Mathematics and Statistics: Brush-up Lecture Notes

Instructor: Albert Rodriguez-Sala*

These lecture notes are designed to provide basic and introductory knowledge in Mathematics and Statistics to start a PhD in Economics. The first two years of a Econ-PhD tend to be mathematically intensive. The purpose of this course is to smooth the transition from undergraduate to a PhD program. This work has benefited from the notes collected in the PhD courses of Mathematics Brush-up, taught by Alessandro Ruggieri; Probability and Statistics, taught by Jordi Caballé; and Optimization, taught by Tomás Rodríguez; at Universitat Autonoma de Barcelona. Conceptual mistakes, typos and all other errors are mine.

Contents

*Universitat Autònoma de Barcelona and Barcelona GSE

1 Preliminaries

In mathematics, a sequence of statements or propositions, each of which is properly formed and correctly justified by those before it, is called a **proof**. Statements that we agree to accept without proof are called **axioms**. The tools which we make compound statements out of simpler ones are called connectives. The most important connective is **implication**, We use implication to join the links in the "chains" that constitute our proofs. We say proposition A implies B, denoted as $A \longrightarrow B$, where A is called the hypothesis and B the conclusion. To express implication we also say "if A then B"; "B follows from A"; B if A", "A only if B"; "A is sufficient for B; B is necessary for A. Implication is defined by the *truth table* 1:

Table 1: Truth table: Implication.

A	B	$A \rightarrow B$
T	T	T
T	F	F
F	T	T
F	F	T

We may think of an implication as a rule that is true if it is being obeyed and false if it is being broken. The rule is not broken if (False $\longrightarrow$ True) or by (False $\longrightarrow$ True) and therefore the implication holds. Such implications are said to be **vacuously true**.

If two statements always have the same truth value, we say that they are **logically equivalent** and we denote them as $A \longleftrightarrow B$. We express logically equivalent conditions with the expressions A *if and only if* B or that $A(B)$ is a *necessary and sufficient* condition for $B(A)$. When one wishes to prove the statement $A \longrightarrow B$ there are four fundamental approaches:

- **Direct proof (proof by construction)**: the conclusion is established by logically combining the axioms, definitions, and earlier theorems. Assume that A is true. Use A to show that B must be true.
- **Proof by contradiction**: consists in showing that if some statement is assumed true, a logical contradiction (a statement that is both true and false) occurs, hence the statement must be false. Assume that A is true and that not B is also true. Then show that a contradiction arises; therefore, A and not B is false which implies A and B are true.
- **Proof by contraposition**: The conclusion is established by the logically equivalent contrapositive statement: "if not A then not B.
- **Proof by induction**: First show that a single "base case" is proved. Then, show that an "induction rule" is proved that establishes that any arbitrary case implies the next case. Since in principle the induction rule can be applied repeatedly (starting from the proved base case), it follows that all (usually infinitely many) cases are provable.

2 Set Theory

2.1 Basic set operators

Let A, B be sets. A is a subset of B, denoted $A \subseteq B$ if $a \in A \longrightarrow a \in B$. Two sets are **equal**, $A = B$ if $A \subseteq B$ and $B \subseteq A$. If $A \subseteq B$ and $A \neq B$, then we write $A \subset B$.

The **union** of 2 sets A, B, defined as $A \cup B$ is the set formed by all the elements present in at least one of the sets (i.e. $A \cup B = \{x : x \in A \text{ and } x \in B\}$. The **intersection** of two sets, denoted $A \cap B$, is the set formed by all the elements present in both sets (i.e. $A \cap B = \{x : x \in A \text{ or } x \in B\}$.

Let A , B and X be sets such that $A, B \subseteq X$. The **difference** of set A minus B, denoted $A - B$ or $A \setminus B$ is defined as $A - B = \{x : x \in Aandx \notin B\}$. The **complement** of a set A, denoted A^c is $A^c = \{x \in X : x \notin A\} = X - A$

Let A,B be sets, then its **Cartesian Product** is $A \times B = \{(a, b) : a \in A, b \in B\}$.

Let A, B be sets in a universe X. The following identities capture important properties of absolute complements:

- De Morgan's Laws

$$(A \cup B)^c = A^c \cap B^c \qquad\qquad (A \cap B)^c = A^c \cup B^c$$

- Complement laws:

$$\begin{aligned} A \cup A^c &= X & A \cap A^c &= \emptyset \\ \emptyset^c &= X & X^c &= \emptyset \\ A \subseteq B \longrightarrow B^c \subseteq A^c & & & \end{aligned}$$

Let A be a set. The **power set** of A, denoted as $\mathcal{P}(A)$ or 2^A, is the set of all subsets of A: $\mathcal{P}(A) = \{x : x \subseteq X\}$.

- Example: The power set of $A = \{1, 2, 3\}$ is $\mathcal{P}(S) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.
- Note that the power set of A always includes the empty set $\emptyset$ and itself A.

2.2 Relations

Let A and B be sets. any **binary relation**, $R \subseteq A \times B$, relates elements of A to elements of B. Some basic relations are:

- The **domain** of R is given by: $dom(R) = \{x \in A : \exists y \in B \text{ and } (x, y) \in R\}$.
- The **range** of R is given by: $range(R) = \{y \in B : \exists x \in A \text{ and } (x, y) \in R\}$
- The **inverse** of R is given by: $R^{-1} = \{(x, y) \in B \times A : \exists y \in B \text{ and } x \in A, \text{ such that } (y, x) \in A \times B$.
- Let X and Y be sets. A **function** is binary relation $f \subseteq X \times Y$ if and only if for each $x \in X$, there exists a *unique* $y \in T$, such that Xfy. The convention is then to say $f : X \to Y$ and $y = f(x)$.

Typical properties of a relation R on a set A:

- R is **complete** if for all $a, b \in R$, aRb or bRa.
- R is **reflexive** if for all $a \in A$, aRa.
- R is **transitive** if aRb and $bRc \iff aRc$.
- R is **symmetric** if $aRb \iff bRa$.
- R is **asymmetric** if $aRb \iff \text{not}bRa$.
- R is **antisymmetric** ifaRb & $bRa \iff a = b$

notice that a binary relation can be simultaneously antisymmetric and symmetric; or antisymmetric and asymmetric. However, a binary relation cannot be both symmetric and anti-symmetric at the same time and, hence a binary relation cannot be symmetric, asymmetric and anti-symmetric at the same time. Thus, these three properties are not mutually exclusive but are not fully compatible either.

Let R be a relation on A, then R it is called an **equivalence relation** if it is reflexive, symmetric and transitive.
Let R be a relation on A, then R is called an **order relation** if it is reflexive, transitive and not symmetric. The pair (A, R) is then said to be an **ordered set**.
Let R be a relation on A, then R is a **weak order** if it is transitive, reflexice and complete. In economics we usually think of *preferences* as weak orders.

3 Topology

3.1 Metric spaces

Let X be a non-empty space. A function d: $X \times X \to \mathbb{R}_+$ is called a **metric** or distance function defined on X if for all $x, y, z \in X$ satisfies the following properties:

1. Positivity: $d(x, y) \geq 0$, for all $x, y \in X$
2. Non-degenerated: $d(x, y) = 0 \iff x = y$
3. Symmetry: $d(x, y) = d(y, x)$
4. Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y)$

If d is a distance function on X, then the couple (d, X) defines a **metric space**. A well-defined distance between pairs of elements in a given space induces what is called a **topological space**, as it is informative of how elements in the space relate spatially to each other. However, notice that different metrics might induce different notions of topologies on the same space.

Some standard metrics for a set $X \in \mathbb{R}^n$ are:

- The **taxi or block distance**: $d_1(x, y) = |x_1 - y_1| + \cdots + |x_n - y_n|$
- The **Euclidean Distance**: $d_2(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- For any real number $p \geq 1$, the **p-distance** is defined by: $d_p(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$
- The **maximum distance**: $d_\infty = \max_i |x_i - y_i|$

Example. *The **discrete metric** on X defines points as either equal or different:*

$$d(x, y) = 0 \quad \text{if } x = y$$
$$d(x, y) = 1 \quad \text{if } x \neq y$$

Show that the discrete metric defines a metric space.

Solution: *Note that properties of positivity, non-degenerated and symmetry are satisfied by definition of the metric. To check triangular inequality property, select three generic points x, y, z and note that*

1. *If* $x = y$ *then* $d(x, y) = 0 \leq d(x, z) + d(z, y)$
2. *If* $x \neq y$ *it must be that*
   - *either* $x \neq z \Longrightarrow d(x, y) = 1 \leq 1 + d(z, y)$
   - *or* $y \neq z \longrightarrow d(x, y) = 1 \leq d(x, z) + 1$

3.2 Notion of proximity: Open and closed sets

Among other properties, metrics endow sets with a notion of proximity. That is, with a precise definition of what it means for points to be close together. Given any metric space (d, X) and any point $x_0 \in X$, the **open ball** of radius ε around x_0 is defined as

$$B_\varepsilon(x_0) = \{x \in X : d(x, x_0 < \varepsilon\}$$

Let (d, X) be a metric space. For all $x \in X$ and $\varepsilon \in \mathbb{R}_{++}$, the **neighborhood** of a point x_0 is

$$N_\varepsilon(x_0) = \{x \in X : d(x, x_0 < \varepsilon\}$$

Equivalently, $N_\varepsilon(x_0)$ is a neighborhood of x_0 if there exists an open ball such that $B_\varepsilon(x_0) \subseteq N_\varepsilon(x_0)$. Notice that the neighborhood on a point x_0 not only depends on x and but also (and especially) on the metric d and the space X.

Given any metric space (d, X); a set $S \subseteq X$ is said to be **open** or **open in** X if it is a neighborhood of all its points. That is, for all $x \in S$, there exists a ε such that$N_\varepsilon(x) \subseteq S$. By negation a set $s \in X$ is **not open** if it exists $x \in S$, for which it does not exist ε such that$N_\varepsilon(x) \subseteq S$. The **interior** of a set $S \subseteq X$, $int(S)$ is the largest open set contained in S. Equivalently, it can be defined as the union of all the open subsets of S. Let's state some propositions for open sets:

- A set S in (d, X) is open if and only if $S = int(S)$.
- Real intervals of the form (a, b) are open sets in the metric space $(d_2, \mathbb{R})$.
- The union $\cup_{i \in I} O_i$ of an arbitrary family of open intervals is open.
- the intersection $\cap_{i \in \{1,2,\ldots,n\}} O_i$ of a finite family of intervals is open.

