Inferential Statistics **First semester, 2021**

Lecture 1: Random vectors and convergence

Instructor: Erlis Ruli (ruli@stat.unipd.it), Department of Statistical Sciences

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Random vectors and their distributions

So far our focus has been on a single r.v. at time, but reality is multivariate. For instance, suppose we need to check the performance of a washing machine taken at random from the production line. There are several variables involved:

- washing performance X_1
- duration X_2
- energy consumption X_3
- water consumption X_4
- spin speed X_5, etc.

Let $X = (X_1, X_2, \dots, X_k)$ be a well-behaved vector-valued function of reals, i.e. $X(s) : \mathcal{S} \to \mathbb{R}^k$ on a triple $(\mathcal{S}, \mathcal{A}, P)$. Then X is called *random* vector or r.ve. for short.

Likewise a r.v., a r.ve. X is characterised by its joint d.f., defined by

$$F(x) = P(X \leq x) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k),$$

and by its joint probability density function $f_X(x)$, when it exists.

The r.ve. X is called discrete if $P(X = x_j) > 0$, $j = 1, 2, \ldots,$ with $\sum_j P(X = x_j) = 1$. In this case, the p.d.f of X is $f_X(x) = P(X = x_j)$ for $x = x_j$ and $f_X(x) \geq 0$. Once again, $P(X \in B) = \sum_{x_j \in B} f_X(x_j)$ for $B \subseteq \mathbb{R}^k$.

The r.ve. X is continuous if $P(X = x) = 0$ for all $x \in \mathbb{R}^k$ and there is a function f_X defined on $\mathbb{R}^k$ such that:

$$f_X(x) \geq 0 \quad \text{for all} \quad x \in \mathbb{R}^k, \quad \text{and} \quad P(X \in J) = \int_J f_X(x) dx_1 dx_2 \cdots dx_k.$$

A similar comment applies regarding absolute continuity and for the ease of exposition, hereafter by continuous we mean absolutely continuous.

1.1.1 Marginal distributions and their moments

We focus on $k = 2$ but note the results can be extended to any positive integer k. Given $X = (X_1, X_2)$ a *bivariate* r.ve, the components X_1 and X_2 are both r.v.'s and their *marginal* distributions are derived from F_X or f_X as we outline below.

The marginal p.d.f. of X_1, is

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_X(x_1, t) dt.$$

In words, the marginal distribution is obtained by marginalising over the nuisance component. The expectation of X is $\mu = (\mu_1, \mu_2)$, where

$$\mu_i = \int_{-\infty}^{\infty} t f_{X_i}(t) dt, \quad i = 1, 2.$$

The variance of X_i is

$$\sigma_i^2 = \int_{-\infty}^{\infty} (t - \mu_i)^2 f_{X_i}(t) dt.$$

The covariance between X_1 and X_2 is defined by

$$\begin{aligned}
\text{cov}(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\
&= E(X_1 X_2) - E(X_1)E(X_2) \\
&= \mu_{12} - \mu_1 \mu_2 \\
&= \sigma_{12}.
\end{aligned}$$

Furthermore, the coefficient correlation is defined by $\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}}$, where $\sigma_{12} \in \mathbb{R}$ and $\rho_{12} \in [-1, 1]$. Note that if X is a r.v. then $\text{cov}(X, X) = \text{var}(X) = \sigma_X^2$.

1.1.2 Conditional distributions and moments

Conditional distribution functions are defined similarly to the conditional probability of events (see L0, § 0.3.1). Let $X = (X_1, X_2)$ be a r.ve. with joint p.d.f $f_X(x) = P(X_1 = x_1, X_2 = x_2)$, then the p.d.f. of X_1 given $X_2 = x_2$ is defined as

$$P(X_1 = x_1 | X_2 = x_2) = \tfrac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)}.$$

If X is a continuous r.ve. with p.d.f. $f_X(x)$, then the conditional p.d.f. of X_1 given $X_2 = x_2$ is defined by

$$f_{X_1|X_2}(x_1|x_2) = \tfrac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}.$$

Remark 1.1 *(i) For any fixed x_2, the conditional distributions are proper probability distributions. This can be seen from the fact that, in the discrete case, summing over all possible x for X_1 leaves us with the probability of the event $X_2 = x_2$, i.e. $\sum_t P(X_1 = t, X_2 = x_2) = P(X_2 = x_2)$. A similar observation applies to the continuous case, for which we have*

$$\int_{-\infty}^{\infty} f_X(x) dx_1 = \int_{-\infty}^{\infty} f_X(x_1, x_2) dx_1 = f_{X_2}(x_2).$$

(ii) The conditional distributions shown above are functions indexed by x_2. This means that for any s, t in the domain of X_2, if $s \neq t$ then, in general $P(X_1 = x_1 | X_2 =$

$s) \neq P(X_1 = x_1|X_2 = t)$. That is, changing the conditioning event could lead to a different conditional distribution.

