L1ne

Zeste de savoir

Statistique descriptive à une dimension

12 août 2019

Table des matières

Derrière ce nom rallongé se cache en réalité un domaine que vous utilisez régulièrement. La **statistique descriptive à une dimension** consiste à étudier des données basées sur une seule observation et d'en émettre des conclusions : les notes obtenues par des étudiants, les tailles d'un groupe de personnes ou encore les températures relevées chaque jour pendant une période. Ici, on parle d'une dimension puisque l'on observe qu'une variable, qu'un aspect à la fois. On ne s'intéresse pas à savoir s'il existe un quelconque lien entre deux phénomènes, on ne regarde qu'une chose à la fois : c'est le rôle de la statistique descriptive à une dimension (ou statistique *unidimensionnelle*).

L'objectif de ce tutoriel est double : vous exposer tout le vocabulaire nécessaire aux études statistiques, et vous introduire les différentes notions de mesures statistiques pour exploiter des données unidimensionnelles. Vous serez alors le roi des statistiques (à une dimension toujours).

> **!** Bien que vous n'ayez pas besoin de grandes connaissances en mathématiques, il est néanmoins utile d'avoir quelques bases pour appréhender correctement ce tutoriel. En particulier, comprendre l'indice de sommation Σ est indispensable. Certains points demandent quelques connaissances avancées, mais l'objectif global du tutoriel est qu'un étudiant en fin de lycée ou début de cycle universitaire puisse découvrir cette discipline sans grandes difficultés.

Je remercie tout particulièrement Looping, Anto59290, KFC et Holosmos pour m'avoir donné plusieurs conseils afin d'aboutir à ce tutoriel.

1. Vocabulaire

Avant de se lancer directement dans l'exploitation des résultats, nous devons fixer quelques mots de vocabulaire. En effet, nous allons avoir beaucoup d'informations et plusieurs types de

variables. Il est donc nécessaire de bien formaliser tout ceci au début. Cette partie est assez dense, mais elle a le mérite de poser les bases et de lever toute ambiguïté quant à la nature des mots employés . Commençons par l'environnement d'une étude statistique, les *Qui ? Quoi ?* indispensables à chaque étude statistique.

— **L'individu** est l'unité d'observation. Par exemple, dans un sondage, un individu est une personne ayant répondu au sondage. Mais cela peut aussi être un objet, comme par exemple un comparatif de performance entre plusieurs voitures.
— La **population** est l'ensemble des individus concernés par l'étude. L'ensemble des entreprises françaises est une population.
— **L'échantillon** est la partie de la population que l'on étudie. L'ensemble de 1000 entreprises françaises tirées au hasard constitue un échantillon de la population précédente.

! Il ne faut absolument pas confondre **échantillon** et **population** ! En statistique descriptive, on **décrit** un échantillon et non une population ! Ce dernier cas est le rôle de la statistique inférentielle dont le but est d'estimer les caractéristiques d'une population à partir d'un échantillon.

Si j'insiste sur ce point, c'est que l'on a souvent tendance à tirer des conclusions trop hâtives lorsque l'on étudie un échantillon sans en considérer la nature. Comme vous pouvez le remarquer, pour effectuer des études statistiques, il ne suffit pas de savoir faire des calculs, il faut également **avoir des connaissances dans le domaine de l'étude** !

Pour vous aider, voici un schéma réalisé par Blackline qui résume très bien ce qui a été dit plus haut

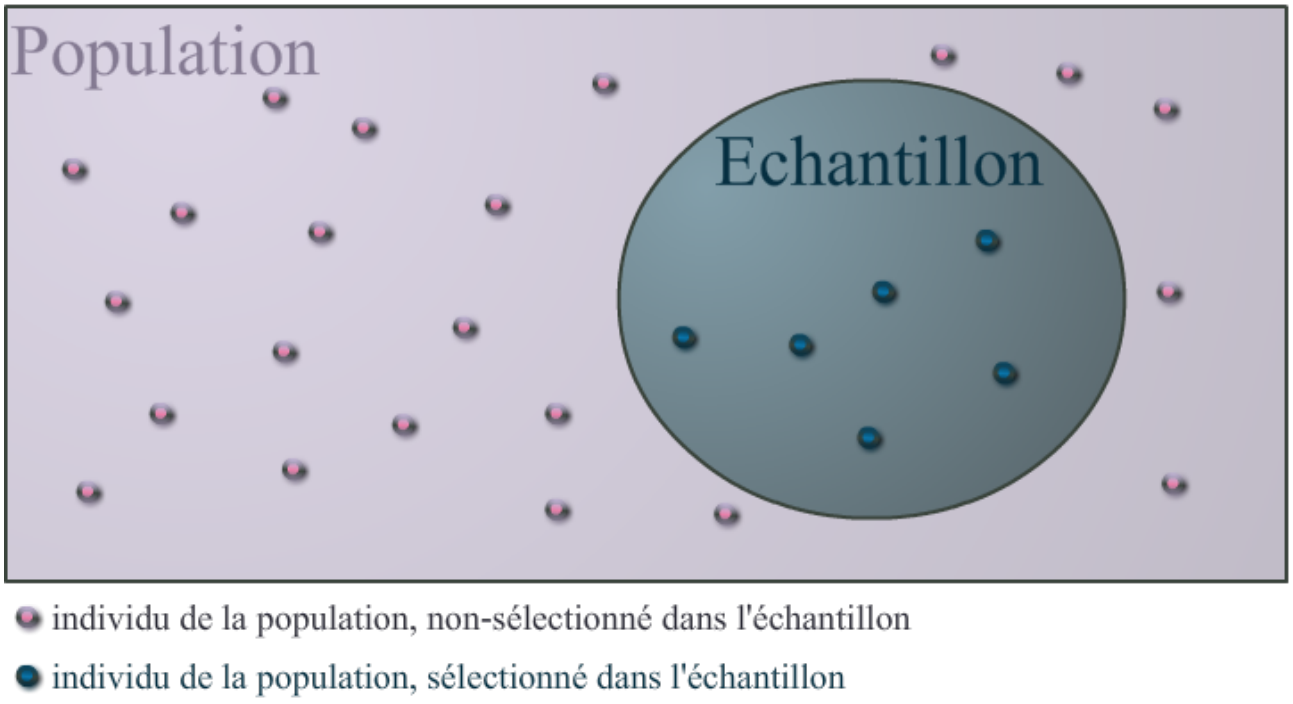


FIGURE 1. – Schéma illustrant une population et un échantillon (Image de Blackline)

À présent, nous allons définir les **variables**, car c'est ce qui nous intéresse tout particulièrement. Une **variable** est une caractéristique (un "aspect") d'un individu. Par exemple, la taille d'une personne, la température de l'air à une date donnée ou la vitesse maximale d'une voiture sont des variables. Le principe d'une variable peut se comprendre à l'aide du schéma suivant.

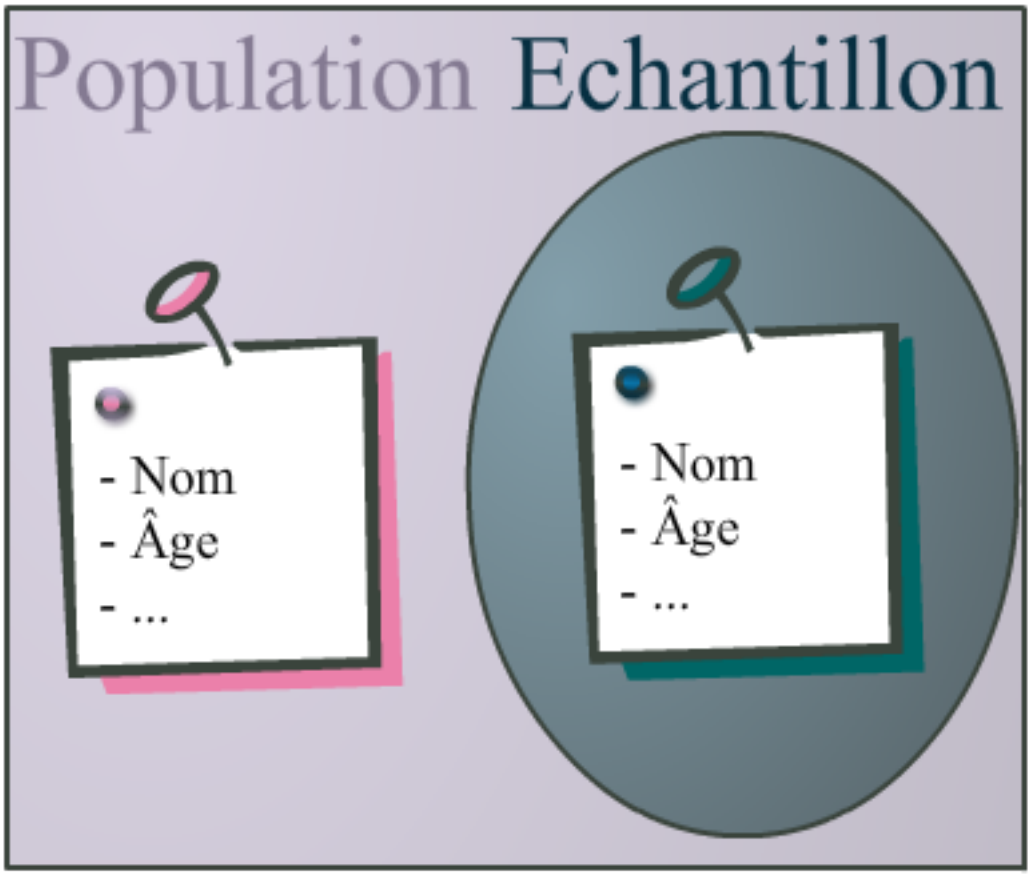


FIGURE 1. – Schéma illustrant des individus avec des variables (Image de Blackline)

En résumé, une **population** est constituée **d'individus**. Tous ces individus possèdent leurs propres caractéristiques (pour des personnes, cela correspond à la taille, l'âge, le nom, ...), que l'on regroupe sous forme de **variables**. Seulement, dans une étude statistique, on ne s'intéresse qu'à une partie de cette population, et cette partie est appelée **l'échantillon**.

1.0.1. Un premier exemple

On souhaite réaliser un sondage à Paris pour connaître le temps de trajet moyen des utilisateurs des transports en commun pour aller travailler. Pour cela, à chaque personne sondée, on recueille :

— Son âge et sa catégorie socio-professionnelle (CSP)
— Son temps moyen passé dans les transports en commun

On a donc trois variables (l'âge, la CSP et le temps moyen). Essayons de définir correctement l'environnement de l'étude :

— Ici, les individus sont les **personnes sondées**.

— La population est l'ensemble des personnes **utilisant** les transports en commun.
— L'échantillon est l'ensemble des personnes ayant **répondu favorablement** au sondage. Lorsque l'on connait l'endroit et l'heure du sondage, il est utile de le préciser.

Une fois que tout ceci est fixé, nous pouvons commencer à travailler avec nos données. Cela permet notamment de bien fixer les idées quant au sujet de l'étude.

Notez qu'en règle général, en statistique descriptive, le choix de la population peut différer pour un même échantillon mais sous une autre étude. En revanche, ici, il est nécessaire de bien définir notre échantillon, car c'est sur cet échantillon que nous allons travailler et émettre des conclusions après l'étude.

2. Les types de variable

Il existe deux types de variables, les unes à valeurs **numériques**, et les autres à valeurs **ordinales**.

— Les **variables quantitatives**, qui sont des variables à valeurs numériques, pour lesquelles les opérations arithmétiques ont un sens. Par exemple, un âge, une distance, un volume, etc.
— Les **variables qualitatives**, où les valeurs possibles sont codées par des modalités (ou catégories). Par exemple, la couleur des yeux, le département, ou tout autre codage où les opérations arithmétiques ne sont pas correctement définies.

Parmi les variables *quantitatives*, on dispose des variables **discrètes** et **continues**. Enfin, parmi les variables *qualitatives*, on retrouve les variables **nominales** et **ordinales**. Nous allons présenter en détail les variables quantitatives plus bas. Pour les variables discrètes, la différence réside dans le fait qu'une variable ordinale est naturellement **ordonnée** (mention au Baccalauréat, niveau d'appréciation d'un produit, ...) alors qu'une variable nominale ne dispose **d'aucun ordre** (on ne peut pas ranger dans un ordre précis la couleur des yeux, ou la localisation d'une entreprise).

Un moyen de faire la différence entre ces deux types est de remarquer que *quantitatif* se rapproche de *quantité*, c'est-à-dire numérique.

Dans le cadre de ce tutoriel, nous n'étudierons que le cas des variables quantitatives. Lorsque l'on étudie une variable qualitative, il y a plusieurs paramètres à prendre en compte, ce qui rend leur manipulation plus délicate.

Il est conseillé de faire des pauses et de lire cette partie plusieurs fois : il y a beaucoup de notions, une multitude de notations et des concepts assez différents de ceux que vous avez l'habitude d'utiliser. Il est tout à fait normal de ne pas tout comprendre à la première lecture, allez-y à votre rythme.

2.0.1. Les variables quantitatives discrètes

Commençons par le cas le plus simple, celui des variables (quantitatives) discrètes. Il s'agit des variables prenant leurs valeurs dans un ensemble fixé dénombrable. Par conséquent, si X est une variable discrète, elle peut prendre k valeurs différentes :

$$\{a_1, ..., a_k\}$$

Où les a_i sont des valeurs numériques, avec i un nombre entier entre 1 et k.

i Très souvent, j'utiliserai i en tant qu'indice se situant entre 1 et k. Par exemple, a_i est le i-ème élément de l'ensemble $\{a_1, ..., a_k\}$.

k est donc le nombre d'éléments de l'ensemble précédent. Par exemple, si l'on reprend le premier exemple, la variable *âge* est à valeurs dans $\{1, 2, ...N, ...\}$ (on veillera à prendre N assez grand mais assez réaliste : il est inutile de prendre $N = 1000$ par exemple). Pour chaque valeur a_i, on note n_i **l'effectif**, soit le **nombre d'individus** dont la variable X a pour valeur a_i. Usuellement, on omet volontairement les valeurs qui ne sont pas prises au moins une fois par un individu. Par exemple, j'observe l'âge de 10 personnes, ce qui me donne la séquence suivante :

$$22, 45, 22, 26, 26, 31, 45, 31, 31, 45$$

On considère alors que les valeurs que peut prendre la variable *âge* sont $\{22, 26, 31, 45\}$.

! Ce qu'il faut bien comprendre, c'est que les valeurs que peut prendre une variable sont **uniquement en rapport** avec l'échantillon ! En effet, ici, nous avons listé toutes les valeurs présentes dans l'échantillon, mais nous sommes bien d'accord que l'on peut avoir 19 ou 52 ans.

? Mais si l'on se trompe et que l'on rajoute une valeur en plus, est-ce que cela change tout ?

Dans une étude statistique, cela signifie que l'on ajoute du "bruit", soit des informations non utiles ou non désirées. Par exemple, dans ma séquence d'âges, inscrire 60 ne serait pas utile puisque aucune des 10 personnes n'a 60 ans. C'est pour cette raison que l'on ne se focalise que sur les valeurs observées, sans rajouter d'autres informations. À présent, précisons quelques données élémentaires sur les variables. Pour mieux les comprendre, nous allons rester sur l'exemple précédent des 10 personnes (pour rappel $22, 45, 22, 26, 26, 31, 45, 31, 31, 45$) :

- **L'effectif**, associé à la valeur a_i, et noté n_i, est le nombre d'individus dont la variable X vaut a_i. Par exemple, l'effectif associé à 22 est 2 puisqu'il y a deux individus ayant 22 ans.
- **L'effectif total** est le nombre d'individus (ici, il s'agit de n). Ici, il y en a 10.
- La **fréquence** (associée à une valeur) est le rapport de l'effectif sur l'effectif total. On note f la fréquence et elle est définie, pour la i-ème valeur a_i :

$$f_i = \frac{n_i}{n}$$

La fréquence pour 22 est donc $\frac{2}{10} = 0,2$.

— La **fréquence cumulée** (ascendante) est la somme des fréquences du début jusqu'à la valeur concernée. On note F la fréquence cumulée et elle est définie, pour la i-ème valeur prise par X :

$$F_i = \sum_{p=1}^{i} f_p = f_1 + ... + f_i$$

La fréquence cumulée pour 22 est donc $0,2$, puisque 22 est la première valeur ordonnée de cette série.

