MOHAMED KHIDER UNIVERSITY OF BISKRA.
FACULTY OF EXACT SCIENCES AND NATURAL AND LIFE SCIENCES
DEPARTMENT OF BIOLOGY

COURSE TITLE

Mathematic and statistic Level 1^{st} year LMD

Dr. AFROUN Faïrouz

Biskra university, 2025/2026

Contents

List of Figures

List of Tables

General introduction

Mathematics and statistics are important for biology students because they provide the tools to analyze data, build models, and conduct research. These subjects are crucial for understanding biological concepts quantitatively, whether it's tracking population changes, analyzing genetic data, or understanding rates of chemical reactions.

Key areas where math and statistics are applied in biology

- **Data analysis**: Statistics are used to describe data, uncover patterns, and determine if results are statistically significant.
- **Modeling**: Mathematics is used to create models that can predict or describe biological phenomena, such as how a disease spreads or how populations change over time.
- **Experimental design**: Statistics is essential for designing experiments that will yield reliable results and avoid common pitfalls in data collection.
- **Specific disciplines**:
  - Biochemistry: Mathematical computer models help in understanding complex reactions and processes.
  - Genetics: Math is used in computer programs for analyzing DNA sequences.
  - Zoology: Mathematical tools can express the relatedness of species and estimate when they diverged from a common ancestor.
  - Medicine: Diagnostic tools like MRI and EEG rely on mathematical principles.
- **Lab work**: Basic math is necessary for practical tasks like preparing solutions with correct concentrations, a fundamental skill in many biology labs.

Chapter 1

Limits and Continuity

Introduction

In this chapter, we study two fundamental notions of calculus: **limits** and **continuity**. These concepts are essential for understanding how functions behave and change, which is useful in many areas of biology such as population dynamics, enzyme kinetics, and growth models.

1.1 What is a function?

A function is a relation that associates each element of a set called the starting set with an element of another set called the ending set.

$$f : D \to A$$
$$x \to f(x).$$

Définition 1.1 *Let f be a function on D_f.*

- *The function f is **odd** iff (if and only if) the following statements are correct.*
  1. *$\forall x \in D_f$ then $-x \in D_f$*
  2. *$\forall x \in D_f$ we have $f(-x) = -f(x)$.*
- *The function f is **even** iff the following statements are correct.*
  1. *$\forall x \in D_f$ then $-x \in D_f$*
  2. *$\forall x \in D_f$ we have $f(-x) = f(x)$.*
- *A **periodic** function is a function that repeats itself in regular intervals or periods. The function f is said to be periodic if $\forall\ x \in D_f\ \exists\ p \in \mathbb{R}^*$:*

$$f(x+p) = f(x)$$

.

Symbols	Explanation
$\forall$	for all
$\exists$	exists
$\in$	in
$\Rightarrow$	implies
$\Leftrightarrow$	equivalane
iff	if and only if

1.2 Limit of a Function

The **limit** of a function describes the value that the function approaches as the variable approaches a given number.

2.2 Formal Definition

Let $f(x)$ be defined around a. We say that $f(x)$ tends to a limit L as x tends to a, and we write:

$$\lim_{x \to a} f(x) = L$$

if for every $\varepsilon > 0$, there exists $\delta > 0$ such that:

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

Left and Right Limits

- The **right limit** of f at a is $\lim\limits_{x \to a^+} f(x)$ (when $x > a$). - The **left limit** of f at a is $\lim\limits_{x \to a^-} f(x)$ (when $x < a$).

If both limits exist and are equal, then the limit at a exists.

2.4 Example

Let

$$f(x) = \begin{cases} x^2, & x < 2, \\ 3x - 2, & x \geq 2. \end{cases}$$

We have:

$$\lim_{x \to 2^-} f(x) = (2)^2 = 4, \quad \lim_{x \to 2^+} f(x) = 3(2) - 2 = 4$$

Therefore, $\lim_{x \to 2} f(x) = 4$.

1.3 Continuous Functions

Definition

A function f is said to be **continuous** at a point a if:

$$\lim_{x \to a} f(x) = f(a)$$

That is, the limit of $f(x)$ as x approaches a is equal to the actual value of f at that point.

Continuity on an Interval

A function f is continuous on an interval I if it is continuous at every point of I.

Examples

1. $f(x) = x^2$ is continuous everywhere on $\mathbb{R}$. 2. $f(x) = \frac{1}{x}$ is continuous on $\mathbb{R} \setminus \{0\}$.

Types of Discontinuity

- **Jump discontinuity:** when the left and right limits exist but are different.

$$f(x) = \begin{cases} 2x+1, & x < 1 \\ x+3, & x \geq 1 \end{cases}$$

  Then,

$$\lim_{x\to 1^-} f(x) = 3, \quad \lim_{x\to 1^+} f(x) = 4$$

  Since the two limits are not equal, f has a **jump discontinuity** at $x = 1$.

- **Infinite discontinuity:** when the function tends to infinity near a point. Example: $f(x) = \frac{1}{x}$ at $x = 0$.

- **Removable discontinuity:** when the limit exists but is not equal to $f(a)$. Example: $f(x) = \frac{x^2-1}{x-1}$ at $x = 1$.

The Intermediate Value Theorem (IVT)

Theorem Statement

If f is continuous on a closed interval $[a, b]$ and N is any number between $f(a)$ and $f(b)$, then there exists at least one $c \in [a, b]$ such that:

$$f(c) = N$$

Example

Let $f(x) = x^3 - x - 2$ on $[1, 2]$.

$$f(1) = -2, \quad f(2) = 4$$

Since 0 is between -2 and 4, by the IVT, there exists $c \in (1, 2)$ such that $f(c) = 0$. Numerically, $c \approx 1.52$.

Chapter 2

Differentiability

Introduction

Before studying differentiability, we must know the concepts of **function**, **limit**, and **continuity**. Differentiability helps us describe how fast a quantity changes, for example, how fast a population grows or how a reaction rate changes with temperature. It tells us when a function can be approximated by a straight line near a point.

2.1 Definition of Differentiability (One Variable)

Definition

Let f be a function defined in the neighborhood of x_0. We say that f is differentiable at a point x_0 if the limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists in $\mathbb{R}$. When this limit exists, it is denoted by $f'(x_0)$ and called the derivative of f at x_0.

Remark *If we put $x - x_0 = h$, the quantity $\frac{f(x) - f(x_0)}{x - x_0}$ becomes $\frac{f(x_0 + h) - f(x_0)}{h}$. So we can define the notion of differentiability of f at x_0 in the following way:*

$$f \textit{ is differentiable at the point } x_0 \Leftrightarrow \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \textit{ exists in } \mathbb{R}$$

Notations:

We can use the notations $f'(x_0)$, $Df(x_0)$, $\frac{df}{dx}(x_0)$ to designate the derivative of f at x_0.

Example

1. The function $f(x) = x^2$ is differentiable at any point $x_0 \in \mathbb{R}$ and the derivative $f'(x_0) = 2x_0$. As an explanation, given $x_0 \in \mathbb{R}$ we have:

$$\lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{h\to 0} \frac{(x_0+h)^2 - x_0^2}{h} = \lim_{h\to 0}(h+2x_0) = 2x_0.$$

2. The function $f(x) = \sin(x)$ is differentiable at any point $x_0 \in \mathbb{R}$ and the derivative $f'(x_0) = \cos(x_0)$. As an explanation, given $x_0 \in \mathbb{R}$ we have:

$$\begin{aligned}\lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} &= \lim_{h\to 0} \frac{\sin(x_0+h)-\sin(x_0)}{h} \\ &= \lim_{h\to 0} \cos\left(\frac{2x_0+h}{2}\right)\frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}} = \cos(x_0)\end{aligned}$$

Definition (Left and right derivative)

1. Let f be a function defined on an interval of type $[x_0, x_0+\alpha[$ with $\alpha > 0$. We say that f is right-differentiable at x_0 iff:

$$\lim_{h\to 0^+} \frac{f(x_0+h)-f(x_0)}{h}$$

exists in $\mathbb{R}$. This limit is denoted by $f'_r(x_0)$ and is called the right derivative of f at x_0.

2. Let f be a function defined on an interval of type $]x_0-\alpha, x_0]$ with $\alpha > 0$. We say that f is left-differentiable at x_0 iff:

$$\lim_{h\to 0^-} \frac{f(x_0+h)-f(x_0)}{h}$$

exists in $\mathbb{R}$. This limit is denoted by $f'_l(x_0)$ and is called the left derivative of f at x_0.

Proposition

Let f be a function defined in the neighborhood of x_0, we have:

$$f \text{ is differentiable at } x_0 \iff \begin{cases} f \text{ is differentiable on the right and left at } x_0 \\ \quad\text{and} \\ f'_r(x_0) = f'_l(x_0) \end{cases}$$

Example

Let $f(x) = |x|$, we have:

$$\lim_{h\to 0^-} \frac{f(0+h)-f(0)}{h} = \lim_{h\to 0^-} \frac{|h|}{h} = \lim_{h\to 0^-} -\frac{h}{h} = -1 = f_l'(0)$$
$$\lim_{h\to 0^+} \frac{f(0+h)-f(0)}{h} = \lim_{h\to 0^+} \frac{|h|}{h} = \lim_{h\to 0^+} \frac{h}{h} = 1 = f_r'(0)$$

$\Longrightarrow$ The function f is differentiable on the right and on the left at $x_0 = 0$ and moreover $f_r'(0) = 1$ and

$f_l'(0) = -1$, so $f_l'(0) \neq f_r'(0) \Longrightarrow f$ is not differentiable at $x_0 = 0$

Geometrical interpretation

The figure below shows the graph of a function $y = f(x)$:

The ratio $\frac{f(x_0+h)-f(x_0)}{h} = \tan(\theta)$ is the slope of the straight line joining point $A(x_0, f(x_0))$ to point $B(x_0+h, f(x_0+h))$ on the graph. When $h \to 0$, this line tends towards the tangent (AC) to the curve at a point $A(x_0, f(x_0))$. So we get:

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} = \tan(\alpha) = \frac{CD}{AD}$$

is the slope of the tangent to the curve at point $A(x_0, f(x_0))$.

