

Review of some probability concepts: random vectors, large-sample results

(A quick tour)

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

Random vectors¹

The multivariate normal distribution²

Statistics³

Complements & large-sample results⁴

¹Agresti, Kateri: sec 2.6

²Agresti, Kateri: sec 2.7

³Agresti, Kateri: sec 3.1-3.2

⁴Agresti, Kateri: sec 3.3-3.4

Random vectors

Random vectors

In statistics multiple variables are usually observed, and vectors of random variables (**random vectors**) are required. The two-dimensional case is useful to illustrate the main concepts, and will be used here.

For continuous r.v., the **joint (probability) density function** extends the one-dimensional case: it is the $f(x, y)$ function such that, for any $A \subseteq \mathbb{R}^2$

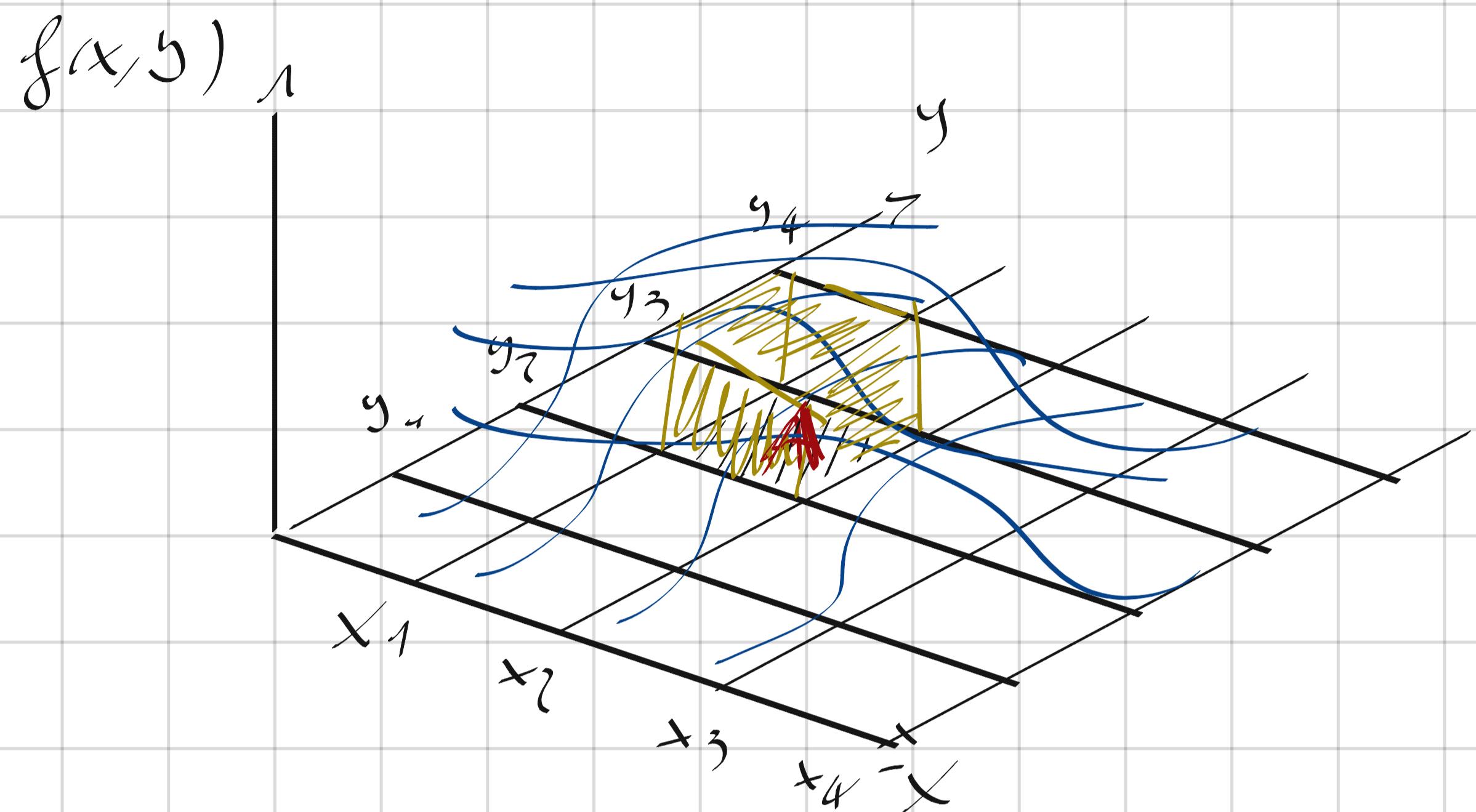
$$\Pr\{(X, Y) \in A\} = \int \int_A f(x, y) dx dy .$$