Voici un petit récapitulatif des symboles et lettres que nous utiliserons systématiquement par la suite :

Nombre de valeurs possibles	Valeur possible	Effectif	Fréquence	Fréquence cumulée
k	a_i	n_i	f_i	F_i

Remarquons ces propriétés immédiates :

$$\sum_{i=1}^{k} n_i = n \qquad \sum_{i=1}^{k} f_i = 1 \qquad F_1 = f_1 \text{ et } F_k = 1$$

2.0.2. Exercice

Pour vous entraîner, commencez par remplir le tableau correspondant à la séquence :

X	n	f	F
22	2	0,2	0,2
26	2	0,2	0,4
31			
45			1

Contenu masqué n°1

2.0.3. Distribution

En théorie des probabilités, la notion de distribution est très importante et revient fréquemment. C'est également le cas en statistiques. Une distribution est un tableau tel que, pour chaque valeur possible a_i, on associe sa fréquence f_i.

i

En théorie des probabilités, on parle plutôt de **loi de probabilité** (loi discrète, continue, ...) alors qu'en statistique, on utilise le terme **distribution** (théorique, empirique, ...).

On a ainsi tendance à regrouper les distributions par des fonctions lorsque l'on travaille sur des données théoriques (lois de probabilités) et par des histogrammes (j'expliquerai les histogrammes en détail plus loin) lorsque l'on travaille sur des données concrètes. La distribution d'une loi normale (loi de référence en probabilités et statistiques) est donnée par cette courbe

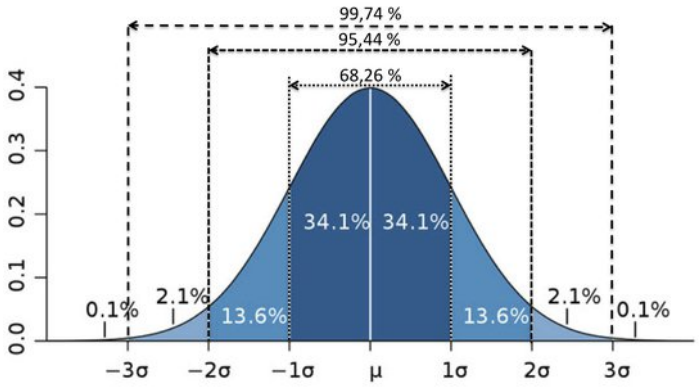


FIGURE 2. – Distribution d'une loi normale (Source : www.ilovestatistics.be)

En particulier, cela signifie que la plupart des observations se situent autour de μ. Je ne rentrerais pas dans les détails sur cette loi, car il me faudrait parler de plusieurs concepts de la théorie des probabilités. L'avantage d'une distribution, c'est que c'est *visuellement parlant*. En un coup d'œil, cela nous renseigne sur la répartition des différentes valeurs.

2.0.4. Les variables quantitatives continues : le problème d'agrégation

Hormis les variables discrètes, nous rencontrons également un autre type de variable : les variables **continues**. Lorsque l'on ne connaît pas à l'avance les différentes valeurs que peuvent prendre une variable ou que l'on travaille sur des nombres réels (à virgule), la variable est alors continue. On distingue une particularité sur les variables continues : celles-ci peuvent être **agrégées**, ce qui change la façon dont nous étudions la variable. Lorsqu'une variable est **agrégée**, c'est qu'elle a été traitée statistiquement.

?

Traitée statistiquement ? Qu'est ce que cela veut dire ?

Il existe des dizaines de procédés de traitement de données. Dans la réalité, on dispose au départ de données dites *brutes*. Ensuite, en fonction de ce que l'on souhaite en faire, nous allons faire

du ménage parmi ces données : supprimer des valeurs aberrantes ou les **regrouper par classes**. Au final, avec des données agrégées, tout est propre, il n'y a plus qu'à les manipuler .

? Dans ce cas, pourquoi est-ce que les variables continues ne sont pas toutes agrégées ?

Et bien deux choses se passent lors de l'agrégation de données :

1. Il y a souvent **perte de l'information**. En effet, à cause du regroupement par classes (que nous allons voir), les valeurs associées aux variables de chaque individu sont rangées "par paquets", et ne sont plus considérés telles quelles.
2. Tout dépend du **contexte** : en fonction de ce que vous souhaitez étudier, vous n'allez pas traiter les données de la même façon.

2.0.5. Les cas agrégé

Dans le cas agrégé, c'est-à-dire où les variables sont rangées par classes, on ordonne d'une manière similaire aux variables discrètes en créant k classes :

$$\{[a_1, a_2[, ..., [a_k, a_{k+1}[\}$$

Avec $a_1 < a_2 < ... < a_k < a_{k+1}$. Nous allons donc définir les effectifs, fréquences et fréquences cumulées par classes :

X	n	f	F
$[a_1, a_2[$	n_1	f_1	$F_1 = f_1$
$[a_2, a_3[$	n_2	f_2	$F_2 = f_1 + f_2$
...	...	...	...
$[a_k, a_{k+1}[$	n_k	f_k	1

Dans le cas des variables agrégées, on introduit c_i, les milieux des classes où, pour tout $1 \leq i \leq k$:

$$c_i = \frac{a_i + a_{i+1}}{2}$$

2.0.6. Le cas non agrégé

Dans le cas non agrégé, on considère chaque valeur associée aux variables de chaque individu de manière unique : il n'y a plus de regroupement dans des classes. Ainsi, on note $x_1, ...x_n$ les valeurs prises par la variable X pour chaque individu (puisqu'il y a n individus).

Données non agrégées ne veut pas dire mauvaises données ! Il s'agit juste de données non traitées, mais rappelez vous que l'on traite des données dans un but précis. Lorsque l'on ne sait pas dans quel objectif sera utilisé des données, mieux vaut les laisser brutes.

3. Mesure de tendance centrale

Dans un premier temps, commençons en douceur avec les mesures de tendance centrale. Il s'agit des mesures les plus classiques et celles qui permettent de résumer le plus simplement un échantillon. En particulier, les mesures de tendance centrale permettent d'obtenir des informations "de groupe", c'est-à-dire vers quelles valeurs les variables ont tendance à se concentrer. Définissons une mesure de tendance que vous connaissez tous et que vous utilisez le plus souvent : la **moyenne arithmétique** !

3.0.1. La moyenne arithmétique

Concrètement, la moyenne (arithmétique) est la mesure permettant de résumer le plus simplement l'échantillon. On la note souvent avec une barre au dessus de la variable.

Discrète	Continue agrégée	Continue non agrégée
$\overline{X} = \sum_{i=1}^{k} f_i a_i$	$\overline{X} = \sum_{i=1}^{k} f_i c_i$	$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Pour rappel, dans le cas continu agrégé, on dispose de k classes dont le milieu de chaque classe est donnée par :

$$\forall 1 \leq i \leq k, c_i = \frac{a_i + a_{i+1}}{2}$$

On utilise couramment le terme *moyenne* pour définir la moyenne arithmétique, mais sachez qu'il existe d'autres moyennes (si vous souhaitez approfondir, je vous laisse vous documenter à ce sujet). Par exemple, pour une variable continue non agrégée, on peut définir plusieurs moyennes :

Arithmétique	Géométrique	Quadratique	Harmonique	Générale d'ordre p
$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$	$G = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$	$Q = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$	$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	$\mathcal{M}_p = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}}$

Par exemple, on utilise la moyenne géométrique lorsque l'on souhaite calculer une moyenne de taux, ou encore la moyenne harmonique pour déterminer la vitesse moyenne de plusieurs trajets.

Pourquoi définissons-nous la moyenne de cette manière ?

Cette mesure permet de résumer grossièrement un échantillon. Dans l'idée, cela signifie que si j'ajoute un individu dans mon échantillon, celui-ci devrait avoir une valeur proche de la moyenne. Autrement dit, il s'agit de la valeur que tous les individus devraient avoir dans le cas le plus hypothétique possible. Par exemple, dans une classe d'étudiants, si la moyenne pour un examen est de 10, cela signifie que si tous les étudiants de cette promotion auraient obtenu 10, la moyenne en serait inchangée.

Mais alors la moyenne peut-être trompeuse ?

Oui ! C'est là un des gros problèmes de la moyenne : en plus d'être sensible aux valeurs extrêmes (c'est-à-dire aux grandes valeurs par rapport aux autres), elle ne **mesure pas la dispersion** des individus. Par exemple, prenons 2 groupes de 6 étudiants ayant passé le même examen et on donne les notes de chaque groupe :

Groupe	Notes	Moyenne
A	$10, 10, 10, 10, 10, 10$	10
B	$0, 0, 0, 20, 20, 20$	10

Voyez-vous le problème ? Nous avons la même moyenne pour les deux groupes, alors que le premier groupe est complètement **homogène**, et le second complètement **hétérogène**. C'est pour cette raison que l'on dit souvent qu'il ne faut pas se fier uniquement à la moyenne car elle est souvent trompeuse. Pour compléter les informations, on utilise ensuite les mesures de dispersions (c'est dans la partie suivante) ou encore la médiane, plus fiable que la moyenne, mais aussi plus restrictive.

3.0.2. La médiane

Il s'agit de la deuxième mesure que l'on calcule très souvent en statistique. La médiane est la valeur telle qu'il y ait autant d'observations au-dessus qu'en-dessous. Mathématiquement, la médiane est définie comme ceci :

$$Med = \text{argmin}_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} |\alpha - x_i|$$

En mathématiques, on cherche souvent à calculer des minimums ou maximums de listes, fonctions ou suites. Par exemple, si $A = \min_{x \in \mathbb{R}} f(x)$ alors A est la plus petite valeur atteinte par f sur $\mathbb{R}$. Maintenant, supposons que je souhaite obtenir non pas la plus petite valeur atteinte par f, mais le x qui fait atteindre f à sa plus petite valeur ? C'est

là qu'intervient argmin, qui a pour rôle de fournir ce fameux x. Pour exemple, les deux équations suivantes sont équivalentes :

$$f(x^*) = \min_{x \in \mathbb{R}} f(x) \Leftrightarrow x^* = \text{argmin}_{x \in \mathbb{R}} f(x)$$

Ainsi, la quantité Med est la valeur telle que

$$\text{argmin}_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} |\alpha - x_i|$$

Soit le plus petit parmi tout les $\alpha \in \mathbb{R}$, ou encore :

$$\sum_{i=1}^{n} |Med - x_i| \leq \sum_{i=1}^{n} |\alpha - x_i|, \forall \alpha \in \mathbb{R}$$

Et un bon moyen de connaître le nombre d'observation situées de part et d'autre d'une valeur, c'est de regarder la **fréquence cumulée** ! En effet, s'il y a autant d'observations au-dessus d'une valeur qu'en-dessous, cela signifie que la fréquence cumulée est égale à $0,5$ au point d'abscisse Med. On en déduit que $Med = F^{-1}(0,5)$. Pour bien comprendre l'utilisation de la fonction réciproque, aidez-vous du graphique suivant :

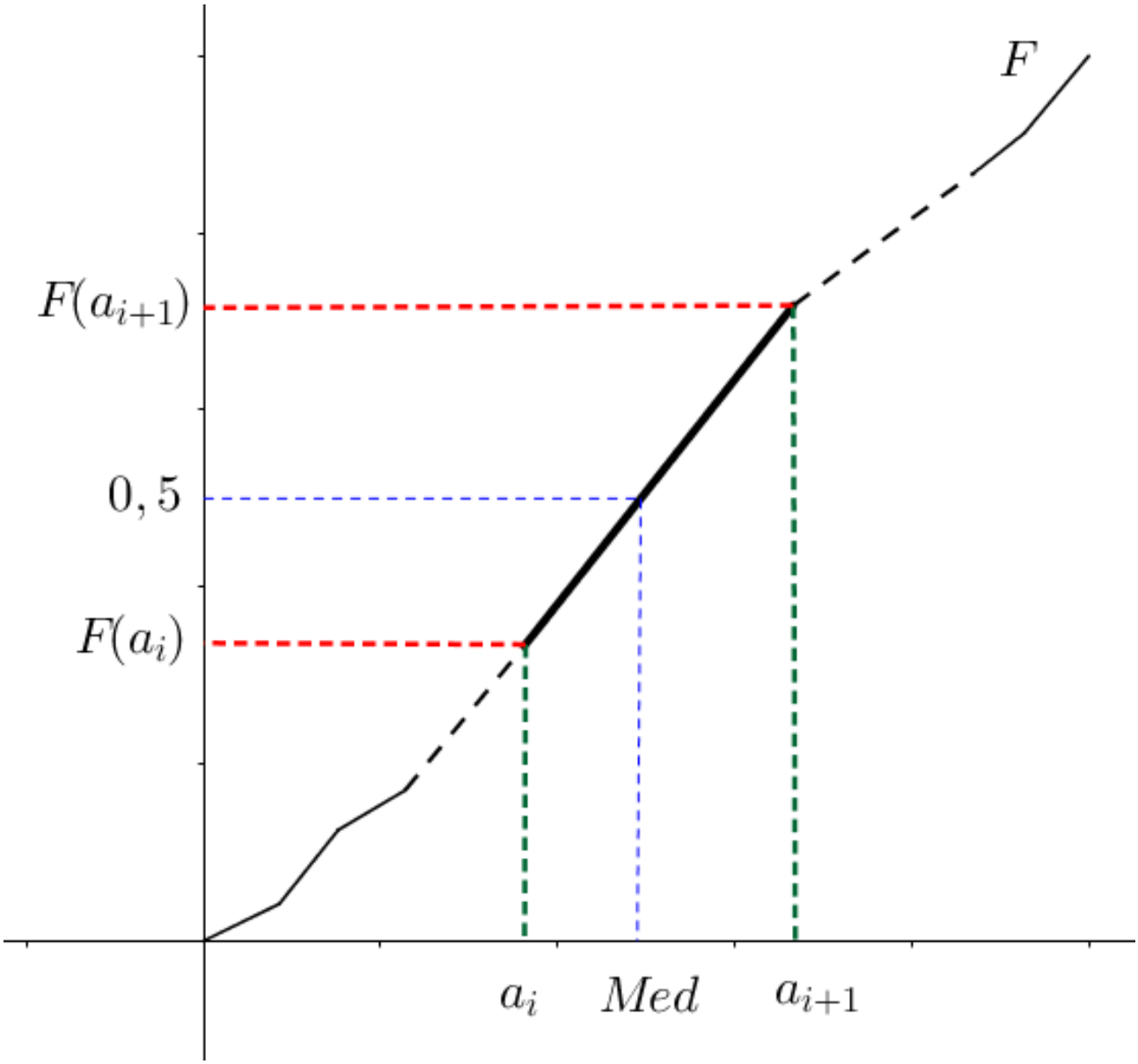


FIGURE 3. – Médiane

Une des différences par rapport à la moyenne, c'est qu'il n'y a **pas de calcul algébrique** général ! En effet, en fonction du type de variable, il peut s'agir de "trouver" la valeur, et non de la calculer.

3.0.2.1. Cas discret Regardons le tableau suivant :

X	F
..	..
a_i	$F_i < 0,5$
a_{i+1}	$F_{i+1} > 0,5$
..	..

D'après ce que l'on a dit auparavant, la médiane se situe entre a_i et a_{i+1}. Usuellement, on considère alors que, dans ce cas, $Med = a_{i+1}$. En revanche, si j'avais eu $F_i = 0,5$ alors on aurait considéré $Med = a_i$.

3.0.2.2. Le cas continu non agrégé

Pour rappel, non agrégé signifie que l'on dispose de n observations x_i pour la variable X. Seulement, on souhaite disposer d'observations **rangées dans l'ordre croissant**. Pour ce faire, on introduit une bijection $\varphi : \{1, ..., n\} \to \{1, ..., n\}$ telle que :

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

Ainsi, on obtient une suite d'observations $(x_{\varphi(i)})$ rangées dans l'ordre croissant. L'intérêt de cette manipulation, c'est que l'on va pouvoir facilement accéder à l'élément "milieu" de cette liste ordonnée, car c'est l'élément le plus au milieu qui nous intéresse pour pouvoir correctement calculer la médiane. Définissons deux cas en fonction de n :

— Si n est impair alors la médiane est la valeur au milieu :

$$Med = x_{\varphi(\frac{n+1}{2})}$$

— Si n est pair alors la médiane est la moyenne arithmétique des deux valeurs centrales :

$$Med = \frac{1}{2}\left(x_{\varphi(\frac{n}{2})} + x_{\varphi(\frac{n}{2}+1)}\right)$$

3.0.2.3. La particularité du cas agrégé

Dans le cas où notre variable est agrégée, la formule est légèrement plus compliquée. En effet, il est rare qu'il existe une fréquence cumulée ayant pour valeur exactement $0,5$. Dans ce cas, nous allons trouver deux fréquences cumulées qui encadrent $0,5$ et effectuer une **interpolation linéaire**. Supposons que $F(a_i) < 0,5$ et $F(a_{i+1}) > 0,5$ (dans le cas où $F(a_i) = 0,5, Med$ vaut alors a_{i+1}), avec $1 \leq i \leq k$ alors la médiane est :

$$Med = a_i + \frac{0,5 - F(a_i)}{F(a_{i+1}) - F(a_i)}(a_{i+1} - a_i)$$

3.0.2.4. Démonstration

Contenu masqué n°2

3.0.2.5. Retour sur notre exemple À présent, je vous propose un exemple (merci à Looping) qui démontre bien l'intérêt de la médiane. Sur une promotion de 20 étudiants, 19 d'entre-eux sont embauchés avec un salaire situé entre 20000 et 30000 dollars par an. En revanche, un étudiant est, quant à lui, embauché à un million de dollars par an. Quelle est la moyenne ici ?

$$m = \frac{19 \times 25000 + 1 \times 1000000}{20} = 73750$$

L'université en question pourrait alors dire "À la sortie de l'université, nos étudiants ont un salaire moyen de 73750 dollars par an !". Et pourtant, s'agit-il de la réalité ? Si l'on regarde le salaire médian, on s'aperçoit qu'il est entre 20000 et 30000 dollars par an (25000 dollars par an si l'on prends le milieu des classes). En conclusion, le salaire médian est ici **beaucoup plus significatif** que le salaire moyen.

i

Globalement, dès que vous détectez une valeur extrême dans votre échantillon (une valeur bien trop importante par rapport aux autres), méfiez-vous de la moyenne.

