Incertitudes...

ou ce que l'on a oublié sur les statistiques

Lucas Voirin

2025-04-06

Table of contents

Avant-propos

Objectifs pédagogiques

L'objectif principal de ce document est de présenter de manière intuitive les concepts fondamentaux des statistiques. Il ne s'agit tout au plus que d'un brouillon d'ébauche de livre.

- Rédiger en français (pour s'inscrire dans une démarche d'ouverture de la science)
- Présenter les concepts fondamentaux des statistiques
- Adopter une approche intuitive et visuelle
- Insister sur la notion de modèle (linéaire) et sa généralisation
- Insister sur la notion d'incertitude
- Insister sur la formulation mathématique des modèles

Public visé

Ce document ne vise pas nécessairement un public n'ayant jamais suivi de cours de statistiques. Au contraire il est écrit pour s'adresser à des personnes ayant déjà eu des cours, qui se rappellent de quelques notions mais de manière floue ou incertaine. Pour autant, considérant les différences de niveau des personnes, il traite de concepts de base et devrait ainsi être accessible à toute personne découvrant les statistiques.

Conseils de lecture

C'est vraiment important de lire les formules mathématiques et de les comprendre. On va commencer par des formules très simples et quand elles seront plus compliquées, on prendra le temps de les détailler.

Les données utilisées

Les données utilisées pour illustrer les propos sont accessibles librement.

Pour illustrer le contenu de ce livre, j'utilise le jeu de données provenant de la librairie `palmerpenguins` (Horst, Hill, and Gorman 2020).

Quelques inspirations

Voici une liste de quelques livres qui m'inspirent pour rédiger ce document.

- *Statistical modeling with R, a dual frequentist and bayesian approach for life scientists* de Pablo Inchausti (2023)
- *Applied Statistical Modelling for Ecologists* de Marc Kéry et Kenneth F. Kellner (2024)
- *The Order of the Statistical Jedi* de Dustin Fife (2022)

1 Variables et distributions

Chapitre pour présenter le concept de variable, de distribution, de jeu de données.

TODO: enlever la partie modélisation

TODO: décrire progressivement la construction de la variable

TODO: décrire la structure d'un jeu données et la façon de le représenter

1.1 C'est quoi une variable ?

En science, c'est courant de mesurer des trucs.

Mesurer un truc, c'est attribuer une valeur numérique à ce truc. Et les méthodes pour associer une valeur numérique à un truc sont nombreuses : on peut prendre une règle pour mesurer une distance, une balance pour mesurer une masse, un thermomètre pour mesurer une température, un sonomètre pour mesurer une intensité sonore...

Une valeur numérique c'est pratique pour décrire un truc, mais on peut aussi "mesurer" ce truc avec des valeurs non numériques. Cela peut être la couleur, l'espèce ou n'importe quelle autre caractéristique du truc.

En général, on ne se contente pas d'une seule mesure.

On répète l'opération de mesure plusieurs fois, souvent dans des conditions différentes. On dit que l'on réalise des **observations**.

Quand on a un ensemble de mesures, on parle alors de **variable**. Pourquoi ? Tout simplement parce que d'une observation à une autre, la valeur de la mesure peut **varier**.

> Par exemple...
>
> On peut mesurer la longueur du bec de manchots : la variable s'appellerait *Taille du bec* et les valeurs de cette variable seraient exprimées de manière numérique en milimètres.
>
> On pourait aussi noter le nom de l'espèce de manchot : la variable serait *Espèce* et les valeurs de cette variable seraient par exemple les noms des différentes es-

pèces.

À chaque fois que l'on mesure la taille du bec et l'espèce sur un manchot différent, on réalise une nouvelle observation.

Et il y a beaucoup de choses qui peuvent faire varier les valeurs de mesure (par exemple l'espèce sur laquelle on fait la mesure, la date à laquelle on fait la mesure...). En fait, les trucs qui peuvent faire varier les mesures d'une variable... sont des variables aussi !

Par exemple...

La longueur du bec de manchots peut varier en fonction de l'espèce de manchot sur laquelle on effectue la mesure ou en fonction du poids du manchot.

1.1.1 Différents types de variables

Il existe différents types de variables. Le type d'une variable est simplement le type de valeur dont la variable est composée.