Note that $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$.

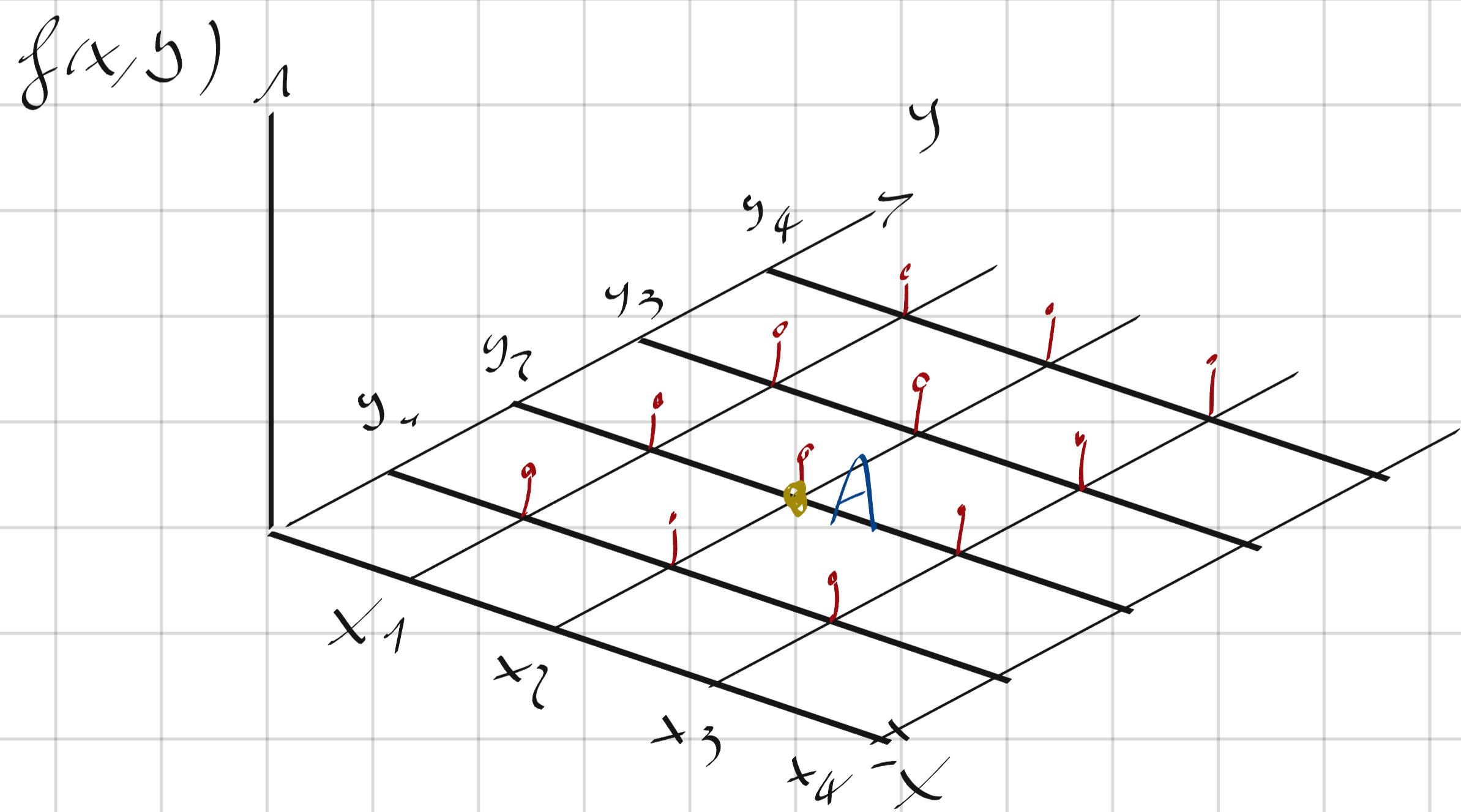
The probability density function defines the **joint distribution** of the random vector (X, Y) .

