Rappels : Analyse Univariée et Bivariée

Paul Chapron [1] & Yann Ménéroux [1]

2021-2022

[1]IGN-ENSG-UGE

ENSG Géomatique
ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

Introduction

Dans les cours précédents ...

Notions pour manipuler les variables aléatoires, et estimer certains descripteurs

- co-variance
- intervalle de confiance
- bootstrap
- ...

Motivation

L'analyse univariée permet de décrire la forme et de quantifier les caractéristiques de la répartition des valeurs d'une variable.

- Notion de distribution
- Visualisation (Histogramme, densité, boxplots, . . .)
- Moments, Quantiles, CV

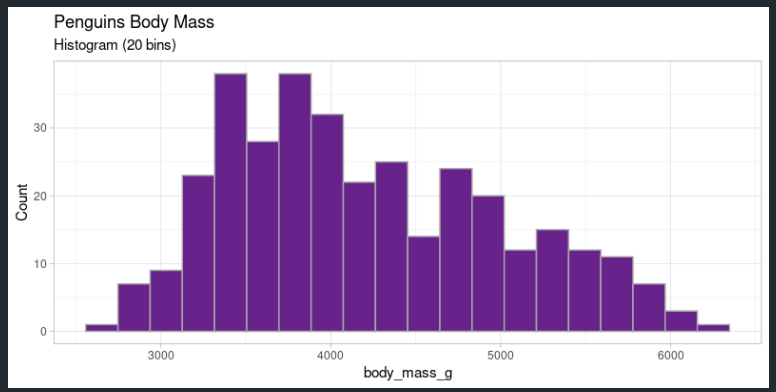
Analyse Univariée

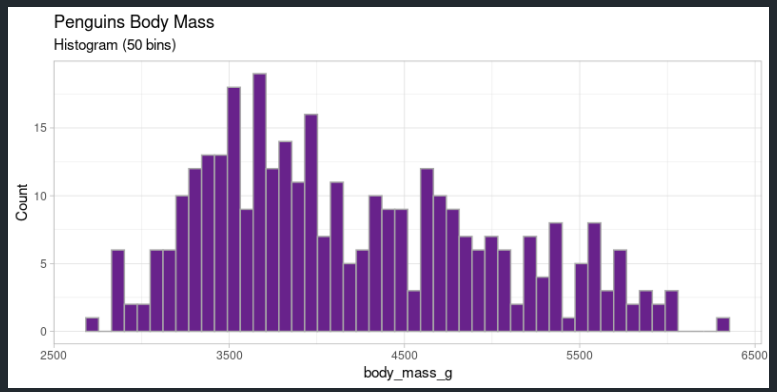
Histogramme

Histogramme d'une variable

Représentation graphique des effectifs associés à des classes de valeurs d'une variable numérique

Le nombre de classes peut varier !



