An Intermediate Course in Mathematics and Statistics

Manuel Lleonart-Anguix

Barcelona School of Economics
Master of Data Science for Decision Making

September, 2023

These are the lecture notes for the Brush-Up course: An Intermediate Course in Mathematics and Statistics of the Master in Data Science for Decision Making at the Barcelona Graduate School of Economics. Written by Manuel Lleonart-Anguix. The year 2021.

Contents

Part I

Mathematics

L. 1 Introduction to mathematical notation

This section shows the basic notation and tools we will work on within the course. This material will not be covered in the lectures and should be read in advance by the students.

L. 1.1 Sets of numbers

A set is a collection of different elements. These collections can be ordered or unordered. For example, the students of DSDM, the natural numbers, or the mountains of Italy are examples of sets. We are interested in a particular type of set, the subsets of the real numbers.

- The numbers that arise of the counting $1, 2, 3, ...$ are called natural numbers, noted $\mathbb{N}$. Some mathematicians decide to include 0 or not.
- Adding the negatives, we have the integer numbers, noted as $\mathbb{Z}$[1]. All the natural numbers are integers, but the opposite is not true.
- The set of all the numbers that we can express as $\frac{p}{q}$ with $p,\ q \in \mathbb{Z}$ is called the set of rational numbers, noted $\mathbb{Q}$. Notice that this set also contains the previous two.
- The largest set of numbers that we will see within this brush-up is the set of the real numbers, $\mathbb{R}$, which are represented by a finite or infinite quantity of decimals. This set is not equivalent to $\mathbb{Q}$. For example, the number $\sqrt{2}$ is a real number but not rational.

Before continuing, I will note two considerations about these sets. First, these are not **all** the sets of numbers that exist. For example, $\sqrt{-5}$ is not in any of these sets. Second, there are other classifications of numbers, but this is the most useful for all non-mathematicians.

In your study of the roots of polynomials (for example, in matrix theory), you can find the previously mentioned roots of negative numbers. Most of the mathematics courses for non-mathematics bachelors don't cover these numbers. For a brief introduction to the world of complex numbers, I recommend the first chapter of (Gamelin, 2000). We can define operations among the elements of a set. With this, we have different algebraic structures as groups or fields (that must fulfill some properties).

[1] For the German word *Zahlen*.

L. 1.2 Some symbols

In mathematics, we don't use the English language for every sentence. Instead, mathematicians use their logic language. Here, I show the symbols that we will use in the course.

To say that one element belongs to a set, we use the operator $\in$. For example, $x \in X$ means that the element x belongs to the set X. To indicate the opposite, we use the symbol $\notin$.

$$1 \in \mathbb{N}, \quad \text{but } \sqrt{2} \notin \mathbb{N}$$

When all the elements of one set, A, belong to another, B, we say that B includes in A. We write $A \subseteq B$. If there is an element in A that doesn't belong to B, we say that B doesn't include A. In this case, we write $A \nsubseteq B$.

The symbols $\cup$ and $\cap$ denote the union and intersection of two sets, respectively. Let A and B be two sets. $A \cup B$ is the set of all the elements that belong to A or B. $A \cap B$ is the set of all the elements that belong to A *and* B. Formally, we write:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$
$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

Finally, we will use logical symbols. The most important are:

- Exists $\exists$.
- Exists, and it is unique $\exists!$
- Implies $\Rightarrow$. Imagine that we have two predicates. For example, $A =$"I go to the grocery" and $B =$ "It's Monday."

  In the sentence, every Monday I go to the grocery, we have that if it's Monday, I will go for sure to the grocery. We can write it as

  $$B \Rightarrow A$$

  However, I might go on other days to the grocery. Therefore, the opposite is not necessarily correct.

$$B \Leftarrow A$$

If I am in the grocery, it might be Monday or not. When both predicates are equivalent, we use $\iff$. For example, in the sentence I will eat chocolate if and only if I am ill.

$$\text{I am ill} \iff \text{I eat chocolate}$$

We can refer to the number of elements a set A has as the cardinality of the set, $|A|$. For example, for the set $A = \{a, b, c\}$, its number of elements is $|A| = 3$.

L. 1.3 Sums and products

Sometimes we want to express sums or products that are too long to write. Imagine that we want to write the sum of the first 100 numbers. Instead of writing

$$1 + 2 + 3 + 4 + 5 + 6 + ... + 100$$

We can write it using the symbol $\sum$. We write

$$\sum_{i=1}^{100} i = 1 + 2 + 3 + 4 + 5 + 6 + ... + 100$$

It reads as *the sum from i equal one to i equal to one hundred of i.* With the product, there is something similar. We use $\prod$.

$$\prod_{i=1}^{100} i = 1 \cdot 2 \cdot 3 \cdot 4 \cdot ... \cdot 100$$

Notice that in both cases, instead of using i, we can use any expression that depends on i.

L. 2 Matrix Algebra

The content that I show in these notes can be found in the appendix of almost every handbook of intermediate econometrics. For a further extension of matrix properties I recommend you to check the **Matrix Cookbook**.

L. 2.1 Linear algebra

Linear algebra begins with the study of linear equations and the seek for their solutions.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & ... & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & ... & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \ddots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & ... & + & a_{mn}x_n & = & b_m \end{cases} \tag{1}$$

where a_{ij} and b_i are known and x_i are unknowns. You might remember from high school that the system (1) can be solved using matrices. The existence and the number of solutions of the system depend both on the independent terms b_i and the coefficients of the system a_{ij}. In linear algebra, we work with different objects such as matrices, vectors, and linear functions.

Linear algebra belongs to the basis of other mathematical fields as statistics and is deeply related to analysis and geometry. In particular, matrices are fundamental in the study of data, as they are a useful tool to store, classify and work with data. That is the reason why we will study matrices in this course.

L. 2.2 Vectors

Vectors are important objects in almost every field of mathematics. In this course, we will limit to study $\mathbb{R}^n$. A vector in $\mathbb{R}^n$, call it $v \in \mathbb{R}^n$ is an ordered n-tuple of real numbers $v = (v_1, v_2, ..., v_n)$, where $v_i \in \mathbb{R}$ for all $i = 1, 2, ..., n$.

We define two operations in $\mathbb{R}^n$

- **Addition of vectors.** Let $v,\ w \in \mathbb{R}^n$, we define the addition of vectors as

$$v + w = (v_1 + w_1, v_2 + w_2, ..., v_n + w_n)$$

where $+$ is the standard addition in $\mathbb{R}$.

With these two operations, $\mathbb{R}^n$ is a space vector.[2]

- **Scalar multiplication.** Let $\alpha \in \mathbb{R}$, we define the scalar multiplication as

$$\alpha v = (\alpha v_1, \alpha v_2, ..., \alpha v_n)$$

L. 2.3 Matrices. Types and properties

Let n and $m \in \mathbb{N}$. A **matrix** is a collection (array) of mn elements organized in rows and columns. A in equation (2) defines a matrix of m rows and n columns.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \tag{2}$$

We will say that A is an $m \times n$ (n by m) matrix, where a_{ij} denotes the element of the i-th row and the j-th column, also known as entries. We can also say that A is of **order** (or size) $m \times n$. We can refer to the i-th row as $\{a_{ij}\}_{i=1}^{m}$ or as $\{a_{ij}\}_{1 \leq i \leq m}$. Similarly for the columns. The set of all the $m \times n$ matrices with elements in K[3] is defined as $M_{m \times n}(K)$.

Matrix is one of the most useful tools in any data-related work. They help us to present data and to work with it. We can use matrices to present the results of surveys, important macroeconomic variables from countries, the nexus between friends in a group, coefficients in a linear system of equations... Therefore, understanding how to work with matrices will facilitate our work as data scientists.

Types of matrices

- **Square matrix.** If in (2) $n = m$, we say that A is a square matrix.

[2] We will not go deep in this concept as escapes from our objectives. However, you can think of a space vector as a set of elements with two operations that meet some conditions.

[3] Here K refers to a group (or field) of elements. You can think of the natural numbers, $\mathbb{N}$, or the real numbers, $\mathbb{R}$. But matrices can be defined over any group. However, we will only work with matrices on $\mathbb{R}$.

- **Column matrix.** If $n = 1$, we say that A is a column matrix.
- **Row matrix.** If $m = 1$, A is a row matrix.

  Notice that column and row matrices correspond to vectors.
- **Null matrix.** A matrix is null if $a_{ij} = 0$ for all $1 \leq i \leq m$ and $1 \leq j \leq m$.

$$0 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

- **Symmetric matrix.** A matrix is symmetric if $a_{ij} = a_{ji}$ for all $1 \leq i \leq m$ and $1 \leq j \leq m$.

  Notice that a necessary condition for a matrix to be symmetric is that it is a square matrix.
- **Diagonal matrix.** A matrix is diagonal if it is a square matrix and $a_{ij} = 0$ whenever $i \neq j$. We say that $(a_{ii})_{i=1}^{n}$ is the main diagonal of the matrix A.

  Notice that every diagonal matrix is also symmetric.

  We define the **trace** of a matrix as the sum of all the elements in the diagonal. Formally,

$$trace(A) = \sum_{i=1}^{n} a_{ii}$$

- **Identity matrix.** The identity matrix I is a square matrix such that $a_{ii} = 1$ and $a_{ij} = 0$ for $i \neq j$.

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

  The identity matrix is an example of a diagonal and symmetric matrix.
- **Triangular matrix.** An upper triangular matrix has all its elements below the diagonal equal to zero. A lower triangular matrix has all its elements above the diagonal equal to zero. A diagonal matrix is both upper triangular and lower triangular.

When working with data, we will both use square matrices and non-square matrices. The GDP of three countries in a hundred years can be presented in a matrix of size 3×100.

Two matrices A and B are equal if $a_{ij} = b_{ij}$ for ever i and j. We can operate with the elements of $M_{m\times n}(\mathbb{R})$.

Transposition of matrices

Given a matrix A, we define its **transpose**, and call it A', as the matrix with entries $a'_{ji} = a_{ij}$. This means that we replace the rows of A by its columns. Notice that a matrix A is symmetric if $A = A'$. If A is a matrix $n \times m$, A' will be a matrix $m \times n$.

We will study three operations with matrices: addition, scalar product, and product. However, there are several operations with matrices that we will not cover in this course and might be useful, as, for instance, the *Kroneker* product, the *Hadamard* or the *vec* operator.

Addition and subtraction of matrices.

As we are used to in the real numbers, we can also add and subtract matrices that are of the same order. Notice that it is impossible to add to matrices when they have different sizes. As with the real numbers, we will note the matrix addition with $+$ and the matrix subtraction with $-$.