GRAPHIC RAPPRESENTATION OF

$$P((X, Y) \in A) = \iint_A f(x, y) dx dy$$



In CONTINUOUS SPACE, A IS
A RECTANGLE IN R^2 AND THE
PROB IS THE VOLUME UNDER
THE FUNCTION



In DISCRETE SPACE, A
IS THE INTERSECTION THAT
RAPPRESENT 2 EVENTS HAPPENING
AT THE SAME TIME.

Marginal distribution

CONDITIONAL

JOINT

→ MARGINAL

	Z	K	W	
X	xz	xK	xw	xzKw
Y	yz	yK	yw	yZKw
	xyz	xYK	xyw	xyzKw

The joint distribution embodies information about each components, so that the distribution of X , ignoring Y , can be obtained from $f(x, y)$.

The *marginal* density function of X is given by

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy,$$

WE INTEGRATE OUT 1 OF THE 2 VARIABLES AND AVOID THE BEHAVIOR OF THE OTHER R.V.

and similarly for the other variable.

(Note: here and elsewhere we always use the symbol f for any p.d.f., identifying the specific case by the argument of the function).

Conditional distribution

The *conditional density function* of Y given $X = x_0$ updates the distribution of Y by incorporating the information that $X = x_0$.

It is given by the important formula

$$f(y|X = x_0) = \frac{f(x_0, y)}{f(x_0)}, \quad \text{provide } f(x_0) > 0.$$

The simplified notation $f(y|x_0)$ is often employed.

The conditional p.d.f. is properly defined, since $f(y|X = x_0) \geq 0$ and $\int_{-\infty}^{\infty} f(y|x_0) dy = 1$.

A symmetric definition applies to X given $Y = y_0$.

Conditional distribution: useful properties

In the two dimensional case, it is readily possible to write

$$f(x, y) = f(x) f(y|x). \quad \Rightarrow \quad \frac{f(x, y)}{f(x)} = \frac{20}{60} = 0.3$$
$$20 = 60 \cdot \frac{20}{60}$$

Extensions to higher dimensions require some care:

	y	K	w
x	xy	xK	xw
z	zy	zK	zw
	xzY	xzK	xzW
	xzY	xzK	xzW

$$f(x, y, z) = f(x, y|z) f(z)$$

$$f(x, y|z) = f(x|z) f(y|x, z)$$

$$f(x, y, z) = f(x|y, z) f(y, z)$$

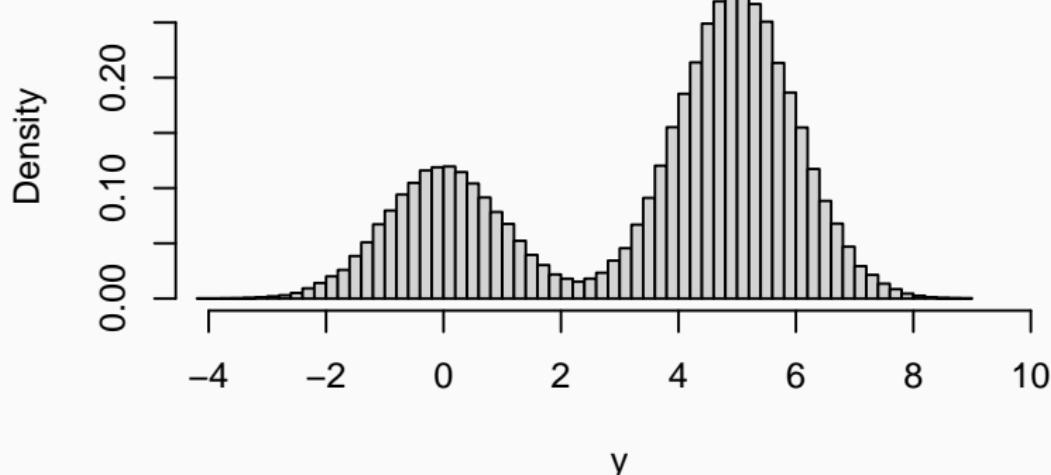
$$f(x, y, z) = f(x|y, z) f(y|z) f(z)$$

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2, x_1) \dots f(x_n|x_{n-1}, \dots, x_2, x_1)$$

R lab: simulation from joint distributions (a mixture)

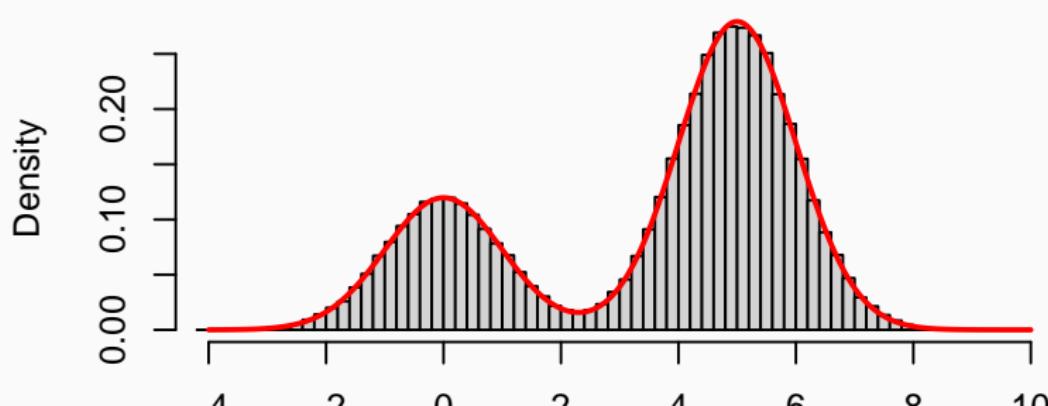
```
x <- rbinom(10^5, size = 1, prob = 0.7)
y <- rnorm(10^5, m = x * 5, s = 1) ### Y| X = x ~ N(x * 5, 1)
hist.scott(y, main = "", xlim = c(-4, 10))
```

WE SIMULATE FROM A BINOMIAL TO A NORMAL DISTRIBUTION



R lab: simulation from joint distributions (cont'd.)

```
xx <- seq(-4, 10, l = 1000)
ff <- 0.3 * dnorm(xx, 0) + 0.7 * dnorm(xx, 5)
### This is a mixture of normal distributions
hist.scott(y, main = "", xlim = c(-4, 10))
lines(xx, ff, col = "red", lwd = 2)
```



Bayes theorem

From the factorization of the joint distribution it readily follows that

$$f(x, y) = f(x) f(y|x) = f(y) f(x|y)$$

from which we obtain the **Bayes theorem**

$$f(x|y) = \frac{f(x) f(y|x)}{f(y)}.$$

This is a cornerstone of statistics, leading to an entire school of statistical modelling.