A set $S \subseteq X$ is **closed** or **closed in** X if its complement S^c is open. A point $x \in X$ is a boundary point of $S \subseteq X$ if every neighborhood of X contains points of S and points of S^c. The **boundary** of a set S, denoted $bn(S)$, is the set of all its boundary points, that is is the set of points which can be approached both from S and from the outside of S. The **closure** of a set S is the set and its boundary $\overline{S} = S \cup bn(S)$. Let's state some propositions for closed sets:

- A set S in a topological space (d, X) is closed if and only if the set S contains its boundary $bn(S)$.
- a set S in (d, X) is closed if and only if it coincides with its closure (i.e. $S = \overline{S}$).
- The set $S = [a, b] \subset \mathbb{R}$ is closed in $(d_2, \mathbb{R})$.
- the set $\{x\}$ is closed in $(d_2, \mathbb{R})$.
- for any metric space endowed with a discrete metric (d, X) any subset is both open and closed in X. This is not true in the case of other metric spaces.

Notice that the notion of openness and closeness are not mutually exclusive: any set might be open, closed, both, or neither. In any topological space (d, X), the entire set X is open by definition, as it is the empty set. On the other hand, the complement of the entire set X is the empty set; since X has an open complement, X must be closed too. It follows that, in any topology, the entire space is simultaneously open and closed, either called **clopen**.

3.3 Sequences and convergence in a metric space

A **sequence** (x^n) in a set X is a function $x : \mathbb{N} \to X$ of finite or infinite *ordered* list of elements of X $(x^1, x^2, x^3, \ldots)$.
Given a sequence (x^n) in a metric space (d, X), (z^m) is a **subsequence** of (x^n) if there exists a strictly increasing function $f : \mathbb{N} \to \mathbb{N}$ such that for all m, $z^m = x^{f(m)}$.
A sequence of real numbers (x^n) is said **bounded** if $\exists M \in \mathbb{R}$ for which every term x_n satisfies $|x_n| \leq M$.
Given the metric space $(d, \mathbb{R})$, a sequence (x^n) is said to **converge** to a **limit point** $\overline{x} \in \mathbb{R}$ under the metric $d(x, y)$ if

$$\forall \varepsilon > 0, \exists \overline{n} \in \mathbb{N} : n > \overline{n} \longrightarrow d(x_n, \overline{x}) < \varepsilon$$

Equivalently, a sequence (x^n) converges of $\overline{x}$ if (x^n) eventually and forever after enters any neigborhood of $\overline{x}$. That is for any neigborhood N_ε of $\overline{x}$ there exists $\overline{n} \in \mathbb{N}$ such that for all $n > \overline{n}, x_n \in N_\varepsilon(\overline{x})$.
Let's state some propositions for sequences:

- A Sequence x^n can converge to at most one limit.
- The sequence $x^n = (x_n^1, x_n^2, \ldots x_n^k)$ converges to $x = (x^1, x^2, \ldots x^k)$ in $(d_p, \mathbb{R})$ if and only if each series x_n^i converges to x_i in $(d, \mathbb{R})$.

Closed set (sequential definition): A set S is closed if and only if any convergent sequence formed by elements of S has limit in S. Otherwise, a set S is not closed if there exists at least one convergent sequence formed by elements of S that has a limit outside S.
A sequence (x^n) is called **Cauchy** if for all $\varepsilon > 0$, there exists $\overline{n}$, such that for all $k, l > \overline{n}$ implies $d(x_k, x_l) < \varepsilon$.

- **Proposition:** Every convergent sequence is Cauchy.

  Proof. Let x^n be a convergent sequence with limit point $\overline{x}$. Let $\varepsilon > 0$. Then there exists an $\overline{n} \in \mathbb{N}$ such that $d(x_n, \overline{x}) < \frac{\varepsilon}{2}$ for all $n \geq \overline{n}$. Thus, for all k, l it must be that:

  $$d(x_k, x_l) \leq d(x_k, \overline{x}), + d(x_l, \overline{x}) < \varepsilon$$

  ■

3.4 Boundedness, compactness, and completeness

A set S is said to be **complete** if every Cauchy sequence of S converges to a point in S.
Let (d, X) be a metric space and $S \subseteq X$. A class of subsets of X is said to **cover** S if $S \subseteq \cup_{i \in I} O_i$. If the sets O_i are open, then we say that O is an **open cover**.
A metric space (d, X) is said to be **compact (total boundedness definition)** if every open cover has a finite subset that also covers X. A subset S of X is said to be compact in X if every open cover of S has a

finite subset that also covers S. By negation, a space is not-compact if there exists at least one open cover of S with no finite subset that cover S.

A metric space (d, X) is said to be **compact (sequence definition)** if every sequence has a subsequence that converges to some element of X. a set $S \subseteq X$ is said to be compact if every sequence in S has a subsequence that converges to some limit in S. Notice that if X is finite, then any sequence must have a convergent subsequence.

- **Proposition:** the real space, $\mathbb{R}$, endowed with the Euclidean norm, is not a compact metric space.
- the interval $X = (0, 1)$ is not compact.

Let (d, X) be a metric space. A subset $S \subseteq X$ is said to be bounded in X if $\exists \varepsilon > 0$ and $x \in S$ such that $S \subseteq N_\varepsilon(x)$.
A set $S \subseteq \mathbb{R}$ is said **bounded from above** if $\exists k \in \mathbb{R}$ such that $k > s$ for all $s \in S$. Analogously, S is said **bounded from below** if $\exists k \in \mathbb{R}$ such that $k > s$ for all $s \in S$. Note that in both definitions k does not need to belong to the set S. A set S is bounded if it has both upper and lower bounds. Note that a set of real numbers is bounded if it is contained in a finite interval. Finite sets are bounded, as we can consider the maximum distance between two objects in the set and fix an ε equal to that and construct a neighborhood.

Suppose that S is a set of real numbers and is bounded above. Then, there is a number $M \in \mathbb{R}$ called **supremum**, **sup** or **least upper bound** such that:

1. M is a lower bound of S.
2. Given any $\varepsilon > 0$, there exists $s \in S$ such that $s < m + \varepsilon$.

Analogously, suppose S is bounded below. Then there is a number $m \in \mathbb{R}$ called **infimum**, **inf**, or **greatest lower bound** such that:

1. M is a lower bound of S.
2. Given any $\varepsilon > 0$, there exists $s \in S$ such that $s < m + \varepsilon$.

The **maximum (minimum)** of a set S is its largest (smallest) element if such an element exists. If a set has a maximum (minimum), that is a supremum (infimum) for that set. The converse is true if and only if the supremum (infimum) belongs to the set. By construction, in any bounded set both the least upper bound and the greatest lower bound are well-defined and finite, though they might not belong to the set itself (for example if the set is open).
Notice that every compact metric space is both closed and bounded. However, closed and bounded sets are not necessarily compact for some metric spaces.

Theorem. ***(Heine-Borel's theorem)*** *Consider the metric space $(d, \mathbb{R}^n)$ where $n \in \mathbb{N}$ and d is a non-discrete metric function. Then, any subset $S \subset \mathbb{N}$ is compact if and only if it is closed and bounded.*

4 Continuity

A map $f : X \to \mathbb{R}$ is said to be **continuous at a point** x if for all $\varepsilon > 0$, there exists a $\delta > 0$ such that:

$$\forall y \in X \quad s.t. \quad d(x, y) < \delta \Longrightarrow d(f(x), f(y)) < \varepsilon$$

If f is not continuous at x, we say that f is discontinuous at x. If f is continuous at every point in a set X, we say that f is continuous on X. Notice that the definition given above is equivalent to:

$$f(N_\delta(x)) \subseteq N_\varepsilon(f(x))$$

- **Proposition:** Let $X \subseteq \mathbb{R}$ and let $f : X \to \mathbb{R}$ and $g : f(X) \to \mathbb{R}$ be continuous functions defined, respectively, on X and $f(X)$. Then the composition function $h = g \circ f$ is a continuous function on X.

Notice that continuity is a local property. Consider a point x in the domain of a function f, and its image $f(x)$ in the co-domain. For any point x, the image of points nearby x under f are close to $f(x)$. That is, given $\varepsilon > 0$, we can find a $\delta > 0$ such that all points in the δ-neighborhood of x are mapped into the ε-neighborhood of $f(x)$. From the definition of continuity, different x might have different δ. To achieve global continuity, a *unique* δ regardless what x is.
A function $f : X \to \mathbb{R}$ is said to be **uniformly continuous** if for all $\varepsilon > 0$, there exists a $\delta > 0$ such that:

$$\forall y, x \in X \quad s.t. \quad d(x, y) < \delta \longrightarrow d(f(x), f(y)) < \varepsilon$$

Note that this is equivalent to sat that for uniform continuity we require a $\delta > 0$ such that δ is not a function of x. Obviously, a function is not uniformly continuous if it is not continuous.

- **Proposition:** If $f : X \to Y$ is continuous and X is compact, then it is uniformly continuous.

A function $f : X \subset \mathbb{R} \to Y \subset \mathbb{R}$ is **continuous (sequences definition)** if and only if for any sequence (x_n) converging to $\bar{x} \in \mathbb{R}$, the sequence $(f(x_n))$ converges to $f(\bar{x}) \in \mathbb{R}$

- **Proposition:** Let $f : X \subset \mathbb{R} \to Y \subset \mathbb{R}$. The following statements are equivalent:

  1. f is continuous.
  2. for all sets $O \subseteq Y$ open, the set called inverse image, $f^{-1}(O) \subseteq X$, is open.
  3. for all sets $S \subseteq Y$ closed, the set $f^{-1}(S) \subseteq X$, is closed.

That is a function f is continuous if and only if the preimatges of open (closed) sets are open (closed).

Theorem. *Intermediate value theorem: If $f : \mathbb{R} \to \mathbb{R}$ is continuous on the closed interval $[a, b]$ and $k \in \mathbb{R}$ is a scalar between $f(a)$ and $f(b)$, then there exists a $c \in [a, b]$ such that f(c)=k.*

4.1 The Weierstrass' theorem

Theorem. *Weierstrass' theorem: If X is a compact metric space and $f : X \to \mathbb{R}$ is a continuous function, then there exists maximum and minimum of f in X.*

Proof. Notice that the set spanned by the images of the whole domain X under f, is closed and bounded by continuity of f. Since $f(X)$ is bounded, both $sup\{f(x) : x \in S\}$ and $inf\{f(x) : x \in S\}$ exist. Since the supremum and the infimum belong to the closure ($cl(f(X))$ and f(X) is closed, then they must belong to f(X) as well. ■

5 Convexity

5.1 Convex sets

Consider initially the real space $\mathbb{R}^n$. For any two points $x, y \in \mathbb{R}^n$, the set $\{\lambda x + (1-\lambda)y : \lambda \in \mathbb{R}\}$ is called **line** through x and y. Similarly, the set $\{\lambda x + (1-\lambda)y : \lambda \in [0,1]\}$ is called **segment** between x and y. Moreover, for all $m \in \mathbb{N}$ and for all $x^1, \ldots, x^m \in \mathbb{R}^n$, their **convex combination** are points of the following form: $\sum_{i=1}^m \lambda_i x^i$, with $\lambda_i > 0$ for all i and $\sum_{i=1}^m \lambda_i = 1$. thus the segment between x and y is the set of all convex combinations of x and y.