If a conditional distribution admits moments, these are called *conditional moments.* The *conditional expectation* of X_1 given $X_2 = x_2$ is denoted by $E(X_1|X_2 = x_2)$ or $\mu_{X_1|X_2}$ is defined by

$$E(X_1|X_2 = x_2) = \int_{-\infty}^{\infty} t f_{X_1|X_2}(t|x_2)dt.$$

The *conditional variance* of X_1 given $X_2 = x_2$ is denoted by $\text{var}(X_1|X_2 = x_2)$ and is defined by

$$\text{var}(X_1|X_2 = x_2) = \int_{-\infty}^{\infty} (t - \mu_{X_1|X_2})^2 f_{X_1|X_2}(t|x_2)dt.$$

1.1.3 Independence of random variables

We remarked in Remark 1.1 that in general the conditional p.d.f. could change if the conditioning event changes. If on the contrary the conditional distribution does not vary with the conditioning event, then the two r.vs. are said to be *independent.* More concretely, two random variables X_1 and X_2 are said to be independent if their events $(X_1 = x_1)$ and $(X_2 = x_2)$ are independent, for any admissible x_1, x_2. This is stated formally by the next definition.

Definition 1.1 *Let X_1 and X_2 be two r.v.s with joint d.f. F_X. Then X_1 is independent from X_2 if*

$$P(X_1 = x_1|X_2 = x_2) = P(X_1 = x_1) \quad \textit{and} \quad P(X_2 = x_2|X_1 = x_1) = P(X_2 = x_2), \quad \textit{for all } x_1, x_2,$$

when X_1, X_2 are both discrete. X_1 is independent from X_2 if

$$f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1) \quad \textit{and} \quad f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2), \quad \textit{for all } x_1, x_2,$$

when X_1, X_2 are both continuous.

Another useful characterisation of independence which applies to discrete as well as continuous random variable is the following.

Theorem 1.1 *The r.v.s $X_1, \ldots, X_n$ are independent if and only if one of the following two (equivalent) conditions hold:*

(i) $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{j=1}^n F_{X_j}(x_j)$, *for all* $x_j \in \mathbb{R}$, $j = 1, \ldots, n$.

(ii) $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{j=1}^n f_{X_j}(x_j)$, *for all* $x_j \in \mathbb{R}$, $j = 1, \ldots, n$.

Furthermore, we have the following properties.

Theorem 1.2 *Let $X_1, \ldots, X_n$ be r.v.s all of the same type, i.e either all discrete or all continuous and having joint d.f. F_X. Then:*

(i) If X_1 *and* X_2 *are independent r.v., then* $E(X_1X_2) = E(X_1)E(X_2)$ *and* $\mathrm{cov}(X_1, X_2) = 0$.

(ii) Given reals a_0, a_1, b_0, b_1 *and* $X_a = a_0 + a_1X_1, X_b = b_0 + b_1X_2$, *then*

$$\mathrm{cov}(X_a, X_b) = \mathrm{cov}(a_0 + a_1X_1, b_0 + b_1X_2) = a_1b_1\mathrm{cov}(X_1, X_2).$$

(iii) If $Y = a + bX_1$ *for any reals* a, b, *then* $|\mathrm{cov}(X_1, Y)| = \sigma_{X_1}\sigma_Y$.

(iv) Let $\mu_1 = E(X_1), \mu_2 = E(X_2)$, $\sigma_1^2 = \mathrm{var}(X_1), \sigma_2^2 = \mathrm{var}(X_2)$, $\mathrm{cov}(X_1, X_2) = \sigma_{12}$ *and set* $X = aX_1 + bX_2$, *for any real* a, b. *Then*

$$E(X) = a\mu_1 + b\mu_2, \quad \mathrm{var}(X) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_{12}.$$

Furthermore, if $\sigma_{12} = 0$ *then* $\mathrm{var}(X) = a^2\sigma_1^2 + b^2\sigma_2^2$.

