Procédures de statistique descriptive
Aspects opératoires

§ 1 Variable aléatoire discrète

Si la variable aléatoire est *discrète*, c'est-à-dire si elle ne prend que des valeurs isolées, alors les données sont représentées par un *diagramme en bâtons*.

Données

Les données se présentent généralement sous la forme d'un tableau des effectifs des modalités :

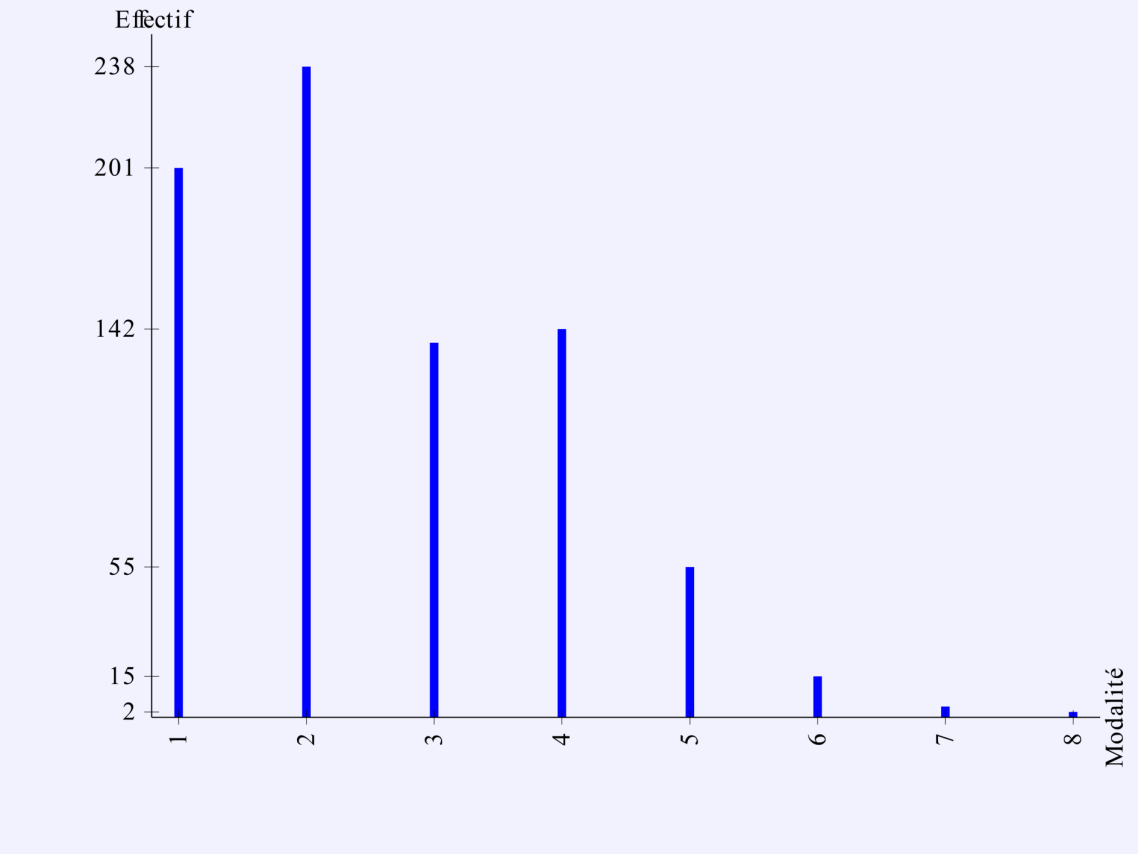
Modalité	Effectif
x_1	n_1
x_2	n_2
x_3	n_3
...	...

On calcule l'effectif total :

$$n_1 + n_2 + n_3 + ... = n$$

Diagramme en bâtons des effectifs

Les modalités sont portées en abscisses.
Les effectifs sont portés en ordonnées.



