

---

# LNG-1100 : Méthodes expérimentales et analyse de données

Intro : principes de base

Guilherme D. Garcia

[fr.gdgarcia.ca](http://fr.gdgarcia.ca) ↗

1



# Mesures d'accommodement

- **Important:**

Les étudiants qui ont droit à des mesures d'accommodement en cours de session doivent procéder à leur activation immédiatement dans [monportail.ulaval.ca/accommodement](http://monportail.ulaval.ca/accommodement) afin que celles-ci puissent être mises en place.

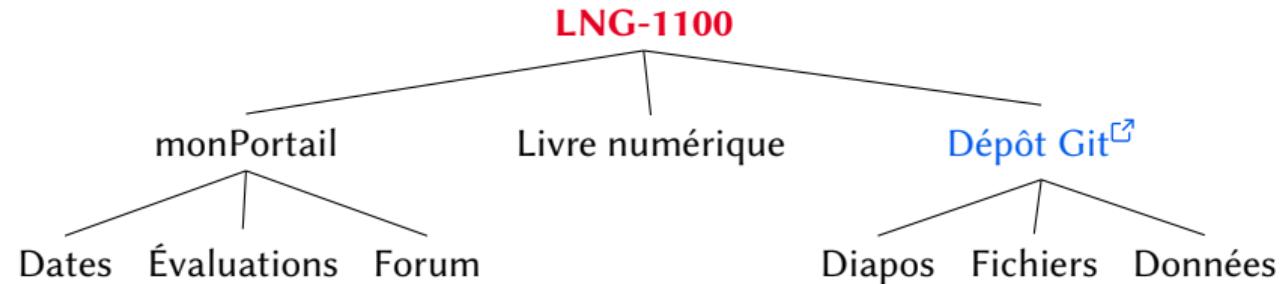
- **Formuler et tester** des hypothèses de recherche [en linguistique]
- **Se familiariser** avec les éléments de base de l'analyse de données quantitatives
- **Interpréter et synthétiser** des résultats statistiques dans un rapport scientifique

### Nos analyses

- Notre cours examine les données de façon **quantitative**
- Une analyse **qualitative** est également possible, bien qu'elle ne fasse pas partie de nos objectifs

# LNG-1100

Familiarisez-vous avec la structure du cours



# Nos matériels

☞ Livre du cours : [Ing1100.quarto.pub](#) ↗

(Garcia 2024)

- Livres additionnels qui seront utilisés :

## Français :

Barnier (2023)

introduction à R  
R + statistique

Larmarange (2023)

## Anglais :

- Wickham *et al.* (2023)
- Garcia (2021)

R (très détaillé)  
R + statistique

# Évaluation et IA

- Vous êtes encouragés à utiliser l'IA pendant le cours ([Claude.ai<sup>↗</sup>](#) est la meilleure option)
- ☞ Toutefois, gardez deux choses à l'esprit :
  1. vos examens (40 %) ne sont **pas** à livre ouvert, et vous n'aurez pas d'ordinateur pour vous aider.  
Donc, si vous dépendez de l'IA pour vos analyses, vous aurez des problèmes (p. ex., vous trouverez l'examen « trop long »)
  2. l'IA combine souvent des fonctions provenant de multiples extensions R, alors que dans ce cours nous utiliserons uniquement la une famille d'extensions. Limiter et harmoniser votre répertoire de fonctions rendra votre apprentissage plus clair, plus cohérent et plus facile à consolider
- ☞ Dans le cours, il faut pratiquer le codage **fréquemment**. C'est la façon la plus facile de réussir

## Un petit sondage

[forms.office.com/r/G700T82tNs](https://forms.office.com/r/G700T82tNs)



# Plan de la séance

1. Les données peuvent être trompeuses
2. Nos outils : **R + RStudio** (Positron)
3. La structure du cours; principes généraux de la recherche (linguistique); hypothèse, expérimentation; comité d'éthique; design expérimental : sondage, ABX/AXB, tâche de jugement à choix forcé, etc.

**Les données peuvent être trompeuses**

# Exemple 1

Quel est le problème ici?

## **Le chômage a baissé de 50 % le mois dernier**

- ☞ Si le taux de chômage est passé de 2 % à 1 %, c'est effectivement une baisse de 50%, mais cela peut induire les gens en erreur en leur faisant penser qu'il y a eu une amélioration massive de l'emploi

## Exemple 2

Quel est le problème ici?

