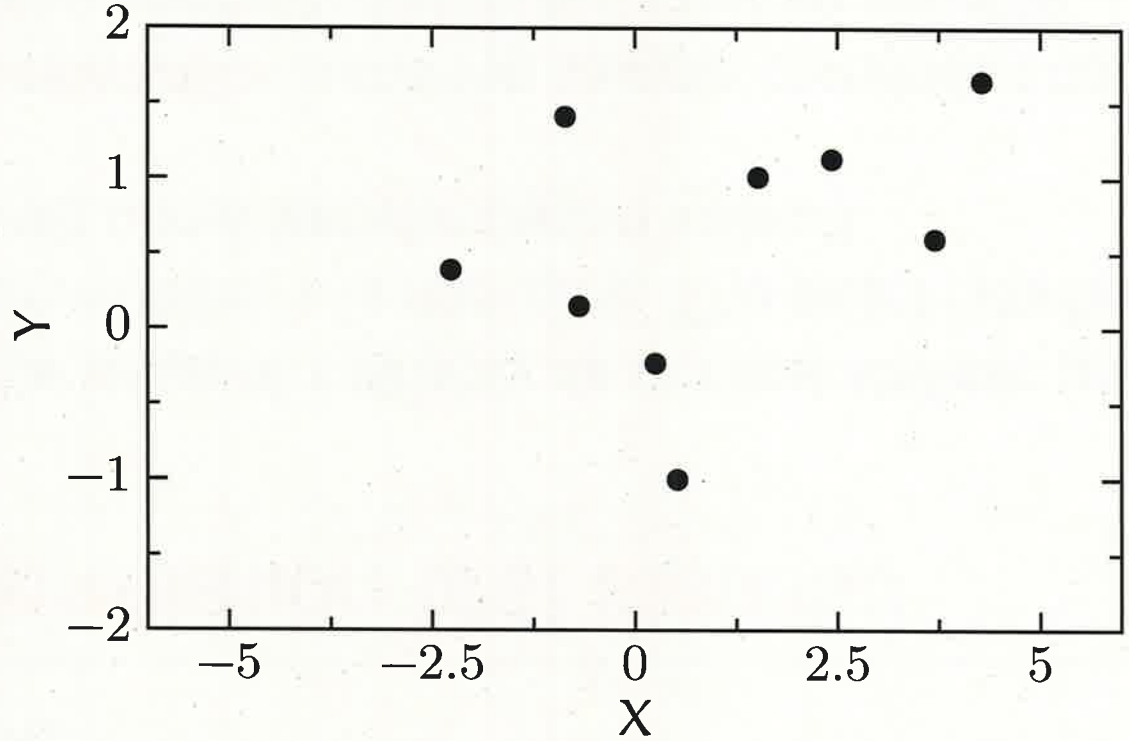
Expanding this out gives (the $n-1$ terms all cancel out)

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{y})^2}}$$

The sample correlation coefficient is a random variable whose distribution is known when the input data is bivariate normal (but it's horrible!).

Hypothesis testing with correlation

Sometimes it is useful to be able to test the hypothesis that data is not correlated — is there correlation in the data below?



is there a link between x and y?

Example, in steel alloy samples, if I measure the random variation in the chromium content and the variation in yield strength, are they correlated?

Sample correlation test statistic

calculate probability of getting that r assuming data was uncorrelated

We could use the value of r directly as our test statistic and do a two-tailed test on whether r is non-zero. The exact distribution of r is horrible and so we use a transformation instead.

The *Fisher transformation* turns the sample correlation coefficient r into an (approximately) normally distributed random variable —

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim N\left(\overbrace{\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}^{\text{mean}}, \overbrace{\frac{1}{n-3}}^{\text{variance}}\right)$$

number of observations

(population correlation coefficient

where ρ is the true (or population) correlation coefficient and n is the number of samples.

Fisher transformation

Transforming this into a standard normal distribution gives

$$\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \overbrace{\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}^{\text{mean}}\right]\underbrace{\sqrt{n-3}}_{1/\text{std}} \sim N(0,1)$$

Example. The 10 data samples previously plotted are given by

$$\begin{bmatrix}0.25\\-0.23\end{bmatrix}\begin{bmatrix}-2.3\\0.39\end{bmatrix}\begin{bmatrix}-0.69\\0.15\end{bmatrix}\begin{bmatrix}4.3\\1.7\end{bmatrix}\begin{bmatrix}2.4\\1.1\end{bmatrix}\begin{bmatrix}1.2\\-2.1\end{bmatrix}\begin{bmatrix}1.5\\1\end{bmatrix}\begin{bmatrix}-0.86\\1.4\end{bmatrix}\begin{bmatrix}3.7\\0.6\end{bmatrix}\begin{bmatrix}0.52\\-1\end{bmatrix}$$

Calculating the sample correlation coefficient yields $r = 0.2466$.
Use H_0: the random variables are uncorrelated ($\rho = 0$). We obtain

$$\frac{1}{2}\ln\left(\frac{1+0.2466}{1-0.2466}\right)\sqrt{10-3} = 0.67 < 1.96$$

can not reject null hypothesis

Hence, to 5% significance this data is not correlated (critical value 1.96).

Exercise

A material used in mobile phone batteries contains an impurity suspected of having an effect on the battery lifetime. A consumer campaign group carries out a small study of seven samples and find a sample correlation coefficient of $r = -0.8890$ between level of impurity and battery lifetime. The company manufacturing the batteries claims the impurity is uncorrelated with battery lifetime (the null hypothesis).

$\rho = 0$

Does the null hypothesis hold to 5% significance? Does the campaign group have a reason to claim that impurities affect battery lifetimes?

$$\frac{1}{2}\ln\left(\frac{1-0.885}{1+0.885}\right)\sqrt{7-3} = -2.83 \leq -1.96$$

reject null hypothesis