Independence and conditional independence

When $f(y|x)$ does not depend on the value of x , the r.v. X and Y are *independent*, and

$$f(x, y) = f(y) f(x)$$

More in general, n r.v. are independent if and only if

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2) \dots f(x_n).$$

Conditional independence arises when two r.v. are independent given a third one:

$$f(y, x|z) = f(x|z) f(y|z)$$

An important part of statistical modelling exploits some sort of conditional independence.

Example of conditional independence: the Markov property

The general factorization defined above

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2, x_1) \dots f(x_n|x_{n-1}, \dots, x_2, x_1)$$

will simplify considerably when the *first order Markov property* holds:

$$f(x_i|x_1, \dots, x_{i-1}) = f(x_i|x_{i-1})$$

which means that X_i is independent of X_1, \dots, X_{i-2} given X_{i-1} . We get

$$f(x_1, x_2, \dots, x_n) = f(x_1) \prod_{i=2}^n f(x_i|x_{i-1}).$$

When the variables are observed over time, this means that the process has *short memory*, a property quite useful in the statistical modelling of **time series**.

Mean and variance of linear transformations

For two r.v. X and Y and two constants a, b we get

$$E(aX + bY) = aE(X) + bE(Y).$$

The result follows from the more general one

$$E\{g(X, Y)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

For variances we need first to introduce the **covariance** between X and Y

$$\text{cov}(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - \mu_x \mu_y,$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. Then

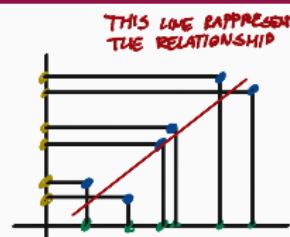
$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \text{cov}(X, Y).$$

Note: for X, Y independent it follows that $\text{cov}(X, Y) = 0$. The reverse is not true, unless the joint distribution of X and Y is multivariate normal.

COVARIANCE CAN CLASSIFY THREE TYPES OF RELATIONSHIPS:

- POSITIVE TRENDS
- NEGATIVE TRENDS
- NO TRENDS (SO NO RELATIONSHIP)

COVARIANCE IS COMPUTATIONAL STEPPING STONE TO SOMETHING THAT IS INTERESTING, LIKE CORRELATION ALONE DOESN'T MEAN TOO MUCH



Mean vector

For a random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$, the **mean vector** is just

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}.$$

The mean vector has the same properties of the scalar case, so that for example $E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$, and for \mathbf{A} and \mathbf{b} a $n \times n$ matrix and a $n \times 1$ vector, respectively, it follows that

$$E(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}E(\mathbf{X}) + \mathbf{b}.$$

Variance-covariance matrix

The variance-covariance matrix of the random vector \mathbf{X} collects all the variances (on the main) diagonal and all the pairwise covariances (off the main diagonal), being the $n \times n$ symmetric semi-definite matrix

THIS REPRESENTS ALL THE
VARIANCE AND COVARIANCE
WE CAN COMPUTE

$$\Sigma = E\{(\mathbf{X} - \mu_x)(\mathbf{X} - \mu_x)^\top\} = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_1, X_2) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_1, X_n) & \text{cov}(X_2, X_n) & \cdots & \text{var}(X_n) \end{pmatrix}$$

Useful properties:

$$\Sigma_{A\mathbf{X} + \mathbf{b}} = A\Sigma A^\top$$

$$\Sigma_{\mathbf{X}^\top A \mathbf{X}} = \mu_x^\top A \mu_x + \text{tr}(A\Sigma)$$

Transformation of random variables and random vectors

Given a continuous r.v. X and a transformation $Y = g(X)$, with g an invertible function, it readily follows that

$$f_y(y) = f_x\{g^{-1}(y)\} \left| \frac{dx}{dy} \right|.$$

The result is extended to two continuous random vectors with the same dimension

$$f_{\mathbf{Y}}(\mathbf{Y}) = f_{\mathbf{X}}\{g^{-1}(\mathbf{Y})\} |\mathbf{J}|,$$

with $J_{ij} = \partial x_i / \partial y_j$.