**Le revenu moyen dans la ville A est de 100 000 \$**

- ☞ Si un petit nombre de personnes très riches vivent dans la ville, la moyenne peut être fortement biaisée. La médiane ou le mode pourraient fournir une image plus précise du revenu typique

## Exemple 3

Quel est le problème ici?

**Le taux d'intérêt a augmenté de 2 % l'année dernière**

- ☞ Si le taux est passé de 5 % à 7 %, c'est une augmentation de 2 points de pourcentage, mais une augmentation de 40 % du taux d'intérêt lui-même

## Exemple 4

Quel est le problème ici?

**Un test médical avec une précision de 95 % a donné un résultat positif. Par conséquent, le patient a 95 % de chances d'avoir la maladie**

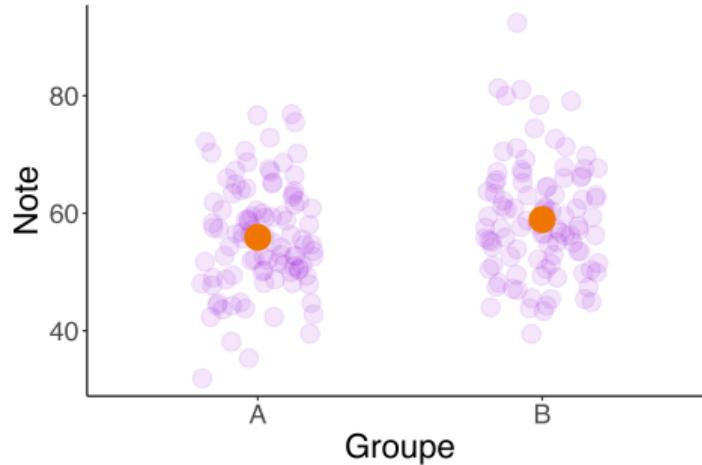
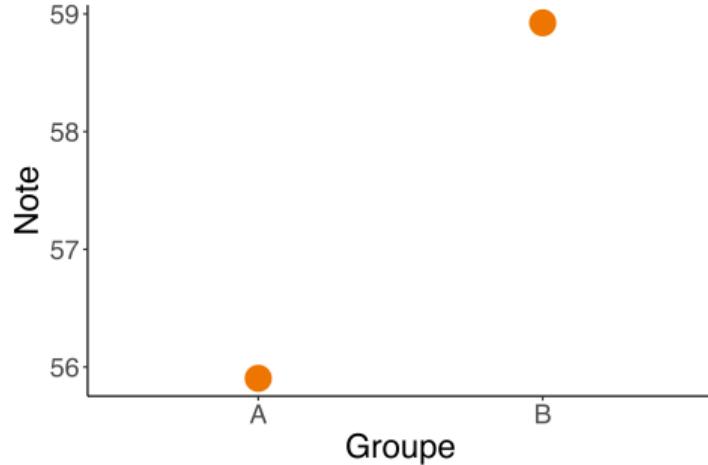
- ☞ Sans tenir compte de la prévalence de la maladie dans la population, cette affirmation peut être trompeuse. Si la maladie est rare, un résultat positif peut toujours être un faux positif

Pour bien comprendre le problème, on utilise [le théorème de Bayes](#)↓

## Exemple 5

Quel est le problème ici?

- Une figure peut-elle influencer vos conclusions?



## Vos connaissances de base

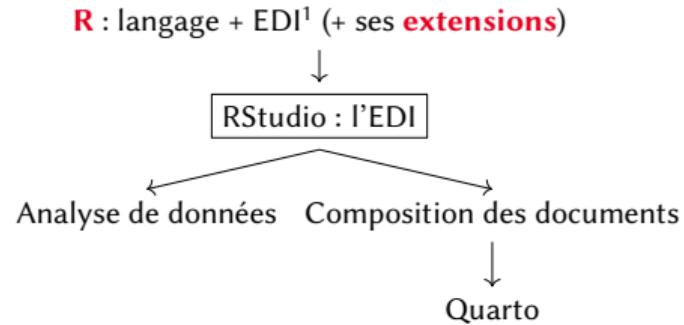
[forms.office.com/r/qqd8t6Tv1G](https://forms.office.com/r/qqd8t6Tv1G)