(v) More generally if $(b_1, \ldots, b_n) \in \mathbb{R}^n$ *and* $T = \sum_{j=1}^n b_jX_j$ *then*

$$E(X) = \sum_{j=1} b_jE(X_j), \quad \mathrm{var}(X) = \sum_{j=1}^n b_j^2\mathrm{var}(X_j) + \sum_{i\neq j} b_jb_j\mathrm{cov}(X_i, X_j).$$

Furthermore, if $\mathrm{cov}(X_i, X_j) = 0$ *for all* $i \neq j = 1, \ldots, n$, *then* $\mathrm{var}(X) = \sum_{j=1}^n b_j^2\mathrm{var}(X_j)$.

Proof: We prove (i) and (ii) assuming X to be continuous.
(i): By Theorem 1.1 $f_X(x) = f_{X_1}(x_1)f_{X_2}(x_2)$, thus

$$E(X_1X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} stf_X(s,t)dsdt = \left(\int_{-\infty}^{\infty} sf_{X_1}(s)ds\right)\left(\int_{-\infty}^{\infty} tf_{X_2}(t)dt\right) = E(X_1)E(X_2).$$

Thus $\text{cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = 0$.
(ii) : By the definition of covariance

$$\begin{aligned}\text{cov}(X_a, X_b) &= E[(X_a - E(X_a))(X_b - E(X_b))]\\ &= E[(a_0 + a_1X_1 - E(a_0 + a_1X_1))(b_0 + b_1X_2 - E(b_0 + b_1X_2))]\\ &= E[a_1(X_1 - \mu_1)(X_2 - \mu_2)b_1] = a_1E[(X_1 - \mu_1)(X_2 - \mu_2)]b_1.\end{aligned}$$

■

1.1.4 A matrix-view of random vectors

So far the r.ve. $X = (X_1, X_2, \ldots, X_k)$ was studied in a component-wise fashion. A matrix/vector approach is also possible, which leads to more intuitive and compact notation.

First some further notation. X is always meant to be column vector, unless stated otherwise. All matrices and vectors hereafter are understood to be real-valued. We denote a matrix A with elements a_{ij} by $A = [a_{ij}]$, with the index i running over the rows of A and j running over its the columns. An alternative notation for a vector a, usually denoted by $a = (a_1, \ldots a_n)$, is $a = [a_i]$. Two matrices $A = [a_{ij}]$, $B = [b_{ij}]$ are equal if they have the same size and $a_{ij} = b_{ij}$ for all i, j.

To work with r.ve.s we need to define the expectation of a matrix-valued function. For $g(X) = [g_{ij}(X)]$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ a matrix-valued function we define

$$E(g(X)) = E[g_{ij}(X)] = \begin{bmatrix} E(g_{11}(X)) & E(g_{12}(X)) & \ldots & E(g_{1n}(X)) \\ E(g_{21}(X)) & E(g_{22}(X)) & \ldots & E(g_{2n}(X)) \\ \vdots & \vdots & \ddots & \vdots \\ E(g_{m1}(X)) & E(g_{m2}(X)) & \ldots & E(g_{mn}(X)) \end{bmatrix},$$

provided all expectations exist and are finite. Here are two noteworthy examples.

- $g(X) = (X_1, \ldots, X_k)$: this leads to $E(g(X)) = (\mu_1, \ldots, \mu_k)$; the vector $\mu = (\mu_1, \ldots, \mu_k)$ is called expected value of X.

- $g(X) = (X - \mu)(X - \mu)^{\mathrm{T}}$: implies that $g(X) = [g_{ij}(X)] = [(X_i - \mu_i)(X_j - \mu_j)]$. Thus

  $$E(g(X)) = [E(g_{ij}(X))] = [E((X_i - \mu_i)(X_j - \mu_j))] = [\mathrm{cov}(X_i, X_j)].$$

  This matrix is called *covariance matrix* of X, and is denoted by the $k \times k$ matrix Σ, i.e. $\Sigma = [\sigma_{ij}]$, with $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$. By construction, Σ is symmetric and positive semi-definite. The latter means that for every vector $a \in \mathbb{R}^k$, $a^{\mathrm{T}}\Sigma a = \mathrm{var}(aX) \geq 0$.

We define the *correlation matrix* $P = [\rho_{ij}]$, with $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}}$. In matrix notation

$$P = \Delta^{-1/2}\Sigma\Delta^{-1/2}, \quad \text{with} \quad \Delta = \mathrm{diag}(\Sigma) = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2).$$

The following two properties extend those of Theorem 1.2 to $k \times 1$ random vectors.

Theorem 1.3 *Let X be a k-dimensional r.ve. with mean vector μ and covariance matrix Σ. Furthermore, let $A = [a_{ij}]$ be a $n \times k$ matrix and $b = (b_1, b_2, \ldots, b_n)$ a vector. If $Y = AX + b$, then Y is a r.ve. and*

(i) $E(Y) = A\mu + b$

(ii) $\mathrm{var}(Y) = A\Sigma A^{\mathrm{T}}$.