Fréquences

À partir du tableau des effectifs des modalités, on dresse le tableau des fréquences :

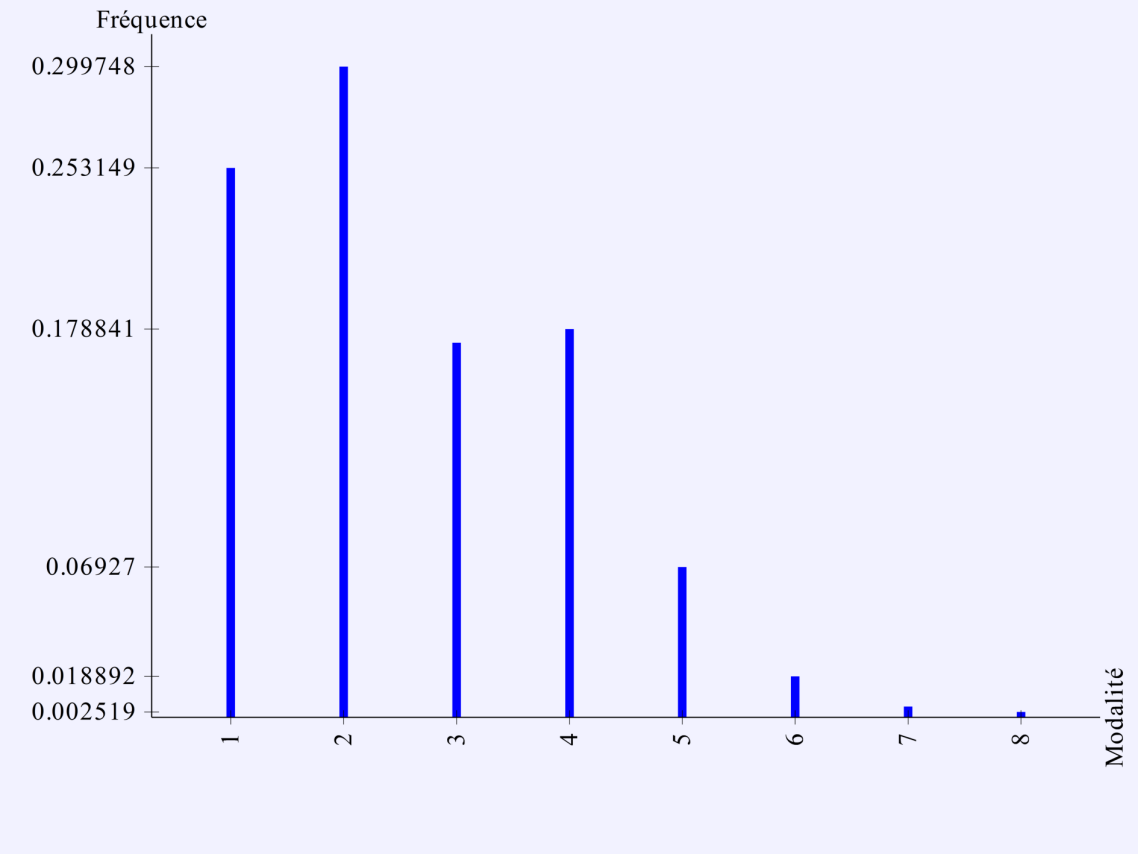
Modalité	Fréquence
x_1	$f_1 = \frac{n_1}{n}$
x_2	$f_2 = \frac{n_2}{n}$
x_3	$f_3 = \frac{n_3}{n}$
...	...

On a :

$$f_1 + f_2 + f_3 + ... = 1$$

Diagramme en bâtons des fréquences

Les modalités sont portées en abscisses.
Les fréquences sont portées en ordonnées.