**Nos outils : R + RStudio (Positron)**

# RStudio

- On commence à travailler avec RStudio à partir de **la semaine prochaine**
- ☞ Lisez le chapitre 1 de notre livre **avant** notre séance



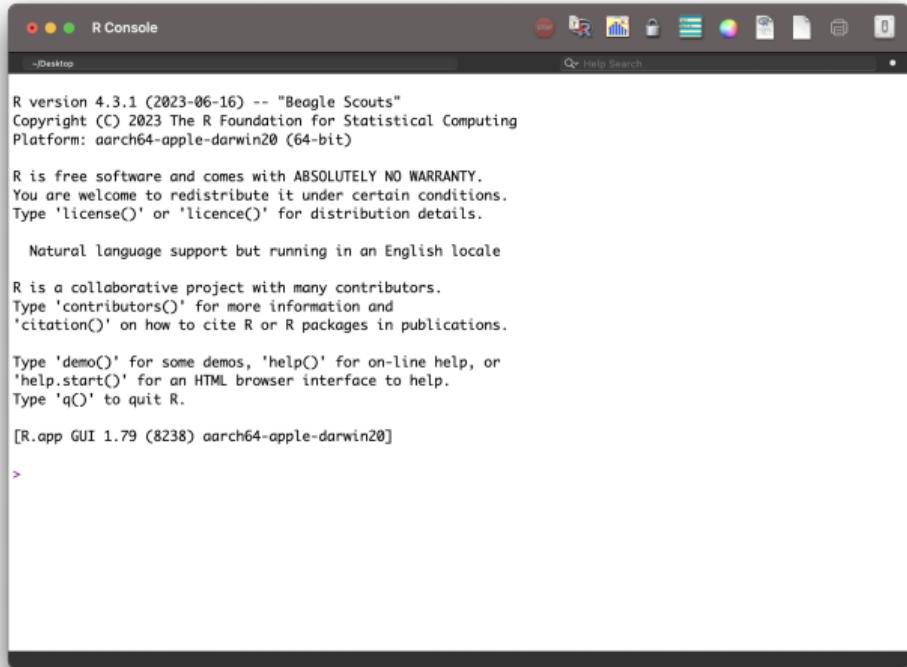
- Tous ces outils sont disponibles dans la version en ligne [posit.cloud](#)↗

---

<sup>1</sup>Environnement de développement intégré. Normalement, on choisit un EDI ou un éditeur de texte pour travailler avec le codage.

# R

## Langage + EDI simple



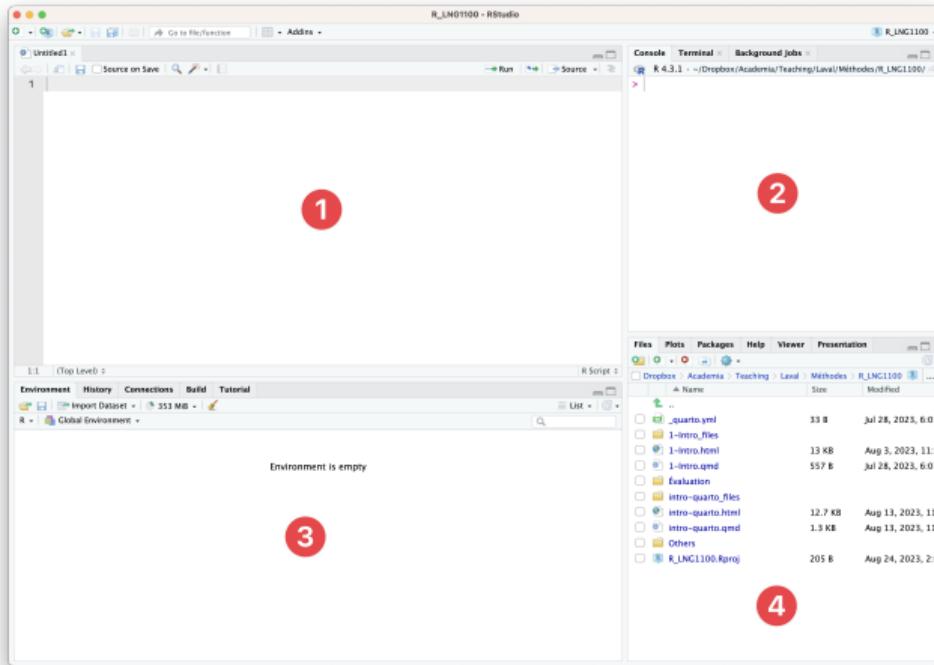
R version 4.3.1 (2023-06-16) -- "Beagle Scouts"  
Copyright (C) 2023 The R Foundation for Statistical Computing  
Platform: aarch64-apple-darwin20 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[R.app GUI 1.79 (8238) aarch64-apple-darwin20]  
>