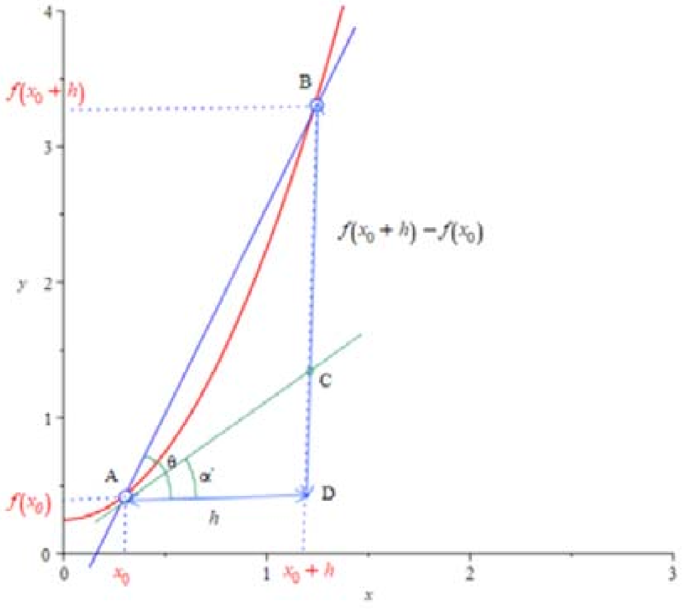


Figure : Geometrical Interpretation of Differentiability at a point x_0

Remark *According to the figure above, the equation of the tangent to the curve $y = f(x)$ at the point $A(x_0, f(x_0))$ is $y - f(x_0) = f'(x_0)(x - x_0)$*

Proposition

Let f be a function differentiable at a point x_0, then f is continuous at x_0.

Proof:

We have: $\lim_{x\to x_0}(f(x)-f(x_0))=\lim_{x\to x_0}\left(\frac{f(x)-f(x_0)}{x-x_0}\right)(x-x_0)$
Since f is differentiable at x_0 we get:
$\lim_{x\to x_0}(f(x)-f(x_0))=\lim_{x\to x_0}f'(x_0)(x-x_0)=0 \implies f$ is continuous at x_0

Remark *The opposite of this theorem is incorrect. A function can be continuous at a point x_0 without being differentiable at the same point. For example, the function $x\mapsto|x|$ is continuous at $x_0=0$ but not differentiable at the same point.*

Differential on an interval. Derivative function.

Definition

Let f be a function defined on an open interval I. We say that f is differentiable on I if: it is differentiable at any point on I. The function defined on I by: $x\mapsto f'(x)$ is called the derivative function or simply the derivative of the function f and is denoted by f' ou $\frac{df}{dx}$.

Remark *let f be a function defined on an interval I and $a,b\in\mathbb{R}\cup\{+\infty,-\infty\}$ then:*

- *We say that f is differentiable on $I=[a,b]$ iff: it is differentiable on the open interval $]a,b[$ and differentiable on the right at a and on the left at b.*
- *We say that f is differentiable on $I=[a,b[$ if: it is differentiable on the open interval $]a,b[$ and differentiable on the right at a.*
- *We say that f is differentiable on $I=]a,b]$ if: it is differentiable on the open interval $]a,b[$ and differentiable on the left at b.*

Operations on differentiable functions

Proposition : (At a point)

Let f,g be two functions differentiable at x_0, then we have:

- $f+g$ is differentiable at x_0 et $(f+g)'(x_0)=f'(x_0)+g'(x_0)$
- $f.g$ is differentiable at x_0 et $(f.g)'(x_0)=f'(x_0).g(x_0)+f(x_0).g'(x_0)$
- If we have: $f(x_0)\neq 0$, alors $\frac{1}{f}$ is differentiable at x_0 et $\left(\frac{1}{f}\right)'(x_0)=-\frac{f'(x_0)}{f(x_0)^2}$
- If we have: $g(x_0)\neq 0$, then $\frac{f}{g}$ is differentiable at x_0 and

$$\left(\frac{f}{g}\right)'(x_0)=\frac{f'(x_0).g(x_0)-f(x_0).g'(x_0)}{g(x_0)^2}$$

Proposition : (On an interval)

Let f and g be two functions differentiable on an open interval I then:

- $f+g$ is differentiable on I and $(f+g)' = f'+g'$
- $f.g$ is differentiable on I and $(f.g)' = f'.g+f.g'$
- If $f \neq 0$ on I, $\frac{1}{f}$ is differentiable on I and $\left(\frac{1}{f}\right)' = -\frac{f'}{f^2}$
- If $g \neq 0$ on I, $\frac{f}{g}$ is differentiable on I and

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

Proposition : Differentiability and composition

Let $f : I \longrightarrow \mathbb{R}$ and $g : J \longrightarrow \mathbb{R}$ be two functions where I and J are two open intervals such that: $f(I) \subset J$

- **Differentiability at a point:** If f is differentiable at x_0 and g is differentiable at $f(x_0)$, then $g \circ f$ is differentiable at x_0 and $(g \circ f)'(x_0) = f'(x_0).g'(f(x_0))$
- **differentiability on an interval:** If f is differentiable on I and g is differentiable on J, then $g \circ f$ is differentiable on I and $(g \circ f)' = f'.(g' \circ f)$

Proposition : Differentiability and inverse function

Let $f : I \longrightarrow J$ be a bijective and differentiable function at $x_0 \in I$. Then f^{-1} is differentiable at $y_0 = f(x_0)$ if and only if $f'(x_0) \neq 0$ and in this case: $(f^{-1})'(y_0) = \frac{1}{f'(x_0)}$.

Proposition

Let $f : I \longrightarrow J$ be a bijective and differentiable function on I. If $f' \neq 0$ on I, then f^{-1} is differentiable on J and we have : $(f^{-1})' = \frac{1}{f' \circ f^{-1}}$

Mean value Theorem

Theorem : (Rolle's theorem)

Let f be a function defined on $[a,b]$. If we have:

1. f is continuous on $[a,b]$.
2. f is differentiable on $]a,b[$
3. $f(a)=f(b)$

then there exists a real number $c\in]a,b[$ such that $f'(c)=0$

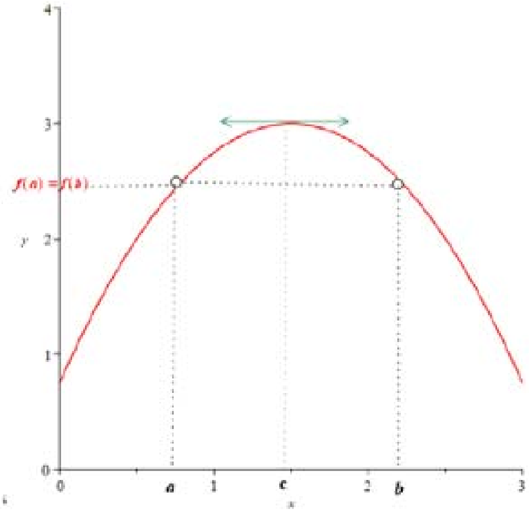


Figure : Geometrical interpretation of Rolle's theorem

Theorem : (Mean value Theorem)

Let f be a function defined on $[a,b]$, if we have:

1. f is continuous on $[a,b]$.
2. f is differentiable on $]a,b[$

then there exists a real number $c\in]a,b[$ such that:

$$f(b)-f(a)=f'(c)(b-a)$$

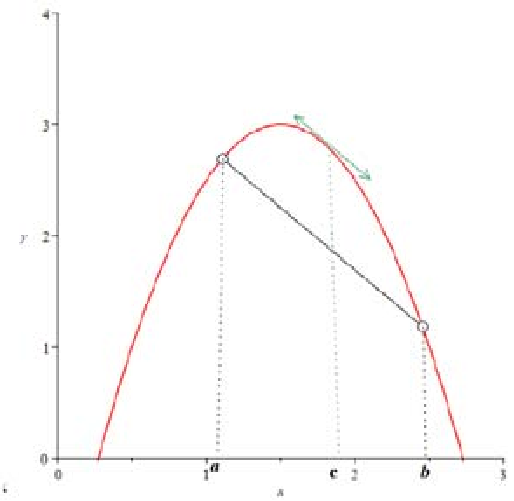


Figure : Geometrical interpretation of the mean value theorem

Consequence:(second form of the mean value theorem)

Let f be a function defined on I, $h > 0$ and $x_0 \in I$ such that $x_0 + h \in I$, then if we have:

1. f is continuous on $[x_0, x_0 + h]$.
2. f is derivable on $]x_0, x_0 + h[$

then there exists a $\theta \in]0, 1[$ such that:

$$f(x_0 + h) - f(x_0) = f'(x_0 + \theta.h)h$$

Example

By using the mean value theorem, show that:

$$\forall x > 0; \sin(x) \leq x$$

By putting $f(t) = t - \sin(t)$ we get:

$$\forall x > 0 \text{ we have:} \begin{cases} f \text{ is continuous on } [0, x] \\ \quad\text{and} \\ f \text{ is differentiable on }]0, x[\end{cases}$$

According to the mean value theorem, there exists $c \in]0, x[$ such that:

$$f(x) - f(0) = f'(c)(x - 0)$$

$$\Longleftrightarrow x - \sin(x) = (1 - \cos(c))x \Longleftrightarrow \sin(x) = \cos(c)x$$

$$\Longrightarrow \sin(x) \leq x \text{ (as } \cos(c) \leq 1)$$

Theorem : Generalized mean value theorem

Let f and g be two real functions defined on $[a, b]$ such that:

1. f and g are continuous on $[a, b]$.
2. f and g are differentiable on $]a, b[$.