The properties of the matrix addition are the following:

- **Neutral element**. The null matrix is the neutral element for the addition of matrices.

$$A + 0 = A$$

- **Symmetry**.

$$A + B = B + A$$

- **Commutativity**. It does not matter the order of the additions.

$$(A + B) + C = A + (B + C)$$

- **Opposed element**. For every matrix A there exist another element $-A$ such that,

$$A - A = 0$$

Exercise 1.

$$A = \begin{pmatrix} 1 & 3 & 4 \\ 3 & 2 & 3 \\ 4 & 3 & 5 \end{pmatrix}, \quad B = \begin{pmatrix} 5 & 8 & 1 \\ 0 & 3 & 4 \\ 0 & 0 & 1 \end{pmatrix}$$

1. *What is the rank of matrices A and B?*
2. *Is it possible to add them? Why or why not? If it's possible, add them. What happens with $A - B$?*
3. *What is the second row of matrix A? And the third column of matrix B?*
4. *Can you identify matrix A with some of the types described in the previous section?*

$$A' = \begin{pmatrix} 1 & 3 & 4 \\ 3 & 2 & 3 \end{pmatrix}$$

Answer to the same questions with matrix A' instead of A.

The product of a matrix and a scalar.

We can multiply a matrix times a real number. Let $\lambda \in \mathbb{R}$. We define the product of the scalar λ times the matrix A as the matrix which entries are λ times the entries of A.

$$\lambda A = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \cdots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \cdots & \lambda a_{mn} \end{pmatrix}$$

Matrix multiplication.

Unlike with the addition of matrices, we can multiply two matrices that do not have the same rank. If the columns of matrix A coincide with the rows of matrix B, we can multiply

$A \cdot B$. However, $A \cdot B$ might be not well defined. In general, we will omit the sign $\cdot$ and just note AB to refer to the product. The entry $i,\ j$ of the matrix AB is equal to

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Here you can see the reason why we need the columns of matrix A to coincide with the rows of matrix B. The resulting matrix AB will have the same rows of A and the same columns of B.

Notice that, in general, $AB \neq BA$. Can you find an example of two matrices such that AB is different from BA? And an example of two matrices such that $AB = BA$?

Observe that given the matrix multiplication, we can calculate the multiplication of a scalar and a matrix as

$$\begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{pmatrix} A = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \cdots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \cdots & \lambda a_{mn} \end{pmatrix}$$

If we only want to multiply the j-th column of a matrix times an scalar, we can do the following:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \vdots & \lambda & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1i} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda a_{j1} & \lambda a_{j2} & \vdots & \lambda a_{ji} & \cdots & \lambda a_{jn} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mi} & \cdots & a_{mn} \end{pmatrix}$$

Can you think of how to multiply a row of a matrix by a scalar using matrix notation?

Exercise 2. *Solve the following questions about matrix multiplication.*

1. *Can you multiply AB? And BA? Why? Why not? Do it whenever it's possible.*

2. *Now solve the previous question with A' instead of A.*

In matrix multiplication, the element I works as a neutral, as $AI = IA = A$. We can try to find an inverse element for a given matrix A as we did with the addition of matrices. However, not every matrix has an inverse for the matrix product. This will be discussed in the following section.

A natural extension to matrix multiplication is the power matrix. We note the n-th power of matrix A as A^n. Where

$$A^n = A \cdot A \cdot ... \cdot A$$

We will say that a matrix is **idempotent** whenever $A^2 = A$. Can you think of an example of an idempotent matrix?

Matrices as linear transformations

A **linear transformation**, T, is an application (a function) between two vector spaces such that:

- $T(u + v) = T(u) + T(v)$.
- $T(\lambda u) = \lambda T(u)$

where u, v are vectors and λ is a scalar. Most of the matrix theory can be built up through linear transformations. We can say that there exists a correspondence one-to-one between linear transformations and matrices.[4] Let $u = (u_1, u_2, u_3)'$ and

$$T(u) = (2u_3 - u_1, u_1 + u_2 + u_3)' \tag{3}$$

a linear transformation. Notice that T maps from $\mathbb{R}^3$ to $\mathbb{R}^2$; it takes a vector of dimension three and gives a vector of dimension two. We can use the matrix product to represent T. If we define matrix A as

$$A = \begin{pmatrix} -1 & 0 & 2 \\ 1 & 1 & 1 \end{pmatrix}$$

[4] This means that given a linear transformation there is a matrix that represents it.

we can say that

$$T(u) = Au$$

Can you prove it? Get a generic vector $u = (u_1, u_2, u_3)'$ and multiply it by matrix A. Then, show that Au gives the same result as $T(u)$ in equation (3).

Hence, matrices not only work as structures of data. They can represent several things. We have seen two more: linear applications and coefficients of a system, which are highly related.

L. 2.4 Determinants and regular matrices

For this section, I will assume that every matrix is square unless otherwise is specified. As we mention in the previous section, is not possible to find an inverse for every matrix. The subset of the matrices that have an inverse is called the general linear group, and its elements, the regular matrices. More formally, we say that a matrix A is regular whenever there exists another matrix B such that

$$AB = I = BA$$

We usually note matrix B as A^{-1}, representing the inverse.

One way to study if a matrix is invertible is by looking for its inverse. For example, we can use the Gauss method. It consists of an augment matrix A with an identity matrix and applies transformations to $A|I$ until we have an identity in the position of A. The matrix in the position of the old I is now A^{-1}.

Exercise 3. *Can you find the inverses of matrix A and B by the Gauss' method?*

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Gauss' method is not the only way to calculate an inverse. We can also use Cramer's Rule. However, it is computationally demanding and is only feasible for a matrix of size 3.

However, there are easier ways to determine when a matrix is regular. To understand this is important to study linear combinations of vectors.

Linear combinations

Let v_1 and v_2 be two vectors (you can think of column matrices) of the same size. We say that v_2 is a linear combination of v_1 whenever exists a real number α such that

$$\alpha v_1 = v_2$$

Equivalently, given a set of n vectors $\{v_1, v_2, ..., v_n\}$ we say that v_{n+1} is a linear combination of the set if there exist real numbers $\alpha_1, \alpha_2, ..., \alpha_n$ such that

$$\sum_{i=1}^{n} \alpha_i v_i = v_{n+1}$$

This means that we can combine the vectors in the set and multiply them by real numbers to obtain the last vector. When, in a set of vectors, any vector is a linear combination of the others, we say that the set is linear independent, or equivalently, that the vectors are linear independent.

Why is it important to understand the linear combination of vectors in the study of regular matrices? Because a matrix is regular if and only if none of its columns (or rows) is a linear combination of the others.

Therefore, to study if a matrix is regular or not we can try to find linear combinations between its rows or columns. Think in the following matrix

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

It is easy to see that the first row plus two times the third row equals the second. Therefore, we don't have to study if it's invertible looking for its inverse, we can just say that its inverse doesn't exist.

We define the **rank** of a matrix as the number of rows that are not a linear combination of the others. Therefore, the maximum rank a square matrix can have is the number of rows it has. The rank is also defined for non-square matrices. In this case, the rank is the maximum number of rows or columns that are linearly independent. Notice that a matrix of size $n \times m$, with $n \leq m$ will have a rank smaller or equal to n. If the rank of a square matrix is maximum, we say that this matrix is invertible or regular, because all its rows are linearly independent. For a matrix A we note its rank as $rank(A)$.

For the example shown above, it is relatively easy to find the linear combination between the rows. Nevertheless, there is another method, that in some cases might be easier, to know when a matrix has or not inverse. This is the study of its determinant.

Determinants

Determinants are only defined for square matrices. Despite almost every student understands what a determinant is, it is not easy to properly define the determinant of a matrix without going deep into mathematical theory. As this is not the objective of our course, I will present an intuitive definition of the determinant developed by induction over the size of the matrix. I will note the determinant of matrix A as $|A|$.

If a matrix A has size one, we will say that the determinant of the matrix is equal to its entry

$$|A| = \left|\begin{pmatrix} a_{11} \end{pmatrix}\right| = a_{11}$$

If a matrix A has size two, the determinant, developed by the first column is

$$|A| = \left|\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\right| = a_{11} \cdot a_{22} - a_{21} \cdot a_{21}$$

For a matrix of size n, we say that its determinant is equal to

$$|A| = \left| \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \right| = a_{11} \cdot \left| \begin{pmatrix} a_{22} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m2} & \cdots & a_{mn} \end{pmatrix} \right| - a_{21} \left| \begin{pmatrix} a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m2} & \cdots & a_{mn} \end{pmatrix} \right| + ...$$

$$... + a_{m1} \left| \begin{pmatrix} a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m-1,2} & \cdots & a_{m-1,n} \end{pmatrix} \right|$$

First, we get the elements in the first column. The first element is a_{11}. We select the sub-matrix that removes column one and row one from matrix A and multiply the element a_{11} by the determinant of this matrix. Then, we do the same with all the other entries. To know what sign we should put, we have to consider the position of the element in the matrix if the number of the column plus the row ad to a pair number, we put a plus. In the other case, we put a minus.

It is important to understand how to calculate determinants, but for big enough matrices, programs will do it for us. We can think of the determinant as a number assigned to a matrix that we can use to know if it is regular (determinant different from zero) or not (determinant equal to zero). To check the properties of determinants, I strongly recommend you to consult the **Matrix Cookbook** (Section 1.2).

We can use the determinant to calculate the rank of a matrix. Let A be a matrix of size $m \times n$. Assume that $n \leq m$. Therefore, we can use the sub-matrices of size $n \times n$ (we call them minors) to calculate the rank. If **one** of the minors of size A, call them A_n, have determinants different from zero the rank of A is n. On the other hand, if $|A_n| = 0$ for **all** the minors of size n, we say that the rank of A is strictly smaller than n. Equally, we can do the same process for all the minors of size $n - 1$. If we find a minor of this size with a determinant different from zero, we say that $rank(A) = n - 1$. If not, we can say that $rank(A) < n - 1$. Proceeding until we find a minor of size $s \in \mathbb{N}$ which determinant is different zero, we will find that $rank(A) = s$.

One of the most important applications of the rank comes from the theorem of **Rouché-Frobenius** (also known as Rouché-Capelli). Assume that $Ax = b$ represents a system of linear equations. Where A is the matrix of coefficients, x a column matrix with the unknowns and b represents the independent terms.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & ... & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & ... & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \ddots & = & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & ... & + & a_{mn}x_n & = & b_m \end{cases}$$

Here,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

The theorem of Rouché-Frobenius states that whenever $rank(A) = rank(A|b)$ the system has at least one solution (compatible system). Moreover, if $rank(A) = n$, the number of variables, there is a unique solution (compatible determinate system). If $rank(A) = rank(A|b) < n$ the system has infinite solutions (compatible indeterminate system). If $rank(A) < rank(A|b)$ the system has no solution (incompatible system).

If the system is compatible determinate, the solution comes by $x = A^{-1}b$.