A set $S \subset \mathbb{R}^n$ is said to be **convex** if for all $x, y \in S$ and for all $\lambda \in [0,1]$ we have that

$$(\lambda x + (1-\lambda)y) \in S$$

This implies that a set is convex if and only if it contains all the segments between any pair of points the set itself. Equivalently, a set S is convex if and only if every combination of points of S lies in S.

Given a set $S \in \mathbb{R}^n$, the smallest convex set that contains S is called **convex hull** of S, denoted by $co(S)$. An analogous definition of convex hull states it is the set of all convex combinations of points in the set S:

$$co(S) = \{\lambda_1 x_1 + \cdots + \lambda_k x_k : x_i \in S, \lambda_i \geq 0, \sum_{i=1}^{k} \lambda_i = 1\}$$

Notice that any closed convex set can be written as the convex hull of itself. Let's state some useful properties of convex sets:

- the intersection of any collection of convex sets is a convex set.
- The linear combination of convex sets is convex: Let S and T be convex subsets in $X \subseteq \mathbb{R}^n$ and α and β be real numbers. Then, the set $Z = \alpha S + \beta T = \{z \in Z : \alpha x + \beta y, \;\; x \in S, y \in Y\}$ is convex.

Theorem. ***Brouwer's fixed-point theorem*** *Every continuous function f from a convex compact set $X \subset \mathbb{R}^n$ to itself ($f : X \to X$) has a **fixed point** x^* such that $f(x^*) = x^*$.*

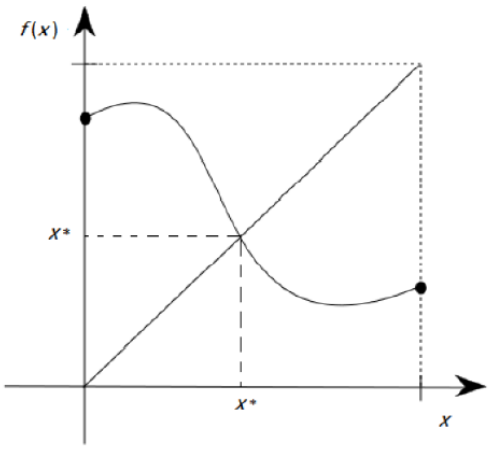


Figure 1: Illustration of the Bouwer's fixed point theorem in 1-D case (Figure from Jordi Caballé's lecture notes).

5.2 Convex functions

Let S be a convex subset in a real vector space. A function $f : X \to \mathbb{R}$ is said to be **convex (strictly convex)** if for all $x, y \in S, \lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) \underset{(<)}{\leq} \lambda f(x) + (1-\lambda)f(y)$$

And a function $f : X \to \mathbb{R}$ is said to be **concave (strictly concave)** if for all $x, y \in S, \lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) \underset{(>)}{\geq} \lambda f(x) + (1-\lambda)f(y)$$

That is a function is convex (concave) if $-f$ is concave (convex).
Let $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$, then:

- The **epigraph** of f is defined as: $epif := \{(x,y) \in \mathbb{R}^{n+1} : x \in S, y \geq f(x)\}$
- The **hypograph** of f is defined as: $hypf := \{(x,y) \in \mathbb{R}^{n+1} : x \in S, y \leq f(x)\}$

Notice that any function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its epigraph is a convex set. Conversely, f is concave if and only if its hypograph is a convex set. We define:

- The **upper contour set** of f at a point $\alpha \in \mathbb{R}$ as $U_f(\alpha) = \{x \in U : f(x) \geq \alpha\}$
- The **lower contour set** of f at a point $\alpha \in \mathbb{R}$ as $L_f(\alpha) = \{x \in U : f(x) \leq \alpha\}$

Let S be a convex subset of $\mathbb{R}^n$, then a real function $f : S \to \mathbb{R}$ is said to be

- **quasi-concave** if for all $x, y \in S$ and $\lambda \in [0,1]$ then:

$$f(\lambda x + (1-\lambda)y) \geq min\{f(x), f(y)\}$$

- **quasi-convex** if for all $x, y \in S$ and $\lambda \in [0,1]$ then:

$$f(\lambda x + (1-\lambda)y) \leq max\{f(x), f(y)\}$$

Notice that a function f is quasi-concave if $U_f(\alpha)$ is a convex set $\forall \alpha \in \mathbb{R}$. Analogously, a function f is said quasi-convex if $L_f(\alpha)$ is a convex set $\forall \alpha \in \mathbb{R}$.

- **Proposition:** Let $f : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then, f is both quasi-concave and quasi-convex.

  Proof. Consider $x, y \in \mathbb{R}, \lambda \in (0,1)$. Assume, without l.o.g, that $x > y$. Then

  $$x > \lambda x + (1-\lambda)y > y$$

  Since f is increasing, then

  $$f(x) \geq f(\lambda x + (1-\lambda)y \geq f(y)$$

  Since $f(x) = max\{f(x), f(y)\}$ then from the first inequality, $f(y) \leq max\{f(x), f(y)\}$, so f is quasi-convex. Then, since $f(y) = min\{f(x), f(y)\}$, then, from the second inequality, $f(x) \geq min\{f(x), f(y)\}$, so f is quasi-concave. ■

6 Differentiation

6.1 The Derivative and its properties

The **derivative** of a function $f : \mathbb{R} \to \mathbb{R}$ at a point $x \in \mathbb{R}$ is a real number that describes the instantaneous change of the value f as x changes:

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

A point x^* is said to be **local maximum** or **maximizer** if there exists $\delta > 0$ such that $f(x^*) \geq f(x) \forall x \in N_\delta(x^*)$, that is $\forall x \in (x^* - \delta, x^* + \delta)$. A **local minimum** or **minimizer** is defined in an analogous way: $x^* := \{\exists \delta : f(x^*) \leq f(x) \forall x \in N_\delta(x^*)\}$

- **Proposition:** Let O be an open subset of $\mathbb{R}$, and let $f : O \to \mathbb{R}$ a function with derivative at x, and let x^* be a local maximum (minimum). Then, $f'(x) = 0$

  Proof. Suppose x^* is a local maximum (for a local minimum is analogous). Then, $f(x^* + h) - f(x^*) \forall h > 0$, small enough, ($|h| < \delta$. Taking limits as h goes to 0 from the right, it must be that

  $$\lim_{h \to 0^+} \frac{f(x^* + h) - f(x^*)}{h} \leq 0 \quad \forall h \in (0, \delta)$$

  Moreover, taking limits from the left

  $$\lim_{h \to 0^-} \frac{f(x^* + h) - f(x^*)}{h} \geq 0 \quad \forall h \in (-\delta, 0)$$

  Thus, since $f(x)$ has a derivative at x, then, it must be that:

  $$\lim_{h \to 0^+} \frac{f(x^* + h) - f(x^*)}{h} = \lim_{h \to 0^-} \frac{f(x^* + h) - f(x^*)}{h} = f'(x^*) = 0$$

  ■

properties of derivatives:

- If $f(x) = c$ for all x. Then, $f'(x) = 0$
- If $f(x) = x$ then, $f'(x) = 1$
- $[f(x) \underset{(-)}{+} g(x)]' = f'(x) \underset{(-)}{+} g'(x)$ (Sum rule)
- $[kf(x)]' = kf'(x)$
- $[f(x) \cdot g(x)]' = f'(x) \cdot g(x) + g'(x) \cdot f(x)$ (Product rule)
- $\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)g(x) - g'(x)f(x)}{[g(x)]^2}$ (quotient rule)
- $(f(g(x))' = f'(g(x)) \cdot g'(x)$ Equivalently, $\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$ (Chain rule)
- $\left[\frac{1}{f(x)}\right]' = -\frac{f'(x)}{[]f(x)]^2}$

Theorem. ***(Mean Value theorem)*** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function and differentiable on (a, b). Then there exists a point $c \in [a, b]$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$*

6.2 Taylor Polynomials

When we approximate a function by its tangent line, we suffer an error. We can reduce such approximation error by approximating the function by a more sophisticated object than the tangent line: a polynomial of order higher than one.

If $f(a), f'(a), \dots f^{(n)}(a)$ all exist, the **nth Taylor polynomial** of the function f at the point a is

$$\begin{aligned} T_n(x) &= \sum_{k=0}^{n} \frac{f^k(a)}{k!}(x-a)^k \\ &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n \end{aligned}$$

If the function f has derivatives of all orders a, the **Taylor series** is a series expansion of a function about a point a given by

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} T_n(x) \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \\ &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots \end{aligned}$$

Then, we say we do a **Taylor approximation of order nth** when we approximate $f(x)$ by computing a Taylor expansion till its nth polynomial.

6.3 Derivatives of real multivariable functions

Now suppose the case of of a real value multivariate function $f : \mathbb{R}^n \to \mathbb{R}$. Then,

- The **directional derivative** of $f : \mathbb{R}^n \to \mathbb{R}$ at point x in the direction of $\mathbf{u}$ is defined by

  $$D_u f(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha \mathbf{u}) - f(x)}{\alpha}$$

  where $\mathbf{u}$ is a vector $\mathbf{u} = (u_1, u_2, \dots, u_n)$, that gives the direction (set of coordinates) to where to move.

- The **Partial derivative** of f with respect to the $i-th$ argument, x_i, is defined by

  $$\frac{\partial f(x)}{\partial x_i} = f_i(x) = D_{x_i} f(x) = Df(x, \mathbf{u_i}) \quad \text{with } \mathbf{u_i} = (\underset{1}{0}, \dots, 0, \underset{i}{1}, 0, \dots, \underset{n}{0})$$

- The **Jacobian** of f is defined as the vector $(1 \times n)$ formed by all the partial derivatives of the function

  $$\mathcal{J}_x(f) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

- the **Hessian** of f is defined as the matrix $(n \times n)$ formed by all the second-order partial derivatives of the function

$$\mathcal{H}_x(f) = \begin{bmatrix} \frac{\partial 2f}{\partial 2x_1} & \cdots & \frac{\partial f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \\ \frac{\partial 2f}{\partial x_m \partial x_1} & \cdots & \frac{\partial f}{\partial^2 x_m} \end{bmatrix}$$

The Hessian allows us to determine the convexity/concavity of the function f. Let f be a twice-differentiable function of many variables on the convex open set S, then:

1. A function f is concave if and only if $\mathcal{H}_x(f)$ is negative semi-definite for all $x \in S$.
2. A function f is convex if and only if $\mathcal{H}_x(f)$ is positive semi-definite for all $x \in S$.
3. A function f is strictly concave if $\mathcal{H}_x(f)$ is negative definite for all $x \in S$.
4. A function f is strictly convex if $\mathcal{H}_x(f)$ is positive definite for all $x \in S$.