Then there exists a real number $c \in]a, b[$ such that:

$$(f(b) - f(a))g'(c) = (g(b) - g(a))f'(c)$$

Proposition : (Variations of a function)

Let f be a continuous function on $[a, b]$ and differentiable on $]a, b[$, we have:

1. If $f'(x) > 0$ on $]a, b[$, then f is strictly increasing on $[a, b]$.
2. If $f'(x) \geq 0$ on $]a, b[$, then f is increasing on $[a, b]$.
3. If $f'(x) < 0$ on $]a, b[$, then f is strictly decreasing on $[a, b]$.
4. If $f'(x) \leq 0$ on $]a, b[$, then f is decreasing on $[a, b]$.
5. If $f'(x) = 0$ on $]a, b[$, then f is constant on $[a, b]$.

L'Hôpital's rule

Theorem

Let f and g be two continuous functions on I (where I is a neighborhood of x_0), differentiable on $I - \{x_0\}$ and satisfying the following conditions:

1. $\lim\limits_{x \to x_0} f(x) = \lim\limits_{x \to x_0} g(x) = 0$
2. $\forall x \in I - \{x_0\}$; $g'(x) \neq 0$

Then:

$$\lim_{x \to x_0} \frac{f'(x)}{g'(x)} = l \implies \lim_{x \to x_0} \frac{f(x)}{g(x)} = l$$

Example

$$\lim_{x \to 0} \frac{\sin(x)}{x} = \lim_{x \to 0} \frac{\cos(x)}{1} = 1$$

Remark *The converse is generally false. For example:* $f(x) = x^2 \cos(\frac{1}{x})$, $g(x) = x$. *We have:* $\lim\limits_{x \to 0} \frac{f(x)}{g(x)} = \lim\limits_{x \to 0} x \cos(\frac{1}{x}) = 0$. *While* $\lim\limits_{x \to 0} \frac{f'(x)}{g'(x)} = \lim\limits_{x \to 0}(2x \cos(\frac{1}{x}) + \sin(\frac{1}{x}))$ *does not exist (since:* $\lim\limits_{x \to 0} \sin(\frac{1}{x})$ *does not exist)*

Remark *Also, the Hopital's rules is true when* $x \to \pm\infty$

2.2 Rules of Differentiation

Rule	**Formula**
Constant	$(c)' = 0$
Power	$(x^n)' = nx^{n-1}$
Sum	$(f + g)' = f' + g'$
Product	$(fg)' = f'g + fg'$
Quotient	$\left(\frac{f}{g}\right)' = \frac{f'g-fg'}{g^2}$
Chain Rule	$(f(g(x)))' = f'(g(x)) \cdot g'(x)$

Common Derivatives:

Function	Derivative
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
e^x	e^x
$\ln x$	$\frac{1}{x}$

Chapter 3

Integrals

Integration is one of the two main operations in calculus (the other is differentiation). While differentiation finds the *rate of change*, integration finds the *total quantity* or the *area under a curve.*

Integration is used in biology to calculate:

- Total bacterial growth over time,
- Total oxygen consumption,
- Accumulated concentrations of a substance.

If we know a function $f(x)$ that represents a rate (for example, growth rate), the **integral** of $f(x)$ between a and b gives the total change:

$$\text{Area under } y = f(x) \text{ between } x = a \text{ and } x = b.$$

3.1 Indefinite Integral

The **indefinite integral** of $f(x)$ is a function $F(x)$ such that:

$$F'(x) = f(x)$$

and we write:

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

where C is the constant of integration.

Examples

In the following c denote a real constant ($c \in \mathbb{R}$).

1. $\int x^n dx = \frac{1}{n+1} x^{n+1} + c$, with $n \neq -1$.
2. $\int x^{-1} dx = \int \frac{1}{x} dx = \ln(|x|) + c$.
3. $\int e^x dx = e^x + c$.
4. $\int a^x dx = \frac{a^x}{\ln a} + c$.

5. $\int \sin(x)dx = -\cos(x) + c.$

6. $\int \cos(x)dx = \sin(x) + c.$

7. $\int \frac{1}{\sin^2(x)}dx = -ctan(x) + c.$

8. $\int \frac{1}{\cos^2(x)}dx = \tan(x) + c.$

9. $\int \frac{1}{\sqrt{1-x^2}}dx = \arcsin(x) + c.$

10. $\int \frac{1}{1+x^2}dx = \arctan(x) + c.$

11. $\int \frac{1}{\sqrt{a^2-x^2}}dx = \arcsin\left(\frac{x}{a}\right) + c.$

12. $\int \frac{1}{a^2+x^2}dx = \frac{1}{a}\arctan(\frac{x}{a}) + c.$

13. $\int \frac{1}{a^2-x^2}dx = \frac{1}{2a}\ln\left(\left|\frac{a-x}{a+x}\right|\right) + c.$

14. $\int \frac{1}{\sqrt{x^2\pm a^2}}dx = \ln\left(\left|x + \sqrt{a^2 \pm x^2}\right|\right) + c$

3.2 Definite Integral

The **definite integral** gives a numerical value representing the area under the curve between a and b:

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

Example

$$\int_0^2 x^2\,dx = \left[\frac{x^3}{3}\right]_0^2 = \frac{8}{3}$$

3.3 Properties of Integrals

$$\int_a^b kf(x)\,dx = k\int_a^b f(x)\,dx$$
$$\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$$
$$\int_a^a f(x)\,dx = 0$$
$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$$

3.4 Integration Techniques

Substitution Method

If $u = g(x)$, then:

$$\int f(g(x))g'(x)\,dx = \int f(u)\,du$$

Example:

$$\int 2xe^{x^2}dx$$

Let $u = x^2 \Rightarrow du = 2x\,dx$

$$\int e^u du = e^u + C = e^{x^2} + C$$

$\int \frac{e^{-x}}{\sqrt{1-e^{-2x}}}dx =??$
To compute this integral we use the following substitution
we put

$$t = e^{-x} \Longrightarrow dt = -e^{-x}dx$$

so

$$\begin{aligned}\int \frac{e^{-x}}{\sqrt{1-e^{-2x}}}dx &= \int \frac{-1}{\sqrt{1-t^2}}dt = \arccos(t) + c = \\ &= \arccos\left(e^{-x}\right) + c = -\arcsin\left(e^{-x}\right) + c \qquad \text{with } c \in \mathbb{R}.\end{aligned}$$

Integration by Parts

$$\int u\,dv = uv - \int v\,du$$

Example:

$$\int xe^x dx$$

Let $u = x \Rightarrow du = dx$, and $dv = e^x dx \Rightarrow v = e^x$

$$\int xe^x dx = xe^x - \int e^x dx = e^x(x-1) + C$$

To compute the integral we use the integration by parts method. For this, we take the following:

$$\begin{cases} u &= x+2 \\ v' &= \cos(2x) \end{cases} \Longrightarrow \begin{cases} u' &= 1 \\ v &= \frac{1}{2}\sin(2x) \end{cases}$$

Thus,

$$\begin{aligned}\int (x+2)\cos(2x)dx &= \frac{1}{2}(x+2)\sin(2x) - \int \frac{1}{2}\sin(2x)dx \\ &= \frac{1}{2}(x+2)\sin(2x) + \frac{1}{4}\cos(2x) + c \qquad \text{with } c \in \mathbb{R}.\end{aligned}$$

Integration of Rational Functions

In this section we will take a more detailed look at the use of partial fraction decompositions in evaluating integrals of rational functions, a technique we first encountered in the inhibited growth model example in the previous section.
We begin with a few examples to illustrate how some integration problems involving rational functions may be simplified either by a long division or by a simple substitution.

Example To evaluate $\int \frac{x^2}{x+1}\,dx$, we first perform a long division of $x+1$ into x^2 to obtain

$$\frac{x^2}{x+1} = x - 1 + \frac{1}{x+1}.$$

Then

$$\int \frac{x^2}{x+1}\,dx = \int \left(x - 1 + \frac{1}{x+1}\right) dx = \frac{1}{2}x^2 - x + \log|x+1| + c.$$

Example To evaluate $\int \frac{2x+1}{x^2+x}\,dx$, we make the substitution

$$\begin{aligned} u &= x^2 + x \\ du &= (2x+1)dx. \end{aligned}$$

Then

$$\int \frac{2x+1}{x^2+x}\,dx = \int \frac{1}{u}\,du = \log|u| + c = \log|x^2+x| + c.$$

Example To evaluate $\int \frac{x}{x+1}\,dx$, we perform a long division of $x+1$ into x to obtain

$$\frac{x}{x+1} = 1 - \frac{1}{x+1}.$$

Then

$$\int \frac{x}{x+1}\,dx = \int \left(1 - \frac{1}{x+1}\right) dx = x - \log|x+1| + c.$$

Alternatively, we could evaluate this integral with the substitution

$$\begin{aligned} u &= x + 1 \\ du &= dx. \end{aligned}$$

With this substitution, $x = u - 1$, so we have

$$\begin{aligned}\int \frac{x}{x+1}\,dx &= \int \frac{u-1}{u}\,du\\ &= \int \left(1-\frac{1}{u}\right)\,du\\ &= u - \log|u| + c\\ &= x+1-\log|x+1|+c.\end{aligned}$$

Note that this is the same answer we obtained above, although with a different constant of integration.