Effectifs cumulés

Il s'agit de dresser le tableau des effectifs cumulés jusqu'à une modalité. Pour ce faire, on utilise le tableau des effectifs. Par exemple, par définition,

$$N_5 = n_1 + n_2 + n_3 + n_4 + n_5$$

Si l'on a déjà calculé les effectifs cumulés précédents, on peut les utiliser :

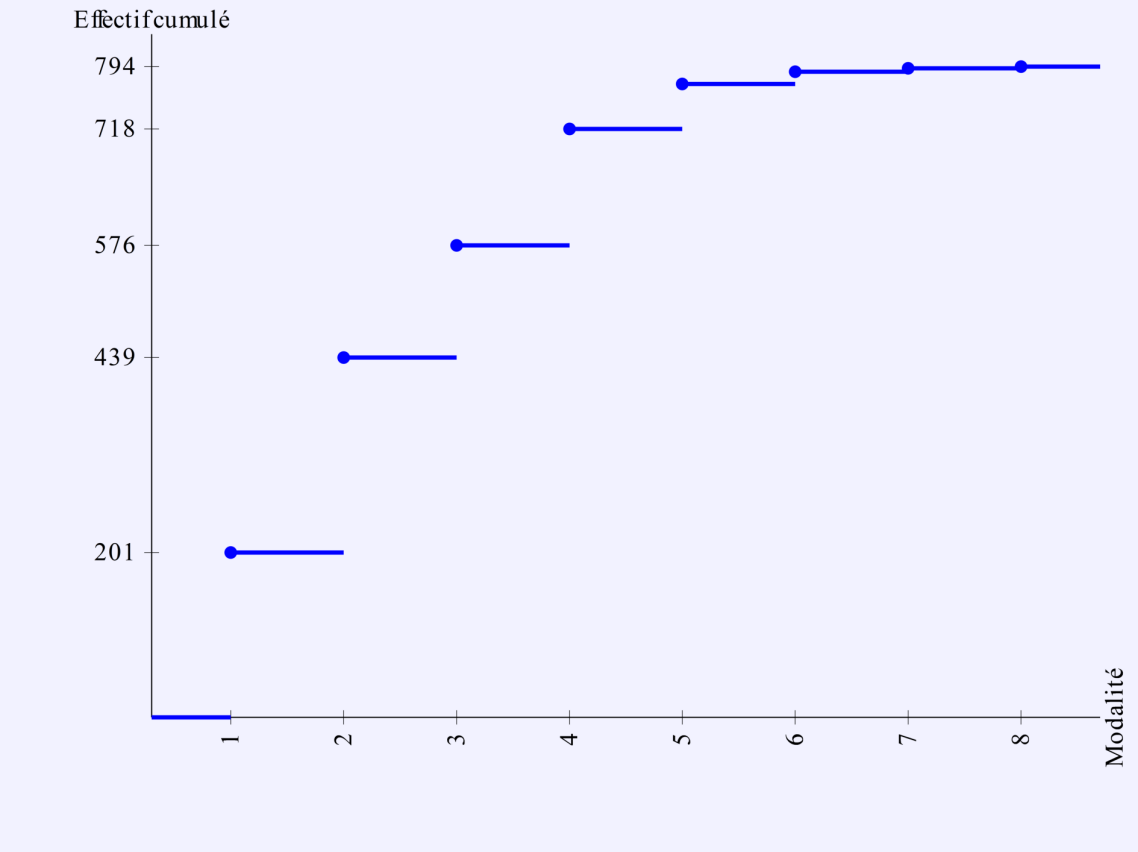
$$N_5 = N_4 + n_5$$

Modalité	Effectif cumulé
x_1	$N_1 = n_1$
x_2	$N_2 = N_1 + n_2$
x_3	$N_3 = N_2 + n_3$
x_4	$N_4 = N_3 + n_4$
x_5	$N_5 = N_4 + n_5$
...	...
	n

Le dernier effectif cumulé a pour valeur l'effectif total n.

Fonction de distribution des effectifs

En abscisses, on porte les modalités.
En ordonnées, on porte les effectifs cumulés.