Proof: (*i*): Let $g(X) = AX + b$, $C = A\mu + b$. Note that $g(X)$ and C are $n \times 1$ vectors with ith element equal to $g_i(X) = a_{i1}X_1 + \ldots + a_{ik}X_k + b_i$ and $c_i = a_{i1}\mu_1 + \ldots + a_{ik}\mu_k + b_i$, respectively. Thus

$$E(g(X)) = [E(g_i(X))] = [E(a_{i1}X_1 + \ldots + a_{ik}X_k + b_i)] = [a_{i1}\mu_1 + \ldots + a_{ik}\mu_k + b_i] = [c_i] = C.$$

(*ii*): Let $g(X) = [AX + b - E(AX + b)][(AX + b - E(AX + b)]^{\mathrm{T}}$ and $D = A\Sigma A^T$. Applying point (*i*) and some algebra we get $g(X) = A(X - \mu)(X - \mu)^{\mathrm{T}}A^{\mathrm{T}}$, an $n \times n$ matrix. Now

$g(X) = [g_{ij}(X)]$ and $D = [d_{ij}]$, and after some algebra we get

$$g_{ij} = \sum_{s=1}^{k}\sum_{t=1}^{k} a_{is}a_{tj}(X_s - \mu_s)(X_t - \mu_t), \quad d_{ij} = \sum_{s=1}^{k}\sum_{t=1}^{k} a_{is}a_{tj}\sigma_{st}.$$

Thus

$$\begin{aligned} E(g(X)) = [E(g_{ij}(X))] &= \left[E\left(\sum_{s=1}^{k}\sum_{t=1}^{k} a_{is}a_{tj}(X_s - \mu_s)(X_t - \mu_t)\right)\right] \\ &= \left[\sum_{s=1}^{k}\sum_{t=1}^{k} a_{is}a_{st}E\left((X_s - \mu_s)(X_t - \mu_t)\right)\right] \\ &= \left[\sum_{s=1}^{k}\sum_{t=1}^{k} a_{is}a_{tj}\sigma_{st}\right] = [d_{ij}] = D. \end{aligned}$$

■

1.1.5 Transformation of random vectors

Also the transformation of a r.ve. X leads to a r.ve. Y which distribution under appropriate smoothness conditions can be easily determined.

We use the notation: if $a_1, a_2, \ldots, a_n$ are $1 \times m$ vectors, then $A = [a_1|a_2|\cdots|a_n]$ denotes the $m \times n$ matrix with $A = [a_{ij}]$, where a_{ij} is the ith element of a_j, i.e. A is formed by stacking the vectors $a_1, \ldots, a_n$ row-wise.

Theorem 1.4 *Let $X = (X_1, \ldots, X_k)$ a r.ve. with p.d.f $f_X(x_1, \ldots, x_k)$ and let $g(x) = (g_1(x), \ldots, g_k(x))$, with $g : R^k \to R^k$ be a bijective and differentiable vector-valued function with inverse $g^{-1}(y) = (g_1^{-1}(y), \ldots, g_k^{-1}(y))$. Then $Y = g(X)$ is a r.ve. with p.d.f.*

$$f_Y(y) = f_X(g^{-1}(y))|\det(J(y))|,$$

where the Jacobian of the transformation is

$$\det(J(y)) = \det\left(\left[\frac{dg^{-1}(y)}{dy_1}\middle|\frac{dg^{-1}(y)}{dy_2}\middle|\cdots\middle|\frac{dg^{-1}(y)}{dy_k}\right]\right),$$

and $\frac{dg^{-1}(y)}{dy_i} = (g_1^{-1}(y)/\partial y_i, \ldots, g_k^{-1}(y)/\partial y_i)$ is a column vector-valued function.

1.2 Two notable examples of random vectors

1.2.1 The multinomial distribution

Consider a *generalised Bernoulli experiments* in which we have a sequence of n independent experiments, and on each experiment, the result is exactly one of the k possibilities $b_1, b_2, \ldots, b_k$. On a given trial let b_i occur with probability θ_i, $i = 1, \ldots, k$, with $\theta_i > 0$ and $\sum_{i=1}^k \theta_i = 1$.