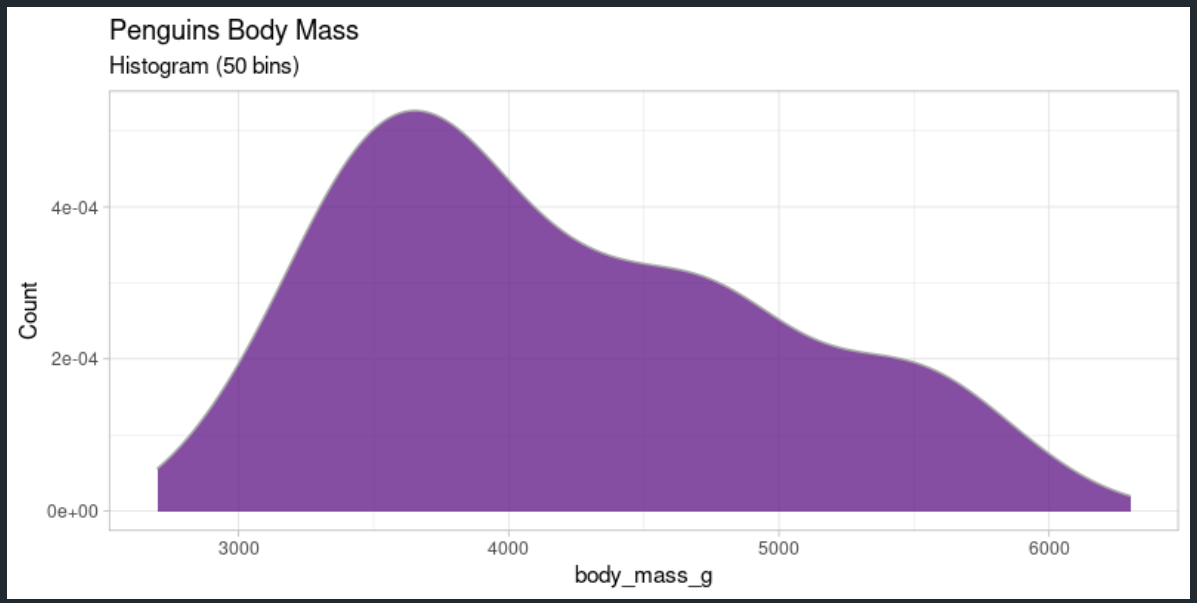


Distribution

Synonymes: distribution empirique, distribution des fréquences, distribution statistique

Tableau ou graphique qui associe les (classes de) valeurs à leur fréquence d'apparition

≈ « Histogramme des fréquences en continu»



Distribution

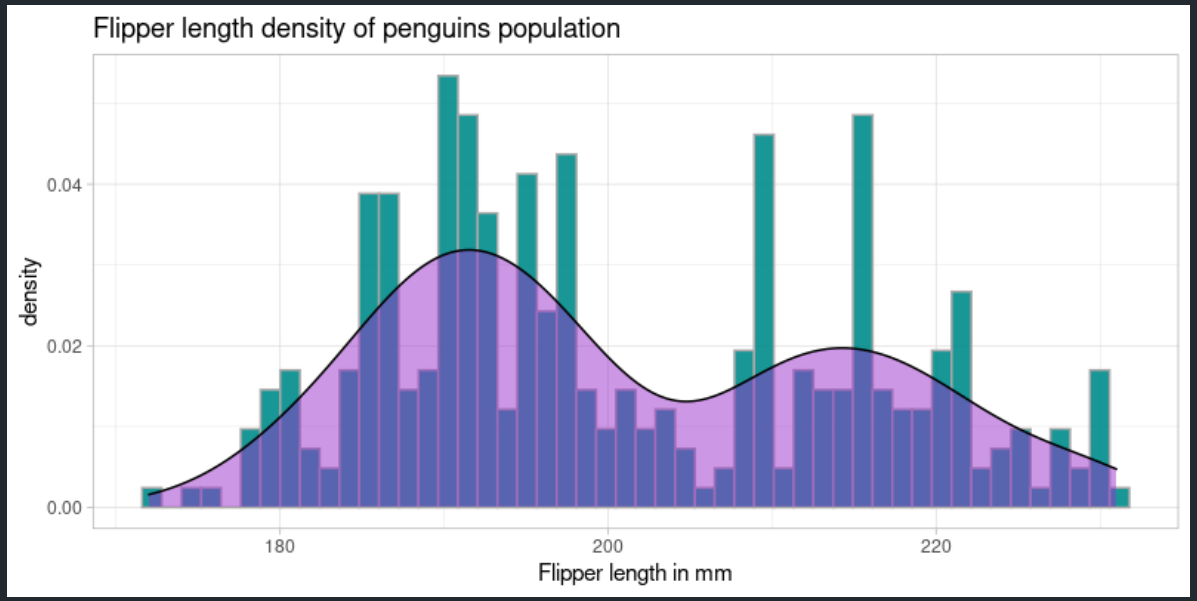
La distribution peut être définie comme une fonction qui donne la probabilité qu'un individu x pris au hasard ait la valeur V_x pour la variable V :

$$distribution(V) \equiv P(V = V_x), \forall V_x \in \Omega_V$$

Avec Ω_V l'ensemble des valeurs que peut prendre V : l'univers de V

Lorsque la variable prend des valeurs réelles, on parle de densité de probabilité, c'est pourquoi on retrouve ce terme "density" sur les axes des ordonnées dans les graphiques de distribution.

Distribution et histogramme



N.B. En toute rigueur, représenter une courbe de distribution de probabilité par dessus un histogramme est impropre : il faudrait deux graphiques distincts, ou au moins deux axes des ordonnées: un pour l'histogramme, représentant un effectif, l'autre pour la distribution , représentant une probabilité

Distribution et lois

Parfois , les distributions empiriques ressemblent à celles de lois de probabilités connues.