- R inclut son propre EDI
- Mais il est trop simple

# RStudio

EDI très puissant et intuitif



1. Notre analyse
2. Notre résultat
3. Nos variables
4. Fichiers locaux, figures

# Tidyverse

## Les extensions (bibliothèques)

- Extension (bibliothèque; *package*) :  
collection de fonctions, de données et de documentation regroupées ensemble
  - R contient plus de **17 000** extensions
- ☞ Dans notre cours, on utilisera principalement l'extension **tidyverse**



- On va explorer en détail [cette famille d'extensions](#) dans deux semaines

## **Principes généraux de la recherche (linguistique)**

# Le processus typique

1. Trouver un sujet/problème approprié
2. Poser une question<sup>2</sup> (et une hypothèse?)
3. Développer une expérience et/ou **collecter et examiner des données**
4. **Évaluer les résultats par rapport à l'hypothèse**
5. **Communiquer** et publier les résultats

☞ Dans notre cours, on met l'accent sur les éléments **mis en évidence**

---

<sup>2</sup>L'une des étapes les plus difficiles.

# Exemple

## Un sujet et une question de recherche

1. **Sujet** : la production et la perception des voyelles /y ø œ/<sup>3</sup> en français
  - o lunette - neutre - jeune ← très difficile pour des apprenants
2. Les apprenants qui prononcent mal ces voyelles sont-ils capables de les distinguer?
  - o Si oui, il s'agit d'un problème de **production**
  - o Si non, Il s'agit d'un problème de **perception** aussi
3. On peut créer une expérience pour examiner le comportement des apprenants
  - o Par conséquent, il faut consulter le site du **CÉRUL** ↗ (Comités d'éthique)
  - o Parfois, le projet sera exempté (c'est rare...)
  - o **Alternative** : utiliser des données déjà disponibles (p. ex., en ligne)

---

<sup>3</sup>Alphabet phonétique international ↗

# Exemple

## Design expérimentale

- Quel type de collecte de données...?
  - Sondage (MS Forms et Google Forms : gratuits et faciles) ← **Notre point de départ cette session**
  - ABX/AXB (PsychoPy, Gorilla, Praat, etc.)
  - Jugement à choix forcé (idem)
  - ...

# Exemple

Design expérimentale : sondage

- Le sondage est probablement le type de collecte de données le plus simple
- ☞ Quel type de question pourrait-on poser dans un sondage sur les voyelles françaises?

« Écoutez le mot suivant et tapez ce que vous entendez dans la casse ci-dessous »

- On utilise plusieurs mots du français : ceux qui contiennent les voyelles ciblées et ceux qui contiennent d'autres voyelles (les **éléments distracteurs**)
- ☞ Enfin, on a de nombreuses réponses qui peuvent être analysées

# Exemple

Design expérimentale : **ABX/AXB**

- Méthode de comparaison entre deux stimulus
- Imaginons un mot non réel : ‘zule’ /zyl/
- Le participant écoute la séquence /zøl/ - /zyl/ - /zyl/

☞ On pose la question :

- Le **troisième** mot est-il identique au premier ou au deuxième mot? 

1	2
---	---
- Le **deuxième** mot est-il identique au premier ou au troisième mot? 

1	3
---	---

**ABX**

**AXB**

- Les réponses nous permettent d'évaluer la précision perceptuelle des participants

# Exemple

Design expérimentale : **choix forcé**

- On **force** le participant à choisir une option (entre deux/trois/etc.)
  - Imaginons une paire minimale comme ‘peu’ /pø/ et ‘pu’ /py/
  - Le participant écoute /pø/
- ☞ On pose la question :
- Quel mot venez-vous d’écouter? **peu** **pu**
- Les réponses nous permettent d’évaluer la précision perceptuelle des participants

# Design expérimentale

- Supposons l'exemple ABX. On peut examiner plusieurs **variables** d'intérêt :
  - La voyelle (3 voyelles ciblées) et la précision de la réponse (0 ou 1)
  - Le temps de réaction pour chaque séquence
  - Le niveau de compétence linguistique du participant
  - L'âge du participant
  - ...
- Finalement, on collecte aussi des données des locuteurs natifs (**groupe de contrôle**) pour s'assurer que l'expérience fonctionne

## Question

☞ Quel serait le problème si l'on ne collectait pas des données du groupe de contrôle?