On n independent experiments, we take $\mathcal{S}$ equal to the set of all k^n ordered sequences of length n with components $b_1, b_2, \ldots, b_k$; for example if $s = b_1 b_3 b_2 b_2 \cdots b_k$, then on trial 1 occurs b_1, on trial 2 occurs b_3, on 3 and 4 occurs b_2 and so on, on the last, i.e. nth, trial occurs b_k. To the point

$$s = \underbrace{b_1 b_1 \cdots b_1}_{y_1} \underbrace{b_2 \cdots b_2}_{y_2} \cdots \underbrace{b_k \cdots b_k}_{y_k},$$

we assign probability $\theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$. This is the probability assigned to any sequence having y_i occurrences of b_i, $i = 1, \ldots, k$. The number of such sequences is given by the multinomial coefficient $\frac{n!}{y_1! y_2! \cdots y_k!}$, with $n = \sum_{i=1}^k y_i$. Letting Y_i denote the r.v. which counts the occurrences of b_i, $i = 1, \ldots, k$, we have that

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = \tfrac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k},$$

and the r.ve. $Y = (Y_1, \ldots, Y_k)$ with the above p.d.f. is called multinomial r.ve., denoted $Y \sim \mathrm{Mn}(n; \theta_1, \ldots, \theta_k)$. The multinomial distribution has $k-1$ parameters $\theta_1, \ldots, \theta_{k-1}$, because $\theta_k = 1 - (\theta_1 + \ldots + \theta_{k-1})$.

To fix ideas consider the following example.

Example 1.1 *In an experiment in which four unbiased dice are thrown independently, find the probability of exactly two 1's and one 2.*

The the trial is given by the throw of a single dice. We first have to figure out k. Since we are only interested in 1, 2 *and everything else that is different from 1 and 2, the possibilities* *for each single trial are,*

$$\begin{aligned} b_1 &= \text{“1 occurs”} \quad \theta_1 = \tfrac{1}{6}, \quad y_1 = 2, \\ b_2 &= \text{“2 occurs”} \quad \theta_2 = \tfrac{1}{6}, \quad y_2 = 1, \\ b_3 &= \text{“3,4,5, or 6 occurs”} \quad \theta_3 = \tfrac{4}{6}, \quad y_3 = 1. \end{aligned}$$

The required probability is thus

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \tfrac{4}{2!1!1!}(1/6)^2(1/6)^1(4/6)^1.$$

The multinomial distribution is useful for modelling categorical variables which can assume one of the k possible values, such as for instance, preference choices against k transport options, or k political parties, etc.

The multinomial distribution has many interesting properties, here are some useful ones.

Theorem 1.5 *Let $Y = (Y_1, \dots, Y_k) \sim \mathrm{Mn}(n; \theta_1, \dots, \theta_k)$, then:*

(i) For $k = 2$, the multinomial distribution coincides with the binomial distribution, i.e. $\mathrm{Mn}(n; \theta_1, \theta_2) = \mathrm{Bin}(n, \theta)$.

(ii) Each component of the r.ve. Y is a binomial r.v., i.e $Y_i \sim \mathrm{Bin}(n, \theta_i)$, for all $i = 1, \dots, k$.

(iii) Every d-subvector $(Y_{i_1}, \dots, Y_{i_d})$ of Y, $d \leq k$ has a multinomial distribution, where $\{i_1, \dots, i_d\} \subseteq \{1, 2, \dots, k\}$.

(iv) If $X \sim \mathrm{Mn}(n_x; \theta_1, \dots, \theta_k)$ and $Z = Y + X$, then $Z \sim \mathrm{Mn}(n_z; \theta_1, \dots, \theta_k)$, with $n_z = n + n_x$.

(v) Let $X_1, \dots, X_k$ be independent Poisson r.v.'s with the d.f. of X_i having parameter λ_i. Then the conditional distribution of $X_1, \dots, X_k$ given their sum is multinomial:

$$P\left(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k \middle| \sum_{i=1}^{k} X_i = n\right) = \mathrm{Mn}(n; \theta_1, \dots, \theta_k),$$

where $\theta_i = \lambda_i / \sum_{j=1}^{k} \lambda_j$, $i = 1, \dots, k$.

(vi) $E(Y) = (n\theta_1, \dots, n\theta_k)$ *and for all* $i, j = 1, \dots, k$

$$\mathrm{cov}(Y_i, Y_j) = \begin{cases} -n\theta_i\theta_j & \textit{if } i \neq j \\ n\theta_i(1-\theta_j) & \textit{if } i = j. \end{cases}$$

1.2.2 The multivariate normal distribution

This is our second probability distribution for random vectors. Just like the multinomial distribution can be seen as a multivariate extension of the binomial distribution, the multivariate normal distribution extends the normal distribution in the random vector case.