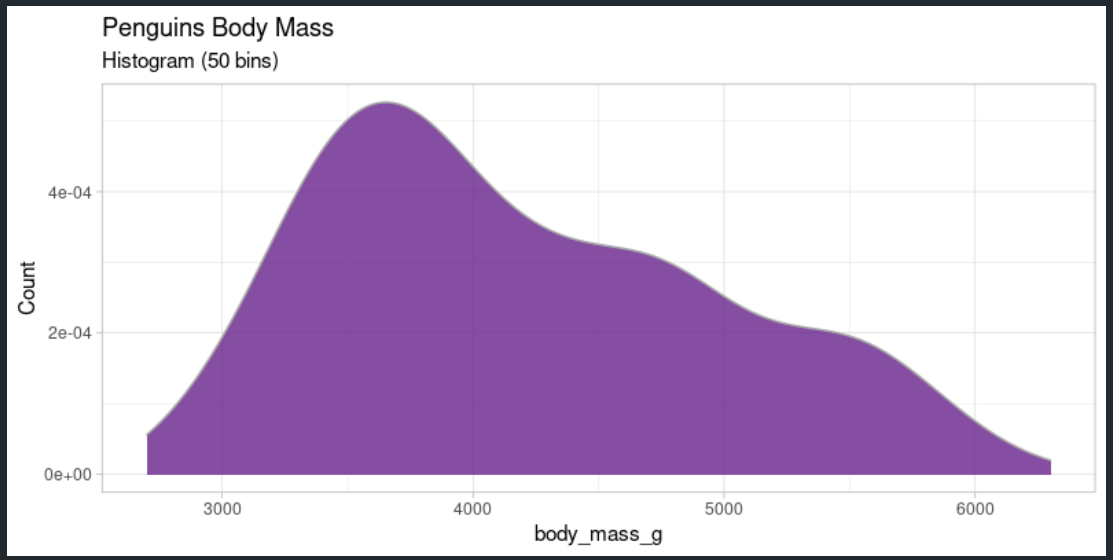
$\rightarrow$ on peut alors modéliser la variable par une variable aléatoire de loi fixée

$\rightarrow$ les paramètres de cette loi doivent être déterminés (ajustement).

Décrire une distribution

La forme d'une distribution donne beaucoup d'informations :

- "pics" : valeurs les plus représentées dans la population
- présence de valeurs extrêmes : la courbe de la distribution est tirée à gauche ou à droite du graphique
- symétrie : les individus se répartissent équitablement de part et d'autre du pic
- aplatissement : la population est plus ou moins resserrée, ou autour de certaines valeurs
- ...



Décrire une distribution : mesures de tendance centrale

Tendance Centrale

La tendance centrale est une valeur qui résume une série de valeurs (quantitative)

- Moyenne
- Médiane
- Mode

Moyenne

$$\bar{x} = \frac{1}{n} \sum_{i=0}^{n} x_i$$

Avantage : chaque valeur compte
Inconvénients :

- sensibilité aux valeurs extrêmes
- pas de signification sur les valeurs discrètes (e.g. 2.5 enfants par foyer)

Pour y remédier (parfois):
$\rightarrow$ exclure les outliers
$\rightarrow$ utiliser un autre estimateur (médiane)
$\rightarrow$ étudier la distribution des valeurs (e.g. cas bimodal) et opérer une classification

Moyenne géométrique

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=0}^{n} x_i}$$

Moins sensible que la moyenne classique aux valeurs extrêmes.

Mode

Le mode d'une variable est la valeur la plus fréquente (d'effectif maximum) d'une variable.

Avantages :

- peu sensible aux valeurs extrêmes
- interprétation simple : cas le plus fréquent

Inconvénient : la valeur du mode ne dépend pas de toutes les observations, la modification d'une valeur n'entraîne pas la modification du mode (ce qui explique sa robustesse aux valeurs extrêmes)

Calcul du mode

Si la variable est quantitative et continue :