3.0.3. Quartiles, déciles et quantiles

Essayons de généraliser la notion de médiane. Par exemple, si maintenant, je souhaite couper non plus en deux mais en quatre ma distribution, de sorte qu'un quart des individus se situent en-dessous d'une certaine valeur, et qu'un quart des individus se situent au-dessus d'une autre valeur (toujours en regardant la fréquence cumulée). Soit $\alpha \in]0, 1[$ alors on note q_α le quantile d'ordre α, c'est-à-dire que l'on a α % d'observations en dessous de q_α et $100 - \alpha$ % observations au-dessus de q_α :

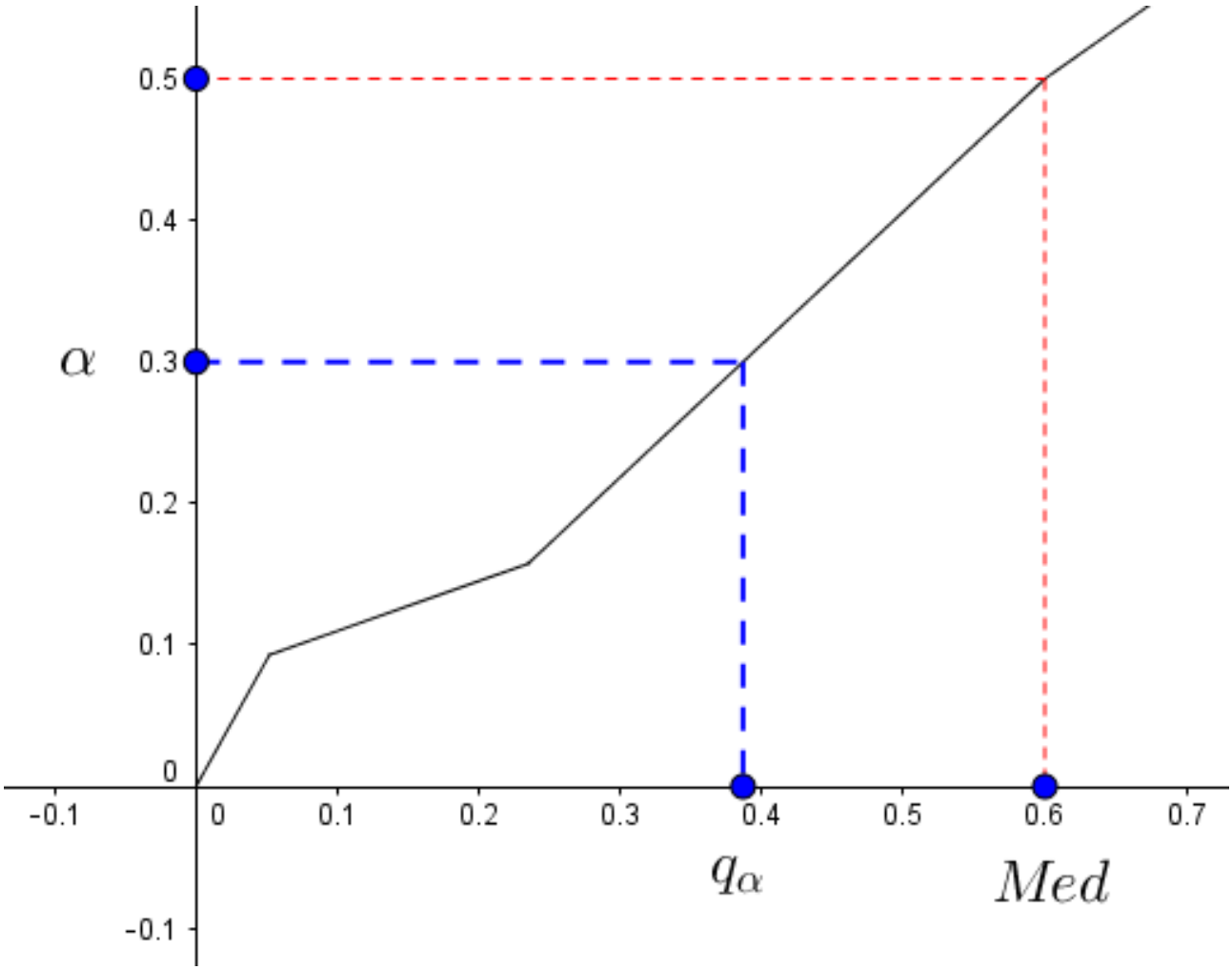


FIGURE 3. – Exemple d'un quantile

Ici, par exemple, on a pris $\alpha = 0,3$, de sorte que 30 % des individus (au sens de la fréquence cumulée) se situent en-dessous de $q_\alpha \approx 0,4$. Dans la suite, on considérera toujours que $0 < \alpha < 1$.

3.0.3.1. Cas discret Comme pour la médiane, on regarde au niveau de la fréquence cumulée.

X	F
..	..
a_i	$F_i < \alpha$
a_{i+1}	$F_{i+1} > \alpha$
..	..

D'après ce que l'on a dit auparavant, le quantile q_α se situe entre a_i et a_{i+1}. On a donc $q_\alpha = a_{i+1}$. Si on aurait eu $F_i = \alpha$ alors on aurait considéré $q_\alpha = a_i$.

3.0.3.2. Cas continu non agrégé On reprend notre bijection $\varphi : \{1, ..., n\} \to \{1, ..., n\}$. Ici, la tâche est plus délicate et il y a plusieurs manières de définir un quantile d'ordre α. On utilisera :

$$q_\alpha = \frac{1}{2}\left(x_{\varphi(\lfloor n\alpha \rfloor)} + x_{\varphi(\lfloor n\alpha \rfloor + 1)}\right)$$

Où $x \mapsto \lfloor x \rfloor$ est la fonction *partie entière*. À noter qu'il faut utiliser les quantiles uniquement dans un environnement propice : inutile de déterminer une quantile d'ordre $0,1$ si vous n'avez qu'une vingtaine d'observations.

3.0.3.3. Toujours ce cas agrégé C'est exactement le même raisonnement que dans le cas de la médiane, sauf qu'au lieu de s'intéresser à $0,5$, on s'intéresse à α. Supposons que $F(a_i) < \alpha$ et $F(a_{i+1}) > \alpha$ (encore une fois, dans le cas où $F(a_i) = \alpha, q_\alpha$ vaut alors a_{i+1}), avec $1 \leq i \leq k$ alors le quantile d'ordre α est :

$$q_\alpha = a_i + \frac{\alpha - F(a_i)}{F(a_{i+1}) - F(a_i)}(a_{i+1} - a_i)$$

La démonstration est la même puisque l'idée consiste encore à effectuer une interpolation linéaire

3.0.3.4. Quelques valeurs importantes En statistique, on utilise très souvent les quantiles pour "partager" un échantillon en plusieurs groupes. Pour cela, les quantiles sont très utiles et on retrouve souvent la notion de **quartile** ou de **décile**. On note très souvent $Q1 = q_{0,25}$ le premier quartile et $Q3 = q_{0,75}$ le troisième quartile (accessoirement, $Q2 = Med$) de sorte que l'échantillon soit partagé en quatre groupes. On utilise quelques fois la notion d'**écart inter-quartile** qui mesure l'écart parmi les 75 % individus les plus proches de la médiane. Cette mesure est définie par :

$$EIQ = Q3 - Q1$$

On résume dans le tableau ci-dessous quelques valeurs usuelles :

Nom	Valeurs possibles
Quartiles	$\alpha \in \{0.25, 0.5, 0.75\}$
Déciles	$\alpha \in \{0.1, 0.2, ..., 0.9\}$
Centiles	$\alpha \in \{0.01, 0.02, ..., 0.99\}$

3.0.4. Mode et classe modale

Le **mode** est, dans le cas d'une variable discrète ou continu non agrégée, la valeur **la plus fréquente** (c'est-à-dire la valeur qui est prise le plus grand nombre de fois par les individus). Dans le cas continu agrégée, on n'utilise non pas le mode mais la **classe modale** (à cause du regroupement par classes) qui est la classe de **fréquence la plus élevée**.

3.0.5. La classe médiane

Pour une variable continue agrégée, lorsque l'on ne souhaite pas réaliser d'interpolation pour déterminer la médiane (par manque d'informations souvent), on utilise la **classe médiane**, notée $CMed$ qui est la première classe dont la fréquence cumulée atteint $0,5$. Par exemple, si $F_3 = 0,4$ et $F_4 = 0,6$, alors $CMed = [a_4, a_5[$.

De manière générale, la moyenne et les quantiles sont les deux principales mesures nécessaires à l'étude statistique, mais il est toujours utile de garder dans un coin de la tête ces mesures qui peuvent toujours nous aider dans certaines situations.

4. Mesure de dispersion

Mesurer la dispersion d'un échantillon, c'est quantifier la manière dont les valeurs prises par la variable se comportent entre-elles, selon qu'elle soient proches ou éloignées les unes des autres.

4.0.1. L'étendue

L'étendue est la mesure de dispersion la plus simple. Comme son nom l'indique, cette mesure précise l'écart entre la plus petite et la plus grande valeur, d'où l'étendue de toutes les valeurs. Il s'agit alors simplement d'une différence de deux valeurs, que l'on note δ :

Discrète	Continue agrégée	Continue non agrégée
$\delta = a_k - a_1$	$\delta = a_{k+1} - a_1$	$\delta = x_{\varphi(n)} - x_{\varphi(1)}$

Où φ est toujours cette bijection de $\{1, ..., n\}$ dans $\{1, ..., n\}$ qui ordonne les (x_i) dans l'ordre croissant. Ce calcul s'effectue à vue d'œil : cela nous donne rapidement une idée sur la façon dont les individus se répartissent. Évidemment, cette mesure est très sensible aux valeurs extrêmes.

Revenons au problème de la moyenne. Rappelez-vous, nous avions mis en évidence le problème d'interprétation de la moyenne dans le cas de nos deux groupes de 6 étudiants. L'objectif est donc de construire une mesure permettant de mesurer cette dispersion des individus autour d'une valeur, habituellement la moyenne.

4.0.2. Écart moyen absolu

Créons ensemble cette mesure. Je prends une réalisation x_i d'un individu, et puisque je souhaite mesurer la dispersion autour de la moyenne, le calcul le plus "naturel" pour la mesurer est la distance en valeur absolue, c'est-à-dire $|x_i - \overline{x}|$. Et si l'on somme pour chaque individu, puisque l'on veut étudier la dispersion de l'ensemble des valeurs, on obtient :

$$\sum_{i=1}^{n} |x_i - \overline{x}|$$

Mais je ne veux que cette mesure soit relatif à la taille de l'échantillon, je devrais donc multiplier chaque distance par un poids, ou plutôt considérer que toutes les observations aient autant de chances de se produire (i.e. qui suit une loi uniforme), ce qui me donnerait :

$$\frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{x}|$$

Et c'est ainsi que nous avons crée notre première mesure de dispersion : **l'écart moyen absolu** !

Discrète	Continue agrégée	Continue non agrégée
$EMA=\sum_{i=1}^{k} f_i\|a_i-\overline{x}\|$	$EMA=\sum_{i=1}^{k} f_i\|c_i-\overline{x}\|$	$EMA=\frac{1}{n}\sum_{i=1}^{n}\|x_i-\overline{x}\|$

Il existe également **l'écart médian absolu**, où cette fois-ci, on mesure la dispersion autour de la médiane ! Elle se construit de la même manière :

Discrète	Continue agrégée	Continue non agrégée
$\sum_{i=1}^{k} f_i\|a_i-Med\|$	$\sum_{i=1}^{k} f_i\|c_i-Med\|$	$\frac{1}{n}\sum_{i=1}^{n}\|x_i-Med\|$

Seulement, dans les faits, on n'utilise pas souvent l'écart moyen absolu, mais plutôt une autre mesure. Je vous donnerai l'explication après avoir présenté cette autre mesure, ne vous inquiétez pas.

4.0.3. Variance et écart-type

Intuitivement, la variance mesure la dispersion (**quadratique**) de la variable autour de sa moyenne. Plus la variance est faible, plus les différentes valeurs prises par les individus sont "proches" de la moyenne. On note v (ou s^2) la variance.

Discrète	Continue agrégée	Continue non agrégée
$v=\sum_{i=1}^{k} f_i(a_i-\overline{x})^2$	$v=\sum_{i=1}^{k} f_i(c_i-\overline{x})^2$	$v=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$

On définit ainsi **l'écart-type** par : $\sigma=s=\sqrt{v}$. On utilise souvent cette mesure puisque, en plus de se ramener à des plus petites valeurs, elle apparaît dans des calculs de statistique où la variance est très souvent inscrite dans une racine carrée.

i

Pourquoi utilise-t-on plus souvent la variance, alors qu'elle semble plus compliquée que l'écart moyen absolu ?

Il faut savoir que les mathématiciens adorent les objets "pratiques". En particulier, lorsque l'on travaille avec des fonctions, on aime bien qu'elles soient continues, dérivables voir de classe C^k. Et c'est ici l'avantage de la variance : cette dernière est différentiable dans $\mathbb{R}^n$ par somme d'applications différentiables, alors que l'écart moyen absolu ne l'est pas (ou en tout cas, pas en un nombre fini de points) dans $\mathbb{R}^n$ (la dérivée de la valeur absolue n'est pas aussi maniable que celle du polynôme du second degré), ce qui rajoute de la complexité dans les problèmes d'optimisation entre autre (on utilise fréquemment la méthode des moindres carrés ordinaires pour la régression, méthode issue de l'optimisation).

Ainsi, dans toute la suite, on préférera s'intéresser à la variance. Néanmoins, dans certains cas, l'écart moyen absolu peut également se révéler utile, à conserver donc dans un coin de la tête.

4.0.3.1. Tout l'intérêt de la variance

Reprenons le tableau des notes des deux groupes d'étudiants :

Groupe	Notes	Moyenne	Variance	Écart-type
A	$10, 10, 10, 10, 10, 10$	10	0	0
B	$0, 0, 0, 20, 20, 20$	10	100	10

Comme vous pouvez le constater, là où la moyenne ne nous donnait pas plus d'informations, la variance et l'écart-type nous donnent toutes les informations pour comprendre la situation. En effet, le premier groupe est bien homogène car sa variance est nulle alors que le second est très hétérogène puisque sa variance est de 100.

i Retenez qu'une variance nulle signifie aucune variation, donc toutes les observations x_i ont la même valeur (ce qui est rare en pratique).

4.0.3.2. Exercice

On donne ci-dessous les notes de trois groupes d'étudiants :

$G_1 = \{10, 12, 11, 13, 14, 12, 7, 15\}$ $G_2 = \{12, 17, 4, 8, 19, 11, 12, 9\}$ $G_3 = \{8, 7, 12, 10, 14, 10, 8, 3, 8, 7\}$

Quelle est le groupe le plus hétérogène ?

Contenu masqué n°3

En conclusion, si l'on devait faire un tableau récapitulatif de toutes nos mesures, et surtout, leurs particularités :

Mesure	Calcul algébrique	Sensibilité aux valeurs extrêmes	Différentiable	Utilisation
Moyenne	Oui	Oui	Oui	Très fréquente

Médiane	Non	Non	Non	Fréquente
Étendue	Oui	Oui	Non	Peu fréquente
Écart moyen absolu	Oui	Oui	Non	Peu fréquente
Variance	Oui	Oui	Oui	Très fréquente

5. Mesure de forme

Nous avons vu comment résumer un échantillon avec une moyenne et comment mesurer sa dispersion avec la variance. Seulement voilà, il nous reste encore quelques mystères : les individus sont-ils plutôt concentrés à gauche de la moyenne, à droite ? La distribution est-elle uniforme, plate ou courbée ? Toutes ces questions motivent la création de mesures de forme, c'est-à-dire de mesures permettant d'identifier plus précisément la répartition des individus.