There are different ways for introducing the multivariate normal distribution. Here we follow a bottom-up approach in which we construct the multivariate normal from independent standard normal r.v.s. Let $Z_i \sim \mathrm{N}(0,1), i = 1, \dots, p$, be independent r.v.. The r.ve. $(Z_1, \dots, Z_p)$ has joint p.d.f

$$\begin{aligned} f_{Z_1,\dots,Z_p}(z_1, \dots, z_p) &= \prod_{i=1}^{n} f_{Z_i}(z_i) \\ &= \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{p} z_i^2} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}z^{\mathrm{T}}z}, \quad z = (z_1, z_2, \dots, z_p). \end{aligned}$$

We say then that the r.ve $Z = (Z_1, \dots, Z_p)$ has *standard multivariate normal distribution*, denoted by $Z \sim \mathrm{N}_p(0, I_p)$, where I_p is the unit diagonal matrix and 0 here denotes the p vector of zeros. Note that by construction $E(Z_i) = 0$ and $\mathrm{cov}(Z_i, Z_j) = 0$ for $i \neq j$ and $\mathrm{cov}(Z_i, Z_i) = \mathrm{var}(Z_i) = 1$, for all $i, j = 1, \dots, p$.

But we are not done yet, what we constructed so far is only one multivariate normal distribution, i.e. the standard one, and we aim at the family of multivariate normals. To this end, let A be $(p \times p)$ matrix for which A^{-1} exists and let $\mu \in \mathbb{R}^p$. Consider the function $g : \mathbb{R}^p \to \mathbb{R}^p$ given by $g(Z) = Y = AZ + \mu$ which has inverse $g^{-1} : \mathbb{R} \to \mathbb{R}$ given by

$g^{-1}(Y) = Z = A^{-1}(Y - \mu)$. Then by Theorem 1.4 we have

$$f_Y(y) = f_Z(g^{-1}(y)) \left| \det \left(\tfrac{dg^{-1}(y)}{dy} \right) \right|,$$

where the now $y \in \mathbb{R}^p$ and $\det \left(\frac{dg^{-1}(y)}{dy} \right)$ denotes the determinant of the Jacobian matrix associated with the transformation. In particular, it holds that

$$\tfrac{dg^{-1}(y)}{dy} = A^{-1}.$$

Putting all the pieces together and letting $\Sigma = AA^{\mathrm{T}}$, such that $\det(\Sigma) = \det(AA^T) = \det(A)^2$ and $\det(\Sigma)^{-1/2} = \det(A)^{-1}$, we have the p.d.f.

$$\begin{aligned} f_Y(y) &= f_Z(g^{-1}(y)) \left| \det \left(\tfrac{dg^{-1}(y)}{dy} \right) \right| \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\left[A^{-1}(y-\mu)\right]^{\mathrm{T}}\left[A^{-1}(y-\mu)\right]} \det(A^{-1}) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)}. \end{aligned}$$

It follows that $E(Y) = \mu$ and $\mathrm{cov}(Y_i, Y_j) = \sigma_{ij}$, where σ_{ij} is the elements on the ith row and jth column of Σ. The r.ve. Y with the above distribution is called the *p-variate normal distribution* with mean vector μ and covariance matrix Σ, denoted $Y \sim \mathrm{N}_p(\mu, \Sigma)$.

Here are some useful properties of the multivariate normal distribution.

Theorem 1.6 *Let* $Y \sim \mathrm{N}_p(\mu, \Sigma)$. *Then*

(i) If $X = BY + b$, *with* B *a* $p \times p$ *matrix and* $b \in \mathbb{R}^p$, *then* $X \sim \mathrm{N}_p(B\mu + b, B\Sigma B^{\mathrm{T}})$.

(ii) $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, *where* $\mu_i = E(Y_i)$ *and* $\sigma_i^2 = \mathrm{var}(Y_i)$, $i = 1, \ldots, p$.

(iii) All the conditional distributions involving components of Y *are normal with suitable parameters.*

1.3 Convergence of random variables

Given a sequence of r.v.s $X_1, X_2, \ldots$, we may be interested in its ”limit”, i.e. the behaviour of the sequence as n diverges. There are different types of convergence, but here we give the definitions of four of them.

Hereafter, the notation $X \sim F_X$ means: the r.v. X has distribution F_X.

The following definition will be extensively used.