Partial fraction decomposition: Distinct linear factors

Now we consider the general problem of evaluating

$$\int \frac{f(x)}{g(x)}\,dx$$

where both f and g are polynomials. We will assume that the degree of g is less than the degree of f. As illustrated in the first and third examples above, if this is not the case, we can first perform a long division to simplify the quotient into the form of a polynomial plus a remainder term which is a rational function with numerator of degree less than the denominator. To begin we will suppose that g factors completely into n distinct linear factors. That is, suppose there are constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ such that

$$g(x) = (a_1x+b_1)(a_2x+b_2)\cdots(a_nx+b_n), \tag{6.4.1}$$

where the factors on the right are all distinct. From a theorem of linear algebra, which we will not attempt to prove here, there exist constants $A_1, A_2, \ldots, A_n$ such that

$$\frac{f(x)}{g(x)} = \frac{A_1}{a_1x+b_1} + \frac{A_2}{a_2x+b_2} + \cdots + \frac{A_n}{a_nx+b_n}. \tag{6.4.2}$$

The expression on the right of (6.4.2) is called the *partial fraction decomposition* of $\frac{f(x)}{g(x)}$. Once the constants $A_1, A_2, \ldots, A_n$ are determined, the evaluation of

$$\int \frac{f(x)}{g(x)}\,dx$$

becomes a routine problem. The next examples will illustrate one method for finding these constants.

Example To evaluate $\displaystyle\int \frac{1}{(x-2)(x-3)}\,dx$, we need to find constants A and B such that

$$\frac{1}{(x-2)(x-3)} = \frac{A}{x-2} + \frac{B}{x-3}.$$

Combining the terms on the right, we have

$$\frac{1}{(x-2)(x-3)} = \frac{A(x-3)+B(x-2)}{(x-2)(x-3)}.$$

Now two rational functions with equal denominators are equal only if their numerators are also equal; hence we must have

$$1 = A(x-3)+B(x-2)$$

for all values of x. In particular, for $x = 2$ we obtain

$$1 = -A,$$

from which it follows that $A = -1$, and for $x = 3$ we have

$$1 = B.$$

Thus

$$\frac{1}{(x-2)(x-3)} = -\frac{1}{x-2} + \frac{1}{x-3},$$

so

$$\begin{aligned}\int \frac{1}{(x-2)(x-3)}\,dx &= -\int \frac{1}{x-2}\,dx + \int \frac{1}{x-3}\,dx\\ &= -\log|x-2| + \log|x-3| + c.\end{aligned}$$

Example To evaluate $\displaystyle\int \frac{3x}{(x+5)(2x-1)}\,dx$, we need to find constants A and B such that

$$\frac{3x}{(x+5)(2x-1)} = \frac{A}{x+5} + \frac{B}{2x-1}.$$

Combining the terms on the right, we have

$$\frac{3x}{(x+5)(2x-1)} = \frac{A(2x-1)+B(x+5)}{(x+5)(2x-1)}.$$

As before, it follows that

$$3x = A(2x-1)+B(x+5)$$

for all values of x. In particular, for $x = -5$ we obtain

$$-15 = -11A,$$

from which it follows that

$$A = \frac{15}{11},$$

and for $x = \dfrac{1}{2}$ we have

$$\frac{3}{2} = \frac{11}{2}B,$$

from which it follows that

$$B = \frac{3}{11}.$$

Hence

$$\frac{3x}{(x+5)(2x-1)} = \frac{15}{11}\frac{1}{x+5} + \frac{3}{11}\frac{1}{2x-1},$$

so

$$\begin{aligned}\int \frac{1}{(x+5)(2x-1)}\,dx &= \frac{15}{11}\int \frac{1}{x+5}\,dx + \frac{3}{11}\frac{1}{2x-1}\,dx \\ &= \frac{15}{11}\log|x+5| + \frac{3}{22}\log|2x-1| + c.\end{aligned}$$

Partial fraction decomposition: Repeated linear factors

Returning to the general problem of evaluating

$$\int \frac{f(x)}{g(x)}\,dx,$$

where f and g are both polynomials and the degree of f is less than the degree of g, we will now consider the case where g factors completely into linear factors, allowing for the possibility that one or more of these factors may be repeated. Specifically, suppose the factor $ax+b$ occurs n times in the factorization of g. Then the partial fraction decomposition of $\dfrac{f(x)}{g(x)}$ must contain a sum of terms of the form

$$\frac{A_1}{ax+b} + \frac{A_2}{(ax+b)^2} + \cdots + \frac{A_n}{(ax+b)^n}, \tag{6.4.3}$$

for some constants $A_1, A_2, \ldots, A_n$, in addition to similar terms for every other factor of g. This is best illustrated in an example.

Example To evaluate $\dfrac{x+1}{(x-1)^3(x-2)}\,dx$, we need to find constants A, B, C, and D such that

$$\frac{x+1}{(x-1)^3(x-2)} = \frac{A}{x-1} + \frac{B}{(x-1)^2} + \frac{C}{(x-1)^3} + \frac{D}{x-2}. \tag{6.4.4}$$

That is, this partial fraction decomposition contains three terms corresponding to the factor $x-1$, since it is repeated three times, and only one term corresponding to the factor $x-2$, since it occurs only once. Moreover, the degrees of the denominators of the terms for $x-1$ increase from 1 to 3. Now combining the terms on the right of (6.4.4), we have

$$\frac{x+1}{(x-1)^3(x-2)} = \frac{A(x-1)^2(x-2) + B(x-1)(x-2) + C(x-2) + D(x-1)^3}{(x-1)^3(x-2)}.$$

Again, it follows that

$$x+1 = A(x-1)^2(x-2) + B(x-1)(x-2) + C(x-2) + D(x-1)^3 \tag{6.4.5}$$

for all values of x. However, because of the repeated factors, we cannot choose values for x which will isolate each of the constants one at a time as we did in the previous examples. Instead, we will illustrate another technique for finding the constants. By multiplying out (6.4.5) and collecting terms, we obtain

$$\begin{aligned} x+1 &= A(x^3-4x^2+5x-2) + B(x^2-3x+2) + C(x-2) + D(x^3-3x^2+3x-1) \\ &= (A+D)x^3 + (-4A+B-3D)x^2 + (5A-3B+C+3D)x - 2A+2B-2C-D \end{aligned}$$

for all values of x. Since two polynomials are equal only if they have equal coefficients, we can equate the coefficients of $x+1$ with the coefficients of the polynomial on the right to obtain the four equations

$$\begin{aligned} A+D &= 0 \\ -4A+B-3D &= 0 \\ 5A-3B+C+3D &= 1 \\ -2A+2B-2C-D &= 1. \end{aligned} \tag{6.4.6}$$

From the first equation we learn that

$$D = -A.$$

Substituting this into the second equation gives us

$$B = A.$$

Substituting both of these values into the third equation results in

$$C = A+1.$$

Finally, substituting for D, B, and C in the fourth equation gives us

$$-2A + 2A - 2(A+1) + A = 1,$$

which gives us $A = -3$. Hence $B = -3$, $C = -2$, and $D = 3$. Thus

$$\begin{aligned} \int \frac{x+1}{(x-1)^3(x-2)}\,dx &= -\int \frac{3}{(x-1)}\,dx - \int \frac{3}{(x-1)^2}\,dx \\ &\quad - \int \frac{2}{(x-1)^3}\,dx + \int \frac{3}{x-2}\,dx \\ &= -3\log|x-1| + \frac{3}{x-1} + \frac{1}{(x-1)^2} + 3\log|x-2| + c. \end{aligned}$$

Example: To evaluate the following integral:

$$\int \frac{1}{(x^2+1)(x-1)}\,dx$$

we need to find constants A, B and C such that:

$$\frac{1}{(x^2+1)(x-1)} = \frac{A}{x-1} + \frac{Bx+C}{x^2+1}.$$

By multiplying both sides by$(x^2+1)(x-1)$, we obtain :

$$1 = A(x^2+1) + (Bx+C)(x-1).$$

Let's develop:

$$1 = A(x^2+1) + Bx(x-1) + C(x-1) = Ax^2 + A + Bx^2 - Bx + Cx - C.$$

Let's group similar terms together:

$$1 = (A+B)x^2 + (-B+C)x + (A-C).$$

By identifying the coefficients :

$$\begin{cases} A+B=0, \\ -B+C=0, \\ A-C=1. \end{cases}$$

We solve the system :

$$C = B, \quad A = -B, \quad \Rightarrow A - C = -B - B = -2B = 1 \Rightarrow B = -\frac{1}{2}.$$

So :

$$A = \frac{1}{2}, \quad B = -\frac{1}{2}, \quad C = -\frac{1}{2}.$$

The decomposition into partial fractions is therefore:

$$\frac{1}{(x^2+1)(x-1)} = \frac{1/2}{x-1} - \frac{1}{2}\cdot\frac{x+1}{x^2+1}.$$