5.0.1. Les moments

L'outil qui sera le plus important pour mesurer la forme de la distribution est le **moment**. Il se définit dans un contexte plus général que la variance. On appelle moment d'ordre $p \in \mathbb{N}^*$ par rapport à $t \in \mathbb{R}$ la quantité suivante :

Discrète	Continue agrégée	Continue non agrégée
$M_p^t = \sum_{i=1}^{k} f_i(a_i - t)^p$	$M_p^t = \sum_{i=1}^{k} f_i(c_i - t)^p$	$M_p^t = \frac{1}{n}\sum_{i=1}^{k}(x_i - t)^p$

Le moment d'ordre p par rapport à 0 s'appelle le **moment simple** d'ordre p et sera noté M_p. Le moment d'ordre p par rapport à $\overline{X}$ s'appelle le moment centré d'ordre p et sera noté μ_p. On utilise souvent le moment centré.

$$\overline{x} = M_1 \qquad \mu_1 = 0 \qquad \mu_2 = v = M_2 - (M_1)^2$$

5.0.1.1. Démonstration

Contenu masqué n°4

5.0.2. Mesures d'asymétrie

Commençons par déterminer s'il y a présence ou non d'une asymétrie : nous allons regarder si la distribution est orientée plutôt à gauche de la moyenne, ou à droite. Pour se faire une

première idée, on a souvent tendance à regarder la position de la médiane par rapport à la moyenne. En effet, si $Med < \overline{X}$, on s'attends à ce que la distribution soit plutôt oblique à gauche, c'est-à-dire étalée à droite (on parle alors **d'asymétrie positive**). Si, au contraire, $Med > \overline{X}$, on s'attends à ce que la distribution soit plutôt oblique à droite et étalée à gauche (on parle alors **d'asymétrie négative**).

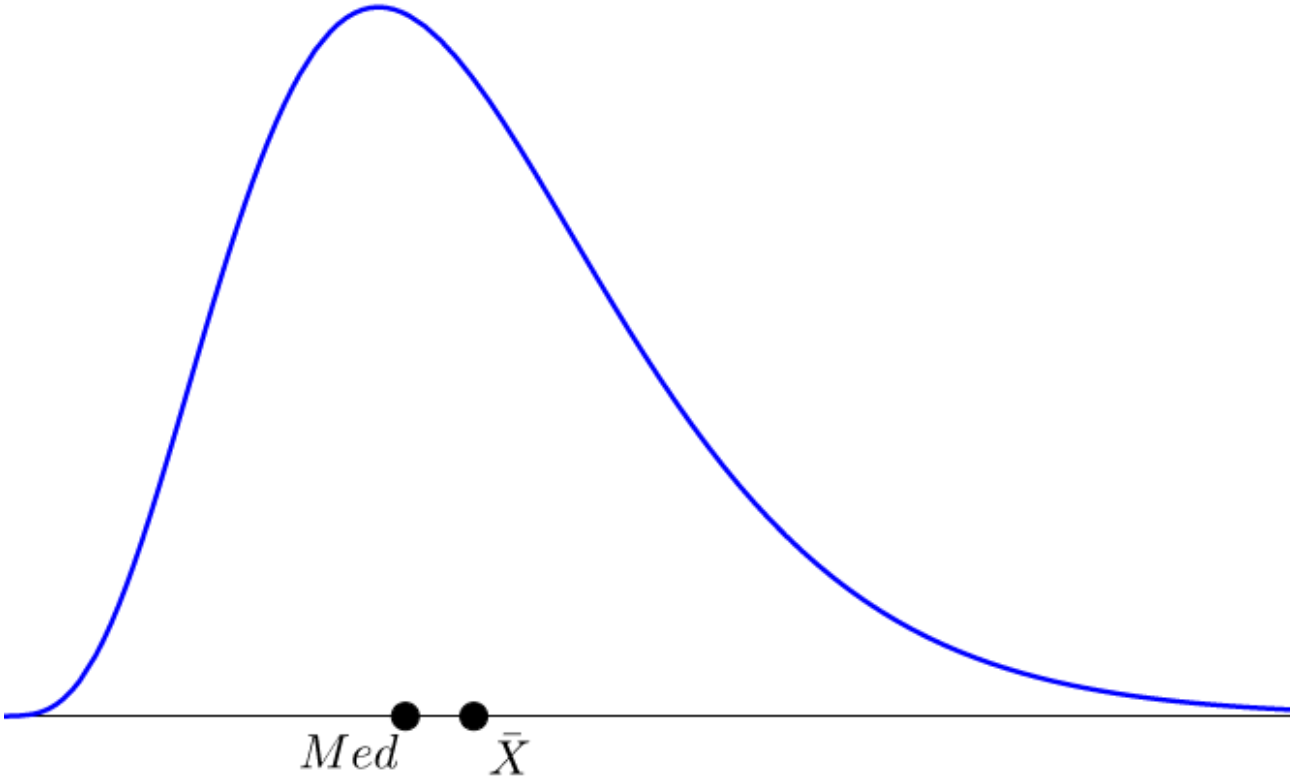


FIGURE 5. – Exemple d'asymétrie positive

Pour comprendre, le graphe ci-dessus représente une distribution théorique (loi du χ^2) qui présente une asymétrie positive : on a l'inégalité $Med < \overline{X}$, que l'on remarque bien sur le graphique puisque la distribution est étalée vers la droite.

5.0.2.1. Coefficient de Yule Introduisons le coefficient de forme le plus intuitif, le **coefficient de Yule** :

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

Étudions cette quantité. On sait déjà, d'après les quantiles, que $Q1 \leq Med \leq Q3$. Ainsi, la quantité au dénominateur est toujours positive. Maintenant, regardons le numérateur : le cas où celui-ci s'annule est atteint lorsque $(Q3 - Med) - (Med - Q1) = 0$ soit

$$Q3 - Med = Med - Q1 \Rightarrow Med = \frac{Q1 + Q3}{2}$$

Or, lorsque Med vaut exactement la moitié de la somme des premier et troisième quartiles, cela signifie qu'il y a autant d'observations en-dessous de $Q1$ qu'au-dessus de $Q3$. La distribution est alors **parfaitement symétrique**. Maintenant, dans le cas où $(Q3 - Med) - (Med - Q1) < 0$ alors

$$Q3 - Med < Med - Q1 \Rightarrow Med > \frac{Q1 + Q3}{2}$$

En particulier, cela signifie que les individus sont plus concentrés vers les fortes valeurs, d'où la distribution est **oblique à droite**, étalée à gauche. Si l'on récapitule :

— Si $S_Y = 0$, alors la distribution est **symétrique**.
— Si $S_Y < 0$, alors la distribution est **oblique à droite**, étalée à gauche.
— Si $S_Y > 0$, alors la distribution est **oblique à gauche**, étalée à droite.

5.0.2.2. Coefficient de Fisher d'asymétrie

Le coefficient de Fisher d'asymétrie est fonction du moment centré d'ordre 3 μ_3 et de l'écart-type σ :

$$\gamma_1 = \frac{\mu_3}{\sqrt{\sigma^3}}$$

Étudions le cas où $\gamma_1 = 0$. Pour plus de facilité, on traite uniquement le cas continu non agrégé et rappelons que :

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^3$$

Résoudre $\gamma_1 = 0$ consiste en réalité à résoudre $\mu_3 = 0$ mais on sait que :

$$\sum_{i=1}^{n}(x_i - \overline{X})^3 = \sum_{i=1}^{n}(x_{\varphi(i)} - \overline{X})^3$$

Où φ est toujours la bijection telle que la suite $(x_{\varphi(n)})$ soit croissante. Posons $q \in \{1, ..., n\}$. On a donc :

$$\sum_{i=1}^{n}(x_{\varphi(i)} - \overline{X})^3 = \sum_{i=1}^{q}(x_{\varphi(i)} - \overline{X})^3 + \sum_{i=q+1}^{n}(x_{\varphi(i)} - \overline{X})^3$$

Puis, par identification :

$$\mu_3 = 0 \Rightarrow \frac{1}{n}\left(\sum_{i=1}^{q}(x_{\varphi(i)} - \overline{X})^3 + \sum_{i=q+1}^{n}(x_{\varphi(i)} - \overline{X})^3\right) = 0 \Rightarrow \sum_{i=1}^{q}(x_{\varphi(i)} - \overline{X})^3 = -\sum_{i=q+1}^{n}(x_{\varphi(i)} - \overline{X})^3$$

Mais cette dernière égalité est valable **pour tout** $1 \leq q \leq n$, c'est-à-dire que l'écart par rapport à la moyenne portée par les q premiers individus et l'opposée de celle portée par les $n-q$ autres individus. En balayant de 1 jusqu'à n, puisque cette égalité est conservée, cela traduit une **parfaite symétrie** de la distribution. Par le même raisonnement, si $\gamma_1 < 0$, on aura :

$$\sum_{i=1}^{q}(x_{\varphi(i)} - \overline{X})^3 < -\sum_{i=q+1}^{n}(x_{\varphi(i)} - \overline{X})^3$$

Et la distribution sera **oblique à droite**, étalée à gauche. Allez, essayons de synthétiser tout ceci :

— Si $\gamma_1 = 0$, alors la distribution est **symétrique**.
— Si $\gamma_1 < 0$, alors la distribution est **oblique à droite**, étalée à gauche.
— Si $\gamma_1 > 0$, alors la distribution est **oblique à gauche**, étalée à droite.

Pour apprécier la symétrie d'une distribution, on utilise la loi normale : c'est la loi de référence en théorie des probabilités, et l'on a souvent recours à la loi normale dite **centrée réduite**, d'espérance nulle (centrée) et de variance 1 (réduite). Il est souvent intéressant de comparer notre distribution par rapport à cette loi, puisque cette dernière est parfaitement symétrique. Pour information, sa densité (i.e. sa loi de probabilité) en tout point x de $\mathbb{R}$ est donné par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Où μ est l'espérance et σ^2 la variance. Par exemple, on dessine, en bleue, une distribution et en rouge, la loi normale centrée réduite.

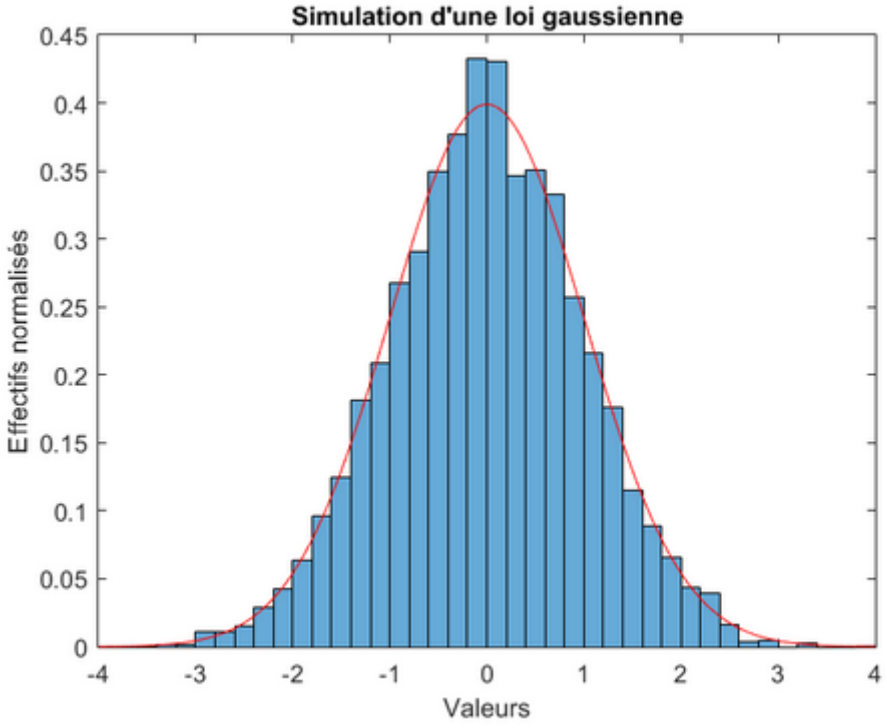
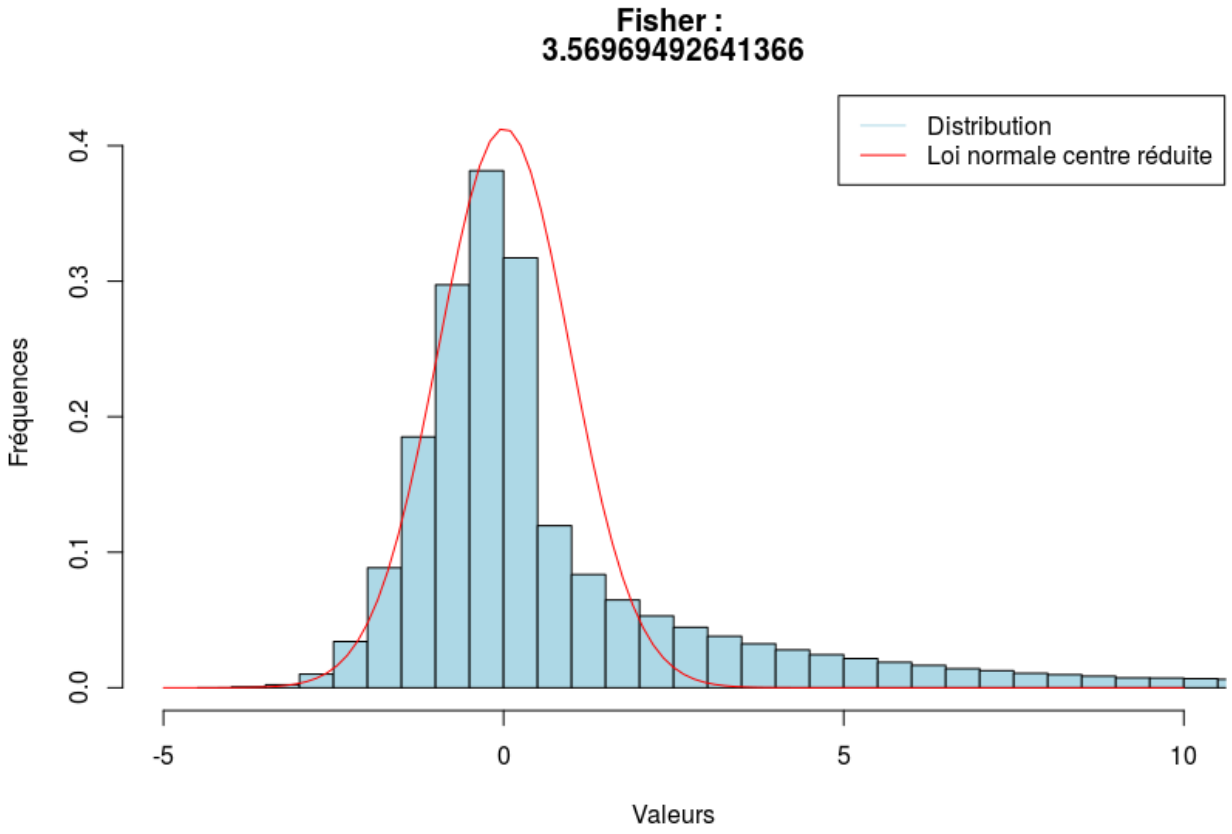
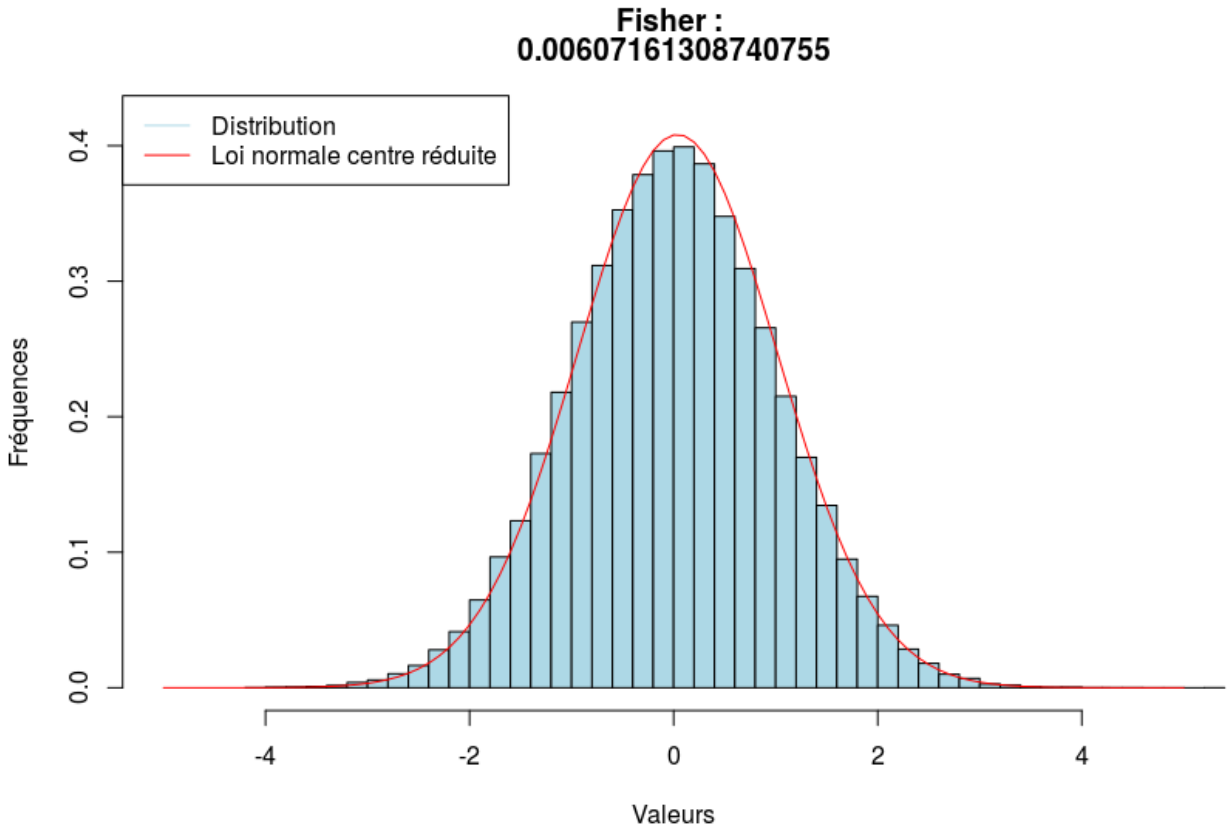
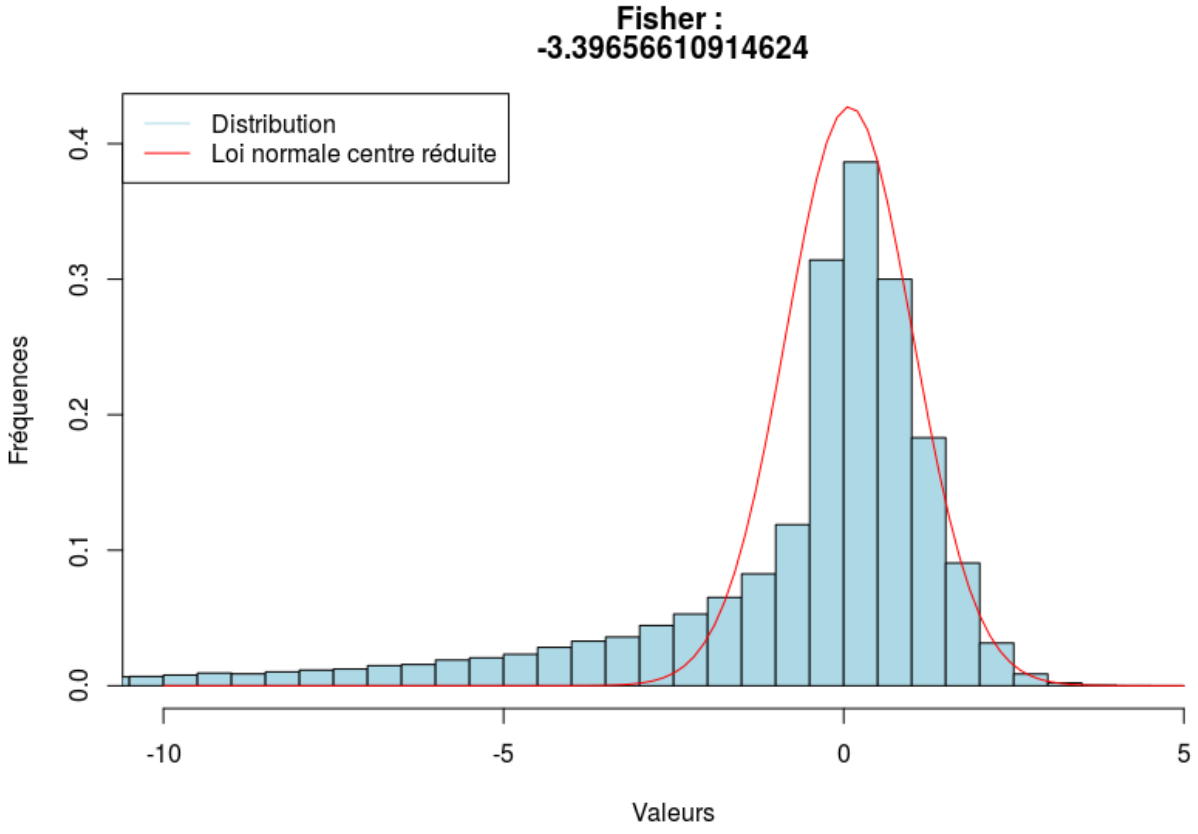