Si les valeurs composant la variable sont des **catégories** (nom d'espèce, couleur, sexe...) alors il s'agira d'une **variable qualitative**. On dit aussi variable catégorique.

Si les valeurs composant la variable sont des **nombres** (distance, température, poids...), alors il s'agira d'une **variable quantitative**. On dit aussi variable numérique.

1.1.2 Représenter une variable

histogramme, boxplot

1.2 C'est quoi une distribution ?

La distribution d'une variable, c'est la la description des probabilités d'obtenir chaque valeur que peut prendre la variable.

En statistique on utilise souvent des lois pour décrire ces distributions.

2 Anatomie d'un modèle

C'est une représentation mathématique de la réalité. Quand on réalise des mesures sur le terrain, on considère que les valeurs mesurées sont la réalité. Le modèle, c'est l'objet mathématique qui va permettre de reproduire ces données.

Un modèle est composé de deux parties :

- une partie déterministe, celle que l'on peut calculer avec des formules mathématiques
- une partie stochastique, que l'on ne sait pas prédire exactement mais dont on essaye définir le comportement général.

$$y = \mu + \epsilon$$

μ, c'est la partie déterministe, on peut la calculer à partir d'une formule mathématique.

ϵ, c'est la partie stochastique. c'est le coeur des statistiques. On ne sait pas prédire cette valeur, mais tout l'enjeu est d'arriver à décrire son comportement.

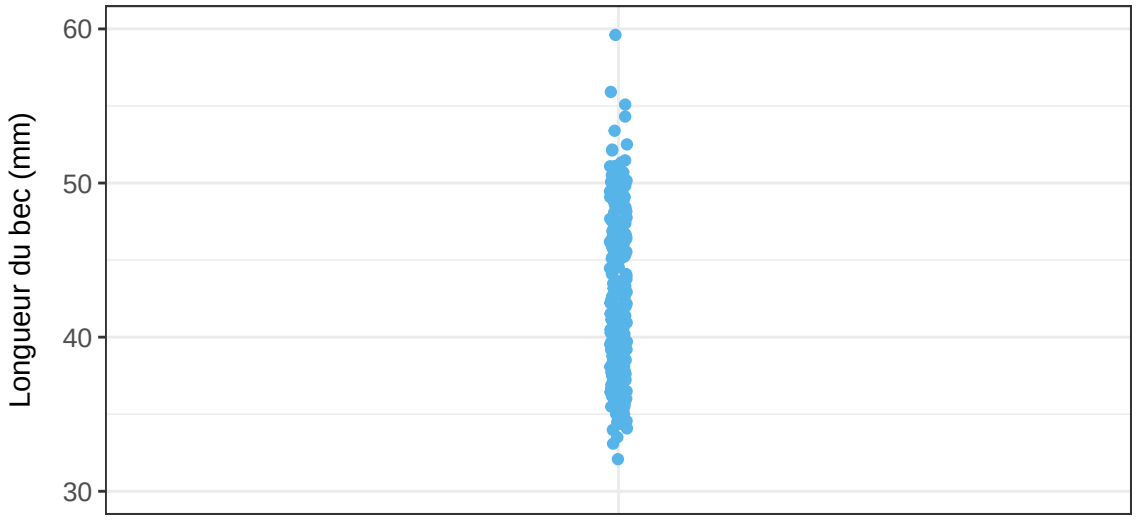
2.1 La moyenne, le plus simple des modèles

2.1.1 La variable à modéliser

On s'intéresse à la taille du bec de manchots. C'est une variable que l'on note y. Il s'agit de notre variable réponse, celle que l'on veut modéliser. y est composée d'un ensemble de valeurs, chaque valeur est une mesure de la longueur du bec d'un manchot. y_1 correspond à la longueur du bec du premier manchot mesuré, y_2 celle du deuxième manchot et ainsi de suite.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