For discrete r.v., the results are simpler, with no need of including the Jacobian of the transformation.

Exercise

$$X \sim N(\mu, \sigma^2)$$

$$Y = e^X := g(Y) \Rightarrow Y \sim \log N(\mu, \sigma^2) := \tilde{g}^{-1}(Y)$$

$$f_Y(y) = f_X \left\{ \tilde{g}^{-1}(y) \right\} \left| \frac{dx}{dy} \right|$$

$$g(x) = y = e^x \Rightarrow x = g^{-1}(y) = \log(y)$$

$$\left(\lim_{x \rightarrow -\infty} e^x = 0, \lim_{x \rightarrow \infty} e^x = +\infty \right)$$

$$\frac{dx}{dy} = \frac{1}{y}$$

$$f_Y(y) = \begin{cases} f_X(\log y) \frac{1}{y} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (\log y - \mu)^2 \right\} \frac{1}{y} & 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

The multivariate normal distribution

The multivariate normal distribution

Start from a set of n i.i.d. $Z_i \sim \mathcal{N}(0, 1)$, so that $E(\mathbf{z}) = \mathbf{0}$ and covariance matrix \mathbf{I}_n . If \mathbf{B} is $m \times n$ matrix of coefficients and $\boldsymbol{\mu}$ a m -vector of coefficients, then the m -dimensional random vector \mathbf{X}

$$\mathbf{X} = \mathbf{B}\mathbf{z} + \boldsymbol{\mu}$$

has a **multivariate normal distribution** with covariance matrix $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\top$.

The notation is

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Joint p.d.f.

Using basic results on transformation of random vectors, starting from the joint p.d.f of Z_1, Z_2, \dots, Z_n we obtain

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\}, \quad \text{for } \mathbf{X} \in \mathbb{R}^m,$$

provided that $\boldsymbol{\Sigma}$ has full rank m . The result can be extended to *singular $\boldsymbol{\Sigma}$* by recourse to the *pseudo-inverse* of $\boldsymbol{\Sigma}$: this is used, for example, in the analysis of *compositional data*.

A useful property which holds only for this distribution: *two r.v. with multivariate normal distribution and zero covariance are independent.*



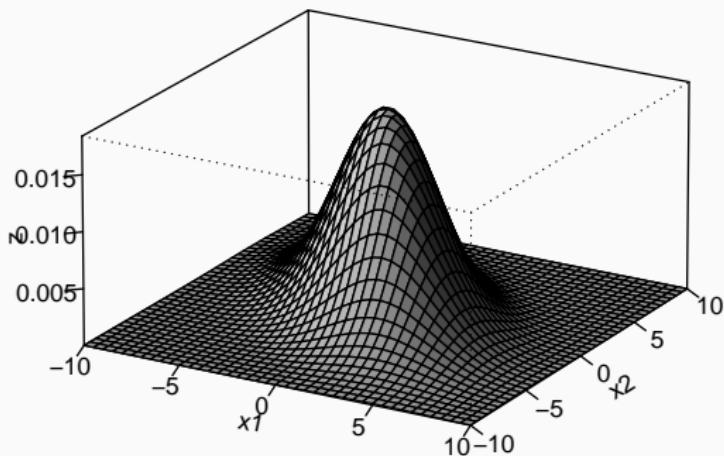
IN GENERAL IS NOT TRUE

INDEPENDENT \Rightarrow ZERO COVARIANCE

ZERO COVARIANCE \Rightarrow INDEPENDENT ONLY FOR
NORMAL DISTRIBUTION

Example: bivariate case

We take $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 10$, $\sigma_2^2 = 10$, $\sigma_{12} = 15$