So:

$$\int \frac{1}{(x^2+1)(x-1)}\,dx = \frac{1}{2}\int \frac{dx}{x-1} - \frac{1}{2}\int \frac{x+1}{x^2+1}\,dx.$$

We separate the second integral:

$$\int \frac{x+1}{x^2+1}\,dx = \int \frac{x}{x^2+1}\,dx + \int \frac{1}{x^2+1}\,dx = \frac{1}{2}\ln(x^2+1) + \tan^{-1}(x).$$

Substitution in the integral :

$$\int \frac{1}{(x^2+1)(x-1)}\,dx = \frac{1}{2}\ln|x-1| - \frac{1}{2}\left(\frac{1}{2}\ln(x^2+1) + \arctan(x)\right) + C.$$

Let's simplify :

$$\boxed{\int \frac{1}{(x^2+1)(x-1)}\,dx = \frac{1}{2}\ln|x-1| - \frac{1}{4}\ln(x^2+1) - \frac{1}{2}\arctan(x) + C.}$$

Example: To evaluate the following integral:

$$\int \frac{x^3+2x^2+3}{x^2+1}\,dx$$

Perform long division:

$$\frac{x^3+2x^2+3}{x^2+1} = x+2+\frac{-x+1}{x^2+1}.$$

Then integrate term by term:

$$\int \left(x+2+\frac{-x+1}{x^2+1}\right)dx = \frac{x^2}{2}+2x+\int \frac{-x}{x^2+1}\,dx+\int \frac{1}{x^2+1}\,dx.$$

$$\int \frac{-x}{x^2+1}\,dx = -\frac{1}{2}\ln(x^2+1), \qquad \int \frac{1}{x^2+1}\,dx = \arctan(x).$$

$$\boxed{\int \frac{x^3+2x^2+3}{x^2+1}\,dx = \frac{x^2}{2}+2x-\frac{1}{2}\ln(x^2+1)+\arctan(x)+C.}$$

Chapter 4

Introduction to descriptive statistical analysis

Introduction

Statistics is a scientific method that consists of: reducing data on large sets, then analyzing, commenting on, interpreting, and finally critiquing this data.

4.1 Basic concepts

In this section we will present some basic concepts and definitions associated with statistical language.

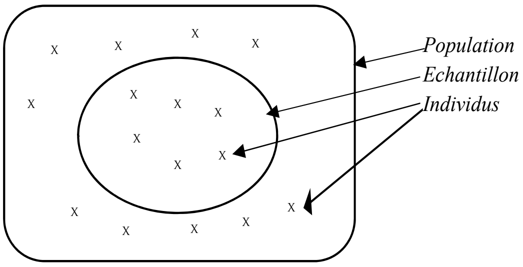
4.1.1 Definitions

Population: The *population*, also called the *universe*, is a well-defined set of homogeneous elements on which a statistical study is to be carried out.

Sample: A *sample* is the subset of the population on which the observations are made.

Individuals: The *individuals*, also called *statistical units*, may be human beings, objects, or animals.

These three concepts (population, sample, and individuals) can be illustrated as follows:



Characteristic: A *characteristic* is a feature or property that allows us to identify individuals and classify them into subsets. Note that each individual may be described by one or several characteristics. Moreover, the modalities must be mutually exclusive, meaning that an individual cannot belong to more than one modality at the same time.

Categories: The categories (modalities) of a characteristic are the different possible situations of that characteristic. For example:

1. The characteristic "Sex" has two categories: {Female, Male}.
2. The characteristic "Marital status" has the following categories: {Married, Single, Widowed, Divorced}.

4.1.2 Character type

We distinguish two types of characteristics:

Qualitative characteristic: A characteristic is said to be qualitative when its categories are not measurable. They are identified by words describing a state. **Example:** Sex, occupation, nationality, ... This type of variable can in turn be classified into two categories:

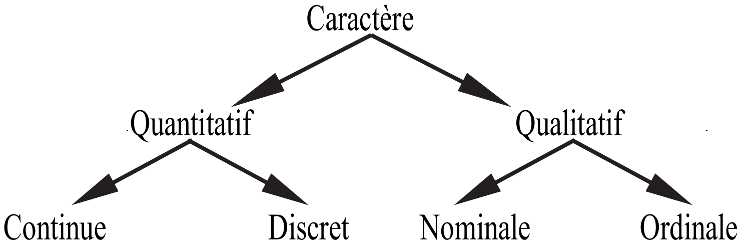
1. **Nominal:** where the categories are measured on a nominal scale, meaning they are expressed by names. Example: eye color, types of plants, ...
2. **Ordinal:** where the categories can be presented on an ordinal scale; they express the degree or level of a state characterizing an individual.

   Example 1: The characteristic "metal resistance to heat" has the following categories: slightly resistant, moderately resistant, highly resistant.
   Example 2: The characteristic "high school diploma grade" has the categories: excellent, very good, good, fair.

Quantitative characteristic: A characteristic is said to be quantitative when its categories are measurable, that is, expressed as numbers. The characteristic is then called a statistical variable, and the different categories are the possible values of the variable.

Example: Age, height, weight, and number of children are quantitative characteristics, statistical variables whose categories are measurable in various specific units. Quantitative variables are of two different types:

1. **Discrete variables (or discontinuous):** A quantitative variable is said to be discrete if it can take only isolated values. A discrete variable that takes only integer values is called discrete. For example, the number of children per household can only be 0, 1, 2, 3, ...; it can never take a value strictly between 0 and 1, or between 1 and 2, or between 2 and 3, ...
2. **Continuous variables:** A quantitative variable is said to be continuous when it can take any value in a finite or infinite interval. For example, the diameter of a tree, its height, or the average grade of a semester, ...



4.2 Statistical tables and graphical representations

The statistical information collected in its *raw* form is practically unusable. In order to give it meaning and usefulness, it must be organized, classified, and processed, mainly by using tables and graphs.

Presenting qualitative or quantitative statistical data in the form of statistical tables is a very important and essential step for subsequent statistical procedures. Afterwards, the statistical tables are represented by graphs in order to visualize the behavior of the statistical variable.

In what follows, after introducing the notion of frequency (absolute and relative), we will highlight the difference between graphs specific to qualitative characteristics, discrete quantitative characteristics, and continuous quantitative characteristics.

4.2.1 Statistical tables: Count and frequency

Let us consider a population composed of N individuals described by a characteristic X, which consists of the categories $x_1, x_2, \dots, x_k$. Presenting this information in a table consists of counting the number of individuals corresponding to each category and then organizing them.

The theoretical table may be presented as follows:

Categories x_i of the characteristic	x_1	x_2	...	x_i	...	x_k
Number of individuals n_i	n_1	n_2	...	n_i	...	n_k

The number n_i of individuals having category x_i of the characteristic X is called the **count** or **absolute frequency**. Thus, we have the notion of the **total count**, denoted by n, defined as

$$n = \sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k,$$

which represents the size of the sample taken for statistical analysis.

The observations organized in the table form a statistical series (or statistical distribution), which consists of all the data and their corresponding counts, denoted by $\left\{(x_i, n_i)\,,\ i = \overline{1,k}\right\}$.

The **frequency** or **relative frequency** of the category x_i is the number $f_i = \frac{n_i}{n}$, which measures the proportion of individuals having category x_i in the sample. Note that the frequency f_i satisfies the following two properties:

1. For all $i \in \{1, \dots, k\}$, we have $0 \leq f_i \leq 1$.
2. $\sum\limits_{i=1}^{k} f_i = f_1 + f_2 + \cdots + f_k = 1$.

Exemple 1 *Let the sample below of size $n = 50$, taken from a discrete quantitative variable:*

$$\begin{array}{ccccccccccccccccccccccccc}
1 & 4 & 3 & 5 & 1 & 6 & 3 & 1 & 6 & 1 & 5 & 2 & 1 & 4 & 6 & 1 & 6 & 6 & 4 & 2 & 1 & 5 & 1 & 3 & 5 \\
2 & 4 & 2 & 6 & 1 & 3 & 2 & 4 & 3 & 1 & 4 & 5 & 6 & 1 & 5 & 2 & 2 & 4 & 4 & 2 & 5 & 1 & 1 & 3 & 2
\end{array}$$

The grouping of the observations (counting of frequencies) provides us with the following table:

X_i	*1*	*2*	*3*	*4*	*5*	*6*
n_i	*13*	*9*	*6*	*8*	*7*	*7*
$f_i = n_i/n$	*0.26*	*0.18*	*0.12*	*0.16*	*0.14*	*0.14*

4.2.2 Graphical representation of a qualitative variable

To graphically present a qualitative variable, several types of diagrams are available. Below are the most commonly used ones in practice.