Then, the following equivalences are true

$$\begin{aligned} f \text{ concave} \iff \text{Hessian semi-definite negative} &\iff \text{its latent roots are } \lambda_i \leq 0 \\ &\longleftrightarrow \text{principal minors order } k, \text{ have sign } -1^k \\ f \text{ convex} \iff \text{Hessian semi-definite positive} &\iff \text{its latent roots are } \lambda_i \geq 0 \\ &\longleftrightarrow \text{principal minors are} \geq 0 \end{aligned}$$

So that to know the convexity/concavity of a function we evaluate its Hessian and the eigenvalues associated with it.

6.4 Differentiability

On the one hand, the derivative of a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ at a point $x \in \mathbb{R}$ is defined as a real number that describes the instantaneous change of the value of f as x changes. On the other hand, the notion of derivative is tightly linked to the line that best approximates f near x.

Let f be a function from a vector space $V \subset \mathbb{R}^n$ to another vector space $W \subset \mathbb{R}^n$. We say that f is a linear map if for any two vectors $\mathbf{x}, \mathbf{y} \in V$ and any scalar $\alpha \in \mathbb{R}$, two conditions are satisfied:

1. $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$ (additivity)
2. $f(\alpha \mathbf{x}) = \alpha f(\mathbf{x})$ (homogeneity of degree 1)

Therefore, we can only consider the map L s.t. $L(t) = f(x) - f'(x)(t - x)$

A function f is said to be **differentiable** at x is there exists such a linear map. For functions on the reals this is equivalent to the existence of a derivative: the linear map will be the one intersecting $f(x)$ with slope $f'(x)$. For functions defined on other spaces the definition of differentiability is more general.

A function $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$ is said **differentiable** at $x \in O$ if there exists a matrix $\mathcal{J}_x(n \times 1)$, the Jacobian, such that

$$\lim_{\|h\| \to 0} \frac{\|f(x+h) - f(x) - J_x h\|}{\|h\|} = 0$$

Notice the following:

1. The derivative $Df : 0 \to$ assigns a jacobian to each x.

2. the differential $df_x : \mathbb{R}^n \to \mathbb{R}$ is a mapping that assigns to each x the linear operator $df_x(h) = \mathcal{J}_x h$

Theorem. *Consider the function $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$. if f is differentiable at $x \in O$ then f is continuous at x. Suppose f has a well-defined partial derivatives and those are continuous. Then f is differentiable.*

A function $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$ is said **continuously differentiable** if it is differentiable and Df is a continuous function. Therefore, a funcion is continuous differentiable if and only if the partial derivatives of the function of f exists and are continuous.

Theorem. *Inverse Function Theorem Let $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and let $x^0 \in O$. If the determinant of the Jacobian of f is not equal to zero at x^0, then there exists an open neighborhood of x^0, U, such that:*

1. *f is a one-to-one correspondence in U and f^{-1} is well defined.*
2. *$V = f(U)$ is an open set containing $f(x^0)$.*
3. *$f^{-}1$ is continuously differentiable with $D[f^{-1}(x^0)] = [Df(x^0)]^{-1}$*

This theorem is very useful when we cannot invert $f(x, y)$. We can find the Jacobian of the inverse, by taking the jacobian and then invert it.

7 Static Optimization

7.1 Convex Constraint Set

Let S be a convex set in $\mathbb{R}^n$, $f : S \to \mathbb{R}$ and consider the problem

$$\underset{x}{optimize} f(x) : x \in S)$$

Suppose S is an open set. Then, a necessary condition for the optimum is to be a **stationary point**:

$$Df(x*) = 0$$

That is all partial derivatives must be zero. In the more general case, this condition is not necessary,for example, if S was closed; nor sufficient, x^* could be a saddle point if f is not differentiable.

- **Sufficient conditions for global optimums**: Suppose $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$ is a differentiable function defined on a convex set S and let x^* be a stationary point int the interior of S. Then:
  1. If f is strictly concave in $S \longrightarrow x^*$ is a global maximum.
  2. If f is strictly convex in $S \longrightarrow x^*$ is a global minimum.
  3. The set of optimizers $\sigma(\theta) :=$ argmax/argmin$\{f(x) : x \in X\}$ is either empty or convex. That is, there cannot be multiple isolated points as maximizers (minimizers).

Thus, to determine whether a stationary point x^* is a maximum or a minimum we need to find the convexity and concavity of the function f.

$$f \text{is concave} \iff \text{Hessian at } x^* \text{ semi-definite negative} \iff \text{its latent roots are } \lambda_i \leq 0$$
$$f \text{is convex} \iff \text{Hessian at } x^* \text{ semi-definite positive} \iff \text{its latent roots are } \lambda_i \geq 0$$

7.2 The Lagrange problem

Consider the problem of maximize or minimize a function $f : S \subseteq \mathbb{R}^n$ knowing that the variables must satisfy a system of J equations $g_j(x) = 0$. To do so, we use the **Lagrange method** defined by the **lagrange function**

$$L(x, \lambda) = f(x) + \sum_{j=1}^{n} \lambda_j g_j(x)$$

Where note that we impose a penalization λ_j from deviating from each constraint g_j.
We compute the following *first order conditions*:

$$\frac{\partial L}{\partial x} = \frac{\partial f}{\partial x} + \sum_{j=1}^{n} \lambda_j \frac{\partial g_j}{\partial x} = 0$$
$$\frac{\partial L}{\partial \lambda_j} g_j(x) = 0 \ \ \forall j = 1, \ldots, J$$

Any stationary point x^* of the Langrangean function $L(x, \lambda)$ is a candidate to global maximum or mininum if the gradient vector of the functions g_i in the stationary point are linearly independent.

7.3 The Kuhn-Tucker problem

Consider the problem of maximize (minimize) a function $f : S \subseteq \mathbb{R}^n$ subject to J constraints $g_j(x) \leq 0$, where both f and g are continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$.

$$optimize_x f(x) \qquad s.t. g(x) \leq 0$$

The constraints can be

1. **binding** or **active** at a feasible point x^0 if they hold with equality, in which case we are in the Lagrangian case. That is there exists $\lambda_j \neq 0$
2. **Inactive** otherwise, in which case they will not have effects on the local properties of the solution. That is $\forall \lambda_j, \lambda_j = 0$

To solve this problem, we define the function

$$L(x, \lambda) = f(x) + \sum_{j=1}^{n} \lambda_j g_j(x)$$

Then, the **Kuhn-Tucker conditions** for the problem are

1. $D_x L(x, \lambda) = Df(x) + \sum_{j=1}^{n} \lambda_j Dg_j(x) = 0$
2. $\lambda_j \underset{(\leq)}{\geq} 0, g_j(x) \leq 0$ and $\lambda_j [g_j(x)] = 0$ for $j = 1, \ldots, J$
   Where
   - If x^* is a maximum $\Longrightarrow$ all $\lambda_i \geq 0$
   - If x^* is a minimum $\Longrightarrow$ all $\lambda_i \leq 0$

Exercise. ***(Optimization in a non-closed region)****. Find the maximum and minimum of the function*

$$f(x,y) = \frac{x}{2} - y \quad s.t. \begin{cases} x + e^{-x} \leq y \\ x \geq 0 \end{cases}$$

Solution*:*
First we define the Lagrangian function

$$L(x,y,\lambda) = x/2 - y - \lambda_1(x + e^{-x} - y) - \lambda_2(-x)$$

the first order conditions are

$$D_x L(x,y,\lambda) = 1/2 - \lambda_1(1 - e^{-x}) + \lambda_2 = 0$$
$$D_y L(x,y\lambda) = -1 + \lambda_1 = 0$$

From which we get $\lambda_1 = 1, e^{-x} + 0 = 1/2$*. Then, we have the following cases to study:*

1. $[\lambda_1 = 0$ *and* $\lambda_2 \neq 0]$
   we do not have any candidate.

2. $[\lambda_1 \neq 0$ *and* $\lambda_2 = 0]$
   We have a candidate given by the conditions $e^{-x} + \lambda_2 = 1/2$ *and* $x + e^{-x} = y$*. So that,*

$$e^{-x} = \frac{1}{2} \qquad x = ln\frac{1}{2}^{-1} \qquad x = ln2$$
$$and$$
$$y = ln2 + e^{-ln2} = ln2 + 1/2$$

   so that we find the point $(ln2, 1/2 + ln2)$ *which is a candidate to a maximum* ($\lambda = 1 \geq 0$)*.*

3. $[\lambda_1 = 0$ *and* $\lambda_2 = 0]$
   We do not have any candidate since $\lambda_1 = 1$

4. $[\lambda \neq 0$ *and* $\lambda \neq= 0]$
   In this case we have that $x = 0$ *and* $y = x + e^{-x} = y$ *so that we find the point* $(0,1)$ *which is neither a maximum or a minimum because* $\lambda_1 = 1 \geq 0$ *and* $\lambda_2 = -1/2$

We found $(ln2, 1/2 + ln2)$ *as a candidate to maximum with* $\lambda_1 = 1, \lambda_2 = 0$*, so that the Lagrangian function becomes* $L(x,y,\lambda) = x/2 - y - (x + e^{-x} - y)$ *and its Hessian is:*

$$\mathcal{H}(f) = \begin{bmatrix} -e^{-x} & 0 \\ 0 & 0 \end{bmatrix}$$

And $det(\mathcal{H}(f)|_{(ln2,1/2+ln2)} = 0$ *so that we cannot determine the convexity/concavity with the Hessian. However, we can do so computing the eigenvalues of* $\mathcal{H}(f)$*. Solving the characteristic equation we find* $\lambda_1 = -1/2$ *and* $\lambda_2 = 0$*. Because* $\lambda \leq 0$*, the Hessian is semi-definite negative and thus* f *is concave. The point* ***(ln2, 1/2 +ln2) is a maximum****. Also note that in this exercise there is no minimum (the region is not bounded!) :*

$$\lim_{x=0, y\to\infty} f(x,y) = \lim_{x=0, y\to\infty} \frac{0}{2} - y = -\infty$$

8 Integration

The concept of integration is strongly interlinked with derivation (*the fundamental theorem of calculus)* and with summation. An integral can be thought of a sum of infinite points. When variables (as time, distribution functions, populations) are discrete we take summations; when variables are continuous, we integrate.