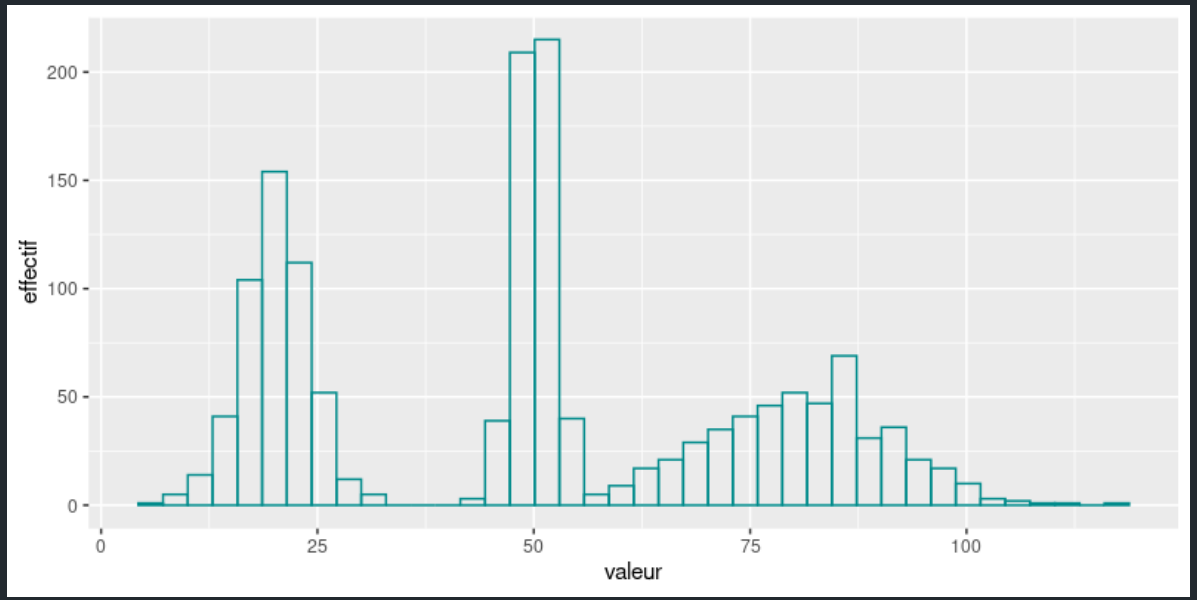
- découper l'étendue de la variable ($max - min$) en intervalle égaux
- compter les effectifs de chaque intervalle
- le mode est la moyenne des valeurs des bornes de l'intervalle de plus grand effectif.

(C'est exactement ce que fait un histogramme graphiquement !)

Modes

Par définition, le mode est unique, mais on peut appeler modes les valeurs des autres pics d'une distribution.
On parle de distribution bi-modale ou tri-modale lorsqu'une distribution présente deux ou trois pics.

Les valeurs modales d'une distribution sont les valeurs correspondant à ces pics.



Médiane

La médiane est la valeur qui sépare une population en deux classes d'égal effectif.
C'est la valeur la plus proche de toutes les autres.

Avantages :

- Souvent plus pertinente que la moyenne
- les valeurs extrêmes ne modifient pas sa valeur
- interprétation facile: un individu sur deux a une valeur inférieure (respectivement supérieure) à la médiane.

Inconvénient : Comme le mode , la médiane ne dépend pas de toutes les observations

N.B. la robustesse de la médiane est bien utile dans le cas de distribution particulièrement asymétriques, où la moyenne est dégradée par les valeurs extrêmes, à droite (valeurs très élevées) ou à gauche (valeurs très faibles).

Moyenne et médiane

Que peut on dire d'une population dont la médiane est inférieure à la moyenne ?

Exemple : revenus mensuels en équivalent temps plein en France en 2016 : le revenu mensuel net moyen est de 2 238 €, le revenu mensuel net médian est de 1 789 € : selon l'`https://www.insee.fr/fr/statistiques/4277680?sommaire=4318291`

Supposons qu'on cherche à évaluer si un salaire mensuel net équivalent temps plein de 2000€ est un bon salaire en France, sans définir trop rigoureusement ce qui signifie «bon».

2000€ est inférieur à la moyenne du pays, on peut le considérer comme trop bas pour être «bon». 2000€ est supérieur au salaire médian, il est supérieur à (au moins) la moitié des salaires du pays, et on peut logiquement le considérer comme un «bon» salaire. Cette double interprétation est due au fait que certains salaires très

Moyenne et médiane

Que peut on dire d'une population dont la médiane est inférieure à la moyenne ?

Décrire une distribution : mesures de dispersion

Dispersion

La tendance centrale est une valeur qui résume une série de valeurs (quantitative)