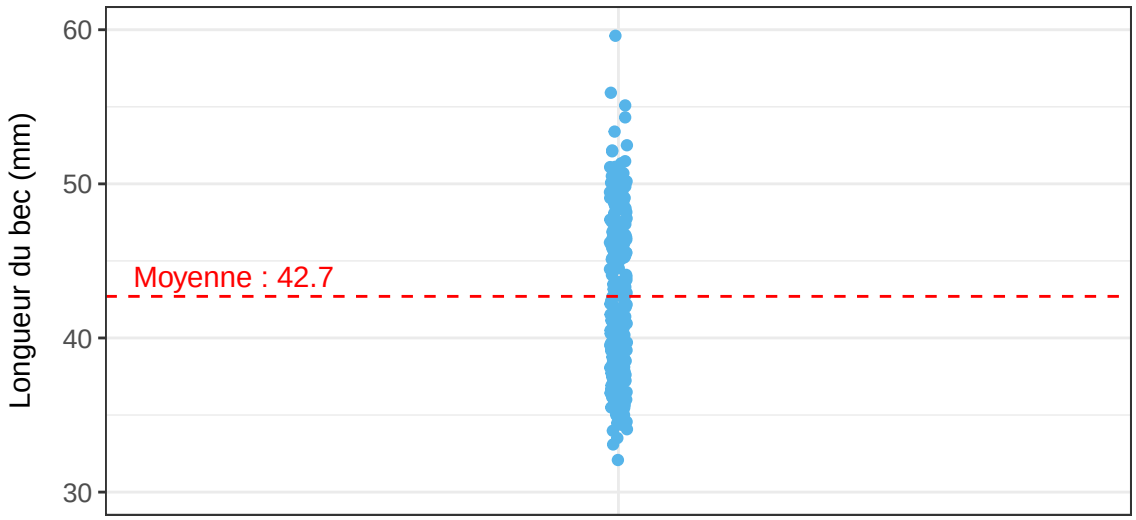
Voici l'ensemble des mesures, représentées par des points dont la position sur l'axe des ordonnées donne la valeur de la mesure de longueur du bec. La position sur l'axe des abscisses est volontairement un peu différente pour chaque point, mais ça c'est juste pour pouvoir mieux voir l'ensemble des points.

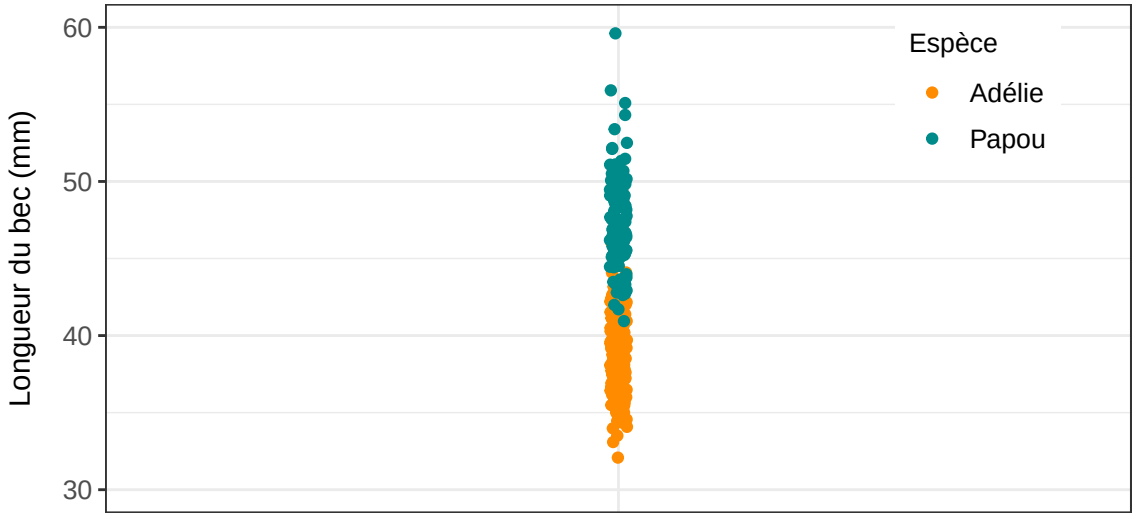


On voit que les mesures de longueurs de bec vont de un peu plus de 30 mm à 60 mm, avec davantage de points entre 35 et 50 mm.

La première chose que l'on pourrait faire avec l'ensemble de ces valeurs, c'est de calculer leur moyenne. Mathématiquement on note cette moyenne $\bar{y}$.

La moyenne des mesures de longueur de bec est de 42.7. Et c'est déjà un modèle de nos données !


Longueur du bec (mm)
60
50
40
30
Moyenne : 42.7


Longueur du bec (mm)
60
50
40
30
Espèce
Adélie
Papou

3 Summary

In summary, this book has no content whatsoever.

References

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.