L. 2.5 Eigenvectors and eigenvalues

A lot of problems can be modeled using linear transformations (or matrices). Eigenvectors and eigenvalues make it easier to understand those transformations. Let A be a $n \times n$ matrix. We say that x is an eigenvector of matrix A iff

$$Ax = \lambda x \tag{4}$$

for some number λ. λ is known as the eigenvalue and x is the eigenvector associated with this eigenvalue. Having coefficients in $\mathbb{R}$ is not a sufficient condition for λ to be a real number. Eigenvalue theory has important implications for the span of sub-vectorial spaces that I will not cover in this course.

Notice that x is a vector of size $n \times 1$ (a column vector). Here it is interesting to study the dimension of both terms in equation (4). (**Do it!**).

$$Ax = \lambda x = \lambda I x \Longleftrightarrow Ax - \lambda I x = 0 \Longleftrightarrow (A - \lambda I)x = 0$$

The previous equality gives us a useful condition to find all the eigenvalues of a matrix. Using the properties of determinants, we have that

$$|(A - \lambda I)|x = 0$$

As this happens for every vector x, we have that

$$|(A - \lambda I)| = 0 \Rightarrow p(\lambda) = |(A - \lambda I)| \tag{5}$$

Equation (5) defines the characteristic polynomial of matrix A. Finding the roots of $p(\lambda)$ means finding the eigenvalues of A. Eigenvalues are also useful to know the rank of matrices (the rank is equal to the number of eigenvalues different from zero).[5]

Exercise 4. *Prove that the eigenvalues of matrix A are*

$$\lambda_1 \approx 6.875, \quad \lambda_2 \approx 3.526, \quad \lambda_3 \approx 1.599$$

and find the eigenvectors associated to it.

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 1 & 3 & 1 \\ -1 & -2 & 5 \end{bmatrix}$$

L. 2.6 Positive definite matrices

A special type of matrix frequently used in statistics is the positive definite or positive semi-definite. By definition, A is a positive definite matrix if it's a square matrix of size n such that for any vector $x \in \mathbb{R}^n$ we have

[5] There is a lot of matrix theory that relates determinants, eigenvalues and ranks, however, we will not cover it in this course. If you are interested in it check the references.

$$x'Ax > 0$$

except if $x = 0$. Equivalently, the matrix A is positive semi-definite if for any vector

$x \in \mathbb{R}^n$ we have

$$x'Ax \geq 0$$

Notice that in general, it's not straightforward to prove that a matrix is positive definite by definition. Instead, we can use some theorems or properties to determine if a matrix is positive definite.[6]

Two important characterizations of positive definite matrices are the following:

PD by eigenvalues. A matrix A is positive definite if and only if all of its eigenvalues are positive. A is positive semi-definite if and only if all of its eigenvalues are non-negative.

Sylvester's criterion. A $n \times n$ symmetric (Hermitian) matrix is PD if and only if all the principal minors are positive.[7]

[6] Again, Matrix Cookbook

[7] The leading principal minors are those which entries coincide with the entries of the matrix.

L. 3 Mathematical analysis

L. 3.1 Functions

A function f is a mathematical element that maps (goes from) a domain, for example, $\mathbb{R}^n$ to a codomain, for example, $\mathbb{R}^m$. This means that f *receives* an element of $\mathbb{R}^n$ and *gives* a unique element in $\mathbb{R}^m$ in exchange. In general, both the domain and the codomain are not limited to any set; we can think of the function that takes the names of mountains and gives their height. In this course, we will cover only real functions (that map to real numbers) of a real variable (that map from real numbers).

To be a bit formal, we define note

$$\begin{aligned} f : \mathbb{R}^n &\leftarrow \mathbb{R}^m \\ x &\mapsto f(x) \end{aligned}$$

Here, three important sets must be defined.

- The domain of f, $Dom(f)$, is defined as the subset (selection of elements) of $\mathbb{R}^n$ such that $f(x)$ is well defined. Formally,

  $$Dom(f) = \{x \in \mathbb{R}^n : \ \exists f(x)\}$$

  Note that, in general, $Dom(f) \neq \mathbb{R}^n$. Can you think of an example where $Dom(f) \subset \mathbb{R}^n$?

- The image set of f, $Im(f)$, is the subset of all the $f(x)$.

  $$Im(f) = \{y \in \mathbb{R}^m : \ \exists x \in \mathbb{R}^n \text{ s.t. } y = f(x)\}$$

  As before, there are cases where $Im(f) \subset \mathbb{R}^m$. Can you say one?

- The last set that is interesting to study is the graph.

  $$Gr(f) = Dom(f) \times Im(f)$$

L. 3.2 Sequences

A sequence is an ordered infinite set $\{a_i\}_{i \in \mathbb{N}}$. In particular, we will work with sequences of real numbers, so $a_i \in \mathbb{R}$. You can also think of a function that maps the natural numbers to the real numbers.

If we can find an expression for the sequence as a function of its position in the sequence, we define the general term. For example, for the sequence

$$a_1 = 1,\ a_2 = 2,\ a_3 = 3$$

we can define the general term as $a_n = n$. In general, not all the sequences have a general term. For example, the sequence of the n first numbers of π does not have a general term.

$$a_1 = 3,\ a_2 = 3.1,\ a_3 = 3,14,\ a_4 = 3,141, ...$$

There is a lot of theory on sequences, but we will only study the concept of limit. Given a sequence, we say that it is convergent if exists a number a such that the distance[8] from a_n to a becomes smaller when n increases. We will say that the sequence $\{a_n\}_{n \in \mathbb{N}}$ converges to a.

L. 3.3 Limits

The concept of limit is used to build up other concepts of analysis as continuity, derivatives, or integrals. Moreover, we use limits to understand the shape of functions. In particular, we will study the limit of functions. The concept of the limit of a real function is similar to the limit of a sequence. However, in this case, we don't study what happens when n goes to infinity (as the function doesn't depend on n in general) but what happens when x converges to a certain number of the domain.

The formal definition of a limit is the following. Given a function $f(x) : \mathbb{R} \to \mathbb{R}$, its limit in y is $L \in \mathbb{R}$ if given $\varepsilon > 0$ exists an $\delta_\varepsilon > 0$ such that

$$|x - y| < \delta \Rightarrow |f(x) - L| < \varepsilon$$

[8] Here, the distance is the absolute value of the difference.

Colloquially speaking, whenever x is close enough to y, $f(x)$ is close to L. It's important to remark that it doesn't matter how x approaches y. If the limit exists, every approach (sequence) of x converging to y should give the same result.

Notice that technically we have two different concepts of limit: one defined for functions and the other for sequences. The first one is necessary to understand several concepts of calculus whereas the second is used in probability among others.

Not all the limits are necessarily a real number. I will show here different cases that appear when calculating limits.

- A limit is a real number. For example in

$$\lim_{x\to 1} \frac{1}{x} = 1$$

- The limit is ±∞. For example in

$$\lim_{x\to\infty} x = \infty$$

- The limit doesn't exist. This can happen due to different reasons, but, in general, we can think that the function converges to a certain point from a side and to another point from the other side. Here we use the lateral limits. If we study the limit when x converges to zero of the function

$$f(x) = \begin{cases} x & \text{if } x < 0 \\ x^2 + 1 & \text{if } x \geq 0 \end{cases}$$

  we should care about the limit of $f(x) = x$ and the limit of $f(x) = x^2$ when studying the point $x = 0$. With this function, we can give the intuition of what a lateral limit is. The left lateral limit of $f(x)$ in the point 0, noted as $\lim_{x\to 0^-}$ is the limit of the function $f(x)$ when x converges to a from the negative numbers (from the left of the real line). The opposite can be explained for the right lateral limit.

  It can also be the case that the function doesn't have an accumulation point as $\sin(x)$. What happens when x goes to infinity? As $\sin(x)$ is a *periodic* function there is no limit.

There is a lot of theory in limits that we will not cover. However, you must know that these three cases exist. We will not enter how to compute a limit.

L. 3.4 Continuity

We say that a function is continuous at a point x whenever the limit of the function at x exists and converges to $f(x)$. This can fail in three different ways.

1. The limit exists but the image of the function is a different point. This is called **avoidable discontinuity**. For example,

$$f(x) = \begin{cases} x & \text{if } x < 0 \\ 1 & \text{if } x = 0 \\ x & \text{if } x > 0 \end{cases}$$

   It is called avoidable because by changing the image to a different point, the function becomes continuous.

2. The lateral limits are different real numbers. This is called finite jump discontinuity.

$$f(x) = \begin{cases} x & \text{if } x < 0 \\ x+1 & \text{if } x \geq 0 \end{cases}$$

3. Finally, it can be the case that the lateral limits are $\pm\infty$. A typical case of this type of discontinuity is

$$f(x) = \frac{1}{x}$$

L. 3.5 Derivatives

We define the derivative of a real function of real variables in $x_0 \in Dom(f)$ as

$$\frac{\partial f}{\partial x}(x_0) = f'(x_0) = \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h} \tag{6}$$

The intuition behind this definition is the following. If we want to study the variation tax of the function $f(x)$ between points x_1 and x_2 we define it as

$$\frac{f(x_1) - f(x_2)}{x_1 - x_2}$$

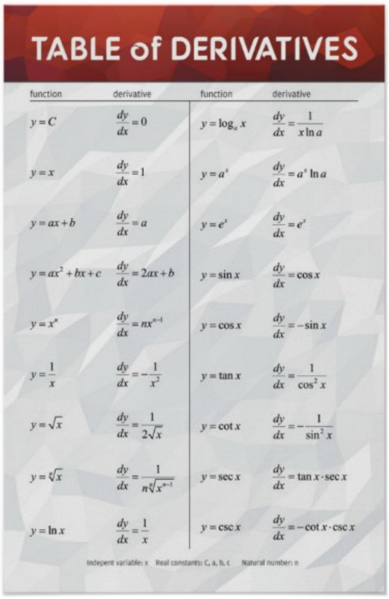

TABLE of DERIVATIVES

function	derivative	function	derivative
$y=C$	$\frac{dy}{dx}=0$	$y=\log_a x$	$\frac{dy}{dx}=\frac{1}{x\ln a}$
$y=x$	$\frac{dy}{dx}=1$	$y=a^x$	$\frac{dy}{dx}=a^x\ln a$
$y=ax+b$	$\frac{dy}{dx}=a$	$y=e^x$	$\frac{dy}{dx}=e^x$
$y=ax^2+bx+c$	$\frac{dy}{dx}=2ax+b$	$y=\sin x$	$\frac{dy}{dx}=\cos x$
$y=x^n$	$\frac{dy}{dx}=nx^{n-1}$	$y=\cos x$	$\frac{dy}{dx}=-\sin x$
$y=\frac{1}{x}$	$\frac{dy}{dx}=-\frac{1}{x^2}$	$y=\tan x$	$\frac{dy}{dx}=\frac{1}{\cos^2 x}$
$y=\sqrt{x}$	$\frac{dy}{dx}=\frac{1}{2\sqrt{x}}$	$y=\cot x$	$\frac{dy}{dx}=-\frac{1}{\sin^2 x}$
$y=\sqrt[n]{x}$	$\frac{dy}{dx}=\frac{1}{n\sqrt[n]{x^{n-1}}}$	$y=\sec x$	$\frac{dy}{dx}=\tan x\cdot\sec x$
$y=\ln x$	$\frac{dy}{dx}=\frac{1}{x}$	$y=\csc x$	$\frac{dy}{dx}=-\cot x\cdot\csc x$

Indepent variable: x Real constants: C, a, b, c Natural number: n

Figure 1: Table of derivatives

the variation in the function divided by the variation in the variable. Is this number is high is because small variations in x produce high variations in $f(x)$. And the opposite, if it is small, high variations in x produce small variations in $f(x)$.