Definition 1.2 *The r.v.s $X_1, \ldots, X_n$ are called identically and independently distributed (i.i.d.) if $X_i \sim F_X$ (dentically distributed) and $X_1, \ldots, X_n$ are independent (see Theorem 1.1). The notation for this is $X_i \overset{\text{iid}}{\sim} F_X$ for all $i = 1, \ldots, n$.*

1.3.1 Modes of convergence

Unless stated otherwise, $X_1, X_2, \ldots$ denotes a sequence of random variables.

Definition 1.3 *$X_1, X_2, \ldots$ is said to be convergent in quadratic mean to a r.v. X if*

$$\lim_{n\to\infty} E(X_n - X)^2 = 0.$$

This type of convergence is denoted by $X_n \xrightarrow{\text{q.m.}} X$.

Definition 1.4 *$X_1, X_2, \ldots$ is said to be almost surely (abbreviated a.s.) convergent to a r.v. X if*

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence is denoted by $X_n \xrightarrow{\text{a.s.}} X$.

Definition 1.5 *$X_1, X_2, \ldots$ is said to be convergent in probability to a r.v. X if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1 \quad or \quad \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

This type of convergence is denoted by $X_n \xrightarrow{P} X$.

Definition 1.6 *$X_1, X_2, \ldots$, with each term having d.f. F_n, is said to be convergent in distribution to a r.v. X with d.f. F if*

$$\lim_{n\to\infty} F_n(t) = F(t), \quad \textit{for all } t \textit{ where } F \textit{ is continuous.}$$

This type of convergence is denoted by $X_n \xrightarrow{d} X$.

1.3.2 Algebra of sequences of random variables

The following result shows connections between the four types of convergence.

Theorem 1.7 *Let $\{X_n\}$ be a sequence of r.v.'s with each term having d.f. F_n, and let X be an r.v. with d.f. F. Then*

(i) If $X_n \xrightarrow{\text{q.m.}} X$ then $X_n \xrightarrow{P} X$.

(ii) If $X_n \xrightarrow{\text{q.m.}} X$ then $X_n \xrightarrow{d} X$.

(iii) If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{P} X$.

(iv) If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{d} X$.

(v) If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{d} X$.

Convergence in probability and convergence in distribution are the most widely used in statistics. The following special case of $X_n \xrightarrow{P} X$ when X is constant is particularly interesting in statistics and we state it as a separate theorem.

Theorem 1.8 *(Weak Law of Large Numbers) Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s with common mean μ and finite variance σ^2. Then*

$$\overline{X}_n = \tfrac{X_1+\cdots+X_n}{n} \xrightarrow{P} \mu, \quad \textit{as } n \to \infty.$$

Proof: In L0, Theorem 0.7 we saw that $E(\overline{X}_n) = \mu/n$ and $\text{var}(\overline{X}_n) = \sigma^2/n$. Applying the Chebyshev inequality we have for any $\epsilon > 0$

$$P(|\overline{X}_n - \mu| \geq \epsilon) = P\left(|\overline{X}_n - \mu| \geq \tfrac{\sigma}{\sqrt{n}}\tfrac{\sqrt{n}}{\sigma}\epsilon\right) \leq \tfrac{\sigma^2}{n\epsilon}.$$

Applying limit on both sides of the inequality leads to the desired result

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \tfrac{\sigma^2}{n\epsilon} = 0.$$

■

Here are some further properties worth knowing.

Theorem 1.9 *(Slutsky's Lemma) Let $X_1, X_2, \ldots$ be such that $X_n \xrightarrow{d} X$ and let $Y_1, Y_2, \ldots$ be another sequence of r.v.'s such that $Y_n \xrightarrow{P} c$, with $c \in \mathbb{R}$ fixed. Then*

(i) $X_n Y_n \xrightarrow{d} cX$.

(ii) $X_n + Y_n \xrightarrow{d} X + c$.

(iii) $X_n / Y_n \xrightarrow{d} X/c$*, provided* $P(Y_n \neq 0) = 1$, $c \neq 0$.

(iv) If $g(\cdot)$ is a continuous function, then $g(Y_n) \xrightarrow{P} g(c)$.

There is also a more general version of Theorem 1.9(iv) which states that if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$ for some suitable function g; a similar results exists for convergence in probability.

The following results will be used in the incoming lectures.