FIGURE 5. – En rouge, la loi normale centrée réduite

Ci-dessous, trois exemples visuels afin de voir le coefficient de Fisher en action.




Fisher :
0.00607161308740755
Distribution
Loi normale centre réduite
Fréquences
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Valeurs


Fisher :
-3.39656610914624
Distribution
Loi normale centre réduite
Fréquences
0.0
0.1
0.2
0.3
0.4
-10
-5
0
5
Valeurs

5.0.3. Mesure d'aplatissement avec les kurtosis

À présent, on souhaite également s'intéresser à l'aplatissement de la distribution, soit la concentration relative des observations autour de la valeur centrale.

? Comment mesure-t-on objectivement l'aplatissement d'une distribution ?

Il a fallu, à un moment donné, définir mathématiquement ce que signifie *aplati*. En effet, tout est relatif : on peut trouver une distribution peu aplatie alors qu'en réalité, elle l'est beaucoup. Pour ce faire, on réalise une **hypothèse de normalité**, on considère que l'on peut estimer la distribution par une loi normale (je vous l'avais dit qu'elle était souvent utilisée).

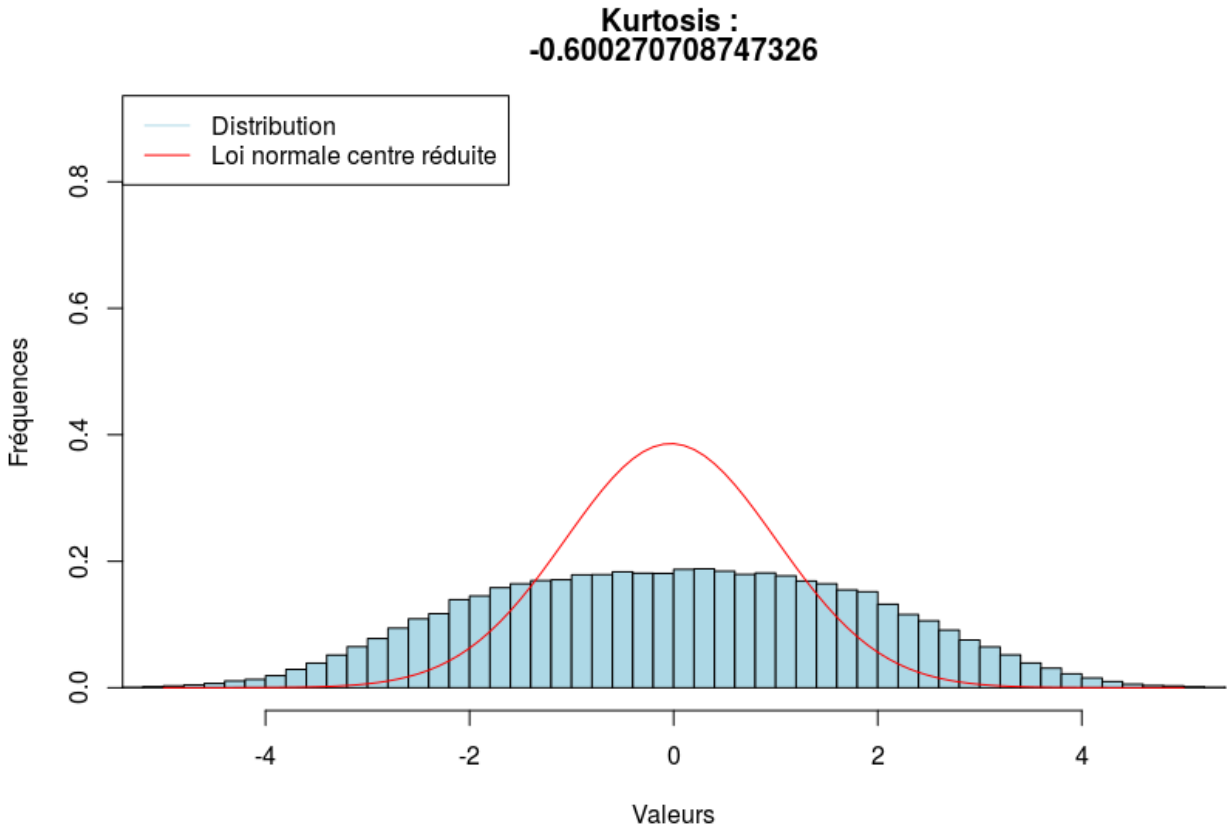
On étudie alors la manière dont la distribution est répartie en fonction de l'hypothèse de normalité. Pour cela, les statisticiens adorent utiliser des mots savants à la place de *plat* ou *concentré* . On dira qu'une distribution est

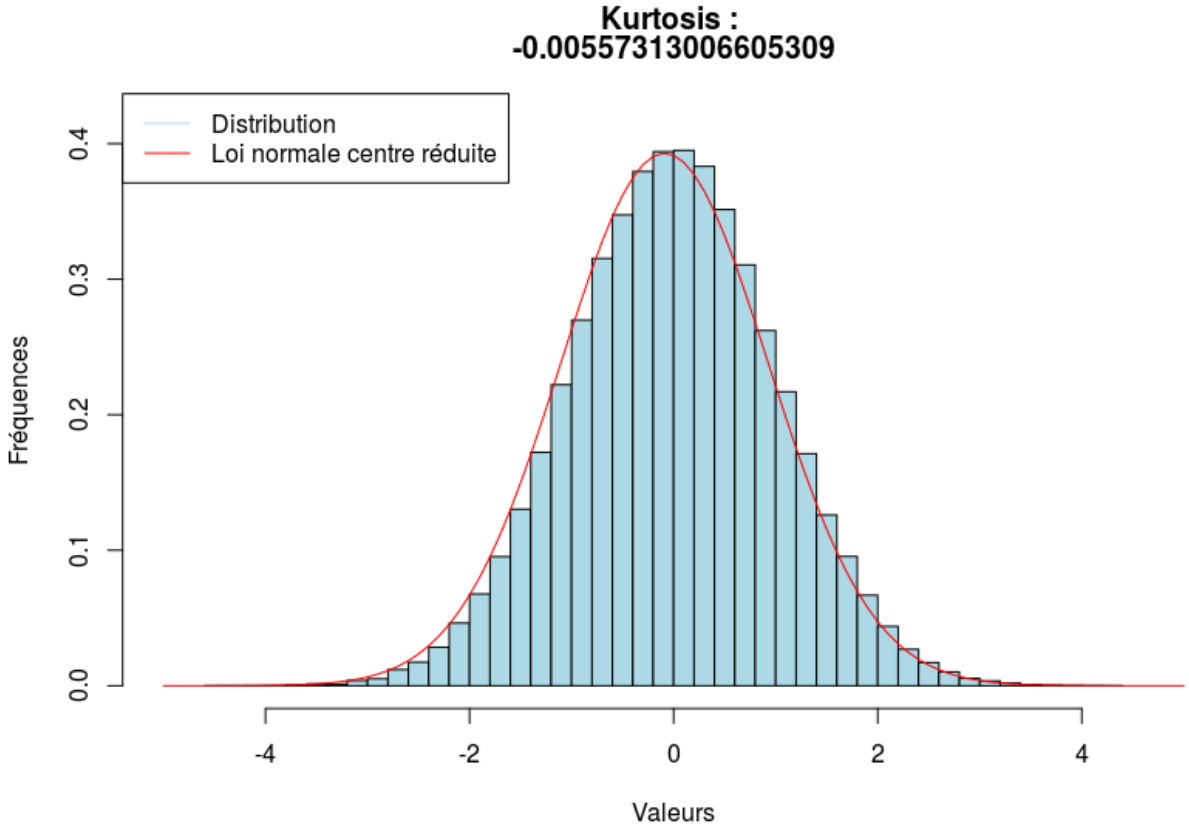
- **Mésocurtique** si les observations sont aussi concentrées que sous l'hypothèse de normalité.
- **Platicurtique** si les observations sont moins concentrées que sous l'hypothèse de normalité.
- **Leptocurtique** si les observations sont plus concentrées que sous l'hypothèse de normalité.

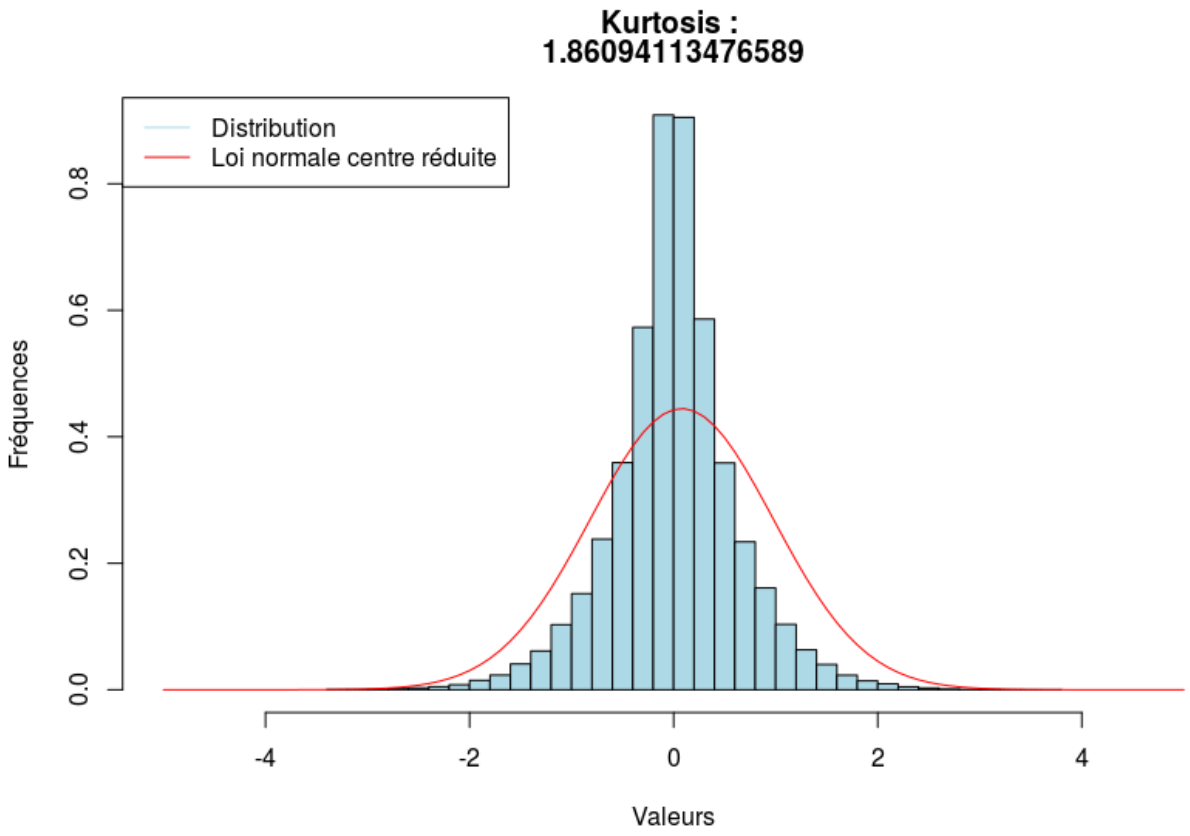
On appelle **kurtosis non normalisé** la quantité β_2 comme étant le rapport entre le moment centré d'ordre 4 μ_4 et la variance élevée au carré :

$$\beta_2 = \frac{\mu_4}{v^2} = \frac{\mu_4}{\mu_2^2}$$

Par exemple, ci-dessous, on donne trois distributions, la première étant platicurtique, la seconde mésocurtique et la dernière, leptocurtique.