Linear transformations

It is simple to verify that if $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ and \mathbf{A} is a $k \times m$ matrix of constants then

$$\mathbf{AX} \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^\top).$$

A special case is obtained when $k = 1$, in that for a m -dimensional vector \mathbf{a}

$$\mathbf{a}^\top \mathbf{X} \sim \mathcal{N}(\mathbf{a}^\top \mu, \mathbf{a}^\top \Sigma \mathbf{a}).$$

Note that for suitable choices of \mathbf{a} (when all the elements 0s or 1s) it follows that **the marginal distribution of any subvector of \mathbf{X} is multivariate normal.**

*THE REVERSE IS NOT ALWAYS TRUE
YOU CAN START WITH MARGINAL THAT ARE NORMAL DISTRIBUTED AND END UP WITH A DISTRIBUTION THAT IS NOT THE NORMAL*

Normality of the marginal distributions, instead, does not imply multivariate normality.

Conditional distributions

Consider two random vectors \mathbf{X} and \mathbf{Y} with multivariate normal joint distribution, and partition their joint covariance matrix as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix},$$

and similarly for the mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)^\top$.

Using results on *partitioned matrices*, it follows that the **conditional distributions are multivariate normal**.

For instance

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} (\mathbf{X} - \boldsymbol{\mu}_x), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}).$$

Statistics

Random sample

The collection of r.v. X_1, X_2, \dots, X_n is said to be a **random sample** of size n if they are *independent and identically distributed*, that is

- X_1, X_2, \dots, X_n are independent r.v.
- They have the same distribution, namely the same c.d.f.

The concept is central in statistics, and it is the suitable mathematical model for the outcome of sampling units from a very large population.
The definition is, however, more general.

(For more details: https://www.probabilitycourse.com/chapter8/8_1_1_random_sampling.php)

Statistics

WHEN WE DO STATISTICS WE WANT TO FIND A WAY TO DESCRIBE AN ENVIRONMENT
A **statistic** is a r.v. defined as a function of a set of r.v.

Obvious examples are the sample mean and variance of data y_1, y_2, \dots, y_n

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Consider a random vector \mathbf{Y} with p.d.f. $f_\theta(\mathbf{Y})$ depending on a vector θ (which is the *parameter*, as we will see).

If a statistic $t(\mathbf{Y})$ is such that $f_\theta(\mathbf{Y})$ can be written as

$$f_\theta(\mathbf{Y}) = h(\mathbf{Y}) g_\theta\{t(\mathbf{Y})\}, \begin{array}{l} \xrightarrow{\text{RELATED TO THE PARAMETER}} \\ \xrightarrow{\text{THROUGH A FUNCTION ON}} \\ \xrightarrow{\text{ONLY RELATED TO DATA}} \end{array}$$

where h does not depend on θ , and g depends on \mathbf{Y} only through $t(\mathbf{Y})$, then t is a **sufficient statistic** for θ : all the *information* available on θ contained in \mathbf{Y} is supplied by $t(\mathbf{Y})$.

The concepts of information and sufficiency are central in statistical inference.

Example: sufficient statistic for the normal distribution

SUFFICIENCY IS RELATED TO THE PARAMETERS OF A DISTRIBUTION

WE WANT TO ESTIMATE 2 PARAMETERS \Rightarrow 2 SUFFICIENT STATISTICS

Given a vector of independent normal r.v. $Y_i \sim N(\mu, \sigma^2)$, it follows that $\theta = (\mu, \sigma^2)$ and

$$\begin{aligned} f_{\theta}(\mathbf{Y}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} \underbrace{(y_i - \mu)^2}_{\text{DENSITY OF}} \right\} \xrightarrow{\text{NORMAL DISTRIBUTION}} \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right\} \xrightarrow{\text{WITH I.I.D. ASSUMPTION}} \\ &= \left[(y_i - \bar{y})(\mu - \bar{y}) \right] \\ &= [(y_i - \bar{y})^2 + (\mu - \bar{y})^2] \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n (\mu - \bar{y})^2 \\ &\quad - 2 \sum_{i=1}^n (y_i - \bar{y})(\mu - \bar{y}) \end{aligned}$$