## Exemples additionnels

ABX/AXB

« Écoutez les trois sons suivants : A, X, B. Le son  X est-il plus semblable à A ou à B? »

Exemple

# Exemples additionnels

choix forcé

« Lequel de ces deux mots sonne plus naturel (en anglais)? »

[kɪ.mɛ.sər]

[kɪ.'mɛ.sər]

## Pratique

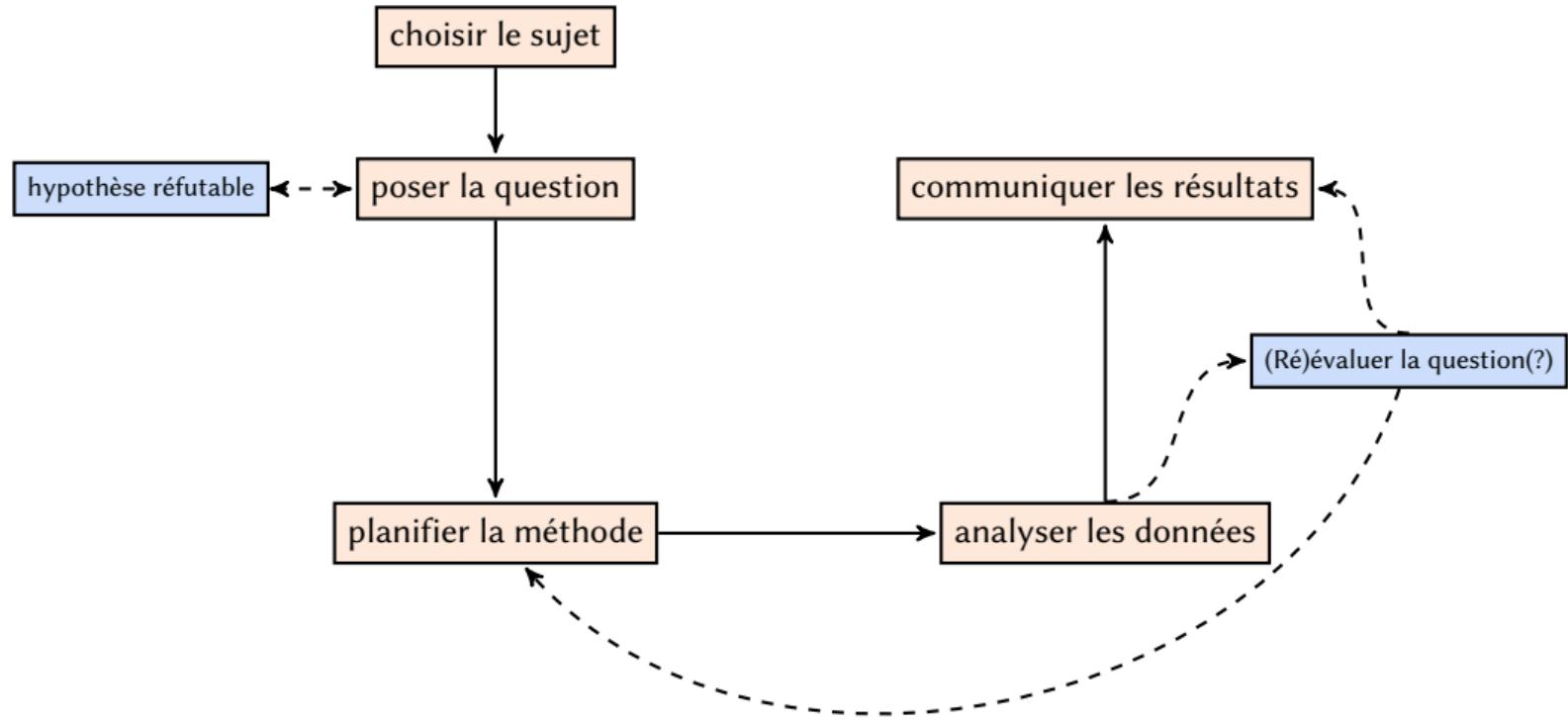
- **Question de recherche** : L'accent régional d'un locuteur influence-t-il la perception de son intelligence dans un contexte professionnel?
- **Méthode** : une enquête où les participants écouteront des enregistrements de différents locuteurs avec différents accents régionaux présentant les mêmes informations professionnelles. Les participants devront ensuite évaluer l'intelligence perçue des locuteurs sur une échelle de 1 à 10.