Using the concept of limit defined before, we can go *infinitesimal* (study little variations). Notice that, for a given h, (6) is equivalent to the definition of variation tax when we take $x_2 = x + h$. Therefore, with the limit, we are studying what happens when x_2 converges to x_1. Derivatives help us to understand how a function behaves in terms of growth.

Let's study the derivative of a polynomial of degree 1 when x converges to zero.

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}=\lim_{h\to 0}\frac{a(x+h)+b-ax-b}{h}=\lim_{h\to 0}\frac{ah}{h}=a$$

Luckily, some wise mathematicians before us developed the proofs for the most used functions in mathematics. Check Figure 1.

Additionally, using the properties of limits, we can define operations with derivatives.

- **Sum**

$$f'(x)+g'(x)=(f+g)'(x)$$

- **Product**

$$(f(x)g(x))' = f'g(x) + fg'(x)$$

- **Division** (for $g(x) \neq 0$)

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'g(x) - fg'(x)}{(g(x))^2}$$

Chain rule

However, not every function is as easy as the ones shown in Figure 1. For example, if we want to differentiate $\ln \sqrt{x}$, we cannot use directly the table to find the solution. Instead, we need to use the chain rule.

Let $f(x)$ and $g(x)$ be two functions. We define $h(x) = f(g(x))$. Therefore

$$h'(x_0) = f'(g(x_0))g'(x_0)$$

This theorem lets us compute the derivatives of a composition of functions. In this example, we have the logarithm of the square root of x. So we can say that $f(x) = \ln(x)$ and $g(x) = \sqrt{(x)}$. Therefore,

$$h'(x) = \frac{1}{\sqrt{(x)}}\frac{1}{2}\frac{1}{\sqrt{(x)}} = \frac{1}{2x}$$

Notice that we can also obtain this result using properties of the logarithms (**do it!**).

Utility of the derivatives

As I said before, derivatives are useful for knowing when functions are decreasing and where increasing. With this, we can also know about its critical points (maximum and minimums). For example, imagine that we have this profit function of a firm.

$$f(q) = -2q^2 + 4q + 500, \quad q \geq 0$$

$f(q)$ corresponds to the profits as a function of the quantity produced. The owner wants to know at which production level he can maximize the profits. Can you say what quantity should she produce and which will be her profits?

Partial derivatives

The previous definition of derivative only works for functions in which the domain is one-dimensional. However, in our work as data analysts, we will deal with multidimensional functions and with their derivatives.

Hence, we need a tool to understand how a function behaves when its domain is in $\mathbb{R}^n$. The most natural extension is to extend the definition of derivative to a vector. Let $v \in \mathbb{R}^n$ and $f(x)$ a function with domain $D \subseteq \mathbb{R}^n$. We define the **directional derivative** of $f(x)$ in the direction of v at point x_0

$$D_v f(x_0) = \lim_{h \to 0} \frac{f(x_0 + hv) - f(x_0)}{h}$$

However, this is a concept that we will not cover in this course. Instead, we will study the **partial derivatives** that can be understood as a special case of directional derivatives. We can define the partial derivative of $f(c)$ for the $k - th$ coordinate as

$$\frac{\partial f}{\partial x_k}(x_0) = \lim_{h \to 0} \frac{f(x_0 + h1_k) - f(x_0)}{h}$$

where 1_k is a vector of zeros with a one in the entry k.

Despite I introduce the concept of derivatives with limits, we will not use them to compute derivatives. This is mainly the reason why we won't work with limits in this course. Instead, we will compute the derivatives using the table shown in Figure 1.

Let's show how to compute the partial derivative of a function with an example.

$$f(x, y) = x^2 y^3$$

Using the definition

$$\lim_{h\to 0}\frac{f((x,y)+(h,0))-f(x,y)}{h}=\lim_{h\to 0}\frac{(x+h)^2y^3-x^2y^3}{h}=\lim_{h\to 0}\frac{((x+h)^2-x^2)y^3}{h}=$$

$$\lim_{h\to 0}\frac{(h^2+2xh)y^3}{h}=\lim_{h\to 0}(h+2x)y^3=2xy^3$$

This example illustrates how to compute the partial derivatives of a function. Treat the components different from the k-th as a constant and take the derivative with respect to x_k as we did with functions in $\mathbb{R}$.

$$\frac{\partial f(x)}{\partial x}=2xy^3$$

L. 3.6 Taylor's theorem

Taylor's theorem is used to approximate *complicated* functions by polynomials, that are much simpler. If $c\in[a,b]$ and the n-th derivative exists in interval (a,b) and is continuous in the closed interval, for every $x\in(a,b)\setminus\{c\}$, exists an x_1 between x and c such that

$$f(x)=f(c)+\sum_{k=1}^{n-1}\frac{f^{(k)}(c)}{k!}(x-c)^k+\frac{f^{(n)}(x_1)}{n!}(x-c)^n$$

There exists also a multidimensional version of Taylor's theorem that I will not mention here. As an example, I will show how to approximate the function e^x by a polynomial of order 2 at 0.

$$f(x)\approx f(0)+f'(0)(x-0)+\frac{f''(0)(x-0)^2}{2}$$

Substituting by the values of e^x, we get

$$e^x\approx 1+x+\frac{x^2}{2}$$

Why is this equation an approximation and not an exact identity? Because there is still needed to add $\frac{f^{(3)}(x_1)}{6}$ where x_1 is some unknown number between x and 0. Notice that if x is far from 0, $(x-0)$ becomes larger and so does the interval for x_1. For this reason, the approximation is worse in this case.

L. 3.7 Integrals

Despite there are different types of integrals (Lebesgue, Riemann,...) and the following statement is not exact, we can say that an integral is the opposite of a derivative.

Let $f(x)$ be a function that maps from a subset of $\mathbb{R}$ to $\mathbb{R}$. We say that $G(x)$ is the integral of $f(x)$ and note

$$\int f(x)dx = G(x)$$

if $G'(x) = f(x)$. Here dx represents the variable with respect to which we are doing the integral.[9]

We will distinguish between two types of integrals: indefinite integrals and definite integrals. As before, let $f(x)$ be a function that maps from a subset of $\mathbb{R}$ to $\mathbb{R}$, we say that $F(x) + K$ is the indefinite integral of $f(x)$ if

$$\int f(x)dx = F(x) + K$$

where K is a constant. Notice that, by the properties of the derivative, the integral is identified only up to a constant. Using the properties of the derivatives, can you say what is equal to the following expressions?

- $\int (f(x) + g(x))dx$
- $\int \alpha f(x)dx$

Different ways of solving integrals

Solving integrals is one of the most difficult parts of analysis. Some mathematicians and engineers might devote years to find ways of solving integrals. Here I show two ways of solving the easiest ones.

[9] Technically, this is not the definition of integral but a result named *Fundamental Calculus Theorem.* However, as we are studying simplified concepts, this will be our beginning point.

Immediate integrals These are the integrals that look like the derivative of a function. So we can just solve them by looking at the table of derivatives. Sometimes, a change of variables might be needed.

Integration by parts If $u(x)$ and $v(x)$ are two functions of x, we have that

$$\int udv = uv - \int vdu$$

Let's see how to apply this way of solving to the function $x \sin x$. We call $u(x) = x$ and $\sin x = dv(x)$. Therefore,

$$u(x) = x \Rightarrow du(x) = dx$$
$$v(x) = -cos(x) \Rightarrow dv(x) = \sin xdx$$

Therefore, using the formula of the integration by parts,

$$\int x \sin xdx = -x \cos x + \int \cos xdx = -x \cos x + \sin x$$

You can check that the integral is correct by computing the derivative of the right-hand side and checking that it works.

Definite integral

Differently from the indefinite integral, the definite integral is a real function between two extremes of an interval a and b. We write

$$\int_a^b f(x)dx$$

The *Barrow's rule* guarantees that, if $f(x)$ is a continuous function in the interval $[a, b]$, there exists a function, also continuous in $[a, b]$ and derivable in (a, b) such that

$$\int_a^b f(x)dx = F(b) - F(a)$$

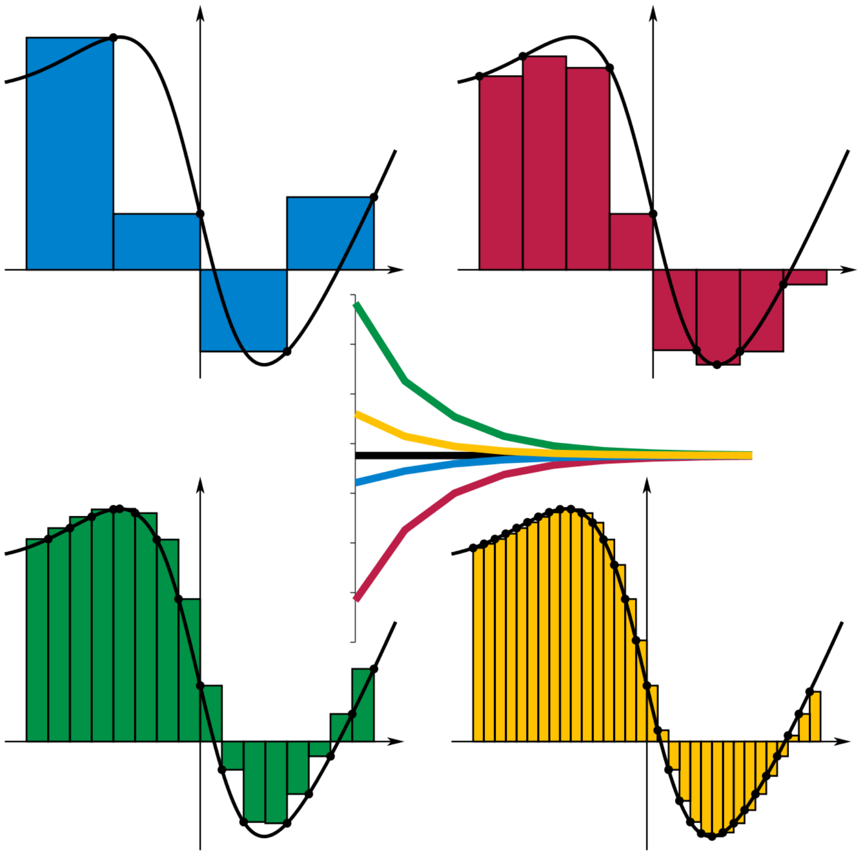

Figure 2: Example of how the Riemann sum works.