WE JUST KEEP THIS PART BECAUSE IT DEPENDS ON y_i (THE DATA)
WHILE THIS DEPENDS ON μ

By some simple algebra, it is possible to show that the two-dimensional statistic $t(\mathbf{Y}) = (\bar{y}, s^2)$ is sufficient for (μ, σ^2) .

but $\sum (y_i - \bar{y})$ is
EQUAL TO 0

Complements & large-sample results

Moment generating function

The **moment generating function** (m.g.f.) characterises the distribution of a r.v. X , and it is defined as

$$M_X(t) = E(e^{tX}), \quad \text{for } t \text{ real}.$$

The name derives from the fact the k^{th} derivative of the m.g.f. at $t = 0$ gives the k^{th} uncentered moment:

$$\frac{d^k M_X(t)}{dt^k} \Big|_{t=0} = E(X^k).$$

Two useful properties:

- If $M_X(t) = M_Y(t)$ for some small interval around $t = 0$, then X and Y have the same distribution.
- If X and Y are independent, $M_{X+Y}(t) = M_X(t) M_Y(t)$.

The central limit theorem

For i.i.d. r.v. X_1, X_2, \dots, X_n with mean μ and finite variance σ^2 , the **central limit theorem** states that for large n the distribution of the r.v. $\bar{X}_n = \sum_{i=1}^n X_i/n$ is approximately

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n).$$

More formally, the theorem says that for any $x \in \mathbb{R}$ the c.d.f. of $Z_n = (\bar{X}_n - \mu)/\sqrt{\sigma^2/n}$ satisfies

$$\lim_{n \rightarrow \infty} F_{Z_n}(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

The proof is simple, and it uses the m.g.f.

The theorem generalizes to multivariate and non-identical settings.

It has a central importance in statistics, since it supports the normal approximation to the distribution of a r.v. that can be viewed as the sum of other r.v.

The law of large numbers

Consider i.i.d. (independent and identically distributed) r.v. X_1, \dots, X_n , with mean μ and $(E|X_i|) < \infty$.

The **strong law of large numbers** states that, for any positive ϵ

$$\Pr\left(\lim_{n \rightarrow \infty} |\bar{X}_n - \mu| < \epsilon\right) = 1,$$

namely \bar{X}_n converges almost surely to μ .

With the further assumption $\text{var}(X_i) = \sigma^2$, the **weak law of large numbers** follows

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| \geq \epsilon) = 0.$$

Proof of the weak law of large numbers

First we recall the *Chebyshev's inequality*: given a r.v. X such that $E(X^2) < \infty$ and a constant $a > 0$, then

$$\Pr(|X| \geq a) \leq \frac{E(X^2)}{a^2}.$$

We apply the inequality to the case of interest, so that

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{E\{(\bar{X}_n - \mu)^2\}}{\epsilon^2} = \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n \epsilon^2},$$

which tends to zero when $n \rightarrow \infty$.

The result may hold also for non-i.i.d. cases, provided $\text{var}(\bar{X}_n) \rightarrow 0$ for large n .

Jensen's inequality

This is another useful result, that states that for a r.v. X and a concave function g

$$g\{E(X)\} \geq E\{g(X)\}.$$

(Note: a concave function is such that

$$g\{\alpha x_1 + (1 - \alpha) x_2\} \geq \alpha g(x_1) + (1 - \alpha) g(x_2),$$

for any x_1, x_2 , and $0 \leq \alpha \leq 1$).

An example is $g(x) = -x^2$, so that

$$-E(X)^2 \geq -E(X^2) \quad \Rightarrow \quad E(X)^2 \leq E(X^2),$$

which is obviously true since $E(X^2) = \text{var}(X) + E(X)^2$.