Fréquences cumulées

Il s'agit de dresser le tableau des fréquences cumulées jusqu'à une modalité. Par exemple, par définition,

$$F_5 = f_1 + f_2 + f_3 + f_4 + f_5$$

Si l'on a déjà calculé les fréquences cumulées précédentes, on peut les utiliser :

$$F_5 = F_4 + f_5$$

Si l'on dispose du tableau des effectifs cumulés, on peut préférer

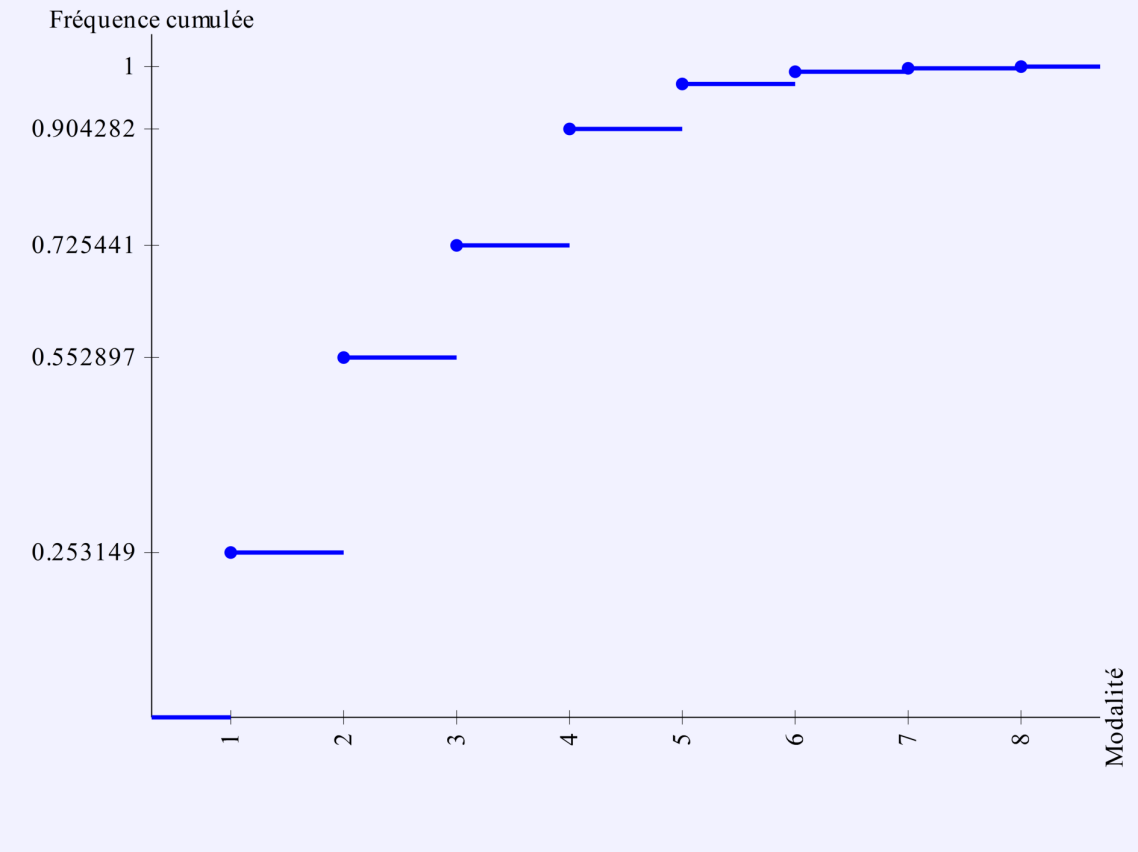
$$F_5 = \frac{N_5}{n}$$

Modalité	Fréquence cumulée
x_1	$F_1 = \frac{N_1}{n}$
x_2	$F_2 = \frac{N_2}{n}$
x_3	$F_3 = \frac{N_3}{n}$
x_4	$F_4 = \frac{N_4}{n}$
x_5	$F_5 = \frac{N_5}{n}$
...	...
	1

La valeur de la dernière fréquence cumulée est 1.

Fonction de distribution des fréquences

En abscisses, on porte les modalités.
En ordonnées, on porte les fréquences cumulées.



Mode

Le mode est la modalité x_i pour laquelle l'effectif n_i ou la fréquence f_i est maximale. Dans le cas où il y a des ex-aequo à la première place des fréquences, le mode est une liste de modalités, et on dit que la distribution est multimodale.

Moyenne

La moyenne peut se calculer comme suit

$$m = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + x_3 \cdot n_3 + ...}{n}$$

ce qui équivaut à

$$m = x_1 \cdot f_1 + x_2 \cdot f_2 + x_3 \cdot f_3 + ...$$

Étendue

L'étendue est l'écart entre les modalités extrêmes.