- **Bar chart**: This type of representation is obtained by constructing as many columns as there are categories of the characteristic. These columns are rectangles with a constant base and height proportional to n_i (or f_i).
- **Pie chart**: The pie chart allows us to visualize the relative share of each category of the characteristic. The base of this representation is a circle divided into as many sectors as there are categories, such that the angle θ_i representing the share of x_i is given by:

$$\begin{array}{ccc} 360^\circ & \rightarrow & n \\ \theta_i = ? & \rightarrow & n_i \end{array} \Rightarrow \theta_i = 360^\circ \times \frac{n_i}{n} = 360^\circ \times f_i$$

Exemple 2 The distribution of workers in a company according to their qualification is summarized as follows:

Qualification	Workers	Employees	Technicians	Engineers	Total
Number of workers n_i	140	30	20	10	200
Relative frequency f_i	0.7	0.15	0.10	0.05	1

The angles corresponding to the distribution of workers according to qualification are:

$$\theta_1 = 360^\circ \times f_1 = 360^\circ \times 0.7 = 252^\circ \qquad \theta_2 = 360^\circ \times f_2 = 360^\circ \times 0.15 = 54^\circ$$
$$\theta_3 = 360^\circ \times f_3 = 360^\circ \times 0.10 = 36^\circ \qquad \theta_4 = 360^\circ \times f_4 = 360^\circ \times 0.05 = 18^\circ$$

The graphical presentation of the distribution of workers can be done using one of the following diagrams:

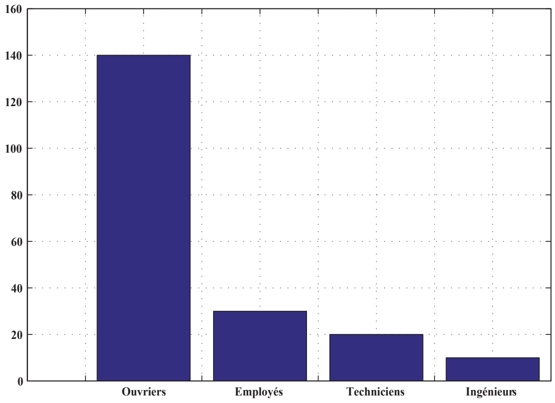


Diagramme en Tuyaux des effectifs

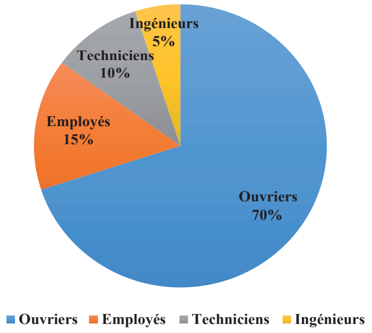


Diagramme circulaire des fréquences

4.2.3 Graphical representation of a discrete quantitative variable

The appropriate graph to represent a statistical series from a discrete quantitative variable is the **bar chart**, where each value x_i of the variable corresponds to a bar whose height is proportional to n_i or f_i.

Suppose that the statistical distribution of the number of rooms per dwelling in a certain locality is given as follows:

Number of rooms x_i	1	2	3	4	5	6	Total
Number of dwellings n_i	5	10	20	30	25	10	100
f_i	0.05	0.1	0.20	0.30	0.25	0.10	1

The corresponding diagram of this series is shown in Figure 4.6 (left). If we connect the tops of the bars, we obtain the **frequency polygon** (see Figure 4.6 (right)).

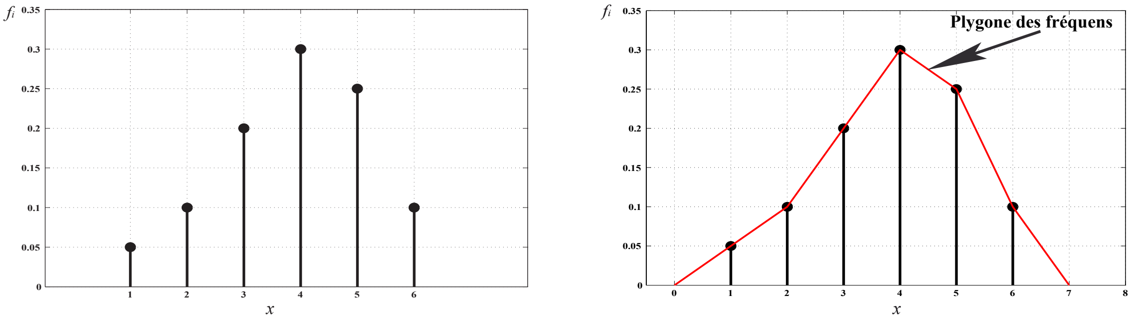


Figure 4.1: *Bar chart of room frequencies per dwelling.*

4.2.4 Cumulative Counts and Frequencies

Définition 4.1 *Cumulative counts (resp. cumulative frequencies) are defined as the number N_i (resp. F_i) such that:*

$$N_i = \sum_{j=1}^{i} n_j \quad (resp.\ F_i = \sum_{j=1}^{i} f_j),$$

The cumulative frequency F_i answers the question: what proportion of individuals have a value less than x_{i+1} or greater than or equal to x_i?

Définition 4.2 *The cumulative curve (or cumulative frequency polygon) (absolute or relative) is the graphical representation of these cumulative frequencies.*

For a discrete statistical variable, the cumulative curve is the representation of a step function whose horizontal steps have coordinates (x_i, F_i). This function is called the empirical distribution function, defined by:

$$\begin{array}{ccl} F & : & \mathbb{R} \longrightarrow [0,1] \\ x & \longrightarrow & F(x),\ such\ that \end{array}$$

$$F(x) = \begin{cases} 0 & if\ x < x_1 \\ f_1 & if\ x_1 \leq x < x_2 \\ f_1 + f_2 & if\ x_2 \leq x < x_3 \\ \sum_{j=1}^{i} f_j & if\ x_i \leq x < x_{i+1} \\ 1 & if\ x \geq x_k \end{cases}$$

Example: distribution of dwellings according to the number of rooms

$$F(x) = \begin{cases} 0 & if\ x < 1 \\ 0.05 & if\ 1 \leq x < 2 \\ 0.15 & if\ 2 \leq x < 3 \\ 0.35 & if\ 3 \leq x < 4 \\ 0.65 & if\ 4 \leq x < 5 \\ 0.90 & if\ 5 \leq x < 6 \\ 1 & if\ x \geq 6 \end{cases}$$

x_i	n_i	f_i	$F_i^{\nearrow}$	$F_i^{\searrow}$
1	5	0.05	0.05	0.95
2	10	0.10	0.15	0.85
3	20	0.20	0.35	0.65
4	30	0.30	0.65	0.35
5	25	0.25	0.90	0.10
6	10	0.10	1	0
Total	100	1		

Statistical table of increasing and decreasing cumulative frequencies for the number of rooms per dwelling.

Remarque 4.1 *The function F is discontinuous at each point of the statistical variable.*

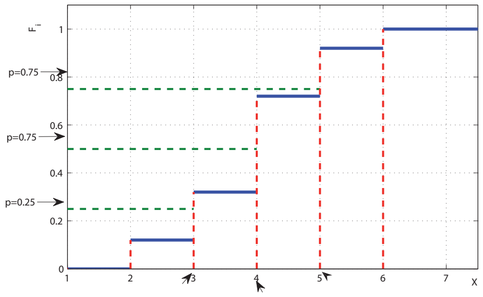


Figure 4.2: *Increasing cumulative curve for the number of rooms per dwelling.*

4.2.5 Graphical representation of a continuous quantitative variable

For a continuous variable, to establish the statistical table, it is necessary first to group the data into **classes**. A **class** is defined by its lower and upper bounds: by convention $[a_{i-1}, a_i[$. It is clear that this requires defining the number of classes and the amplitude associated with each.

Let the i^{th} class be given by $[a_{i-1}, a_i[$, which is fully defined by:

- ⊵ The lower bound of the class: a_{i-1}.
- ⊵ The upper bound of the class: a_i.
- ⊵ The **amplitude** of the class: $A_i = a_i - a_{i-1}$.
- ⊵ The class midpoint: $x_i = \frac{a_i + a_{i-1}}{2}$.

Remarque 4.2 *In practice:*

1. *Generally, classes of equal amplitude are chosen.*
2. *The choice of the number of classes and their amplitude depends on the total number of observations n.*
3. *Any reduction in the number of classes and any increase in amplitude leads to a loss of information.*

4. *Sturges' rule: Number of classes* $\simeq 1 + 3.3 \log N$*, and the amplitude is given by*

$$A = \frac{\max x_i - \min x_i}{\textit{number of classes}}.$$

5. *The values of a continuous statistical variable are the class midpoints.*

The statistical table of a continuous variable series is generally presented as follows:

X	$[a_0, a_1[$	$[a_1, a_2[$	$\cdots$	$[a_{i-1}, a_i[$	$\cdots$	$[a_{m-1}, a_m[$
n_i	n_1	n_2	$\cdots$	n_i	$\cdots$	n_m

where n_i represents the number of observations in the i^{th} class, while the frequency is defined as in Section 4.2.1, i.e., $f_i = \frac{n_i}{n}$ with n the total sample size.

To graphically represent a continuous variable, we use the **histogram**, which is a generalization of the bar chart to the notion of classes. Two situations are possible: series with equal class widths and series with unequal class widths, as illustrated below.

Each class is represented by a rectangle whose **base** is the class amplitude and whose **height** is proportional to the **frequency** or **number of observations** (see Figure 4.3).