Notice that for a given pair a, b, the integral is a real number and not a function as in the case of the indefinite integral.

It is also possible to calculate the integral when a or b are equal to infinite. Notice that in general, not every function has a well-defined integral. For example, $\int_0^1 \frac{1}{x} dx$ gives problems. We say that this integral doesn't converge or that it doesn't exist.

Interpretation of the integral

Despite I will be explaining it as an interpretation of what's an integral, this is the correct definition of the integral. However I find more difficult to create the relation between derivatives and integrals from that initial point.

The integral gives the area behind a function. How do we calculate this area? We can narrow it down by using rectangles. Clearly, if we draw a big rectangle bigger than the function, its area will be bigger than the function. And if we draw a rectangle smaller than the function, the area of the function will be bigger than the area of the triangle. We can, instead of drawing one triangle, draw 2, 4,... The most rectangles we draw, the most close the sum of the area of rectangles will be to the area of the function. If both the small and the big rectangles' areas converge to the same number, we say that's the Riemann Integral.

Part II

Probability

Probability deals with random phenomena. These, in contrast with deterministic ones, are those characterized by the impossibility to know with certainty the result of experiments. For example, if we throw a dice we can say for sure that it will fall until the floor stops it. Even more, knowing the resistance to the air and the initial speed of the dice, we can know how much time will it take to reach the floor. However, it's not so easy to know which will be the number the die shows when it stops. This second phenomenon is called a random experiment.

Probability seeks to study via models the behavior of random phenomena. In general, we will say that the nature of these experiments doesn't allow us to predict its result.

L. 4 Some concepts of probability

Despite not being able to determine the exact result of an experiment, we might be able to determine some set where the result should be. This collection with all the possible results is called *sample space*, in general, noted as Ω.

We will call *random experiment, experiment* or *trial* to the actions that can be repeated under the same conditions to obtain a result. An outcome of a random experiment will be noted as $\omega \in \Omega$. Let's see some examples

- If flipping a coin is a random experiment, the possible outcomes are head (H) or tails (T). Hence, the sample space is $\Omega_1 = \{H,\ T\}$.
- Rolling a die gives as sample space $\Omega_2 = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$.

As the outcomes of the random experiments are subsets of a bigger set, we can use set theory with them. For example, we can consider the Cartesian product of Ω_1 and Ω_2. This set is the sample space of the experiment that consists of flipping a coin **and** rolling a die.

$$\Omega_1 \times \Omega_2 = \{(H,1),\ (H,2),\ (H,3),\ (H,4),\ (H,5),\ (H,6),\ (T,1),\ (T,2),\ (T,3),\ (T,4),\ (T,5),\ (T,6)\}$$

We can also talk about the union or the intersection of events as we do in group theory. Usually, we can talk of the events as subgroups.

Among all the possible outcomes we can differentiate some special cases:

- Sure outcome.
- Impossible outcome.
- Complementary outcomes. If Ω_1 happens, Ω_2 cannot happen. Moreover, $\Omega_1 \cup \Omega_2 = \Omega$. We can note the complementary outcome of A as A^c.

L. 4.1 Laplace formula

Up to this point, we have seen more set theory than probability. If you think about probability, you might have in mind some function that assigns to each event a number between 0 and 1 (or between 0 and 100, which is equivalent) determining how much *likely* is each event to happen. This idea is, in fact, pretty accurate.

The probability is a function defined over a set, Ω, know as the sample space, that has to meet three conditions. It has to be positive for every element of the sample space, the probability of Ω must be one (it's the sure outcome) and

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right)$$

where $\{A_i\}_{i=1}^{n}$ is a sequence of disjoint sets.[10]

The *Laplace* method is a formula that determines the probability of each event when the sample space is finite.

Hence, in this case, we can define the probability of an outcome $A \in \Omega$ as

$$P(A) = \frac{|A|}{|\Omega|}$$

which is the division between the favorable cases and the possible causes. Being strict, this is not a definition of a probability but a way of computing it when $|\Omega| < \infty$. To give the formal definition of probability, we would need to understand some concepts of algebra and measure theory, so we will work with the *intuitive* definition of probability. Notice that from the definition we can get that

[10] Formally, we say that with the set Ω we define a $\sigma-$algebra, $\mathcal{A}$, that is the set with all the possible combinations of subsets of Ω that we can make under some conditions. Then $A_i \in \mathcal{A}$.

$$P(!A) = 1 - P(A)$$

Some properties of the probability.

- The probability of the empty event is zero $P(\varnothing) = 0$.
- If $A_1 \cap A_2 = \varnothing \Rightarrow P(A_1 \cup A_2) = P(A) + P(B)$.
- In general, $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.
- $P(A^c) = 1 - P(A)$.[11]
- If $A \subset B$, then $P(A) \leq P(B)$.

L. 4.2 Conditional probability

This is one of the most important concepts of probability and it is highly related with the role of information; what we know about a certain series of events affects the probability of another event happening. Let's use an example,

We have two boxes with balls, one full of red balls and the other full of white balls (let's say 50 balls per box). We toss a coin and select one of the two boxes. If the coin shows a head, we will pick a ball of the red box, otherwise, we will pick a ball from the white box. **What is the probability of picking a red ball? What is the probability of picking a red ball if we know that the coin was a head?** Obviously the event of the coin will affect to the probability that we assign to a red ball appearing.

Let A and B be two events such that $P(B) > 0$. We define the probability of A condition to B as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Notice that using this formula we can say equivalently that

$$P(A \mid B)P(B) = P(A \cap B)$$

[11] This formula can be extended to any finite union of events.

Can you say what will happen with a higher number of events? It's called the factorization theorem.

Theorem of Total Probability. The following result is called the theorem of total probability. Assume we have a set of events $A_1, ..., A_n$ that are disjoint, have $P(A_i) > 0$ and, $\cup A_i = \Omega$ (we call this a partition). We can write the probability of another event B as

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

Why is useful this theorem? It can be used to define surveys about *delicate* topics. Imagine that we want to know how much students of the UAB take drugs. Some students my be insecure about the anonymity of their answers. Therefore, the results that we get from the survey might not be exact. We can do something else to ensure a higher precision.

Assume we ask seventy students about their drugs consumption (have you ever consumed drugs: yes or no?), and thirty students if their birthday was in the first six months of the year. Hence, we get a hundred answers where 25 are *Yes* and 75 are *No.* If we assume that half of the people is born between January and June and half of the people between July and December, **what can we say about the drugs consumption?**

$$\begin{aligned} P(Yes) &= P(Yes \mid \text{Drugs Question})P(\text{Drugs Question}) \\ &\quad + P(Yes \mid \text{Birthday Question})P(\text{Birthday Question}) \end{aligned}$$

Bayes Theorem. The final of the *important* theorems coming from the conditional probability definition is the Bayes theorem. It studies the probability of an event of the partition given that the event B happened.

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)}$$

L. 4.3 Independent probabilities

One important concept we deal with in probability is the independence. We say that two events are independent if one happening doesn't affect the probabilities of the other happening. For example, if I pick a random number between 1 and 100, the probabilities of each number to be the selected would be $\frac{1}{100}$, in particular, the probability of 27, $P(25)$. Assume that you can select 10 numbers, and you select all the numbers between 20 and 29. Your probability of winning is

$$P(W) = \frac{10}{100} = \frac{1}{10}$$

However, if I announce that the number ends in 5, how do you probabilities change? Now only 10 numbers can be elected, 5, 15, 25,...,95. Among those you have one number elected

$$P(W \mid \text{Number ends in 5}) = \frac{1}{10}$$

Therefore, the fact of knowing that the number ends in 5 doesn't affect your chances of winning. We say that winning and ending in 5 are two independent events. Formally, if

$$P(A \mid B) = P(A)$$

A and B are two independent events. Notice that using the definition of conditional probability,

$$P(A \mid B) = P(A) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A)P(B) = P(A \cap B)$$

If we have a bigger set of events, let's say n events, $\{A_1, A_2, ..., A_n\}$ we can have what it's called pair independence, when

$$P(A_i \cap A_j) = P(A_i)P(A_j)$$

If the probability of the intersection of any subset of this set of events is equal to the product of the events in the subset, $\prod_{i=1}^{n} P(A_i)$, we say that the family is mutually independent.

L. 5 Random variables

In general, when we develop our analysis of data, we don't care about the realizations of the events, but about the properties of these elements. For example, assume that we pick at random one individual from the BSE. We are interested in knowing the distribution or wealth among all the students. Therefore, when we pick a concrete individual, we want to know her wealth.

To the function that associates a numerical value to a result we call it random variable.

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega) = x \end{aligned}$$

We use $\mathbb{R}$ because it is a well known space, so it's easier to work with it.

L. 5.1 Distribution function of a random variable

Definition 1. Given a random variable X we define its **distribution function** as

$$F_X(x) = P(X \leq x), \quad \forall x \in \mathbb{R}$$

A distribution function meets the following properties:

1. $F_X(x) \geq 0 \quad \forall x \in \mathbb{R}$.
2. $F(x)$ is monotonous increasing.
3. Right continuous.[12] **Can you think of a function that is right continuous but not left continuous?**
4. $\lim\limits_{x\to\infty} F_X(x) = 1$
5. $\lim\limits_{x\to-\infty} F_X(x) = 0$

[12] This means that the right limit exists but not necessarily the left one.

With a function of probability (a distribution function) we can define a concept that is frequently used in both probability and statistics: the percentile. Given a random variable, we call the percentile of order $p \in [0, 1]$ as the value x_p which verifies:

$$P(X \leq x_p) \geq p \quad \text{and} \quad P(X \geq x_p) \geq 1 - p$$

So, the percentile 0.5 (or 50) is expected to be in the middle of all the observations that we make. It's also called median.

L. 5.2 Discrete variables

A discrete random variable is a random variable that only take discrete values. It's not necessarily a discrete variable. For example, our weight, or age are continuous variables. But we usually study them in kg or years. So I will not say that I am $\frac{235.75}{24}$ years old. We can treat them as discrete because we will define them in a discrete set: one year, two years,...

We can note $\mathbb{D}$ as the *possible set* of our variables. This set is known as the support. We are interested in knowing for each element of that set,

$$P(X = x_i), \quad x_i \in \mathbb{D}$$

If we know this, we can obtain any other type of probability as, $P(X \in \mathcal{D})$, where $\mathcal{D} \subseteq \mathbb{D}$. **Can you guess $P(\mathbb{D})$?**

For **discrete** variables, we define the probability function as

$$f_X(x) = \begin{cases} P(X = x), & x \in \mathbb{D} \\ 0, & \text{otherwise} \end{cases}$$

Notice that $f_X(x)$ meets two conditions.