Variance

La variance est la moyenne des carrés des écarts à la moyenne :

$$V = \frac{(x_1 - m)^2 \cdot n_1 + (x_2 - m)^2 \cdot n_2 + (x_3 - m)^2 \cdot n_3 + ...}{n}$$

$$V = (x_1 - m)^2 \cdot f_1 + (x_2 - m)^2 \cdot f_2 + (x_3 - m)^2 \cdot f_3 + ...$$

Il s'agit ici de la variance empirique non corrigée (à distinguer de l'estimateur de la variance théorique obtenu en multipliant par $\frac{n}{n-1}$).

Écart-type

L'écart-type est égal à la racine carrée de la variance :

$$s = \sqrt{V}$$

Il s'agit ici de l'écart-type non corrigé (à distinguer de l'estimateur de l'écart-type théorique obtenu en multipliant par $\sqrt{\frac{n}{n-1}}$).

Inégalité de Bienaymé-Tchebychev

L'inégalité de Bienaymé-Tchebychev donne un minorant de la probabilité d'un intervalle centré sur la moyenne :

$$P([\mu - k\sigma, \mu + k\sigma[) \geq 1 - \frac{1}{k^2}$$

Avec les valeurs empiriques $\mu = m$, $\sigma = s$, et pour $k = \sqrt{\frac{1}{1-t}}$, on a

$$P([m - k \cdot s, m + k \cdot s[) \geq t$$

L'inégalité de Bienaymé-Tchebychev n'est pas une estimation. Par exemple, pour $k = 2.23607$, elle donne 0.8 comme minorant alors que, si la distribution est normale, la probabilité de l'intervalle est 0.974653. Cependant, elle a l'avantage de s'appliquer à toutes les distributions, qu'elles soient normales ou non.

Comparaison avec la distribution normale

Pour comparer une distribution discrète avec la distribution normale, il faut préalablement la convertir en une distribution continue. Par exemple, si les modalités sont $0, 1, 2, 3, ...$, on les remplace par les classes $[-0.5; 0.5[$, $[0.5, 1.5[$, $[1.5; 2.5[$, $[2.5; 3.5[$, ...

§ 2 Variable aléatoire continue

Si la variable aléatoire est *continue*, c'est-à-dire si elle peut prendre n'importe quelle valeur d'un intervalle, alors les données sont représentées par un *histogramme*.

Données

Les classes sont des intervalles délimités par leurs bornes. Les données se présentent généralement sous la forme d'un tableau des effectifs des classes :

Classe	Effectif
$[b_0; b_1[$	n_1
$[b_1; b_2[$	n_2
$[b_2; b_3[$	n_3
...	...

On calcule l'effectif total :