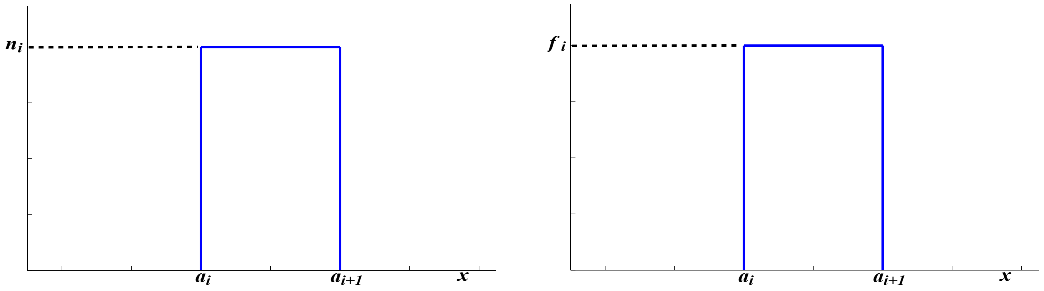


Figure 4.3: *Illustration of graphical representation of a class.*

The frequency polygon in this case is the line joining the midpoints of the top sides of the rectangles (see Figure 4.6).

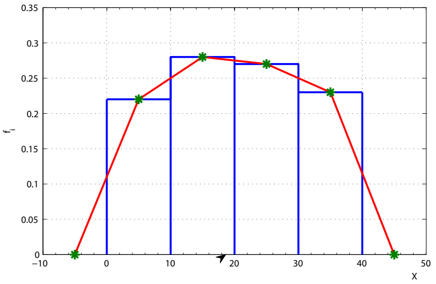


Figure 4.4: *The frequency polygon.*

Exemple 3 *The data below come from an experiment in which the plasma calcium concentration was measured in 40 individuals who received a hormonal treatment.*

X	*[10,16[*	*[16,22[*	*[22,28[*	*[28,34[*	*[34,40[*	*[40,46[*
n_i	*4*	*6*	*17*	*8*	*4*	*1*
f_i	*0.100*	*0.150*	*0.425*	*0.200*	*0.100*	*0.025*

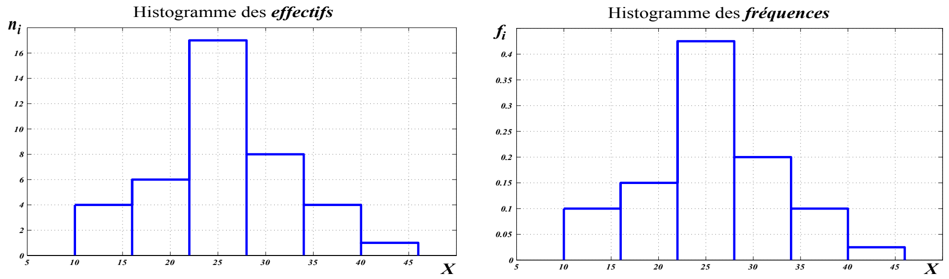


Figure 4.5: *Histogram of counts and Histogram of frequencies of the plasma calcium concentration.*

Exemple 4 *The distribution of a group of individuals according to their height (cm) is given in the following table:*

Height (cm)	n_i	f_i	$F_i^{\nearrow}$	$F_i^{\searrow}$
$[149.5, 159.5[$	*14*	*0.08*	*0.08*	*0.92*
$[159.5, 169.5[$	*32*	*0.18*	*0.26*	*0.74*
$[169.5, 179.5[$	*65*	*0.37*	*0.63*	*0.37*
$[179.5, 189.5[$	*47*	*0.27*	*0.9*	*0.1*
$[189.5, 199.5[$	*17*	*0.10*	*1*	*0*
Total	*175*	*1*		

Statistical table of increasing and decreasing cumulative frequencies for the distribution of the height(cm) of a group of individuals.

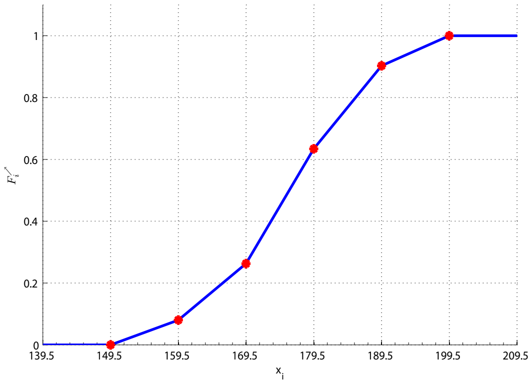


Figure 4.6: *Increasing cumulative curve.*

4.3 Numerical Characteristics of a Statistical Variable

When observing a graphical representation of a statistical series, two impressions may be noted:

1. The order of magnitude of the statistical variable, characterized by the values located at the center of the distribution; this is called the "central tendency characteristic" (or position measure).
2. The fluctuations of the observations around the central value; this is called the "measure of dispersion".

4.3.1 Measures of Central Tendency

They are intended to define the central values of the statistical series. These are: the mode, the median, and the arithmetic mean.

4.3.1.1 The Mode

a) Discrete Case

The mode, denoted by M_o, of a discrete statistical variable is the value with the highest frequency.

Exemple 5 *In Example 2 (Number of rooms per dwelling),* $M_o = 4$ *(because value 4 has the highest count).*

Remarque 4.3 *On a bar chart, the mode corresponds to the bar with the greatest height.*

b) Continuous Case

In this case we refer to the “modal class”, which is the class with the highest frequency per unit width.

In Example 3, the calcium concentration in plasma, the modal class is $[22, 28[$.

Remarque 4.4 *In the continuous case, the mode may be taken as the midpoint of the modal class.*

For an approximate calculation, use the formula: for the modal class $[e_{i-1}, e_i[$,

$$M_o \in [e_{i-1}, e_i[$$

$$M_o = e_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2},$$

where a_i is the class width,
Δ_1: excess of the modal class over the preceding class,
Δ_2: excess of the modal class over the following class.
Graphically: $M_o = e_{i-1} + d$ (d calculated using the scale).

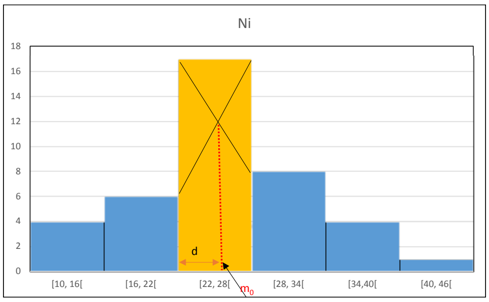


Figure 4.7: *The mode graphically*

4.3.1.2 The median

The median M_e is the value of the statistical variable that divides the individuals, arranged in ascending (or descending) order, into two equal groups.

Exemple 6 *Consider the dataset:* $\{12; 26; 6; 3; 32; 15; 21\}$.
Ordered: $\{3; 6; 12; 15; 21; 26; 32\}$.
The median is $M_e = 15$.

In general, the median of a statistical variable is the value M_e such that $F(M_e) = \frac{1}{2} = 0.5$, where F is the empirical distribution function.

Calculation of the median

a) Discrete case:

Let $x_{(1)}, x_{(2)}, ...x_{(n)}$ be the n observations arranged in increasing order:

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}.$$

1st case: n odd, $n = 2p + 1$

$$\underbrace{x_{(1)} \leq x_{(2)} \leq ... \leq x_{(p)}}_{\text{p observations}} \leq \boxed{x_{(p+1)}} \leq \underbrace{x_{(p+2)} \leq ... \leq x_{(2p+1)}}_{\text{p observations}}.$$

$M_e = x_{(p+1)}$: the value at position $(p + 1)$.

2nd case: n even, $n = 2p$

$$\underbrace{x_{(1)} \leq x_{(2)} \leq ... \leq x_{(p)}}_{p \text{ observations}} \leq \boxed{x_{(p+1)}} \leq \underbrace{x_{(p+2)} \leq ... \leq x_{(2p)}}_{\text{p observations}}.$$

$M_e = \frac{x_{(p)}+x_{(p+1)}}{2}$, with $\left[x_{(p)}, x_{(p+1)}\right]$ the median interval.

In Example 2: number of children per household.

$n = 10 = 2 \times 5 \Longrightarrow M_e = \frac{x_{(5)}+x_{(6)}}{2} = \frac{1+1}{2} = 1.$

$$0 \quad 0 \quad 1 \quad 1 \quad \underset{X(5)}{\underset{\uparrow}{1}} \quad \underset{X(6)}{\underset{\uparrow}{1}} \quad 1 \quad ...$$

Graphical determination of the median:

1. No horizontal plateau of the cumulative curve has ordinate (0.5).

Figure.

$$\begin{aligned} 0.2 &< 0.5 < 0.7. \\ F(0) &< 0.5 < F(1). \\ &\Rightarrow M_e = 1 \text{ child}. \end{aligned}$$

2. If a horizontal plateau of the cumulative curve has ordinate 0.5, the median is undetermined between two consecutive possible values:

$M_e = \frac{x_{i-1}+x_i}{2}$

b) Continuous case:

Since the function F is continuous and monotonic ($\nearrow$) between 0 and 1, the median is the unique solution of the equation $F(x) = \frac{1}{2}$.

Graphical determination of the median (interpolation method):

Let $[e_{i-1}, e_i[$ be the median class.

Figure.

$$\begin{aligned} F(e_{i-1}) &< 0.5 < F(e_i) \\ \tan\alpha &= \frac{F(e_i)-F(e_{i-1})}{e_i-e_{i-1}} = \frac{0.5-F(e_{i-1})}{M_e-e_{i-1}} \\ &\Rightarrow \frac{f_i}{a_i} = \frac{0.5-F(e_{i-1})}{M_e-e_{i-1}} \end{aligned}$$

Hence,

$$f_i(M_e-e_{i-1}) = a_i\,(0.5-F(e_{i-1}))$$

And therefore,

$$M_e = e_{i-1} + a_i\frac{0.5-F(e_{i-1})}{f_i}$$

a_i : class width
f_i : frequency of the median class.
Graphically:
$M_e = e_{i-1} + d$ (where d is calculated from the scale).