Kurtosis :
-0.600270708747326
Distribution
Loi normale centre réduite
Fréquences
0.0
0.2
0.4
0.6
0.8
-4
-2
0
2
4
Valeurs


Kurtosis :
-0.00557313006605309
Distribution
Loi normale centre réduite
Fréquences
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Valeurs



Sachez qu'en règle général, on utilise plutôt le kurtosis normalisé. Je ne rentrerais pas dans les détails, puisque les calculs font intervenir les cumulants .

6. Mesure de concentration

Nous avons presque terminé notre panel de mesures. Seulement je souhaitais vous faire découvrir une dernière catégorie de mesures, appelées **mesure de concentration**, qui sont souvent utiles notamment lorsqu'il s'agit de comparer des richesses au sein d'un échantillon. Seulement, pour comparer cette *richesse*, la fréquence cumulée à elle seule ne suffit plus : il nous faut introduire la fréquence cumulée $\tilde{F}$ relative aux $n_i x_i$. En effet, l'objectif sera de comparer F_i et $\tilde{F}_i$. Plus l'écart sera grand, plus la richesse sera répartie de manière inégale.

Lorsque l'on emploie le terme *richesse*, il peut s'agir de richesse en terme d'argent mais pas uniquement ! Comme nous le verrons dans l'exercice d'application, une richesse peut également désigner une surface d'exploitation ou encore une part.

Il faut garder en tête que richesse signifie plutôt "ce que possède" un individu. Définissons ainsi cette fréquence cumulée $\tilde{F}$.

Discrète	Continue agrégée	Continue non agrégée
$\tilde{F}_i = \sum_{k=1}^{i} \frac{f_k a_k}{M}$ avec $M = \sum_{i=1}^{k} f_i a_i$	$\tilde{F}_i = \sum_{k=1}^{i} \frac{f_k c_k}{M}$ avec $M = \sum_{i=1}^{k} f_i c_i$	$\tilde{F}_i = \sum_{k=1}^{i} \frac{n_k x_k}{M}$ avec $M = \sum_{i=1}^{n} n_i x_i$

Ici, notre masse M représente donc la somme de toutes les valeurs prises par la variable pondérées par les effectifs. N'oubliez pas que dans les cas discret et continue agrégée, on dispose de k classes alors que dans le cas continu non agrégé, nous avons n individus.

6.0.0.1. L'indice de Gini et courbe de Lorenz

L'indice de Gini G est défini comme **deux fois** la surface comprise entre la droite d'équi-répartition et la courbe de concentration.

i Je traite uniquement le cas d'une variable continu non agrégée, mais le résultat est similaire pour les deux autres cas : on fera attention à considérer k et non plus n.

Pour ce faire, on représente la succession de points $(F_i; \tilde{F}_i)$ par une fonction f continue de $[0, 1]$ dans $[0, 1]$, affine par morceaux et telle que, pour tout $0 \leq i \leq n, f(F_i) = \tilde{F}_i$. On appelle alors f la **courbe de Lorenz**. On veut donc calculer :

$$G = 2 \int_0^1 (x - f(x))dx = 1 - 2 \int_0^1 f(x)dx$$

Car $x \geq f(x), \forall x \in [0, 1]$. C'est ainsi que l'on subdivise l'intervalle $[0, 1]$ en n subdivisions :

$$0 = F_0 \leq F_1 \leq F_2 \leq ... \leq F_{n-1} \leq F_n = 1$$

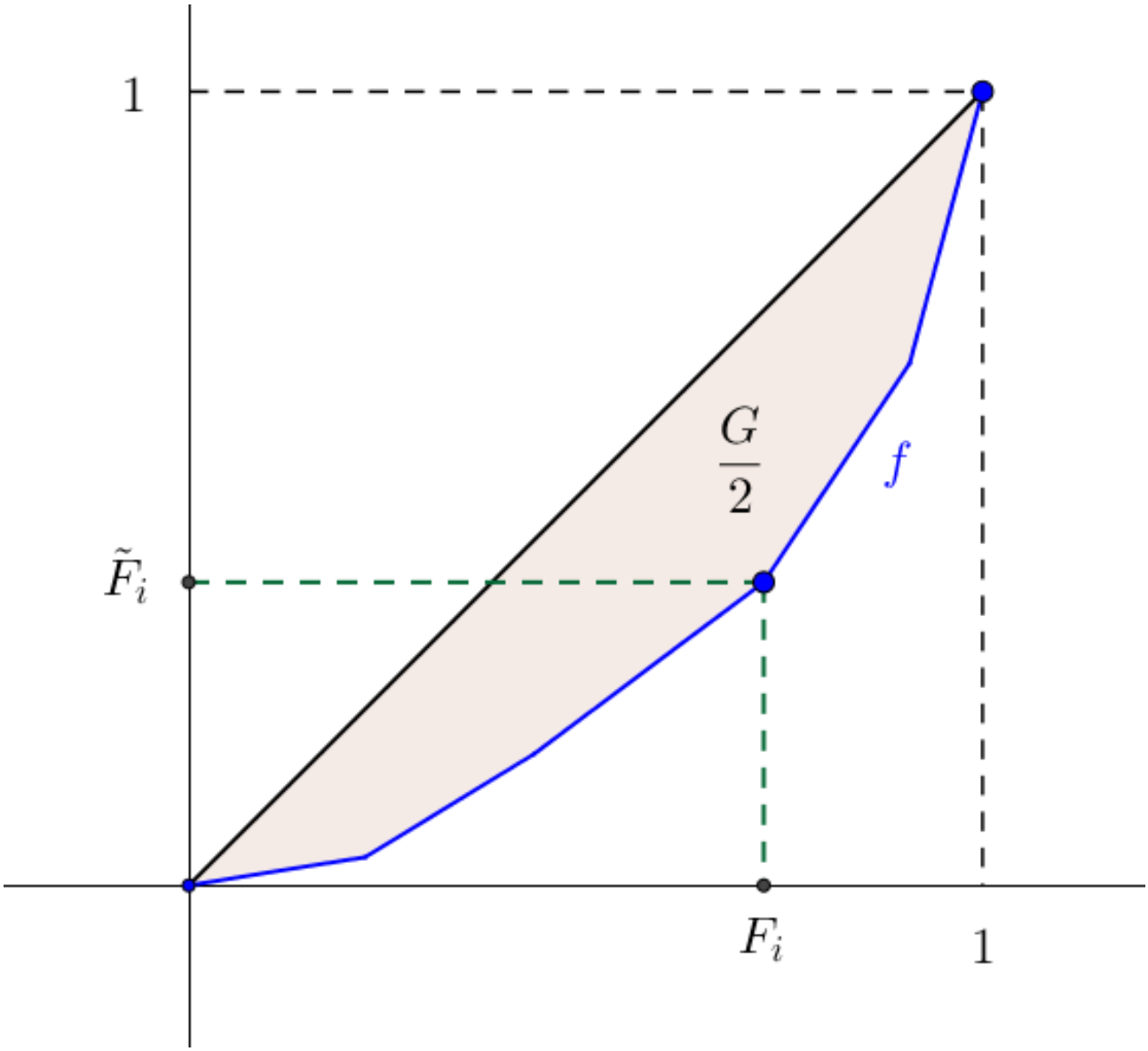


FIGURE 6. – Fréquence cumulée et indice de Gini

L'idée est, dans un premier temps, de déterminer l'intégrale de f sur $[0, 1]$ par la méthode des trapèzes pour ensuite en déduire l'indice de Gini. Pour tout $0 \leq i \leq n$, la surface comprise entre la courbe f et délimitée par les droites d'abscisse $x = F_i$ et $x = F_{i+1}$ est :

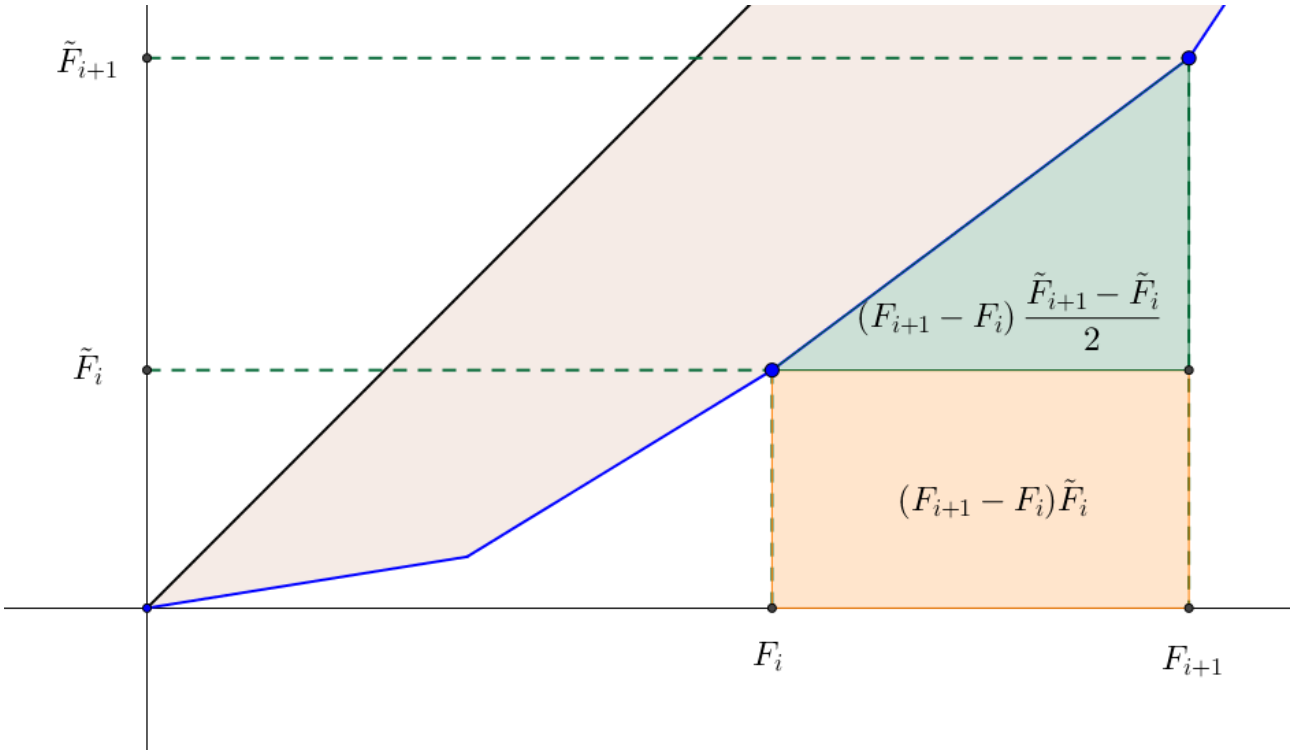


FIGURE 6. – Méthode des trapèzes

$$\int_{F_i}^{F_{i+1}} f(x)dx = (F_{i+1} - F_i)\left(f(F_i) + \frac{f(F_{i+1}) - f(F_i)}{2}\right) = (F_{i+1} - F_i)\left(\frac{f(F_{i+1}) + f(F_i)}{2}\right)$$

À noter qu'ici, la méthode des trapèzes fournit la valeur exacte puisque f est affine par morceaux. Mais $F_{i+1} - F_i = \frac{n_{i+1}}{n}$ et $f(F_i) = \tilde{F}_i$, pour tout $0 \leq i \leq n$ d'où

$$\int_{F_i}^{F_{i+1}} f(x)dx = \frac{n_{i+1}}{2n}(\tilde{F}_{i+1} + \tilde{F}_i)$$

Et donc, la surface totale est vue comme la somme des aires :

$$\int_0^1 f(x)dx = \sum_{i=0}^{n-1} \int_{F_i}^{F_{i+1}} f(x)dx = \sum_{i=0}^{n-1} \frac{n_{i+1}}{2n}(\tilde{F}_{i+1} + \tilde{F}_i)$$

Or l'aire sous la courbe d'équi-répartition vaut $\frac{1}{2}$, l'indice de Gini étant défini comme le double de cette différence avec l'aire sous la courbe des fréquences cumulées $\tilde{F}_i$:

$$G = 1 - \frac{1}{n}\sum_{i=0}^{n-1} n_{i+1}(\tilde{F}_{i+1} + \tilde{F}_i) = 1 - \frac{1}{n}\sum_{i=1}^{n} n_i(\tilde{F}_i + \tilde{F}_{i-1})$$

Avec $F_0 = \tilde{F}_0 = 0$. Pour résumer, l'indice de Gini s'exprime ainsi de cette manière :

Discrète	Continue agrégée	Continue non agrégée
$G = 1 - \sum_{i=1}^{k} f_i(\tilde{F}_i + \tilde{F}_{i-1})$	$G = 1 - \sum_{i=1}^{k} f_i(\tilde{F}_i + \tilde{F}_{i-1})$	$G = 1 - \frac{1}{n}\sum_{i=1}^{n} n_i(\tilde{F}_i + \tilde{F}_{i-1})$

Pour interpréter ce coefficient, il suffit de regarder sa valeur qui se situe entre 0 et 1 :

— Si $G = 0$, nous sommes face à un cas où la distribution des richesses et parfaitement égalitaire : chaque individu possède la même part de richesse. Dans une entreprise, cela signifie que tous les employés ont exactement le même salaire.
— Si $G = 1$, il y a alors une parfaite inégalité : un seul individu possède exactement la richesse de tous les autres individus. Par exemple, dans une entreprise, cela signifierai qu'un employé posséderait exactement les salaires de tous les autres employés, et que ces derniers ne recevraient rien (bien entendu, il s'agit d'un exemple abstrait, ce cas est figure est heureusement impossible dans la réalité).
— Si $0 < G < 1$, il faut alors interpoler par rapport aux deux valeurs précédentes. Plus G est proche de 0, plus la distribution des richesses est égalitaire. À l'inverse, plus G est proche de 1, plus la distribution des richesses est inégale.

7. Représentations graphiques

Dans cette dernière partie, nous allons explorer différentes méthodes pour représenter graphiquement une distribution. Bien sûr, je ne vous présente pas toutes les représentations existantes, il me faudrait un tutoriel en entier pour vous les présenter. En revanche, je vous propose de découvrir les représentations graphiques les plus utilisées, et surtout les plus commodes.

7.0.1. Diagramme en bâtons

Ce diagramme est sans doute celui que vous utilisez le plus souvent. En plus d'être facile à comprendre, il permet de déterminer rapidement la forme et la répartition d'une distribution. Pour chaque modalité a_i (ou chaque classe $[a_i, a_{i+1}[$), on fait correspondre un rectangle de largeur constante et de hauteur proportionnelle à la fréquence :

$$h_i = c \times f_i$$

Où $c \in \mathbb{R}_+^*$ est une constante. Par exemple, si l'on souhaite afficher les effectifs plutôt que les fréquences, on choisira $c = n$.