8.1 The fundamental theorem of calculus

The **fundamental theorem of calculus** is a theorem that links the concept of differentiating a function with the concept of integrating a function. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and non-negative ($f \geq 0$). We want to find the area $A(x)$ between the graph of the function and the horizontal axis on $[a, x] \subseteq [a, b]$

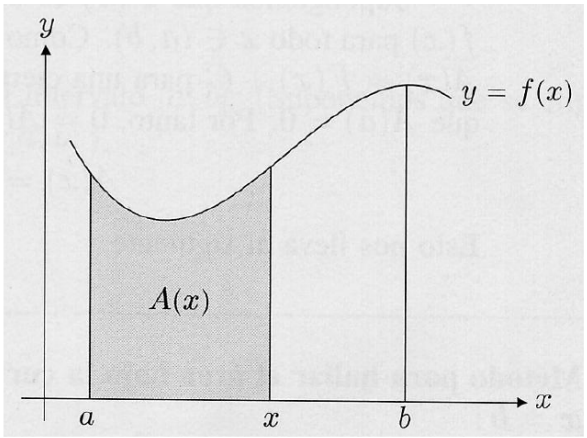


Figure 2: Finding the area below a function (Figure from Jordi Caballé's lecture notes)

Let's increase x by an infinitesimal amount Δx. The increase in the area is $A(\Delta x) = A(x + \Delta x) - A(x)$. note that:

- if $f(x)$ is increasing:
$$f(x)\Delta x \leq A(x + \Delta x) - A(x) \leq f(x + \Delta x)\Delta x$$
- if $f(x)$ is decreasing:
$$f(x)\Delta x \geq A(x + \Delta x) - A(x) \geq f(x + \Delta x)\Delta x$$

In an increasing (decreasing) function, divide the 3 terms in the previous expression by Δx

$$f(x) \underset{(geq)}{\leq} \frac{A(x + \Delta x) - A(x)}{\Delta x} \underset{(\geq)}{\leq} f(x + \Delta x)$$

And take the limit when Δx goes to 0

$$\lim_{\Delta x \to 0} f(x) \underset{(\geq)}{\leq} \lim_{\Delta x \to 0} \frac{A(x + \Delta x) - A(x)}{\Delta x} \underset{(\geq)}{\leq} \lim_{\Delta x \to 0} f(x + \Delta x)$$

Note that the middle term is the definition of a derivative (i.e. $\lim_{\Delta x \to 0} \frac{A(x+\Delta x)-A(x)}{\Delta x} = A'(x)$. Also note that since f is continuous it holds that $\lim_{\Delta x \to 0} f(x + \Delta x) = f(x)$. Thus, we arrive to the following expression

$$f(x) \underset{(\geq)}{\leq} A'(X) \underset{(\geq)}{\leq} f(x)$$

That is

$$A'(x) = f(x) \qquad \text{(Fundamental Theorem of Calculus I)}$$

■

the first part of *the fundamental theorem of calculus* and tells us how we can find the area under a curve using antidifferentiation: Finding the area between the graph of f and the horizontal axis on the interval $[a, b]$ is equivalent to find the F such that $F' = f$.

a function F is a **primitive** or **antiderivative** or **indefinite integral** of a function f if $F' = f$. We write the indefinite integral of f as

$$\int f(x)dx$$

- **Lemma 1:** If F is a primitive of f, then $G = F + C$ where C is a constant, is also a primitive of f (i.e. $G' = f$).
- **Lemma 2:** If F and G are primitives, then $\forall x$, $G(x) - F(x) = C$

From lemma 1 note that

$$A(x) = F(x) + C \qquad \text{where } F' = f$$

The area -the integral- of a point is equal to 0; $A(a) = 0$. Substituting point a for x in previous equation

$$0 = A(a) = F(a) + C \longrightarrow C = -F(a)$$

and therefore,

$$A(x) = F(x) - F(a), \qquad \text{where } F' = f$$

So that the area A in interval $[a, b]$ is

$$A = F(b) - F(a) \qquad F' = f \qquad \text{(Fundamental Theorem of Calculus II)}$$

■

The second part of the *fundamental theorem of calculus* states that the integral of a function f over some interval $[a, b]$ can be computed by using any one, say F, of its infinitely many antiderivatives.

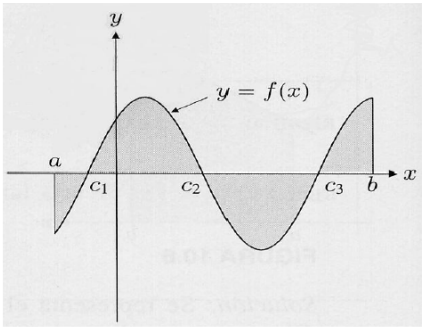
The **definite integral** of a continuous function f on the interval $[a, b]$ is

$$\int_a^b f(x)dx = F(b) - F(a) \equiv [F(x)]_a^b \equiv F(x)|_a^b, \qquad \text{where} F' = f$$

As for now, we work for the case $f(0) \geq 0$. We define the area or integral A of a continuous function $f(x)$ on $[a, b]$ where for all points in the interval $f(x)$ is negative as

$$A = \int_a^b [-f(x)]dx = -\int_a^b f(x)dx \geq 0/$$

Note that an area cannot be negative. To find the area A between the graph of a continuous function f and the horizontal axis on the interval $[a, b]$ where the function takes both positive and negative values; we will sum the integrals from the positive areas and the "inverted" integrals from the negative ones.



then the area of the shaded region will be

$$A = \underbrace{-\int_a^{c_1} f(x)dx}_{\text{negative}} + \underbrace{\int_{c_1}^{c_2} f(x)dx}_{\text{positive}} \underbrace{-\int_{c_2}^{c_3} f(x)dx}_{\text{negative}} + \underbrace{\int_{c_3}^{b} f(x)dx}_{\text{positive}} \geq 0.$$

8.2 Properties of integrals, immediate integrals, and integration by parts

Properties of the indefinite integral of a continuous function f:

1. $\int cf(x)dx = c\int f(c)dx$, where c is a constant.
2. $\int [f(x) + g(x)]dx = \int f(x)dx + \int g(x)dx$

First note that in a definite integral the variable appearing as the argument of the function f is a *mute* variable, that is:

$$\int_a^b f(x)dx = \int_a^b f(z)dz = F(b) - F(a), \text{ where F'=f}$$

Properties of the definite integral of a continuous function f:

1. $\int_a^b f(x) = -\int_b^a f(x)dx$
2. $\int_a^a f(x)dx = 0$
3. $\int_a^b cf(x)dx = c\int_a^b f(x)dx$, where c is a constant.
4. $\int_a^b [f(x) + g(x)]dx = \int_a^b f(x)dx + \int_a^b g(x)dx$
5. $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$, when $a \leq c \leq b$
6. $\frac{\partial}{\partial x}\int_a^x f(z)dz = f(x)$ and $\frac{\partial}{\partial x}\int_x^b f(z)dz = -f(x)$
7. If $f(x) \geq g(x)$ for all $x \in [a, b]$, then:
   $\int_a^b f(x)dx \geq \int_a^b g(x)dx$
8. $\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|dx$, for $a > b$

9. Cauchy-Schwartz Inequality:
$\left[\int_a^b |f(x)g(x)|dx\right]^2 \leq \left[\int_a^b [f(x)]^2\right] \cdot \left[\int_a^b [g(x)]^2\right]$

$\left[\int_a^b |f(x)g(x)|dx\right]^{1/2} \leq \left[\int_a^b [f(x)]^{1/2}\right] \cdot \left[\int_a^b [g(x)]^{1/2}\right]$

Some immediate integrals useful to remember:

- $\int f(x)dx = F(x) + C$
- $\int 0dx = C$
- $\int 1dx = x + C$
- $\int x^n dx = \frac{x^{n+1}}{n+1} + C$ for $x \neq -1$
- $\int \frac{1}{x}dx = ln|x| + C$ for $x \neq 0$
- $\int e^x dx = e^x + C$
- $\int a^x dx = a^x \frac{1}{lna} + C$
- Inverse chain rule
  $\int f'(x)[f(x)]^n dx = \frac{[f(x)]^{n+1}}{n+1} + C$
- $\frac{f'(x)}{f(x)}dx = ln|f(x)| + C$
- $\int f'(x)e^x dx = e^{f(x)+C}$
- $\int sinxdx = -cosx + C$
- $\int cosxdx = sinx + C$
- $\int lnxdx = xlnx - x + C$

Example. ***(using the chain rule)***

$$\int \underset{f(x)}{x(x^2+4)^{1/2}}dx = \int \frac{1}{2}\underset{f'(x)}{2x}\underset{[f(x)]^{1/2}}{(x^2+4)^{1/2}}dx = \frac{1}{2}\left[\frac{(x^2+4)^{3/2}}{3/2}\right] + C =$$
$$= \frac{1}{3}(x^2+4)^{3/2} + C$$

Example. ***(using the chain rule)***

$$\int e^{3x+2}dx = \int \frac{1}{3}\cdot 3e^{3x+2} = \frac{1}{3}\int 3e^{3x+2} = \frac{1}{3}e^{3x+2} + C$$

The **integration by parts** for definite integrals is defined by the following equation

$$\int_a^b F(x)g(x)dx = F(x)G(x)|_a^b - \int_a^b G(x)f(x)dx$$

Proof. Let F and G be primitive functions, then

$$\frac{\partial F(x)G(x)}{\partial x} = f(x)G(x) + F(x)g(x)$$

Computing the indefinite integral of both

$$F(x)G(x) + C = \int f(x)G(x)dx + \int F(x)g(x)dx$$

So that

$$\int F(x)_a^b g(x)dx = F(x)G(x)|_a^b - \int_a^b f(x)G(x)dx$$

■

Example. ***(integration by parts)*** *Let $x > 0$, compute*

$$\int (lnx)dx$$

Note that we can define $F(x) = lnx$ and $g(x) = 1$, so that $f(x) = \frac{1}{x}$ and $G(x) = x$. then,

$$\begin{aligned}\int (lnx)dx &= F(x)\cdot G(x) - \int f(x)G(X)dx + C = \\ &= lnx\cdot x - \int \frac{1}{x}\cdot x dx + C = lnx\cdot x - \int 1dx + C = \\ &= x(lnx - 1) + C\end{aligned}$$

8.3 Improper integrals

We call **improper integrals** the integrals of a continuous function f defined on non-closed intervals.