4.3.1.3 Arithmetic Mean

It is the sum of all observations divided by the total number of observations.

a) Discrete case:

Let X be a discrete variable taking the values $x_1, x_2, \ldots, x_k$ with corresponding frequencies $n_1, n_2, \ldots, n_k$ such that $\sum_{i=1}^{k} n_i = n$.

The arithmetic mean is denoted by

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{n_1x_1+n_2x_2+\ldots+n_kx_k}{n} \\ \overline{x} &= \sum_{i=1}^{k} f_i x_i \quad \text{(Weighted mean).} \end{aligned}$$

b) Continuous case:

The values are grouped into classes; by convention, we choose x_i as the midpoints of each class, and we use the same formula:

$$\begin{aligned} \overline{x} &= \sum_{i=1}^{k} f_i x_i \quad (x_i = \text{midpoint of the } i^{\text{th}} \text{ class}). \\ x_i &= \frac{e_{i-1}+e_i}{2} \end{aligned}$$

Algebraic properties of the arithmetic mean:

Property 1: (Change of origin)

Let $x_1, x_2, \ldots, x_k$ be the observed values of a statistical variable X and $n_1, n_2, \ldots, n_k$ their frequencies.

Let x_0 be a new origin.

Define the new variable
$y_i = x_i - x_0,\ \forall i = \overline{1,n}$, then $\overline{x} = \overline{y} + x_0$ with $\overline{y} = \sum_{i=1}^{k} f_i y_i$.
Indeed, since $x_i = y_i + x_0$,

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{n}\sum_{i=1}^{k} n_i(y_i + x_0) \\ &= \frac{1}{n}\sum_{i=1}^{k} n_i y_i + \frac{1}{n}\sum_{i=1}^{k} n_i x_0 = \overline{y} + x_0.\end{aligned}$$

Remarque 4.5 *This change of variable is used to simplify calculations; in practice, one often chooses $x_0 = M_o$ or M_e.*

Property 2: (Change of scale and origin)
If we choose $y_i = \frac{x_i - x_0}{a}$, where x_0 and a are constants,
then $\overline{x} = a\overline{y} + x_0$.

Remarque 4.6 *In practice, we choose x_0 as: the median, the mode, or the class center.*
$a = gcd(a_i)$ (continuous case),
and the spacing between x_i (discrete case).

Proposition 4.1 *The sum of deviations from the arithmetic mean is zero:*
$\sum_{i=1}^{k} n_i(x_i - \overline{x}) = 0$.
Indeed: $\sum_{i=1}^{k} n_i(x_i - \overline{x}) = \sum_{i=1}^{k} n_i x_i - \sum_{i=1}^{k} n_i \overline{x} = n\overline{x} - n\overline{x} = 0$.

Proposition 4.2 *The sum of squared deviations from the arithmetic mean is minimal:*
$\varphi(a) = \sum_{i=1}^{k} n_i(x_i - a)^2$ is minimal for $a = \overline{x}$.

Relative position of the mode, median, and mean
Consider a unimodal statistical distribution.
1. When the distribution is symmetric, the three measures of central tendency coincide.

Figure

2. When the distribution is asymmetric, the median is generally between the mode and the mean, and is usually closer to the latter.

Figure

4.3.2 Measures of Dispersion

Exemple 7 *Consider the two statistical series:*
$X = \{6; 6; 7; 7; \boxed{8}; 9; 9; 10; 10\}$.
$Y = \{1; 2; 4; 6; \boxed{8}; 10; 12; 14; 15\}$.
We notice that X and Y have the same mean and the same median $\overline{x} = \overline{y} = M_e = 8$, but they are different: the first series X is less dispersed than the second.

Exemple 8 *Consider the statistical series X of k values arranged in increasing order:*
$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(k)}$.

4.3.2.1 Range

The range, denoted by "e", is the difference between the two extreme values: the smallest and the largest observed value.

$$e = \max_{1\leq i\leq k}(x_i) - \min_{1\leq i\leq k}(x_i) = x_{(k)} - x_{(1)}.$$

4.3.2.2 Variance and Standard Deviation

The empirical variance of the statistical variable X taking values x_i, $1 \leq i \leq k$, with frequencies n_i, $1 \leq i \leq k$, and $\sum_{i=1}^{k} n_i = n$, is:

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i(x_i - \overline{x})^2.$$

The empirical standard deviation, denoted by σ_X, is:

$$\sigma_X = \sqrt{V(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2} = \sqrt{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}.$$

Properties of the variance:

1. $V(X) \geq 0$.
2. $V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2 = \sum_{i=1}^{k} f_i x_i^2 - \overline{x}^2$.
3. Let $\overline{x}$ and $V(X)$ be the mean and variance of the statistical variable X.

Define a new variable X' with mean $\overline{x}'$ and variance $V(X')$ such that:

$$x_i' = \frac{x_i - x_0}{a}, \quad i = \overline{1,k}$$

where x_0 and a are constants. Then:

$$V(X) = a^2 V(X') \quad \text{and} \quad \sigma_X = a\,\sigma_{X'}.$$

Remarque 4.7 *When comparing two statistical series of the same nature, the one with the larger σ_X is the more dispersed.*

Coefficient of Variation:
It is a measure of relative dispersion defined by:

$$Cv = \frac{\sigma_X}{|\overline{x}|}.$$

Properties:

1. Cv is a dimensionless quantity.
2. Cv does not depend on the units used.
3. Cv makes it possible to compare two series expressed in different units.

4.3.2.3 Interquartile Range

a) Quantiles: these generalize the median.

· A quantile of order α $(0 \leq \alpha \leq 1)$, denoted by x_α, is the solution of the equation $F(x) = \alpha$. That is, a proportion α of the individuals have the characteristic X less than x_α.

· Quartiles are commonly used.

These are the values of the variable x_i that divide the series into 4 equal parts. There are 3 quartiles, denoted Q_1, Q_2, Q_3, with Q_1 being the quantile of order $\frac{1}{4}$, Q_2 the quantile of order $\frac{1}{2}$, and Q_3 the quantile of order $\frac{3}{4}$.

That is, $F(Q_1) = \frac{1}{4}, F(Q_2) = \frac{1}{2} = F(M_e)$, and $F(Q_3) = \frac{3}{4}$.

· Interquartile Range:

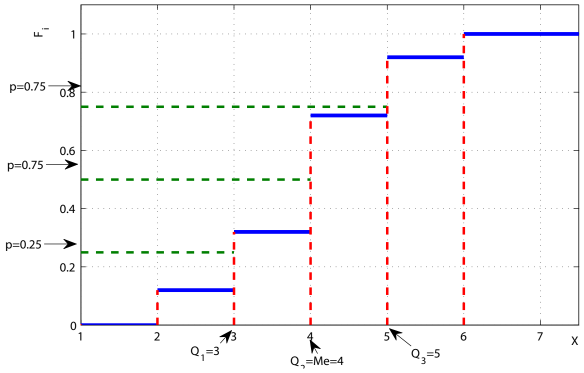
It contains 50% of the population, leaving 25% on each side.

The interquartile range is given by: $Q_3 - Q_1$.

- Practical Determination:

To determine $Q_3 - Q_1$, first calculate Q_1 and Q_3 using the same method used for finding the median.

- Discrete Case (graphically)



If $F(x_{i-1}) < 0.5 < F(x_i) \Rightarrow M_e = x_i$.

If $F(x_{i-1}) < 0.25 < F(x_i) \Rightarrow Q_1 = x_i$.

If $F(x_{i-1}) < 0.75 < F(x_i) \Rightarrow Q_3 = x_i$.

If $\forall x \in]x_{i-1}, x_i[, F(x) = 0.5 \Rightarrow M_e = \frac{x_i + x_{i-1}}{2}$.

Continuous Case:

Same method as for the median.

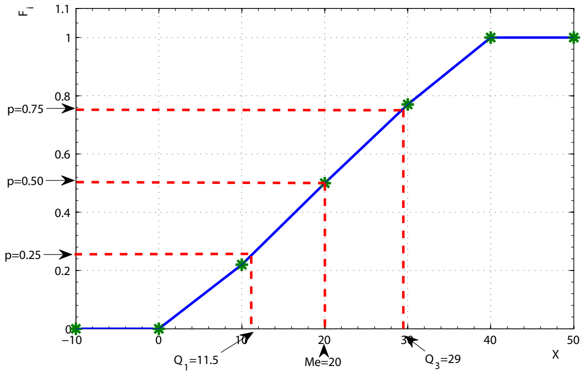
If $F(e_i) = 0.5$ (resp. 0.25, 0.75), then $e_i = M_e$ (resp. Q_1, Q_3).

If $F(e_{i-1}) \leq 0.5 \leq F(e_i)$, then:

$$M_e = e_{i-1} + a_i \frac{0.5 - F(e_{i-1})}{f_i}, \quad \text{with } [e_{i-1}, e_i[\text{ being the median class.}$$

$$Q_1 = e_{i-1} + a_i \frac{0.25 - F(e_{i-1})}{f_i}, \quad Q_1 \in [e_{i-1}, e_i[.$$

$$Q_3 = e_{i-1} + a_i \frac{0.75 - F(e_{i-1})}{f_i}, \quad Q_3 \in [e_{i-1}, e_i[.$$


F_i
p=0.75
p=0.50
p=0.25
X
Q_1=11.5
Me=20
Q_3=29