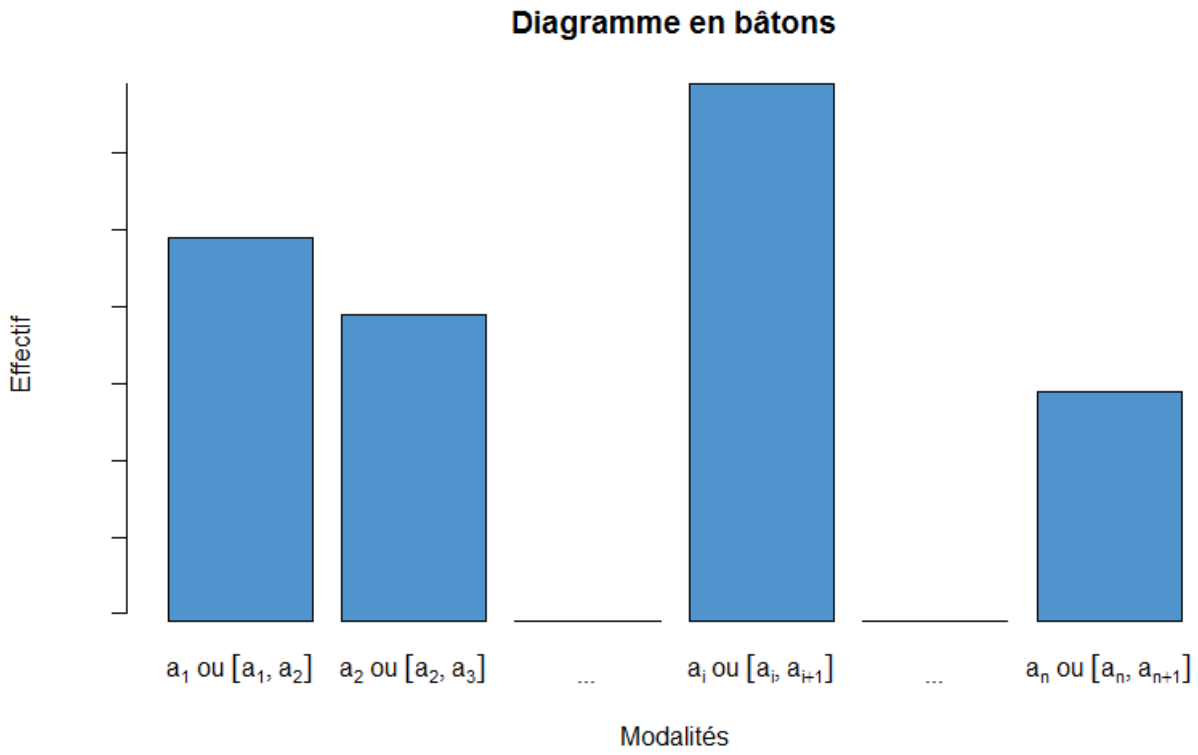


FIGURE 7. – Diagramme en bâtons

7.0.2. Diagramme en secteurs

Aussi appelé *diagramme circulaire*, il joue le même rôle que le précédent, à la différence où cette fois-ci, les fréquences ne sont plus affichées selon des rectangles mais par des secteurs d'angles proportionnels à la fréquence. Ainsi, chaque secteur a pour angle (en radians) :

$$\alpha_i = 2\pi f_i$$

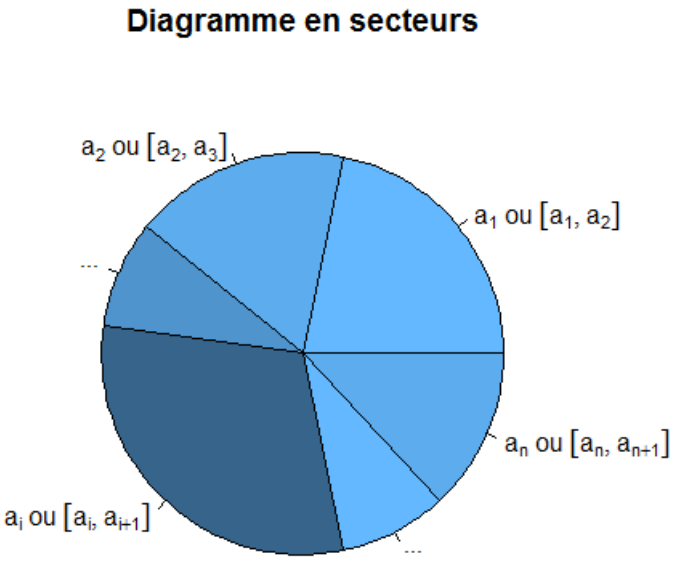


FIGURE 7. – Diagramme en secteurs

Dans les faits, je vous déconseille d'utiliser cette représentation si les données "ne s'y prêtent pas".

Par exemple, je dispose d'un ensemble de données, puis je décide d'afficher un diagramme en secteurs pour visualiser les proportions. J'obtiens ce diagramme

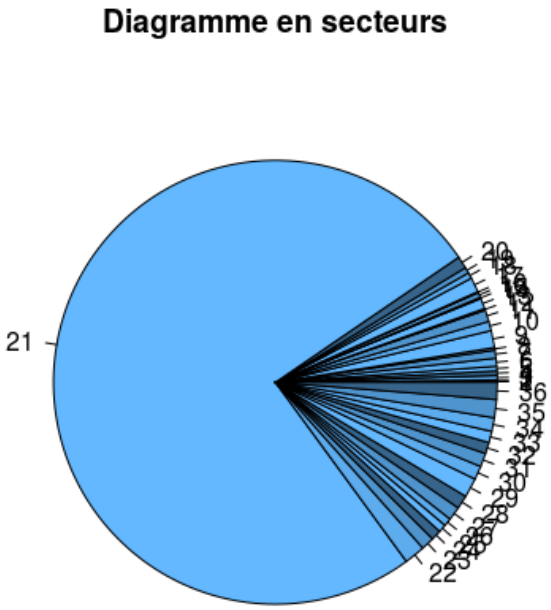


FIGURE 7. – Mauvais choix de représentation

Comme vous pouvez le remarquer, cette représentation est **très mauvaise** dans ce cas : près de 75 % du disque ne représente qu'une seule fréquence, alors que toutes les autres sont représentées sur 25 % du disque. Comment comparer toutes les autres fréquences ? C'est là qu'interviennent les limitations de cette représentation. Ainsi, n'utilisez le diagramme en secteurs que si

— Il n'y a pas beaucoup de classes ou fréquences à considérer
— Les fréquences sont globalement bien réparties (on ne veut pas qu'une seule classe monopolise tout le diagramme)

7.0.3. Histogramme

On utilise souvent l'histogramme pour des variables continues agrégées. Le principe de construction ressemble à celui du diagramme en bâtons, mais la différence réside dans le fait que la largeur des rectangles n'est plus forcément constante : elle dépend de la longueur des classes. En effet, chaque rectangle aura pour largeur L_i et pour hauteur h_i :

$$L_i = a_{i+1} - a_i \qquad h_i = c \times \frac{f_i}{L_i}$$

De sorte que la surface $S_i = c \times f_i$ soit proportionnelle à la fréquence, avec $c \in \mathbb{R}_+^*$.

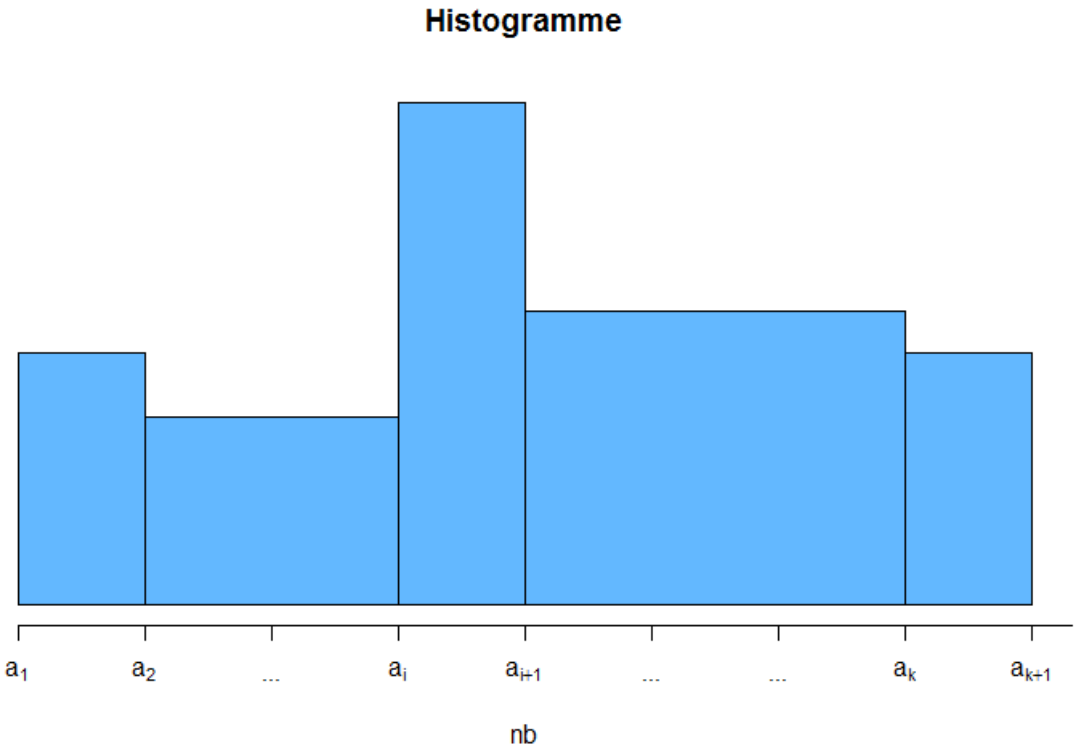


FIGURE 7. – Histograme

i Il n'existe pas de méthodes absolues pour définir le nombre de classes k à utiliser. Néanmoins, on utilise certains critères comme la règle de Sturges permettant de déterminer le nombre de classes k à utiliser pour n individus :

$$k = \lfloor 1 + \log_2 n \rfloor$$

7.0.4. Boîtes à moustache

Derrière ce petit nom sympathique se cache une représentation très intéressante lorsqu'il s'agit de détecter les **valeurs aberrantes**. L'idée de construction est la suivante :

- On construit une boîte entre $Q1$ et $Q3$ (donc de longueur EIQ).
- On détermine à présent *les moustaches* : l'extrémité de la moustache inférieure A est la plus petite valeur x_i telle que $x_i \geq Q1 - 1,5EIQ$.
- L'extrémité de la moustache supérieure B est la plus grande valeur x_i telle que $x_i \leq Q3 + 1,5EIQ$.

$$A = \min\{x_i : x_i \geq Q1 - 1,5EIQ\} \qquad B = \max\{x_i : x_i \leq Q3 + 1,5EIQ\}$$

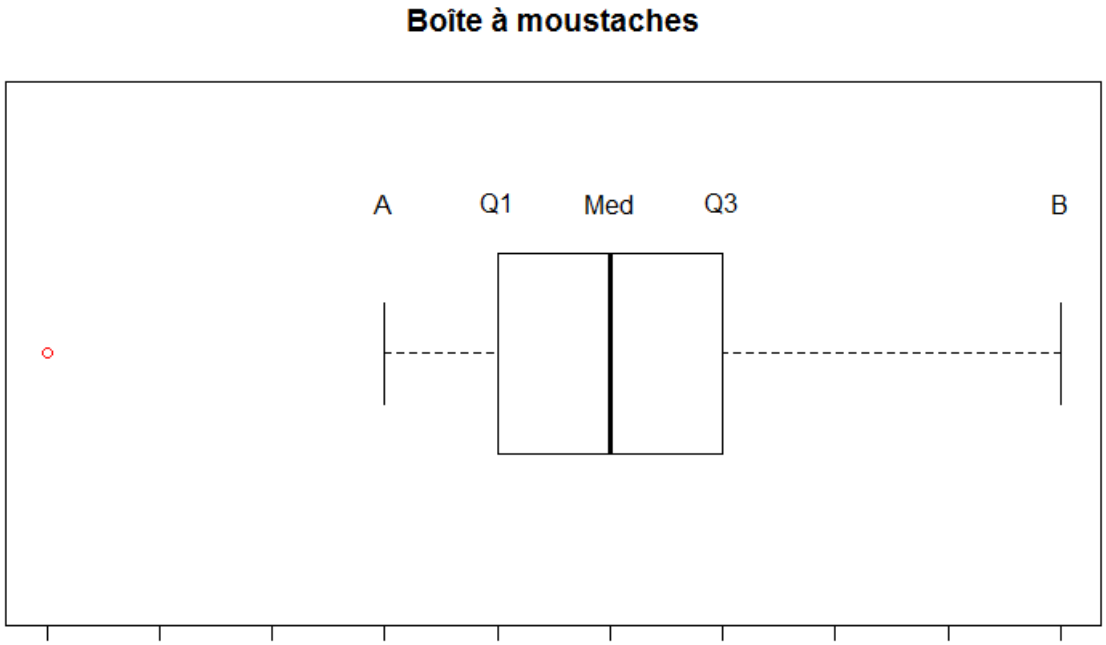


FIGURE 7. – Boîte à moustaches

Comme vous pouvez le constater sur cet exemple, le point rouge est un point aberrant, car il est situé en dehors du rectangle et des moustaches sous l'hypothèse de normalité.

? Pourquoi utilisons-nous $1,5$ dans les extrémités des moustaches ?

Cela vient du fait que le modèle est basé sur la distribution d'une loi normale. Si une variable suit une loi normale, alors l'intervalle $[A, B]$ devrait contenir $99,3$ % des observations, c'est-à-dire que l'on devrait trouver 0.7 % d'observations en dehors de l'intervalle $[A, B]$, que l'on considère alors comme des valeurs aberrantes. Pour le créateur de cette représentation, John Tukey , $1,5$ était un bon compromis pour observer assez de points aberrants, sans pour autant en être débordé.

8. Exercice d'application

Enfin, toute la théorie est terminée ! Afin de mettre en pratique tout ce que vous avez appris au cours de ce tutoriel, je vous propose ainsi un exercice qui va faire appel à la plupart des notions introduites précédemment. Bien sûr, n'hésitez pas à remonter si vous avez oublié une formule, je vous y encourage

8.0.1. Le sujet

Dans une certaine région, on donne le nombre d'exploitations agricoles par surface agricole utile , exprimée en hectares.

Surface agricole utile (SAU)	< 5	**5 à 10**	**10 à 20**	**20 à 50**	**50 à 100**	> 100
Nombre d'exploitations	7800	9600	13200	20400	7200	1800

1. De quel type de variable s'agit-il ? Représentez graphiquement la distribution et calculez le milieu des classes c_i.
2. Déterminez la moyenne, la variance et l'écart-type de l'échantillon.
3. Calculez la médiane et l'écart inter-quartile. D'après ces deux informations, la distribution semble-t-elle symétrique ?
4. Vérifiez votre réponse avec le coefficient de Yule.
5. Pourrait-on dire que la distribution est aussi concentrée que sous l'hypothèse de normalité ?
6. Déterminez l'indice de Gini. Que peut-on dire de la répartition des richesses dans cet échantillon ?

8.0.2. Correction

En plus d'effectuer les calculs, je vous donne le code R associé pour que vous puissiez vous même essayer de développer un programme sous R.

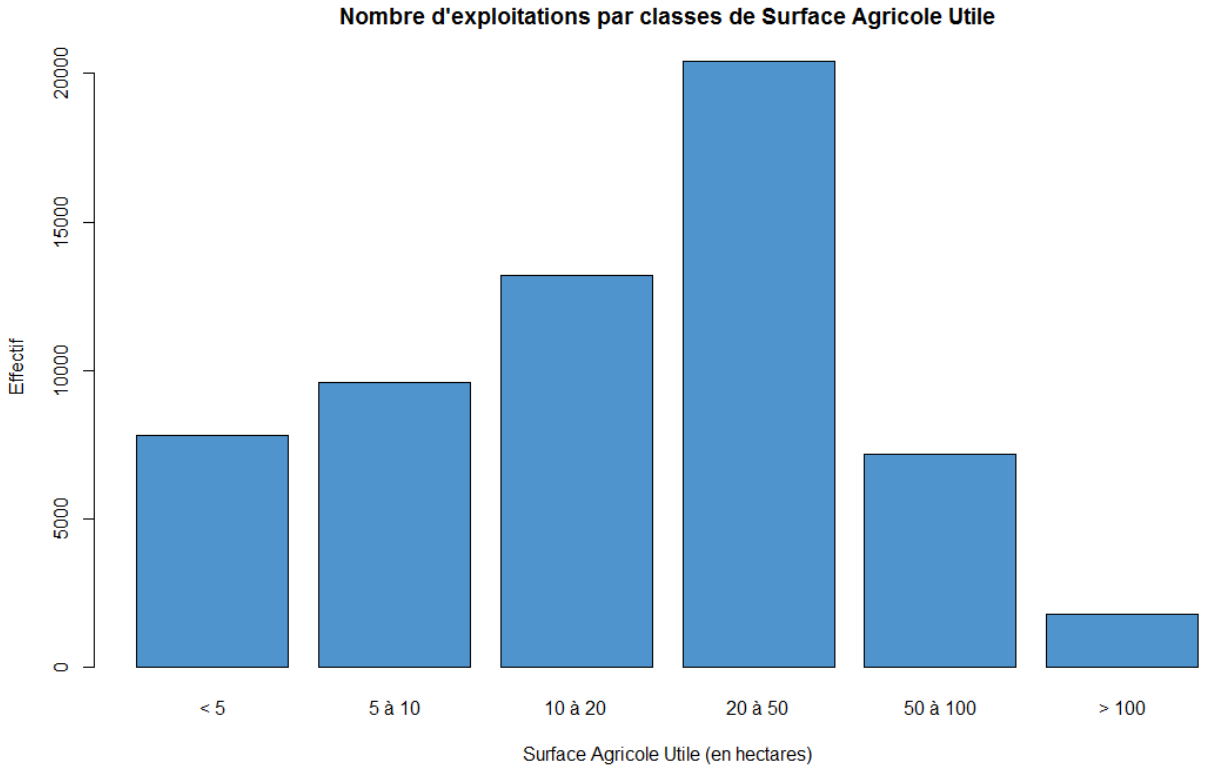
1) Ici, notre variable est la surface agricole utile (SAU). Il s'agit donc d'une variable **quantitative** (ici, un nombre réel) et qui est également agrégée, car nous disposons des classes suivantes :

$$[0, 5[, [5, 10[, [10, 20[, [20, 50[, [50, 100[, [100, 200[$$

À noter que pour la dernière classe, j'ai moi-même utilisé la valeur 200 puisque la dernière classe n'est pas correctement définie. On veillera ainsi à prendre une borne cohérente. En calcule alors le milieu de chacune de ces classes :

$$c_1 = 2,5 \quad c_2 = 7,5 \quad c_3 = 15 \quad c_4 = 35 \quad c_5 = 75 \quad c_6 = 150$$