- integral on the interval $[a, \infty)$: $\int_a^\infty f(x)dx = \lim_{b\to\infty} \int_a^b f(x)dx$
- integral on the interval $(\infty, b]$: $\int_\infty^a f(x)dx = \lim_{a\to\infty} \int_a^b f(x)dx$
- integral on the interval (∞, ∞): $\int_\infty^\infty f(x)dx = \lim_{a\to\infty} \int_a^0 f(x)dx + \lim_{b\to\infty} \int_0^b f(x)dx$
- Integral on the right-semiclosed interval $(a, b]$: $\int_{a^+}^b f(x)dx = \lim_{z^+} \int_z^b f(x)dx$, where $z > a$.
- Integral on the open interval (a, b): $\int_{a^+}^{b^-} f(x)dx \equiv \lim_{z^+} \int_z^c f(x)dx + \lim_{z\to b^-} \int_c^z f(x)dx$, with $c \in (a, b)$

All the previous limits might fail to exist: They could be equal to $\infty - \infty$ or be equal to $\pm\infty$. In the latter case, we sat that the improper integral **diverges**.

Examples:

1. $\int_{0+}^1 \frac{1}{x}dx = \lim_{z\to 0^+} [ln|x|]_z^1 = ln1 - \lim_{z\to 0^+} ln|x| = 0 - (-\infty) = \infty$, So that the improper integral diverges.
2. $\int_1^\infty \frac{1}{x}dx = \lim_{b\to\infty} [ln|x|]_1^b = \lim_{b\to\infty} ln|x| - ln1 = \infty - 0 = \infty$, So that the improper integral diverges.
3. $\int_1^\infty \frac{1}{x^2}dx = \lim_{b\to\infty} \left[-\frac{1}{x}\right]_1^b = \lim_{b\to\infty} \left(-\frac{1}{x}\right)\ \left(-\frac{1}{1}\right) = 0 + 1 = 1$

8.4 Integration with respect to several variables

We worked for the case of a function $f : \mathbb{R} \to \mathbb{R}$, where the integration of a function f is the area between the graph of the function and the x-axis on an interval $[a, b]$. Now, consider functions $f : \mathbb{R}^n \to \mathbb{R}$ defined on higher dimensional spaces. First, suppose a function $y = f(x_1, x_2)$ on the three-dimensional Cartesian plane (i.e. $n = 2$). Then, its integral will be the volume of the region between the surface defined by the function and the plane $A = [a_1, b_1] \times [a_2, b_2]$ which contains its domain. If there are more variables, a multiple integral will yield hypervolumes of multidimensional functions.

The **multiple integral** of a continuous function $f : \mathbb{R}^n \to \mathbb{R}$ defined on a **rectangular** or **hyperrectangular** domain A such that $A = [a_1, b_1] \times [a_2, b_2] \times \dots [a_n, b_n]$ is defined as:

$$\begin{aligned}\int_A f(x_1, x_2, \dots, x_n) d(x_1, x_2, \dots, x_n) &= \int_{a^n}^{b^n} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \\ &= \int_{a^n}^{b^n} \dots \left[\int_{a_2}^{b_2} \left[\int_{a_1}^{b_1} f(x_1, x_2, \dots, x_n) dx_1 \right] dx_2 \right] \dots dx_n\end{aligned}$$

Multiple integration of a function in n variables: $f(x_1, x_2, ..., x_n)$ over a domain A is commonly represented by nested integral signs in the reverse order of execution (the leftmost integral sign is computed last), followed by the function and integrand arguments in proper order (the integral with respect to the rightmost argument is computed last).

Multiple integrals over a rectangle have many properties common to those of integrals of functions of one variable (linearity, commutativity, monotonicity, and so on). One important property of multiple integrals is that the value of an integral is independent of the order of integrands under certain conditions (*Fubini's theorem*). To minimize notation, Let's state the properties for a function $z = f(x, y)$ on the rectangular $A = [a, b] \times [c, d]$ (yet they generalize for any $n \in 2.3, 4, 5, \dots, N$).

- $\int_a^b \int_c^d f_1(x) f_2(y) dx dy = \left[\int_a^b f_1(x) dx \right] \cdot \left[\int_c^d f_2(x_2) dx_2 \right]$
- $\int_c^d \int_a^b f(x, y) dx dy = \int_a^b \int_c^d f(x, y) dy dx$ *(Fubini's theorem)*
- $\frac{\partial^2}{\partial x \partial y} \int_c^y \int_a^x f(t_1, t_2) dt_1 dt_2 = f(x, y)$

Example.

$$\int_D 2x + 3y + 4d(x, y) \quad \textit{where } D = [0, 1] \times [0, 2]$$

$$\begin{aligned}\int_0^2 \left[\int_0^1 2x + 3y + 4dx \right] dy &= \int_0^2 \left[x^2 + 3yx + 4x \right]_0^1 dy \\ &= \int_0^2 5 + 3y dy = \left[\frac{3}{2} y^2 + 5y \right]_0^2 = \frac{3}{2} 4 + 10 = 16\end{aligned}$$

8.5 Integration in non-rectangular areas

We have been working for the case A is a rectangular region (i.e $A = [a_1, b1] \times [a_2, b_2] \dots [a_n, b_n]$). However, in many applications we might need to **integrate on non-rectangular regions**. To perform an integral over a region that is not rectangular, we have to express each of the bounds of the inner integral as a function of the outer variable.

Example: Integral over a non-rectangular region
Consider the following non-rectangular region A

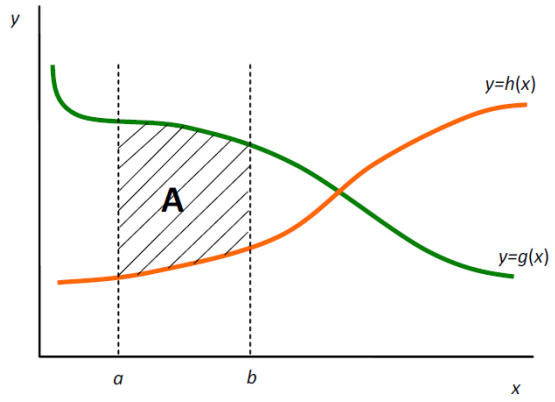


Figure 3: A non-rectangular region A (Figure from Jordi Caballé's lecture notes)

then, we compute the integral of $f(x,y)$ on A in the following manner:

$$\begin{aligned}\int_A f(x,y)d(x,y) &= \int_a^b \int_{h(x)}^{g(x)} f(x,y)dydx \\ &= \int_a^b \left[\int_{h(x)}^{g(x)} f(x,y)dy\right] dx\end{aligned}$$

Note that we express the non-linear bounds in the inner integral as a function of the outer variable.

Exercise.

$$\int_C \frac{1}{y}d(x,y) \qquad \textit{where } C := \{(x,y) : 0 \leq x \leq y, 0 \leq y \leq 1, x+y \geq 1/2\}$$

8.6 Differentiation of integrals

On the differentiability and continuity of an integral

$$H(z) := \int_a^z f(t)ft, \quad \text{Where } z \in [a,b]$$

1. If f is continuous at x, then H is differentiable at x and $H'(x) = f(x)$.
2. If f is discontinuous at x, then H is not differentiable at x
3. The function H is continuous on $[a,b]$,

$$\lim_{z\to x} H(z) = \lim_{z\to x} \int_a^z f(t)dt = \int_a^x f(t)dt = H(x) \quad \forall x \in [a,b]$$

- **Proposition:** Let $f(x,y)$ be a function such that the partial derivative $\frac{\partial f(x,y)}{\partial y}$ exists and is continuous. then

$$\frac{d}{dy}\int_a^b f(x,y)dx = \int_a^b \frac{\partial f(x,y)}{\partial y}dx$$

thus, one may interchange the integral and partial differential operators.

Proof.

$$\begin{aligned}\frac{d}{dy}\int_a^b f(x,y)dx &= \lim_{h\to 0}\frac{\int_a^b f(x,y+h)dxint_a^b f(x,y)dx}{h} = \\ &= \lim_{h\to 0}\frac{\int_a^b [f(x,y+h) - \int_a^b f(x,y)]dx}{h} = \\ &= \int_a^b \lim_{h\to 0}\left[\frac{f(x,y+h)-f(x,y)}{h}\right]dx = \\ &= \int_a^b \frac{\partial f(x,y)}{\partial y}dx\end{aligned}$$

■

- **Proposition: Liebniz rule.** Let $f(x,y)$ be a function such that the partial derivative $\frac{\partial f(x,y)}{\partial y}$ exists and is continuous, and $a(y)$ and $b(y)$ be differentiable functions. Then,

$$\frac{d}{dy}\int_{a(y)}^{b(y)} f(x,y)dx = \int_{a(y)}^{b(y)} \frac{\partial f(x,y)}{\partial y}dx + f(b(y),y)\cdot b'(y) - f(a(y),y)\cdot a'(y)$$

9 Linear Algebra

A **scalar** a is a single number. A **vector** a is a $k \times 1$ list of numbers, typically arranged in a column. Equivalently, a vector a is an element of Euclidean k space, written as $\mathbf{a} \in \mathbb{R}^k$.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

A **matrix** is a $k \times r$ rectangular array of numbers, written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \vdots & & \ddots & \\ ak1 & a_{k2} & \dots & a_{kr} \end{bmatrix}$$

The **transpose** of a matrix $A(k \times r)$ denoted A', A^r is obtained by flipping the matrix on its diagonal. If a matrix A is $k \times r$ then its transpose A' is $r \times k$. Properties of the transpose operator:

$$\begin{aligned}(A+B)' &= A' + B' \\ (cA)' &= cA' \\ (AB)' &= B'A'\end{aligned}$$

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $A = A'$. A matrix is **diagonal** if the off-diagonal elements are all zero. An important diagonal matrix is the **identity matrix** $I_k(k \times k)$ which

has ones on the diagonal. Let A be a $k \times r$ matrix, then:

$$AI_r = A$$
$$I_k A = A$$

A **partitioned matrix** is a matrix such that its elements are matrices, vectors and/or scalars.