1. It's positive in all its domain.
2. $\sum_{i \in \mathbb{D}} f(x_i)$=1.

Notice that, when the support of variable X is ordered, we can calculate $P(X \leq x)$ for any $x \in \mathbb{D}$. Equivalently,

$$F_X(x) = P(X \leq x) = P(x_0) + P(x_1) + ... + P(x) = \sum_{x_i = x_0}^{x} f_X(x_i) \tag{7}$$

L. 5.3 Discrete distributions

The discrete distribution functions are the distribution functions of the discrete variables. Here I present three examples. There are more, as the geometric, hyper-geometric distributions or negative binomials.

Bernouilli

We toss a coin and we want to study if we get a success (heads) or a failure (tails), we pick a student from our class and check if she has diabetes or not. The question that a Bernouilli distribution can answer is: yes or no, success or failure. Hence, we have to assign a one if the event that we want to consider happens and zero otherwise.

Let X be a random variable that is equal to one whenever our coin shows a head, and zero otherwise. Then

$$P(X = x) = p^x(1-p)^{1-x}, \quad \text{for } x = 0, 1$$

Notice that the range of the variable is $\{0, 1\}$. Here $p = P(X = 1)$ and $1-p = P(X = 0)$. We write

$$X \sim Bi(1, p)$$

Binomial

You might be thinking that the Bernouilli is giving us information that we already know. If we know p we don't need to calculate $P(X = 1)$, as we know that it's equal to p. Binomial distribution is the Bernouilli but repeated n times. For example, tossing n coins. We write

$$X \sim Bi(n, p)$$

where n stands for the number of repetitions and p is the Bernouilli parameter. Notice that the support of X is not 0 or 1 anymore. Now X represents the number of success or failures that we can make. Therefore is a natural number smaller than n.

We say that X is a binomial random variable with parameters n and p if

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Notice that the main condition for the repeated Bernouilli to be a Binomial distribution is that there is no relation between the trials that we do. We say that the realizations of the random variable must be independent. Can you express $P(X = k + 1)$ as a function of $P(X = k)$.

Poisson

This distribution is used to study certain cases where a certain event happens related to a finite time interval. Our random variable must meet two conditions to be a Poisson:

1. The probability of the event happening is proportional to the size of the interval.
2. The probability of the event happening more than once is almost zero.

We say that a random variable X follows a Poisson distribution with parameter λ, and we note

$$X \sim Po(\lambda)$$

if the probability function of X is

$$f_X(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \text{for } x = 0, 1, 2, ...$$

Can you write an expression for $F_X(x)$? You might want to use equation (7)

L. 5.4 Continuous variables

In this case, we abandon the discrete set up. So now, our domain for the random variables won't be a countable set.

We say that a random variable is continuous if there is a no-negative integrable function, $f_X(x)$, such that its integral is equal to one in the real line.

$$P(X \in (a,b]) = P(a < X \leq b) = \int_a^b f_X(x)dx$$

We can technically substitute the interval $(a,b]$ by any other type of interval. The function $f_X(x)$ is called the **density function** of probability. Notice that the function itself is not giving us any information about it's probability. You **must not** confuse the discrete case, where $f(x) = P(x)$, with the continuous case.

Now, we define the distribution function (or probability function) of a random continuous variable, as

$$F_X(x) = \int_{-\infty}^x f_X(x)dx$$

L. 5.5 Continuous distributions

The continuous distribution functions are the distribution functions associated with continuous random variables.

Uniform distribution

We will say that a random variable X follows a uniform distribution with parameters a, b (or in the interval $([a,b])$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

We note $X \sim Un(a,b)$. **Can you guess $F_X(x)$?**

Normal distribution

We say that X is a random variable following a normal distribution with parameters μ σ^2 if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

We note

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

Notice that $\mu \in \mathbb{R}$ and $\sigma^2 \in [0, \infty)$. The function $F_X(x)$ is not easy to compute manually, we usually write it as in the following equation

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

Several years ago people used to have tables with all the probabilities of the normal (0,1), known as the standard normal. Hence, having some random variable distributed as a normal μ, σ^2, we can say

$$P(X \leq x) = P\left(X' \leq \frac{x-\mu}{\sigma^2}\right), \quad \text{with } X' = \frac{X-\mu}{\sigma^2}$$

X' is distributed as a normal (0,1). Why? Because we use the following property of the normal distribution: let $X \sim \mathcal{N}(\mu, \sigma^2)$. If we define $Y = aX + b$, then

$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

The normal distribution has some other useful properties. You might want to check its **Wikipedia** page.

Exponential distribution

We say that the random variable X follows an exponential distribution of parameter λ, $X \sim Exp(\lambda)$ for $\lambda > 0$, if the density function of the random variable X is

$$f_X(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ \lambda e^{-\lambda x}, & \text{if } x > 0 \end{cases}$$

With an easy integration we can check that

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ 1 - e^{-\lambda x}, & \text{if } x > 0 \end{cases}$$

As you might have noticed, this function is related with the Poisson distribution of parameter λ. It appears in problems about wait time. This distribution is a special case of the Gamma distribution (that is not covered in this course).

L. 6 Random vectors

Now that we know how to work with both discrete and continuous variables we must go a step further. What happens when we see two variables acting together? Imagine that we can only can observe the distribution of a random variable conditional on another. For this, we must work with vectors of random variables (or random vectors).

We might observe the position of a certain object in the space. This position is determined by three dimensions that are, in general, correlated.

L. 6.1 Some definitions

As we did with variables, we must jump from the sample space to the real numbers. However, in this case we will go to another dimension and instead of studying our variables in $\mathbb{R}$ we will work with $\mathbb{R}^k$.

Let (X, Y) be a random vector. Therefore we will want to know

$$P(a_x \leq X \leq b_x, a_y \leq Y \leq b_y)$$

which will obviously depend on the distribution of X, Y and their conditional distribution. In general, we will work with intervals in $\mathbb{R}^k$ which are the Cartesian product of intervals in $\mathbb{R}$.

The distribution function that describes the probability of the random vector (X, Y) is called joint distribution function of X and Y.

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

We can extend this theory to $\mathbb{R}^n$, however, during the course we are only going to work in $\mathbb{R}^2$. If (X, Y) is a random vector, then X and Y must be random variables. The opposite is also true.

L. 6.2 Integrals in $\mathbb{R}^2$

This section must be understood as a ten minutes guide to integration in $\mathbb{R}^n$ with examples in $\mathbb{R}^2$. You are supposed to complement it at home.

Assume we have a function $f : \mathbb{R}^n \to \mathbb{R}$ that is continuous. We define it's integral over the interval (or rectangle) in $\mathbb{R}^n$, $A = [a_1, b_1] \times [a_2, b_2]$ as

$$\int_A f(x_1, x_2) d(x_1, x_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) dx_1 dx_2 \tag{8}$$

As usual in this course, this is not the exact definition of the integral, but the main consequence of the Fubini's theorem. How do we interpret equation (8)? The interpretation is similar to the one that we did for derivatives in $\mathbb{R}^n$. First we fix all the variables but one, they will be constants for us. We study the integral with respect to x_1 (for example). Then, we *de-fix* one of the previous variables, say x_2, and study the integral with respect to it,...

What happens if the region where we are studying the integral is not the product of intervals? We must somehow convert it into an interval $A' = [a_1, b_1] \times [g_1(x_1), g_2(x_1)]$.

$$\int_{A'} f(x_1, x_2) d(x_1, x_2) = \int_{a_1}^{b_1} \int_{g_1(x_1)}^{g_2(x_1)} f(x_1, x_2) dx_2 dx_1$$

Let's see some examples to understand what is going on. First, we will study the integral of xy in the rectangle $R = \{(x, y) : \ 0 < x < 2, \ 0 < y < 2\}$.

$$\int_A xy\ d(x,y) = \int_0^2 \int_0^2 xy\ dxdy = \int_0^2 \frac{x^2}{2}y\Big]_0^2 dy = \int_0^2 2ydy = y^2\Big]_0^2 = 4$$

The next example shows a case in which out set of interest is not square. We will study again the function $f(x,y) = xy$ but in the set $B = \{(x,y) :\ 0 < x < 2, 0 < y < \frac{x}{2}\}$. We must express y in a function of x for the set B.

Notice that, if $0 < y < \frac{x}{2}$, we can operate at both sides of the inequality to get $0 < 2y < x$.

$$\int_B xy\ d(x,y) = \int_0^1 \int_{2y}^2 xy\ dxdy = \int_0^1 \left(\frac{x^2y}{2}\right)_{2y}^2 dy = \int_0^1 \left(\frac{2^2y}{2} - 2y^3\right)\ dy = 0.5$$

We can do the same for y and take as interval $\left[0, \frac{x}{2}\right]$.

$$\int_B xy\ d(x,y) = \int_0^2 \int_0^{\frac{x}{2}} xy\ dydx = \int_0^2 \left(\frac{xy^2}{2}\right)_0^{\frac{x}{2}} dx = \int_0^2 \frac{x^3}{8}\ dx = \frac{1}{32x}2^4 = 0.5$$

L. 6.3 Joint discrete distribution

For this section we will be considering random vectors that can take values in a countable set. I will be calling X to the random vector and $X_1, X_2, ...X_n$ to the different entries of this vector. We define the probability (joint) function of a random discrete vector in a support $\mathbb{D}$ as

$$f_X(x) = \begin{cases} P(X_i = x_i, i = 1, ..., n), & \text{if } x \in \mathbb{D} \\ 0, & \text{otherwise} \end{cases}$$

As you might imagine, $f_X(x) \geq 0$ and $\sum_{x \in D} f_X(x) = 1$.

Marginal probability function

Can we know the probability of one variable if we know the probability of the random vector? The answer is yes.

Y\ X	**0**	**1**
0	0.2	0.3
1	0.4	0.1

$$f_{X_i}(x_i) = \sum_{x_j \in \mathbb{D}_j, j \neq i} f_X(x_1, ..., x_k)$$

Let's see an example with a vector of dimension two.

We know the probabilities in the following table:

First notice that the probabilities of $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ add to one. Therefore,

$$\begin{aligned}
P(X=0) &= P(X=0,Y=0) + P(X=0,Y=1) = 0.6 \\
P(X=1) &= P(X=1,Y=0) + P(X=1,Y=1) = 0.4 \\
P(Y=0) &= P(X=0,Y=0) + P(X=1,Y=0) = 0.5 \\
P(Y=1) &= P(X=0,Y=1) + P(X=1,Y=1) = 0.5
\end{aligned}$$

L. 6.4 Joint continuous distribution

The joint continuous distribution is the joint distribution of a continuous random vector. As you might imagine, this means integrals.

Let X be a continuous random vector, then its probability function (or distribution function) is defined as

$$P(X \in D) = \int_D f_X(x_1, ..., x_n) d(x_1, ..., x_n)$$

Notice that $D \subseteq \mathbb{R}^n$. $f_X(x)$ is known as the joint density function. Notice that theoretically, up until this point, we only added the word joint to every definition that we had before. The difficulty becomes in the application of this theory.