$$n_1 + n_2 + n_3 + ... = n$$

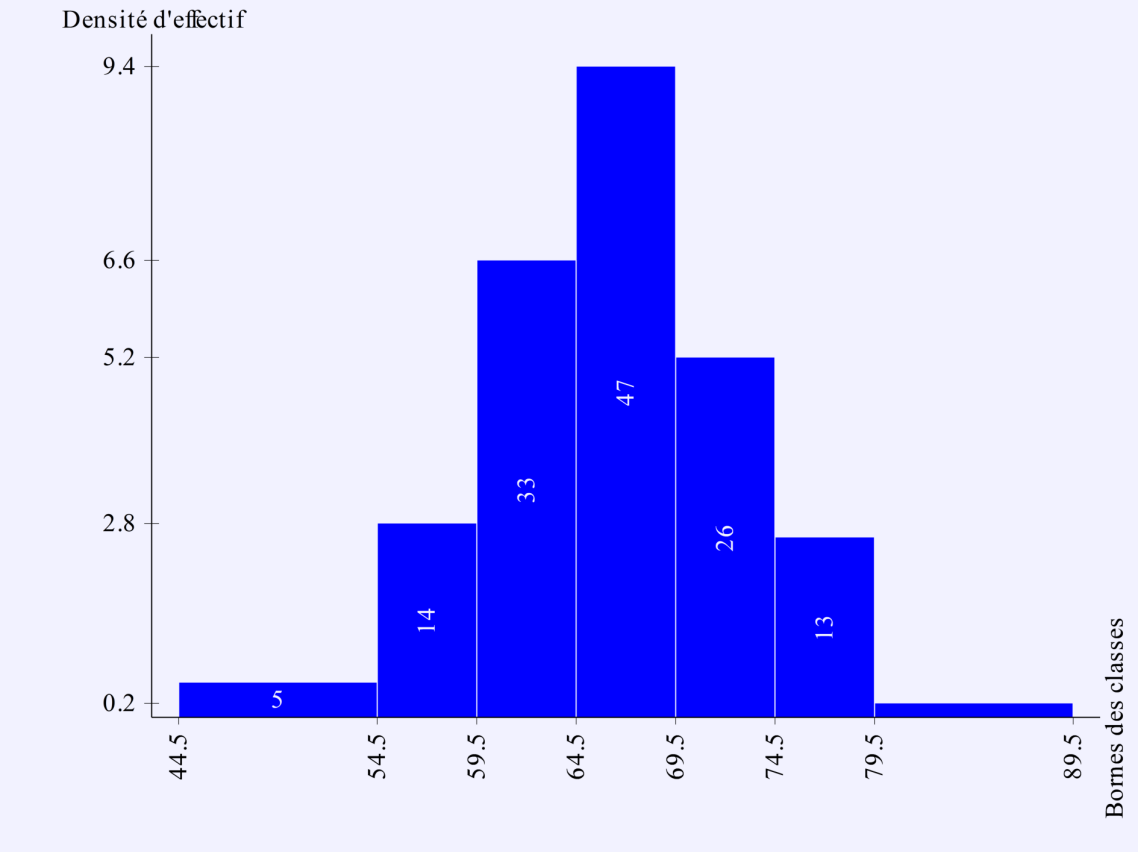
Histogramme des effectifs

Dans un histogramme, les *effectifs* ne sont pas représentés par les hauteurs des rectangles, mais leurs *aires*. Pour chaque rectangle, le côté horizontal est l'amplitude de la classe. La *hauteur* de chaque rectangle, appelée *densité*, est égale à

$$\text{densité d'effectif} = \frac{\text{effectif de la classe}}{\text{amplitude de la classe}}$$

Classe	Densité
$[b_0; b_1[$	$h_1 = \dfrac{n_1}{b_1 - b_0}$
$[b_1; b_2[$	$h_2 = \dfrac{n_2}{b_2 - b_1}$
$[b_2; b_3[$	$h_3 = \dfrac{n_3}{b_3 - b_2}$
...	...

Pour tracer l'histogramme, en abscisses, on reporte les bornes des classes.
En ordonnées, on reporte les densités d'effectif des classes.



Fréquences

À partir du tableau des effectifs des classes, on dresse le tableau des fréquences des classes :

Classe	Fréquence
$[b_0; b_1[$	$f_1 = \frac{n_1}{n}$
$[b_1; b_2[$	$f_2 = \frac{n_2}{n}$
$[b_2; b_3[$	$f_3 = \frac{n_3}{n}$
...	...

On a :