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1r} \\ A_{21} & A_{22} & \dots & A_{2r} \\ \vdots & & \ddots & \\ A_{k1} & A_{k2} & \dots & A_{kr} \end{bmatrix}$$

9.1 Matrix addition

if matrices A, B are of the same order:

$$\underset{(k\times r)}{A} + \underset{(k\times r)}{B} = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1r} + b_{1r} \\ \vdots & \ddots & \\ ak1 + b_{k1} & \dots & akr + b_{kr} \end{bmatrix}$$

Properties:

$$A + B = B + A \text{(Commutative law)}$$
$$A + (B + C) = (A + B) + C \text{(associative law)}$$

9.2 Matrix multiplication

Scalar multiplication: Let $c \in \mathbf{R}$ be a scalar and, A a matrix $(k \times r)$. The scalar product is defined st:

$$cA = Ac = (a_{ij}c) \qquad \forall ij \in k \times r.$$

Inner product (vectors): If **a** and **b** are both $k \times 1$, then their inner product is

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \ldots + a_k b_k = \sum_{j=1}^{k} a_j b_j$$

Note that $a'b = b'a$. We say that two vectors **a** and **b** are **orthogonal** if $\mathbf{a}'$=0.
Matrix product: If the number of columns in A is equal to the number of rows in B, then A and B are **conformable**. In this event the matrix product AB is defined as:

$$\underset{(k\times r)}{A}\underset{(r\times s)}{B} = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_k \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \dots & \mathbf{b}_k \end{bmatrix} = \begin{bmatrix} \mathbf{a_1}'\mathbf{b_1} & \dots & \mathbf{a_1}'\mathbf{b_s} \\ \vdots & \ddots & \\ \mathbf{a_k}'\mathbf{b_k} & \dots & \mathbf{a_k}'\mathbf{b_s} \end{bmatrix}$$

Properties:

- Matrix multiplication is not commutative in general:

$$AB \neq BA$$

- Matrix multiplication is associative and distributive:

$$A(BC) = (AB)C$$
$$A(B+C) = AB + AC$$

the $(k \times r)$ Matrix A, $r \leq k$, is called orthonormal if $A'A = I_r$:

9.3 The trace, the rank, and the inverse matrix

The **trace** of a square matrix A is the sum of its diagonal elements:

$$tr(A) = \sum_{i=1}^{k} a_{ii}$$

The **rank** of the matrix $A(k \times r)$ with $r \leq k$ is the number of linearly independent columns $\mathbf{a}_j$ and is written as rank(A). We say A has full rank if rank(A)=r. A square matrix A is said to be **nonsingular** if it has full rank. This means that there is no vector $\mathbf{c}$ $(k \times 1)$, with $\mathbf{c} \neq 0$ such that $A\mathbf{c} = 0$.

If a square matrix $(k \times k)$ A is nonsingular then there exists a unique matrix $(k \times k)$ A^{-1} called the **inverse** of A which satisfies:

$$AA^{-1} = A^{-1}A = I_k$$

Properties for non-singular matrices A and B include:

- $AA^{-1} = A^{-1}A = I_k$
- $(A^{-1})^{-1} = A$
- $(A^{-1})' = (A')^{-1}$
- $(AB)^{-1} = B^{-1}A^{-1}$ (Exercise: Prove it)
- $(A+B)^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}$
- $det(A^{-1}) = det(A)^{-1}$

Also if A is an **orthonormal** matrix then $A^{-1} = A^{'}$.

9.4 The determinant

The **determinant** is a measure of the volume of a matrix $(k \times k)$. Let A be a (2×2) matrix, then its determinant is:

$$det(A) = |A| = a_{11}a_{22} - a_{12}a_{21}$$

Properties of the determinant:

- $det(A) = det(A')$
- $det(cA) = c^k det(A)$
- $det(AB) = det(A)det(B)$
- $det(A^{-1} = det(A)^{-1}$
- $det(A) \neq 0 \iff A$ is non-singular.

9.5 Eigenvalues

The **characteristic equation** of a $(k \times k)$ matrix is:

$$det(A - \lambda I_k) = 0$$

It has k roots not necessarily distinct and real or complex. the **eigenvalues** or **latent roots** or **characteristic roots**; λ, of A are the roots of the characteristic function.

If λ_i is an eigenvalue of A, then $A - \lambda_i I_k$ is singular so that there exists a non-zero vector $\mathbf{h_i}$ such that

$$(A - \lambda_i I_k)\mathbf{h_i} = 0$$

the vector h_i is called a **latent vector** or **characteristic vector** or **eigenvector** of A corresponding to λ_i. Let Λ be a diagonal matrix with the eigenvalues in the diagonal and let $H = [\mathbf{h_1} \dots \mathbf{h_k}]$. Some useful properties are:

- $det(A) = \prod_{i=1}^{k} \lambda_i$
- $tr(A) = \sum_{i=1}^{k} \lambda_i$
- A is non-singular $\iff \lambda_i \neq 0 \forall i = 1 \dots k$
- if A has distinct eigenvalues, then there exists a nonsingular matrix P such that $A = P^{-1}\Lambda P$ and $PAP^{-1} = \Lambda$.
- If A is symmetric, then $A = H\Lambda H'$, which is called the **spectral decomposition** of A, and $H'AH = \Lambda$. In that case the eigenvalues are all real.
- the eigenvalues of A^{-1} are $\lambda_1^{-1}, \dots, \lambda_k^{-1}$,
- The matrix H has the orthonormal properties: $H'H = I$, $HH' = I$, or alternatively, $H^{-1} = H'$ and $(H')^{-1} = H$

Note that if an eigenvalue has multiplicity 2, it can have 1 or 2 latent vectors linearly independent. If it has multiplicity 3, it can gave 1, 2, 3 latent vectors linearly independent.

Exercise. *Compute the eigenvalues and eigenvectors of the following matrices. Indicate clearly the multiplicity of each vector.*

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

9.6 Positive definiteness

A square matrix is **positive (semi-)definite**, $A \underset{(\geq)}{>} 0$, if and only if $\forall c \neq 0$:

$$c'Ac \underset{(\geq)}{>} 0$$

A matrix A is larger than B if $c'(A - B)c > 0$.

Properties:

- $A > 0 \iff$ A is symmetric and $\lambda_i \geq 0 \forall i$

- $A > 0 \longrightarrow$ A is non-singular and $A^{-1}\exists$ and $A^{-1} > 0$
- $A = G'G \longrightarrow A \geq 0$ and if G has full rank then, $A > 0$
- $A > 0 \longrightarrow \exists B$ st$A = BB'$ where B is **the matrix square root**.

9.7 Matrix Calculus

Let $\mathbf{x} = [x_1 \dots x_k]$ be a $(k \times 1)$ vector and $g(\mathbf{x}) : \mathbf{R}^k \to \mathbf{R}$ a vector function. Then, the **vector derivative** is defined as:

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial g(\mathbf{x})}{\partial x_k} \end{bmatrix} \qquad \text{and} \qquad \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial g(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g(\mathbf{x})}{\partial x_k} \end{bmatrix}$$

Some properties:

$$\begin{aligned} \frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{x}'\mathbf{a}}{\partial \mathbf{x}} = a \\ \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} &= A \\ \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &= (A + A')\mathbf{x} \\ \frac{\partial^2 \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}\partial \mathbf{x}'} &= A + A' \end{aligned}$$

Exercise. *Show that the OLS estimator in matrix notation is* $\hat{\beta} = (X'X)^{-1}X'y$

Proof. Take the linear model $y = X\beta + u$ where y is a $(n \times 1)$ vector of observations; X is a $(n \times k)$ matrix of regressors; β is a $(k \times 1)$ vector of coefficients; and u is a $(n \times 1)$ vector of errors. Then, the OLS estimator $\hat{\beta}$ is the vector of coefficients that minimize squared residuals:

$$\begin{aligned} \min_{\hat{\beta}} \{u'u\} \qquad \{u'u\} &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - y'X\hat{\beta} - \hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

note that the second and third term in the second line are scalars [1] and the fact that the transpose of a scalar is the scalar i.e. $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$.

The FOC is:

$$\frac{\partial u'u}{\partial \hat{\beta}} = \frac{\partial y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}}{\partial \hat{\beta}} = 0$$

Using matrix calculus properties (1) and (3):

$$\frac{\partial 2\hat{\beta}'X'y}{\partial \hat{\beta}} = 2X'y \qquad \frac{\partial \hat{\beta}'X'X\hat{\beta}}{\partial \hat{\beta}} = 2X'X\hat{\beta}$$

[1] the size of the second term is $(1 \times n)(n \times k)(k \times 1) = (1 \times 1)$ while the size of the third is $(1 \times k)(k \times n)(n \times 1) = (1 \times 1)$

so that the FOC becomes:

$$\frac{\partial u'u}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

Isolating $\hat{\beta}$:

$$(X'X)\hat{\beta} = X'y$$
$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y$$

Hence, The OLS estimator is:

$$\hat{\beta} = (X'X)^{-1}X'y$$

■

10 Statistics

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data. We typically work with a **statistical population** or **population** of interest, from which we associate a statistical model, a model under statistical assumptions concerning the generation of the data. Then, the **statistical method** is represented as follows

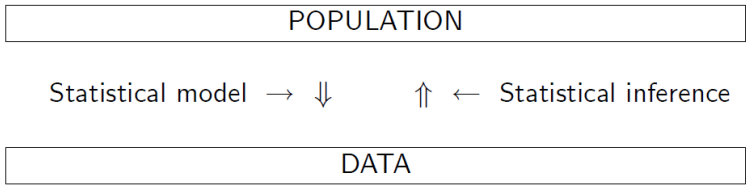


Figure 4: The statistical method (Figure from Jordi Caballé's lecture notes)

Two main statistical methods are used in data analysis:

1. **descriptive statistics**: studies how to analyze and present the data.
2. **statistical inference**: by "inverting" the statistical model, allows us to say something (make inference) about the population from the data.

The statistical model is under the framework of **probability theory**, which deals with the analysis of random phenomena.

10.1 Descriptive statistics

A **population** is a set of people or objects of interests. the elements of a population are a called **individuals**. A **sample** is a subset of a population. Population and sample are relative concepts. A **variable** is a characteristic of a population which can take different values. Variables can be

1. **Qualitative (or categorical) variables**, which are the ones that cannot be measured numerically.
2. **Quantitative variables** which are the ones that can be measured numerically. Quantitative variables can be of 2 types:

(a) **Discrete** or **countable**, which are the ones that can take values from a countable set of numbers (like any list from the natural numbers $\mathbb{N}$). Discrete variables can be **finite** or **infinite**.

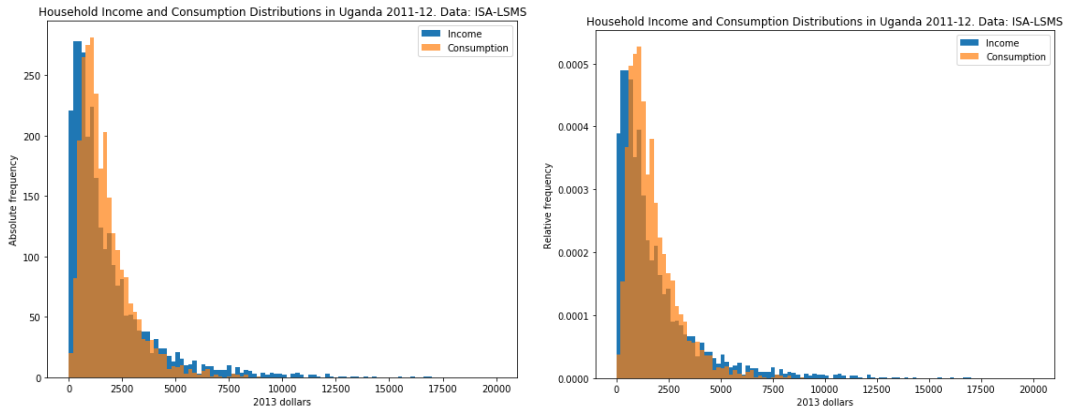
(b) **Continuous** which are the ones that can take values from an uncountable set of numbers (like the set of real numbers $\mathbb{R}$).