1. Quelles variables sont pertinentes dans l'étude en question?
2. Quels sont les problèmes potentiels?

## Méthodes → Analyse

- Notre méthode de collecte de données détermine le type de données dans notre analyse :
    - Temps de réaction = données continues → p. ex., **régression linéaire**
    - Choix forcé = données catégoriques (binaires) → p. ex., **régression logistique**
    - Échelle de classification = données scalaires → p. ex., **régression ordinale**
    - ...
  - Il y toujours plusieurs méthodes appropriées d'analyse pour n'importe quel type de données
- ☞ Par contre, on trouve aussi plusieurs choix d'analyse qui sont **incorrects**

# Synthèse



## Semaine prochaine

- ☞ Lisez attentivement les chapitres 1 et 2 de notre livre numérique
  - Lisez Barnier (2023, ch. 1–3; 5–7)

# Références I

- Julien BARNIER : *Introduction à R et au tidyverse*. 2023. Available at <https://juba.github.io/tidyverse/>.
- Guilherme D GARCIA : *Data visualization and analysis in second language research*. Routledge, New York NY, 2021.
- Guilherme D GARCIA : Méthodes expérimentales et analyse de données. <https://lng1100.quarto.pub/>, 2024. Livre numérique du cours LNG1100 de l'Université Laval.
- Joseph LARMARANGE : *Introduction à l'analyse d'enquêtes avec R et RStudio*. 2023. Available at <https://larmarange.github.io/analyse-R/analyse-R.pdf>.
- Hadley WICKHAM, Mine Çetinkaya RUNDEL et Garrett GROLEMUND : *R for data science*. O'Reilly Media, Inc., Sebastopol, CA, 2023. Available at <https://r4ds.had.co.nz/index.html>.

# Théorème de Bayes et tests médicaux

Si vous vous intéressez à la question

- **On ne va pas étudier ce théorème ce semestre. Mais si vous êtes curieux :**

On considère un test avec une précision de 95%, et une maladie rare ( $\frac{1}{1000}$ ). Quelle est la probabilité d'avoir la maladie si le résultat du test est **positif** ?

Retourner<sup>↑</sup>

# Définition des Probabilités

- La probabilité d'avoir la maladie  $P(M)$ :

$$P(M) = \frac{1}{1000}$$

$$P(\neg M) = 1 - P(M) = \frac{999}{1000}$$

$$P(+) | M) = 0,95$$

$$P(+) | \neg M) = 0,05$$

Retourner<sup>↑</sup>

# Théorème de Bayes

Le théorème de Bayes :

$$P(M|+) = \frac{P(+|M) \times P(M)}{P(+)}$$

$$\begin{aligned} P(+) &= P(+|M) \times P(M) + P(+|\neg M) \times P(\neg M) \\ &= 0,95 \times \frac{1}{1000} + 0,05 \times \frac{999}{1000} \\ &= 0,0509 \end{aligned}$$

$$P(M|+) = \frac{0,95 \times \frac{1}{1000}}{0,0509}$$

$$\boxed{\approx 1,87\%}$$

Retourner↑

# Théorème de Bayes

Le théorème de Bayes est formulé comme suit :

$$\begin{aligned}\text{Probabilité a posteriori : } & P(M|+) \\&= \frac{\text{Vraisemblance} \times \text{Probabilité a priori}}{\text{Vraisemblance marginale}} \\&= \frac{P(+|M) \times P(M)}{P(+)} \\&= \frac{P(+|M) \times P(M)}{P(+|M) \times P(M) + P(+|\neg M) \times P(\neg M)}\end{aligned}$$

Où  $P(M)$  est la probabilité a priori de la maladie,  $P(+|M)$  est la vraisemblance du test positif sachant la maladie, et  $P(+)$  est la vraisemblance marginale d'un test positif.

Retourner↑