Marginal density functions

Here the process is equivalent to the discrete case, but, instead of doing the sum over all possible values of the variables that we don't care about, we want to study its integral, as we are in the continuous case. Hence, given a random continuous vector X, if we want to study the marginal distribution of the random variable X_i, we must calculate

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} f_X(x) dx_{-i}$$

If we want to study the joint density of a sub-vector, let's say $X_k = (X_{i1}, ..., X_{ik})$, we will have

$$f_{X_k}(x_k) = \int_{\mathbb{R}^{n-k}} f_X(x) dx_{-k}$$

Here, the notation dx_{-i} means with respect to all the variables but i and x_{-k} with respect to all the variables but those that are in the vector X_k.

L. 6.5 Independence of random vectors

In the previous section we said that two variables were independent whenever

$$P(A \cap B) = P(A)P(B)$$

Similarly, we can give a definition for two independent random vectors. In the discrete case, the variables that conform a random vector are independent iff

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i), \forall x_1, ..., x_n \in \mathbb{R}$$

With continuous random variables, we say that the variables that conform a random vector are independent iff

$$f_X(x_1, ..., x_n) = f_{X_1}(x_1)...f_{X_n}(x_n), \quad \forall x_1, ..., x_n \in \mathbb{R}$$

Think about a uniform distribution of (x, y) in $[0, 1] \times [0, 1]$, $f_{XY}(x, y) = 1$.

L. 6.6 Conditional distributions

It's important that we think of a random vector as two (or n) random variables that happen simultaneously. Therefore, we cannot study the separate distributions, we must use mathematical theory to understand how these distributions work. What happens if we want to know the distribution of a variable knowing something about the other? For example, we want to know how evolve the wage of individuals that are women, or the hearth rate of diabetics.

We define the discrete conditional probability function of Y given $X = x$ as

$$f_{Y|X}(y \mid x) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

In the future, we will infer how statistics evolve given the data that we observe. And therefore we want to understand what is the evolution of those statistics.

What happens if the vectors are continuous? We define it in a similar way. The conditional density function of a continuous random variable reads as:

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

L. 7 Expectation

In this chapter we will cover two important concepts in probability: the expectation and the variance of a random variable.

The expectation of a discrete random variable is defined as

$$E(X) = \sum_{x_i \in \mathbb{D}} x_i P(X = x_i)$$

whenever this expression exists. **Can you compute the expectation of a Bernoulli distribution?**

What happens when the variable is continuous? Remember that in the continuous case we always substitute the sums by integrals. Therefore,

$$E(X) = \int_{\mathbb{R}} xf(x)\ dx$$

Can you find the expectation of an exponential variable?

The expectation is not only defined for random variables, but also for functions of random variables. The same than we define $E(X)$ we can define $E(g(X))$ for discrete variables.

$$E(g(X)) = \sum_i g(x_i)P(X = x_i)$$

And for continuous variables,

$$E(g(X)) = \int_{\mathbb{R}} g(x_i)f_X(x)\ dx$$

Using the properties of integration, can you guess $E(aX)$? And the $E(X + b)$?

Three interesting properties of the expectation.

- $P(a \leq X \leq b) = 1 \Rightarrow a \leq E(X) \leq b$.
- $P(g(X) \leq h(X)) = 1 \Rightarrow E(g(X)) \leq E(h(X))$.
- $|E(g(X)) \leq E(|g(X)|)$.

The mean (or the expectation) is a measure of the location of the random variable. It's a simple measure that usually doesn't give enough information about the random variable. We might want to study more *numbers* of measure of our random variables that can give more information about its behavior.

We call the moment of order k of a random variable X to

$$\mu_k = E(X^k)$$

Similarly, we define the central moment of order k of a random variable to

$$E((X - E(X))^k$$

In particular, the variance is the central moment of order 2.

Check that $E((X - E(X))^2) = E(X^2) - E(X)^2$. The third and the fourth central moments are known as the coefficient of skewness (symmetry) and kurtosis (equi-distribution).

Instead of finding the integral every moment when we want to compute the k-th moment, we can calculate the moment generating function. We define the moment generating function of a random variable as

$$M_X(t) = E(e^{tX})$$

which is a function of t but not X. Studying the n-th derivative of the moment generating function at $t = 0$, we can get the n-th moment. Notice that we can characterize a random variable by its density or distribution functions, by its moments or by its moment generating functions.

Compute the moment generating function of the exponential distribution. Check that it makes things easier. I will not cover in this part Markov, Jensen, and Chebyshev inequalities, you might want to check them!

L. 7.1 Expectation of a function of a random vector

And then, how we compute the expectation of a random vector? It is exactly as you imagine. If the random vector is discrete,

$$E(g(X)) = \sum_{x \in D_X} g(x) f_X(x)$$

and if it's continuous

$$E(g(X)) = \int_{\mathbb{R}^n} g(x) f_X(x) dx$$

In particular, if we substitute $g(X) = X_1^k X_2^k, ..., X_n^k$ we will get the joint moment of order k. However, as we are dealing with vectors, we can compute the moment of order $(k_1, k_2, ..., k_n)$, substituting $g(X) = X_1^{k_1} X_2^{k_2}, ..., X_n^{k_n}$.

Again, the joint central moment of order $(k_1, k_2, ..., k_n)$ is defined as $E((X_1 - E(X))^{k_1} + E((X_2 - E(X))^{k_2} + ... + E((X_n - E(X))^{k_n}$.

Can you write the joint moment of order $(1, 1)$?

In particular, if X_1 and X_2 are independent variables, we can say that $E(X_1X_2) = E(X_1)E(X_2)$. If two variables are independent,

$$cov(X_1, X_2) = E[(X_1 - E(X_1))(X_2 - E(X_2))] = 0$$

Notice that, with random variables, we had the variance of our variable of interest. Now, if we want to know the *variance* of the whole vector, we must also take into account the covariances of the different random variables, we need a matrix.

L. 7.2 Conditional expectation

If (X, Y) is discrete, we define

$$E(g(X) \mid y) = \sum_{x \in D_x} g(x) P(X = x \mid Y = y)$$

where $D_x = \{x : \ (x, y) \in D\}$. For continuous variables, we will have:

$$E(g(X) \mid y) = \int_{\mathbb{R}} g(x) f_{X|Y}(x \mid y) \ dx$$

One last property that we will see about expectations is the Law of Total expectations of Law of Iterated expectations. If $E(X)$ exists,

$$E(X) = E(E(X \mid Y))$$

Part III

Statistics

L. 8 Basic concepts

Statistics is defined as the compilation, presentation and analysis of data with the objective of making decisions and solving problems.

We will work with the two main branches of statistics:

- Descriptive statistics, the part that compiles and organizes data.
- Statistical inference, which includes the methods of analysis and decision-making using the data.

In general we will work with two sets: the sample and the population. The sample is the part of the population that we observe. With the descriptive statistics we can say things about the behavior of the sample. With the statistical inference we can make decisions trying to extrapolate what we observe in the sample to the population.

Population is a term used to describe the set of individuals that we want to study. Notice that those individuals can be people, animals or bacteria. As we know from probability, the characteristics associated with the individuals are random variables, which can be:

- Qualitative. Either nominal or ordinal.
- Quantitative. Either discrete or continuous.

We can transform any qualitative variable into a quantitative variable assigning a number to each of the possible outcomes. The distribution function that defines how the random variables behave in the population is called distribution function of the population and, in general, it is unknown. The statistician can know the family to which this distribution belongs (normal, exponential, bernouilli,...) in the best cases. Out objective will be to discover the parameters that define the distribution function.

On opposition to the population, we have the sample. The sample is the part of the population that we can observe, and therefore, measure. We call sample size to the amount of individuals that we have in our sample. It's important that our sample is representative of the population. Here, the way in which we construct our sample is key.

Therefore, if we have a sample of size n, we can say that we are taking n observations from a random variable X distributed with a certain distribution function that we don't know. If we construct our sample in a correct way, the variables that we observe will be independent among them and identically distributed (as they are taken from the same probability distribution). We say that they are *iid* for obvious reasons. In particular, we can say that we observe $\{X_1, X_2, ..., X_n\}$ independent random variables. And, as they are independent,

$$f(x_1, ..., x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Here θ represents the parameters that we want to infer but we don't know. **Can you compute the joint density function of n observations iid of a $N(\mu, \sigma^2)$?** We call to this function the joint density function of the sample.

The sample distribution can be manually computed assuming that the variable is discrete and assigning a probability of $\frac{1}{n}$ to each possible value of X_i. Notice that this is also the relative frequency of each variable. The sample distribution function is therefore

$$F_n(x) = \frac{|\{X_i : \ X_i \leq x\}|}{n}$$

This function reduces the amount of information that the whole sample gives. This is what we call an statistic. An statistic is a function of the sample (not of the population).

$$t(X) : t(X_1, .., X_n)$$

As an statistic is a function of a random vector, it will also be a random variable (or vector). The statistic is a summary of the sample. In principal, it will give less information than the sample. Let's see some examples:

- The ordered sample.
- The median or any percentile of order α.
- The range of the sample (the difference between Q_3 and Q_1).

As we did with probability, now, with the sample, we can compute the sample moments. We will work principally with the first and the second moment of the sample, which are the mean and the variance:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$V(X) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Sometimes we will prefer to work with the quasivariance, because it has better properties. The quasivariance is defined as

$$S^2(X) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

Law of large numbers. If we have n iid variables with finite variance and common expectation $E(X_1)$, then

$$\frac{1}{n} \sum_{k=1}^{n} X_k \xrightarrow{p} E(X_1)$$

This means that the probability of that the sample mean and the expectation are different becomes close to zero when we increase the sample size. Using the same logic, we can say that the sample moments converge to the population moments when we increase the sample size.

L. 9 Statistics

L. 9.1 Sample distribution

As the statistics that we calculate with our sample depend on the behavior of a random vector X, we can say that the statistics are also random variables and, therefore, they will have a probability distribution. We call it the sampling distribution. The question that naturally appears is, how do we compute the sampling distribution from the sample? There are three ways.

The direct way of obtaining the sampling distribution.

If we can observe all the possible variables that the statistic of interest gets we can assign to them a probability of happening. This probability is the sampling distribution.

Imagine that we have four balls each one with numbers from 1 to 4. We take two balls with replacement and write its value (X_1, X_2). **What is the sampling distribution of $U = X_{\max} - X_{\min}$?**

First we need to create the table with all the possible outcomes of taking to balls and calculate U. Then, we can compute the sampling distribution of U.

Analytic way of obtaining the sampling distribution.