In descriptive statistics we apply indexes in one or more variable to describe the data. To start with, assume we care about one single variable X of the population. Then,

- The **absolute frequency** $n(x)$ of the value x is the number of times that the value appears in the data.
- The **relative frequency** $f(x)$ of the value x is the fraction or percentage of times that the value x appears in the data
$$f(x) = \frac{n(x)}{N}$$
- The **cumulative absolute frequency** $N(x)$ is the number of times that the variable X takes values smaller or equal than x
$$N(x) = \sum_{y \leq x} n(y)$$
- The **cumulative relative frequency** $F(x)$ is the fraction of times that the variable X takes values smaller or equal than x.
$$F(x) = \frac{N(x)}{N} = \frac{\sum_{y \leq x} n(y)}{N} = \sum_{y \leq x} \frac{n(y)}{N} = \sum_{y \leq x} f(y)$$

Note that cumulative frequencies are only well-defined for quantitative (discrete or continuous) variables. In continuous variables and in discrete ones (with a "large" number of values) we partition the set of values into classes, intervals or bins. that is, we work with grouped data. When we work with grouped data we use **histograms** for the corresponding graphical representation.

Example. *The histogram of income and consumption variables for households in Uganda in 2011/12. Data from the Uganda National Panel Survey (ISA-LSMS umbrella), a nationally representative sample.*



(a) Absolute Frequency

(b) Relative Frequency

Measures of central tendency:

the **Mean, average value or arithmetic mean** denoted $\overline{X}$ or $\overline{X}_n$, of a variable X is

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \sum_x x \cdot f(x)$$

where x_i is the value of the variable X for individual i. Properties of the mean, $\overline{X}$:

1. $\sum_{i=1}^{N}(x_i - \overline{X}) = 0$
2. $\overline{kX} = k\overline{X}$, where k is a constant.

   Proof. $\overline{kX} = \frac{\sum_{i=1}^{N} kx_i}{N} = k \cdot \left(\frac{\sum_{i=1}^{N} x_i}{N}\right) = k\overline{X}$ ■

3. Let $Z = \alpha X + \beta Y$, then:

$$\overline{Z} = \alpha\overline{X} + \beta\overline{Y}$$

The **Median** is a value separating the higher half from the lower half of a data sample, a population or a probability distribution. To compute it, we order all values of the variable X taken by N individuals from the smallest to the largest, so that

$$x_1 \leq x_2 \leq \cdots \leq x_{N-1} \leq x_N$$

then, the median of X is

$$Median(X) = \begin{cases} x_{N/2+1/2} & \text{if } N \text{ is odd} \\ \frac{x_{N/2}+x_{N/2+1}}{2} & \text{if } N \text{ is even} \end{cases}$$

The median is in general less sensitive to outliers and errors.

The **Mode** is the value that appears a larger number of times. The number with a larger absolute (and relative) frequency.

Measures of variability:

The **variance** of a variable X, denoted by $Var(X)$ or S_X^2, measures the average of the square of the deviations from the mean

$$Var(X) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N}$$

Properties of the variance:

1. $Var(X) = \overline{X^2} - \overline{X}^2$
2. $Var(kX) = k^2 Var(X)$

   Proof. $Var(kX) = \frac{\sum_{i=1}^{N}(kx_i - \overline{kX})^2}{N} = \frac{\sum_{i=1}^{N}(kx_i - k\overline{X})^2}{N} = \frac{k^2\sum_{i=1}^{N}(x_i - \overline{X})^2}{N} = k^2 Var(X)$ ■

3. Let $Z = \alpha X + \beta Y$, then:

$$Var(Z) = \alpha^2 Var(X) + \beta^2 Var(Y) + 2cov(X, Y)$$

The **Standard deviation** denoted by $sd(X), S_X$ is $:= +\sqrt{Var(X)}$. Note that $sd(kX) = ksd(X)$ if $k \geq 0$.

The **coefficient of variation** is

$$CV(\overline{X}) = \frac{sd(X)}{|\overline{X}|} \qquad \text{when } \overline{X} \neq 0$$

Note that the coefficient of variation is *immune* to the units of measurement:

$$CV(kX) = \frac{sd(kX)}{|\overline{kX}|} = \frac{ksd(X)}{k|\overline{X}|} = CV(X) \qquad \text{if } k > 0$$

The **range** of the variable X is the diference between the largest and the smallest value taken by the variable. The **interquartile range** is the difference between the 75% value and the 25% one.

Other measures summarizing the shape of a distribution.

The **Coefficient of skewness** or **Coefficient of asymmetry** is a measure of the asymmetry of the probability distribution of the variable X about its mean.

$$CA(X) = Skewness(X) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^3}{N \cdot sd(X)^3}$$

A negative value indicates that the tail is on the left side of the distribution, and positive a positive indicates that the tail is on the right. If the coefficient is zero, then the distribution of X is *symmetric*.

The **coefficient of Kurtosis** is a measure of the thickness of the tails of the distribution and is given by

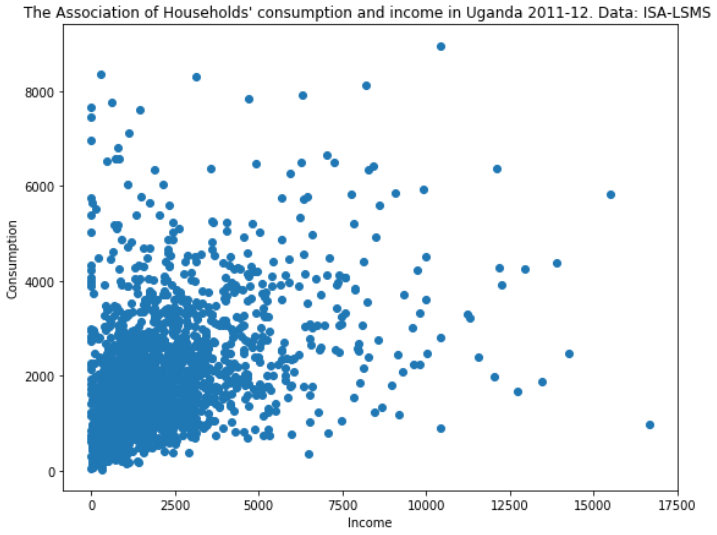
$$CK(X) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^4}{N \cdot sd(X)^4}$$

Multivariate frequency distributions.

The multivariate frequency distribution or joint distribution gives us the distribution of several variables. For instance, the joint distribution of absolute frequencies of two variables X and Y gives us the number of times that each pair of values (x, y) corresponding to the pair of variables (X,Y) appears in the data.

We can summarize these values in a **scatter plot**, which gives us an idea of the type of association or correlation between two quantitative variables.

Example. *The scatter plot of consumption and income variables in Uganda 2011/12. Data from the Uganda National Panel Survey (ISA-LSMS umbrella), a nationally representative sample.*



Similarly to the case of a single variable, we can define the relative frequency $f_{X,Y}(x,y)$ of the pair (x,y) as the fraction or percentage of times that this pair appears in the data:

$$f_{X,Y}(x,y) = \frac{n_{X,Y}(x,y)}{N}$$

the **distribution of (absolute/relative) marginal frequencies of a variable** X is the frequency distribution of this variable with independency of the values taken by the other variables.

1. The **absolute marginal frequency** is

$$n_x(x) = \sum_y n_{X,Y}(x,y)$$

2. The **relative marginal frequency** $f_x(x)$ is

$$f_X(x) = \frac{n_X(x)}{N} = \sum_y f_{X,Y}(x,y)$$

The **distribution of conditional frequencies of the variable** X **given** $Y = y$ is the *relative* frequency distribution of the variable X for all the observations where $Y = y$.

- the **conditional frequency** $fX|Y(x,y)$ of the value x taken by the variable X given $Y = y$ is

$$fX|Y(x,y) = \frac{n_{X,Y}(x,y)}{n_Y(y)} = \frac{fX,Y(x,y)}{f_Y(y)}$$

The **covariance** between X and Y, denoted $Cov(x,y)$ or $S_{X,Y}$, measures the strength of the association between these two variables and is given by

$$cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{N}$$

Properties of the covariance:

1. $S_{X,X} = S_X^2$
2. $S_{X,Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y}$

3. $S_{\alpha X,\beta Y} = \alpha \cdot \beta \cdot S_{X,Y}$ where α and β are scalars.

4. $S_{X,Y} = S_{Y,X}$

The **coefficient of correlation**, denoted by $corr(X, Y)$ or ρ_X, Y, is given by

$$Corr(X, Y) = \frac{S_{X,Y}}{S_X \cdot S_Y}$$

Note that $Corr(\alpha X, \beta Y) = Corr(X, Y)$, so that the coefficient of correlation is *immune* to the units of measurement. Also note that $|Corr(X, Y)| \leq 1$.

When we study several variables it is useful to use the mean vector and the variance covariance matrix. Consider the following vector of K variables:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix}$$

Then, the **mean vector** or **vector of means** of variables $\mathbf{X}$ is

$$\overline{\mathbf{X}} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix}$$

and, the **variance covariance matrix** or **covariance matrix** of the vector $\mathbf{X}$ of variables is the following $K \times K$ matrix

$$\mathbf{S} = \begin{bmatrix} S_1^2 & S_{12} & \dots & S_{1K} \\ S_{21} & S_2^2 & \dots & S_{2K} \\ \vdots & & \ddots & \\ S_{K1} & S_{K2}^2 & \dots & S_K^2 \end{bmatrix}$$

Note that since $\mathbf{S}_{ij} = \mathbf{S}_{ji}$ for all (i, j), the covariance matrix is symmetric.

Now, let us define the vector of scalars $\alpha = [\alpha_1, \dots, \alpha_K]'$ and define the variable $\mathbf{Z}$ as a linear combination of the variables appearing in $\mathbf{X}$: $\mathbf{Z} = \sum_j \alpha_j X_j = \alpha' \mathbf{X}$. Then,

1. $\overline{\mathbf{Z}} = \sum_j \alpha_j \overline{\mathbf{X}} = \alpha' \overline{\mathbf{X}}$

2. $\mathbf{S}_Z^2 = \sum_j \alpha_j^2 S_j^2 + 2 \sum_i \sum_j \alpha_j \alpha_i S_i j = \alpha' \mathbf{S} \alpha$

Also note that $\mathbf{S}_\mathbf{Z}^2 = \alpha' \mathbf{S} \alpha \geq 0$ for all vector of scalars α. Therefore, the covariance matrix is positive semi-definite.