$$f_1 + f_2 + f_3 + ... = 1$$

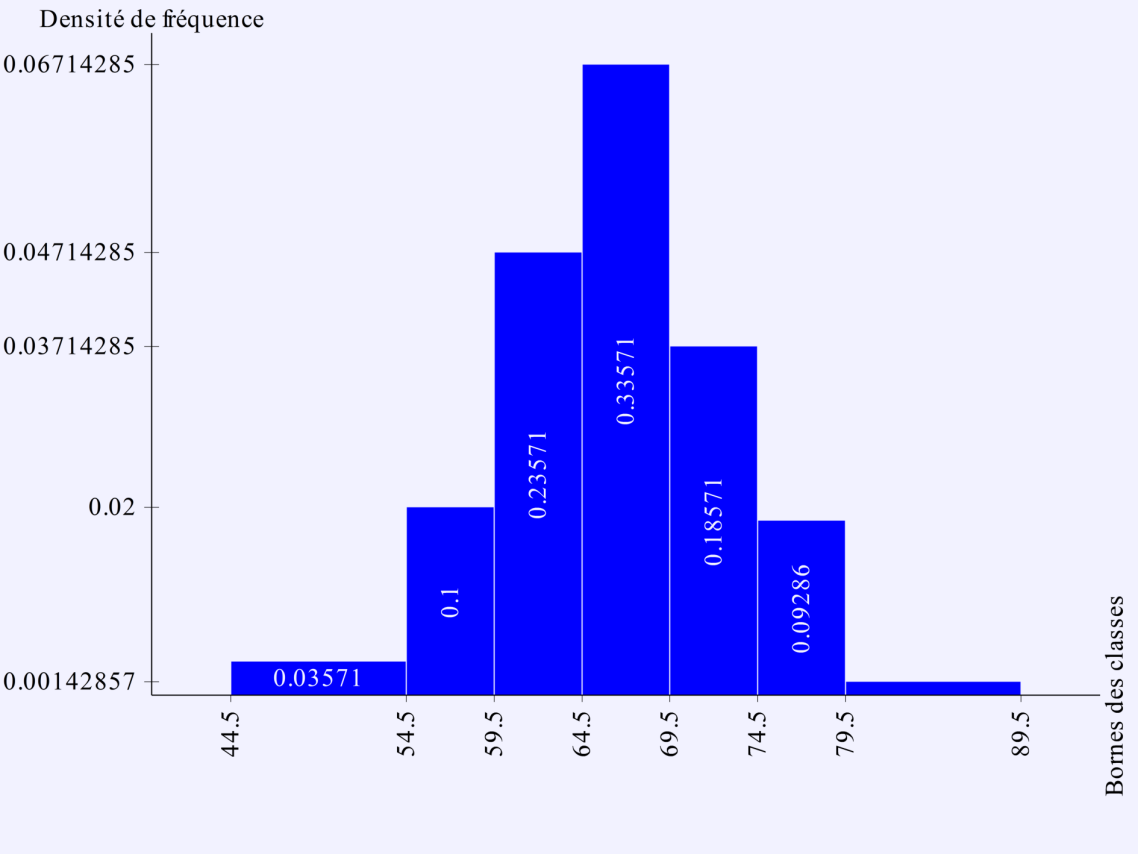
Histogramme des fréquences

Dans un histogramme, les *fréquences* ne sont pas représentées par les hauteurs des rectangles, mais leurs *aires*. Pour chaque rectangle, le côté horizontal est l'amplitude de la classe. La *hauteur* de chaque rectangle, appelée *densité*, est égale à

$$\text{densité de fréquence} = \frac{\text{fréquence de la classe}}{\text{amplitude de la classe}}$$

Classe	Densité
$[b_0; b_1[$	$h_1 = \dfrac{f_1}{b_1 - b_0}$
$[b_1; b_2[$	$h_2 = \dfrac{f_2}{b_2 - b_1}$
$[b_2; b_3[$	$h_3 = \dfrac{f_3}{b_3 - b_2}$
...	...

Pour tracer l'histogramme, en abscisses, on reporte les bornes des classes.
En ordonnées, on reporte les densités de fréquence des classes.



Effectifs cumulés

Il s'agit de dresser le tableau des effectifs cumulés jusqu'à une borne. Pour ce faire, on utilise le tableau des effectifs. Par exemple, par définition,

$$N_5 = n_1 + n_2 + n_3 + n_4 + n_5$$

Si l'on a déjà calculé les effectifs cumulés précédents, on peut les utiliser :