On représente alors la distribution :




```r
nb<-c(7800,9600,13200,20400,7200,1800) # Le nombre d'exploitations par classes
bornes<-c(0,5,10,20,50,100,200) # Les bornes de nos classes
ci<-(bornes[-7]+bornes[-1])/2 # Le milieu des classes
ci
barplot(nb,main="Nombre d'exploitations par classes de Surface Agricole Utile",names.arg=c("< 5","5 à 10","10 à 20","20 à 50","50 à 100","> 100"),col="steelblue3",xlab="Surface Agricole Utile (en hectares)",ylab="Effectif")
```

2) Notons m la moyenne de l'échantillon et v sa variance. Ici, notre effectif total est $n = 7800 + 9600 + 13200 + 20400 + 7200 + 1800 = 60000$, ce qui nous permet de calculer les fréquences f_i :

$$f_1 = \frac{7800}{60000} = 0,13 \qquad f_2 = 0,16 \qquad f_3 = 0,22 \qquad f_4 = 0,34 \qquad f_5 = 0,12 \qquad f_6 = 0,03$$

$$m = \sum_{i=1}^{6} f_i c_i = 2,5 \times 0,13 + 7,5 \times 0,16 + 15 \times 0,22 + 35 \times 0,34 + 75 \times 0,12 + 150 \times 0,03 = 30,225$$

$$v = \sum_{i=1}^{6} f_i (c_i - m)^2 = (2,5 - 30,225)^2 \times 0,13 + (7,5 - 30,225)^2 \times 0,16 + (15 - 30,225)^2 \times 0,22 + (35 - 30,225)^2 \times 0,$$

En notant σ l'écart-type, on calcule alors $\sigma = \sqrt{v} = 30.20367$. Pour récapituler :

$$m = 30,225 \qquad v = 912,2619 \qquad \sigma = 30,20367$$

```
n<-sum(nb) # Le nombre total d'individus
f<-nb/n # Les fréquences
f
m<-sum(ci*nb)/n # La moyenne
m
v<-sum(f*(ci-m)^2) # La variance
v
sigma<-sqrt(v) # l'écart-type
sigma
```

3) Notons F la fréquence cumulée alors :

$$F_1 = 0,13 \qquad F_2 = 0,29 \qquad F_3 = 0,51 \qquad F_4 = 0,85 \qquad F_5 = 0,97 \qquad F_6 = 1$$

On remarque alors que $10 < Med < 20$ car $F_2 < 0,5 < F_3$. Pour cela, on applique la formule d'interpolation linéaire :

$$Med = 10 + \frac{0,5 - 0,29}{0,51 - 0,29}(20 - 10) = 10\left(1 + \frac{0,5 - 0,29}{0,51 - 0,29}\right) \approx 19,55$$

En notant $Q1$ et $Q3$ le premier et le troisième quartile, on remarque que $5 < Q1 < 10$ et $20 < Q3 < 50$ donc :

$$Q1 = 5 + \frac{0,25 - 0,13}{0,29 - 0,13}(10 - 5) = 8,75 \qquad Q3 = 20 + \frac{0,75 - 0,51}{0,85 - 0,51}(50 - 20) \approx 41,18$$

On en déduit alors l'écart inter-quartile $EIQ = Q3 - Q1 \approx 32,43$. On remarque que $Med < \frac{Q1+Q3}{2} \approx 24,965$, ce qui indique que la distribution serait plutôt **oblique à gauche**, aplatie à droite.

4) On calcule le coefficient de Yule S_Y de la distribution :

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)} = \frac{(41,18 - 19,55) - (19,55 - 8,75)}{(41,18 - 19,55) + (19,55 - 8,75)} \approx 0,33$$

On a $S_Y > 0$, ce qui valide notre hypothèse précédente affirmant que la distribution est **oblique à gauche**. Ce calcul permet de confirmer ce que l'on voit sur le graphique de la distribution.

5) Pour répondre à cette question, il nous faut calculer le kurtosis normalisé de la distribution, que l'on note γ_2. Pour rappel,

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{v^2} - 3$$

On connait déjà la variance, alors calculons le moment centré d'ordre 4 μ_4 :

$$\mu_4 = \sum_{i=1}^{6} f_i(c_i - m)^4 = (2,5-30,225)^4 \times 0,13 + (7,5-30,225)^4 \times 0,16 + (15-30,225)^4 \times 0,22 + (35-30,225)^4 \times 0$$

Et donc

$$\gamma_2 = \frac{6788063}{912.2619^2} - 3 \approx 5.16$$

En conclusion, la distribution est **beaucoup plus concentrée** que sous l'hypothèse de normalité.

```
mu4<-sum(f*(ci-m)^4) # Moment centré d'ordre 4
mu4
kurtosis<-mu4/v^2-3
kurtosis
```

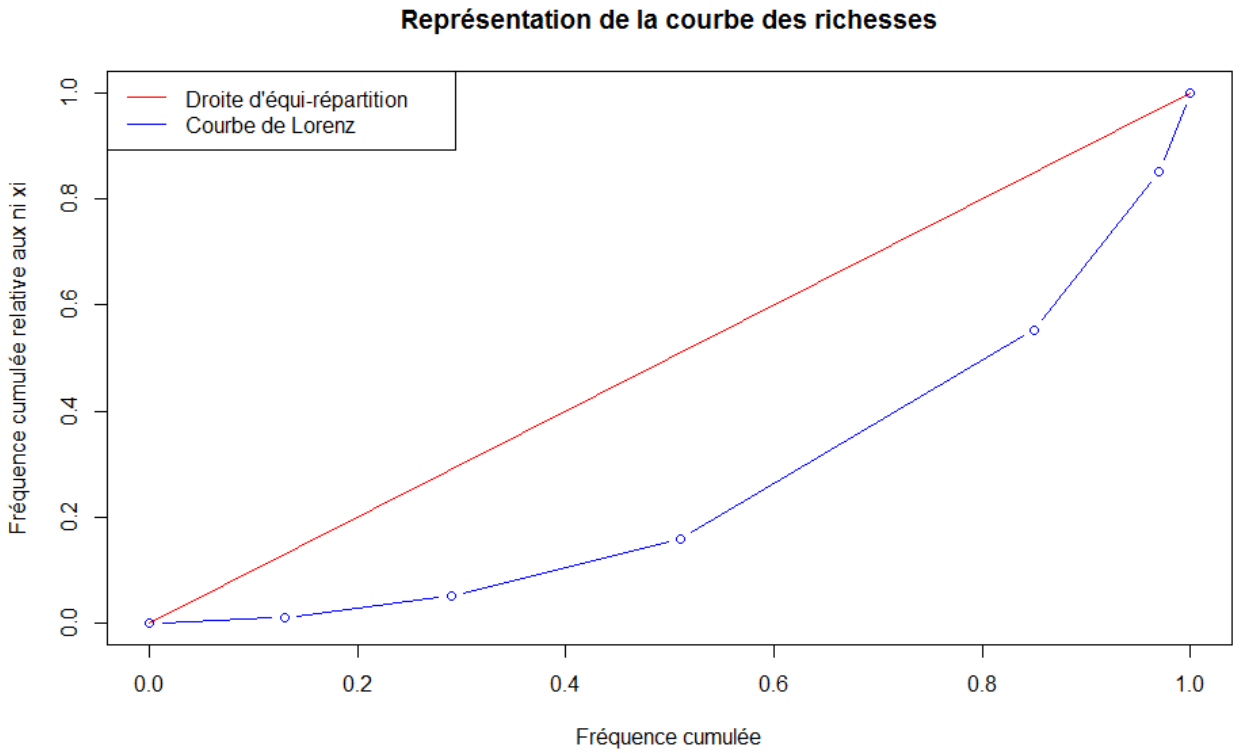
6) On note $\tilde{F}$ la fréquence cumulée relative aux $n_i x_i$ ce qui nous donne :

$$\tilde{F}_1 = 0,01075269 \qquad \tilde{F}_2 = 0,05045492 \qquad \tilde{F}_3 = 0,15963606 \qquad \tilde{F}_4 = 0,55334988 \qquad \tilde{F}_5 = 0,85111663$$

On applique alors la formule pour déterminer le coefficient de Gini :

$$G = 1 - \sum_{i=0}^{5} f_{i+1}(\tilde{F}_{i+1} + \tilde{F}_i) \approx 0,476$$

En somme, cela signifie qu'une certaine partie des exploitations comptabilise, à elle seule, une majeure partie de la SAU. Et cela se remarque très bien sur le graphique de la distribution : seulement une petite quantité d'exploitations dispose plus de 50 hectares de SAU ! Alors d'un grand nombre d'exploitations (51000 exactement) possèdent moins de 50 hectares de SAU. Globalement, on peut conclure qu'il existe une certaine inégalité en terme de SAU. Toutefois, nous ne disposons pas assez d'informations pour parler *réellement* d'une inégalité : on ne connait ni la production de chaque exploitation, ni la localisation, qui sont des facteurs importants. On se contera alors d'émettre l'hypothèse d'une inégalité moyenne selon nos observations. On fournit, ci-dessous, le graphique représentant cette courbe des richesses.




```r
Fc<-cumsum(f) # La fréquence cumulée
Fc
fm<-f*ci/sum(f*ci) # Fréquence relative aux nixi
Fm<-cumsum(fm) # Fréquence cumulée relative aux nixi
Fm
IG<-1-sum(f*(c(0,Fm[-6])+Fm)) # Coefficient de Gini
IG

plot(c(0, Fc), c(0, Fm),type="b",col="blue",xlab="Fréquence
    cumulée",ylab="Fréquence cumulée relative aux ni xi",
    main="Représentation de la courbe des richesses")
lines(c(0, 1), c(0, 1), col="red")
legend("topleft", legend=c("Droite d'équi-répartition", "Courbe de
    Lorenz"), col=c("red", "blue"), lty=1)
```

Félicitations pour être arrivé jusqu'ici . Même si ce tutoriel ne présente pas toutes les notions de la statistique descriptive à une dimension, vous aurez tout de même une solide base pour explorer le merveilleux monde de la statistique. En résumé de ce cours :

- Définition d'un vocabulaire rigoureux régulièrement utilisé en statistique
- Liste des types de variables existantes, même si nous avons uniquement étudié les variables quantitatives
- Beaucoup de mesures ont été présentées : tendance centrale, forme des distributions, dispersion autour de la moyenne, ou encore de richesse.

— Plusieurs représentations des données ont été montrées, et notamment celles qui sont particulièrement utilisées
— Pour finir, un exercice d'application pour clôturer ce cours.

Sachez qu'à présent, de nouvelles voies s'offrent à vous :

— La statistique descriptive à **deux dimensions**, qui s'inscrit dans la continuité de ce tutoriel. Vous l'aurez compris, il s'agit d'étudier non pas une mais **deux** variables à la fois.
— La statistique descriptive **multidimensionnelle** : dès que l'on commence à étudier trois variables ou plus, on fait ce que l'on appelle de la statistique descriptive multidimensionnelle. En particulier, on utilise des méthodes **d'analyse factorielle** (ACP, AFC, ...) et de **classification** pour résumer au mieux l'information de p variables sur n individus.
— Les **séries temporelles**, où l'on étudie les phénomènes qui **dépendent du temps**, et notamment ceux disposant de cycle saisonniers.
— La statistique **inférentielle**, où l'objet des études est **d'effectuer des hypothèses** sur une population à partir d'un échantillon.

Si vous remarquez la moindre erreur ou tout simplement si vous avez une remarque concernant ce tutoriel, n'hésitez pas à me contacter par message.

Contenu masqué

Contenu masqué n°1

X	n	f	F
22	2	0,2	0,2
26	2	0,2	0,4
31	3	0,3	0,7
45	3	0,3	1

Retourner au texte.

Contenu masqué n°2

On cherche à résoudre $Med = F^{-1}(0,5)$. On suppose que F **est affine** sur l'intervalle $[a_i, a_{i+1}]$. On a donc :

$$\forall x \in [a_i, a_{i+1}], F(x) = \alpha x + \beta$$

Avec $\alpha, \beta \in \mathbb{R}$. En particulier, on a :

$$F(a_i) = \alpha a_i + \beta F(a_{i+1}) = \alpha a_{i+1} + \beta$$

$En soustrayant la deuxime ligne la premire$:

Contenu masqué

Il nous reste alors à réinjecter α dans la première ligne :

$$F(a_i) = \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i} a_i + \beta \Rightarrow \beta = F(a_i) - \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i} a_i = \frac{a_{i+1}F(a_i) - a_i F(a_{i+1})}{a_{i+1} - a_i}$$

Et donc :

$$\forall x \in [a_i, a_{i+1}], F(x) = \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i} x + \frac{a_{i+1}F(a_i) - a_i F(a_{i+1})}{a_{i+1} - a_i}$$

Pour résoudre $F^{-1}(0,5)$, il nous reste alors à trouver l'expression de F^{-1} qui est également affine car F l'est et $\alpha \neq 0$. On a donc, pour tout $x \in [F(a_i), F(a_{i+1})]$:

$$F^{-1}(x) = \frac{x - \beta}{\alpha} = \frac{x(a_{i+1} - a_i)}{F(a_{i+1}) - F(a_i)} - \frac{a_{i+1}F(a_i) + a_i F(a_{i+1})}{F(a_{i+1}) - F(a_i)}$$

En ajoutant $a_i F(a_i) - a_i F(a_i)$ au numérateur :

$$F^{-1}(x) = \frac{x(a_{i+1} - a_i)}{F(a_{i+1}) - F(a_i)} - \frac{a_{i+1}F(a_i) + a_i F(a_{i+1}) + a_i F(a_i) - a_i F(a_i)}{F(a_{i+1}) - F(a_i)} = a_i + \frac{x - F(a_i)}{F(a_{i+1}) - F(a_i)}(a_{i+1} - a_i)$$

Il reste alors à remplacer x par $0,5$ et on trouve bien la valeur de la médiane. Retourner au texte.

Contenu masqué n°3

On commence par calculer les moyenne de chaque groupe, que l'on notera $\overline{G1}, \overline{G2}$ et $\overline{G3}$:

$$\overline{G1} = \frac{10 + 12 + 11 + 13 + 14 + 12 + 7 + 15}{8} = 11,75 \qquad \overline{G2} = 11,5 \qquad \overline{G3} = 8,7$$

Calculons la variance de $G1$:

$$v_1 = \frac{(10 - 11,75)^2 + (12 - 11,75)^2 + (11 - 11,75)^2 + (13 - 11,75)^2 + (14 - 11,75)^2 + (12 - 11,75)^2 + (7$$

$$v_1 = \frac{1,75^2 + 0,25^2 + 0.75^2 + 2,25^2 + 0,25^2 + 4,75^2 + 3,25^2}{8} \approx 6.214 \qquad v_2 = 23.143 \qquad v_3 = 9.122$$

On remarque que $v_2 > v_3 > v_1$, le groupe le plus hétérogène est donc le groupe $G2$. Ce que l'on remarque, c'est que les groupes $G1$ et $G2$ ont des moyennes assez proches, et pourtant, leur variance est complètement différente. À l'avenir donc, ne vous fiez jamais à la moyenne, et effectuer un calcul de variance quasi-systématiquement dans vos études statistiques. Retourner au texte.

Contenu masqué n°4

On ne démontre que dans le cas discret, car on procède de la même manière pour les autres cas. Remarquons que, en écrivant M_1, il s'agit de la moyenne de X. Pour le second terme :

$$\mu_1 = \sum_{i=1}^{k} f_i(a_i - \overline{X}) = \sum_{i=1}^{k} f_i a_i - \overline{X} \sum_{i=1}^{k} f_i = \overline{X} - \overline{X} = 0$$

Contenu masqué

Enfin, la dernière équation qui est très importante en probabilité et apparaît sous le nom de théorème de König-Huygens :

$$\mu_2 = \sum_{i=1}^{k} f_i(a_i - \overline{X})^2 = \sum_{i=1}^{k} f_i a_i^2 - 2\overline{X}\sum_{i=1}^{k} f_i a_i + \overline{X}^2 \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} f_i a_i^2 - 2\overline{X}^2 + \overline{X}^2 = M_2 - (M_1)^2$$

Retourner au texte.