Theorem 1.10 *(Delta Method) Let $X_1, X_2, \ldots$ be such that*

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Suppose that $g : \mathbb{R} \to \mathbb{R}$ is differentiable at μ and $g'(\mu) \neq 0$. Then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2 (g'(\mu))^2).$$

Theorem 1.11 *(Multivariate Delta Method) Let $X_1, X_2, \ldots$ be a sequence p-dimensional random vectors such that*

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N_p(0, \Sigma).$$

Suppose that $g(x) = (g_1(x), \ldots, g_k(x))$*, with* $g : \mathbb{R}^p \to \mathbb{R}^k$*,* $k \leq p$*, is such that the the* $k \times p$ *matrix of partial derivatives*

$$B = \left[\frac{\partial g_i(x)}{\partial x_j} \right],$$

are continuous and do not vanish in a neighbour of μ*. Let* $B_\mu = B$ *evaluated at* μ*. Then*

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, B_\mu \Sigma B_\mu^{\mathrm{T}}).$$

1.3.3 The Central Limit Theorem

We now formulate the celebrated Central Limit Theorem (CLT) in its simplest form.

Theorem 1.12 *Let* $X_1, \ldots, X_n$ *be i.i.d. r.v.'s, with finite mean* μ *and finite variance* $\sigma^2 > 0$*. Then*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1). \qquad (1.1)$$

If we let $G_n = P\left[\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq x \right]$ *and* $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$*, then* (1.1) *is equivalent to*

$$\lim_{n \to \infty} G_n(x) = \Phi(x), \quad \textit{for every } x \in \mathbb{R}.$$

The CLT says that for large n, the d.f. of the standardised sums $\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}$ is close to $\Phi(x)$, where $S_n = \sum_{j=1}^{n} X_j$. What "large" and "close" mean is not easy to specify as the value of n and a degree of closeness depend on the nature of the common distribution of the X_n.

Example 1.2 *Let* $Y \sim \mathrm{Bin}(n, \theta)$*. Since* $Y = \sum_{i=1}^{n} X_i$*, where* $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$*, then we can apply the Central Limit Theorem in order to approximate the d.f. of* Y*. For instance suppose* $Y \sim \mathrm{Bin}(8, 0.5)$ *and let us calculate* $P(Y \leq 5)$*. From the binomial distribution we have*

$$P(Y \leq 5) = \sum_{i=0}^{5} \binom{8}{i} 0.5^i (1 - 0.5)^{(8-i)} = 0.8555.$$

By CLT, $\frac{Y-n\theta}{n\theta(1-\theta)} \xrightarrow{d} \mathrm{N}(0,1)$*, thus*

$$\begin{aligned} P(Y \leq 5) &= P\left(\frac{Y-E(Y)}{\sqrt{\mathrm{var}(Y)}} \leq \frac{5-E(Y)}{\sqrt{\mathrm{var}(Y)}}\right) \\ &= P\left(\frac{Y-E(Y)}{\mathrm{var}(Y)} \leq \frac{1}{\sqrt{2}}\right) = G_n\left(\frac{1}{\sqrt{2}}\right) \\ &\to \Phi\left(\frac{1}{\sqrt{2}}\right) = 0.7604, \quad as \quad n \to \infty, \end{aligned}$$

since $E(Y) = 8 \cdot 0.5 = 4$*,* $\mathrm{var}(Y) = 8 \cdot 0.5 \cdot 0.5 = 2$*. Here the approximation in not particularly accurate since n is quite low. The quality of the approximation depends also on* θ*;* θ *close to the boundary makes convergence slower.*

Example 1.3 *Let* $Y_n \sim \mathrm{Poi}(n)$ *and note that,* $Y_n = \sum_{i=1}^n X_i$*, where* $X_i \overset{\mathrm{iid}}{\sim} \mathrm{Poi}(1)$*. It follows that*

$$Y_n \xrightarrow{d} \mathrm{N}(n, n).$$

Example 1.4 *Let* $Y_i \overset{\mathrm{iid}}{\sim} \mathrm{Ga}(\alpha_0, \lambda)$*,* $i = 1, \ldots, n$*. Then*

$$\frac{\sum_{i=1}^n Y_i - n\frac{\alpha_0}{\lambda}}{\frac{\sqrt{n\alpha_0}}{\lambda}} \xrightarrow{d} \mathrm{N}(0,1)$$

and thus $\sum_{i=1}^n Y_i \xrightarrow{d} \mathrm{N}\left(\frac{n\alpha_0}{\lambda}, \frac{n\alpha_0}{\lambda^2}\right)$*. Note that the exact distribution of the sum of* Y_i *has exact form valid for any n, given by*

$$\sum_{i=1}^n Y_i \sim \mathrm{Ga}(n\alpha_0, \lambda).$$

References

[HMC20] Hogg, R. V., McKeen, J. W. and Craig, A. T. (2018) *Introduction to Mathematical Statistics* (8th edition, global ed.), Pearson Education, Chapp. 3, and 5.