We can use probability tricks to save calculations. For example, if we want to know the distribution of the sum of success in a Bernouilli distribution, we know that that sum behaves as a Binomial random variable. The following theorem is often used to compute the sampling distribution of statistics. Let $Y = g(X)$ be our statistic of interest. Therefore, if g is differentiable, with differential different from zero and monotonic and $f_X(x)$ is continuous,

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left|\frac{\partial g^{-1}}{\partial y}\right| & \text{for } y \in g(\{D_X\}) \\ 0 & \text{otherwise} \end{cases}$$

Example. Let X be a random variable and

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x \leq 1 \\ \frac{1}{2x^2} & \text{if } x > 1 \end{cases}$$

If we define $Y = \frac{1}{X}$, what is the distribution function of Y? Prove that it is

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2(1/y)^2}\frac{1}{y^2} & \text{if } 0 \leq y \leq 1 \\ \frac{1}{2}\frac{1}{y^2} & \text{if } y > 1 \end{cases}$$

L. 9.2 Sufficient statistics

Notice that when I defined a statistic, I said that it mean a reduction of the sample. However, it is not necessarily a reduction of the information that the sample gives about the population. The statistic that transform the sample into an ordered set, the ordered sample, gives exactly the same information as the original sample.

Imagine that we have a sample $\mathbf{X} = \{X_1, X_2, ..., X_{10}\}$ of an iid sample distributed as a $Be(p)$. Let $x = (0, 1, 0, 1, 1, 0, 1, 0, 1, 1)$ be the vector of data. Knowing that we have had $t(x) = \sum x = 6$ successes, we are having the same information about p than with the sample. We say that the sum is a sufficient statistic.

A sufficient statistic must keep all the information over the parameters that $\mathbf{X}$ contains and must be a reduction of $\mathbf{X}$. Notice that every statistic defines a partition of the sample space. We will say that two statistics are equivalent if both give the same partition of the sample space. We say that a partition P_1 is a reduction of another, P_2, if each element from P_1 is the union of elements of P_2.

With the information in the previous paragraph we can give a better definition of sufficient statistic. Given $\mathbf{X} = \{X_1, ..., X_n\}$ a random sample of a random variable X distributed as $f(x \mid \theta)$ we say that $t(\mathbf{X})$ is a sufficient statistic for the distribution functions $f(x \mid \theta)$ if and only if $f(x \mid t(X), \theta)$ doesn't depend on θ. **Can you show that $X_1 + X_2$ is a sufficient statistic for the Bernouilli distribution?**

Factorization criteria

An statistic $t(\mathbf{X})$ is a sufficient statistic for the family of distributions $f(x \mid \theta)$ iff

$$f_X(x \mid \theta) = g(t(x), \theta)h(x)$$

with g and h two non negative functions. So, if we can split the sample density function in two different functions one that depends on the statistic and the parameter and another that only depends on the sample, we can prove that our statistic is sufficient.

Finally, we say that an statistic is minimal sufficient if any reduction of the partition defined by $t(x)$ is not sufficient.

L. 9.3 Likelihood function

Let $X \sim f(x \mid \theta)$, we first want to create the likelihood function of the parameter θ. We have information about X, but what we want to know is how the statistic behaves, as the statistic in general can helps us infer information about the population. Therefore, given a sample, we can assign a value to each possible θ and score those values by how likely they are. To do this we need the likelihood function.

Let $\mathbf{X} = \{X_1, ..., X_n\}$ a random sample iid of a random variable distributed $X \sim f(x \mid \theta)$. We say that the likelihood function of the parameter θ is

$$l(\theta) = f_X(x \mid \theta)$$

Notice that the likelihood function of the parameter depends on the sample that we observe.

- If X is a discrete random variable,

$$l(\theta) = P(\mathbf{X} = x \mid \theta) = \prod_{i=1}^{n} P(X_i = x_i \mid \theta)$$

- If X is a continuous random variable,

$$l(\theta) = f_{\mathbf{X}}(x \mid \theta) = \prod_{i=1}^{n} f_{X_i}(x_i \mid \theta)$$

The likelihood principle states that given two samples x and y the likelihood function obtained with one sample will be proportional to the likelihood function obtained with the other. We say that the two functions are equivalent. This is why we use to define the likelihood function as a function that is proportional to $f_{\mathbf{X}}(x \mid \theta)$.

Can you compute the likelihood function of a Bernouilli parameter p given that the sample is $\{1, 0, 1, 1, 0, 1, 1, 1, 1, 0\}$? And of the normal distribution $N(\mu, 1)$? It's important that we have the maximization theory in our minds when working with likelihood functions.

An statistic is sufficient if and only if it's likelihood function depends only on the statistic and the parameters, but not on the sample. We usually work with the log-likelihood function, can you see why in the normal example? The final step is to maximize the likelihood (or log-likelihood) function and obtain the maximum likelihood statistic.

L. 10 Interval estimation

We know how to find estimators for certain variables. However, how can we estimate the quality of those parameters? We need to find a certain criteria that allows us to distinguish between *good* and *bad* estimations. It is proposed to add to the point estimate the possibility of presenting a range of values that give some assurance of finding the value of the parameter among these values. This range of values is called the confidence interval of the parameter.

Definition 2. Given a random sample X, a confidence region $S(X)$ with α as confidence level for the parameter θ is a region of the sample space such that

$$P(\theta \in S(X)) > 1-\alpha, \quad \forall \theta \in \Omega$$

We will define this region for the real numbers as a confidence interval. Therefore, we will look for two boundaries

$$[l_\alpha(X), L_\alpha(X)]$$

Therefore, the confidence interval can be redefined as a pair $l_\alpha(X)$, $L_\alpha(X)$ such that

$$P(l_\alpha(X) > \theta > L_\alpha(X)) > 1-\alpha, \quad \forall \theta \in \Omega$$

L. 10.1 How to build an interval

This procedure is composed by three steps:

1. Find a statistic to be used as pivot, $p(X, \theta)$.
2. Select a confidence level, $1-\alpha$ and with it find the critical values for p, c and C, as a function of α.

$$P(c \leq p(X, \theta), C) \geq 1-\alpha$$

3. Solve for $l_\alpha(X)$ and $L_\alpha(X)$.

Exercise 5. *Let X be a random sample, with $X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known. Calculate a confidence interval for μ.*

- *We will consider as a pivot $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. The distribution of Z is known.*
- *Given α, we know that*

  $$P\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

  where z_x is the percentile x of a standard normal.
- *Solve for μ.*

Exercise 6. *What happens if σ^2 is unknown in the previous exercise? We need to use the fact that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ and $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. In this case, we will use the pivot*

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

where S^2 is the sample variance. Now we need to repeat the process of the previous exercise with $t_{n-1,\alpha/2}$.

Exercise 7. *What happens if μ is unknown and we want a confidence interval for sigma^2? The pivot that we must choose is given by $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Repeat the process of the previous exercises with $\chi^2_{n-1,\alpha/2}$.*

L. 11 Hypothesis testing

Hypothesis testing is the application of interval estimation to the problems that we want to test (with, of course, innumerable amounts of twists, difficulties, and other stuff). In this type of problem, we will elaborate a hypothesis and test if it is true or not with some confidence. A hypothesis, in mathematics, is a statement of the type $\theta \in \Theta_0$. For example, the mean height of the population is $1.75m$ and the paper packages that I buy usually have less than 500 hundred sheets.

To build a contrast, we will need a null hypothesis (the one that we want to reject) and an alternative hypothesis.

$$H_0 : \theta \in \Theta_0, \quad H_A : \theta \notin \Theta_0$$

We assume that the null hypothesis is true unless the opposite is proven.In the simplest case, using the tools of interval estimation, we will need a confidence interval for our parameter. If the parameter that we want to test lies inside of the confidence interval we will not reject the null hypothesis.

With this type of procedure we can commit two different types of errors: not reject something that is false (type II error), or reject something that is true (type I error). This two errors are complementary, in the sense that if we decrease the probability of committing type I error, we will increase the probability of committing type II error.

L. 11.1 p-value and intervals

Despite this topic is long and complex, we will reduce the contrast to those done by p-values. The p-value is defined as

$$\sup_{\theta \in \Theta_0} P(T > t_0 \mid \theta)$$

where T is some statistic measuring the discrepancy between the data and the null hypothesis with value t_0 given our sample. If T is high, H_0 is more likely to be false. Intuitively, it is the highest probability among all the possible values of θ in the null hypothesis of having a certain statistic value t. If the p-value is low, we will reject the null hypothesis. This is easier to see with an example.

Exercise 8. *Let $X_1, \ldots, X_n$ be an iid (independent and identically distributed) random sample of size $n = 100$ from a distribution $N(\mu, \sigma^2 = 2.52)$. We want to test the hypothesis $H_0 : \mu \leq 10$. Suppose we have $\bar{x} = 10.8$. Calculate the p-value associated with this observed statistic. Perform the same calculation to test $H_0 : \mu = 10$.*

Since $\bar{X}$ is the maximum likelihood estimator of the parameter μ, we choose $|\bar{X} - 10|$ as a measure of discrepancy, considering only those samples where $\bar{X} > 10$. Note that if $\bar{X} \leq 10$, we would have evidence in favor of the hypothesis H_0. Therefore, $|\bar{X} - 10| = \bar{X} - 10$.

$$\begin{aligned} \textit{p-value} &= \sup_{\mu \leq 10} P(\bar{X} - 10 \geq 10.8 - 10 | H_0 \textit{ is true}) = P(\bar{X} - 10 \geq 10.8 - 10 | X \sim N(\mu = 10, \sigma^2 = 2.52)) \\ &= \frac{\bar{X} - 10}{2.5} \geq \frac{\sqrt{100} \cdot (10.8 - 10)}{2.5} = 1 - \Phi(3.2) = 0.0007 \end{aligned}$$

Therefore, it is very unlikely that with H_0 true ($\mu \leq 10$), we could obtain a sample of *size 100 with* $\bar{X} = 10.8$*. However, we have it; the sample has been obtained. Thus, we have strong evidence against* H_0*. Note that* $\Phi(\cdot)$ *is the cumulative distribution function of* $N(0,1)$*.*

Now, let's consider the second case. Here, we also choose $|\bar{X} - 10|$ *as the discrepancy statistic. Now, we consider samples where* $\bar{X} > 10$ *and also where* $\bar{X} < 10$*.*

$$\begin{aligned} p\textit{-value} &= P(|\bar{X} - 10| \geq 10.8 - 10) = 1 - P(|\bar{X} - 10| \leq 10.8 - 10) \\ &= 1 - P(-10.8 + 10 \leq \bar{X} - 10 \leq 10.8 - 10) = 0.0014 \end{aligned}$$

In this case as well, we have strong evidence against H_0*.*

References

Axler, S. (2015). *Linear algebra done right.* Springer.

Gamelin, T. W. (2000). *Complex analysis.* Departament of Mathematics, UCLA, Los Angeles, USA: Springer.

Pedersen, K. B. P. M. S. (2012). *The matrix cookbook.* Springer. Retrieved from `http://matrixcookbook.com`

Roy, S. B. . A. (2014). *Linear algebra and matrix analysis for statistics.* CRC Press.

Vandenberghe, S. B. . L. (2018). *Introduction to applied linear algebra vectors, matrices, and least squares.* Cambridge Press.