$$N_5 = N_4 + n_5$$

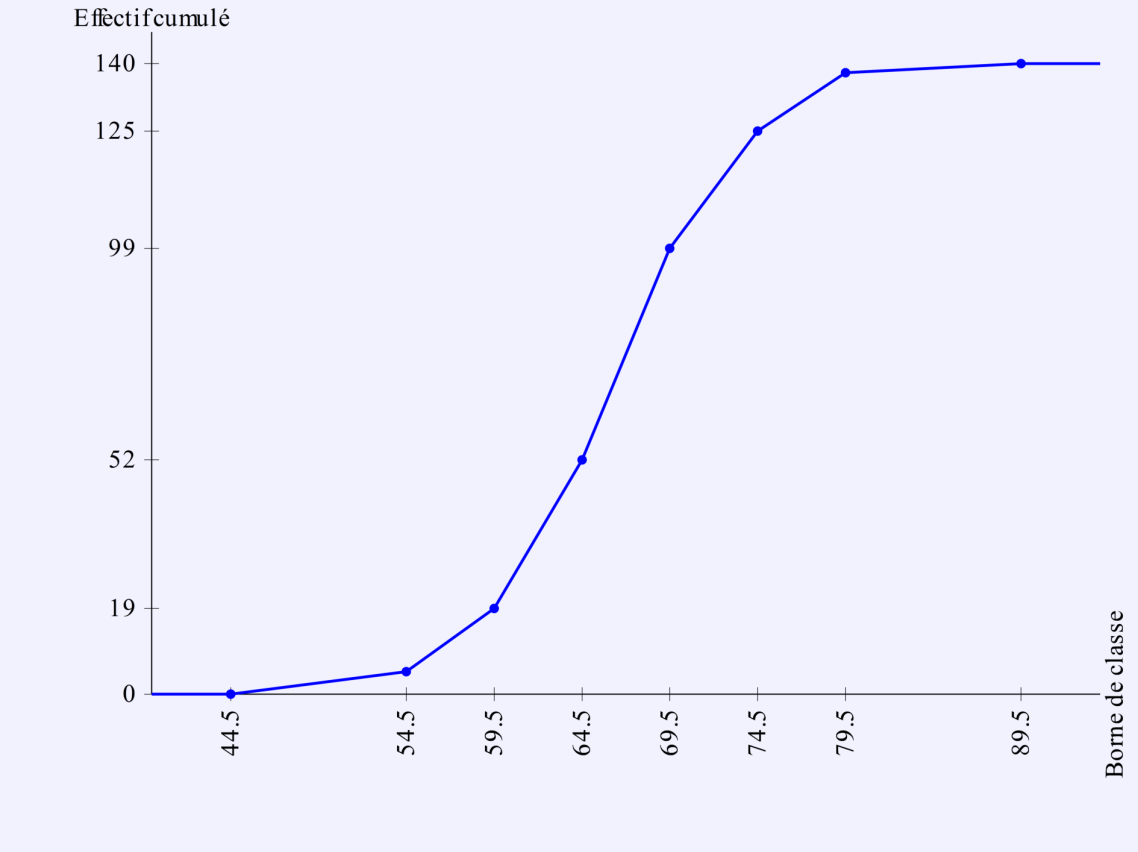
Borne	Effectif cumulé
b_0	$N_0 = 0$
b_1	$N_1 = n_1$
b_2	$N_2 = N_1 + n_2$
b_3	$N_3 = N_2 + n_3$
b_4	$N_4 = N_3 + n_4$
b_5	$N_5 = N_4 + n_5$
...	...
	n

Le dernier effectif cumulé a pour valeur l'effectif total n.

On remarquera que ce tableau comporte une ligne de plus que celui des effectifs. On dira donc «l'effectif d'une classe» et «l'effectif cumulé jusqu'à la borne».

Fonction de distribution des effectifs

En abscisses, on porte les bornes des classes.
En ordonnées, on porte les effectifs cumulés.



Fréquences cumulées

Il s'agit de dresser le tableau des fréquences cumulées jusqu'à une borne. Par exemple, par définition,

$$F_5 = f_1 + f_2 + f_3 + f_4 + f_5$$

Si l'on a déjà calculé les fréquences cumulées précédentes, on peut les utiliser :

$$F_5 = F_4 + f_5$$

Si l'on dispose du tableau des effectifs cumulés, on préférer :

$$F_5 = \frac{N_5}{n}$$

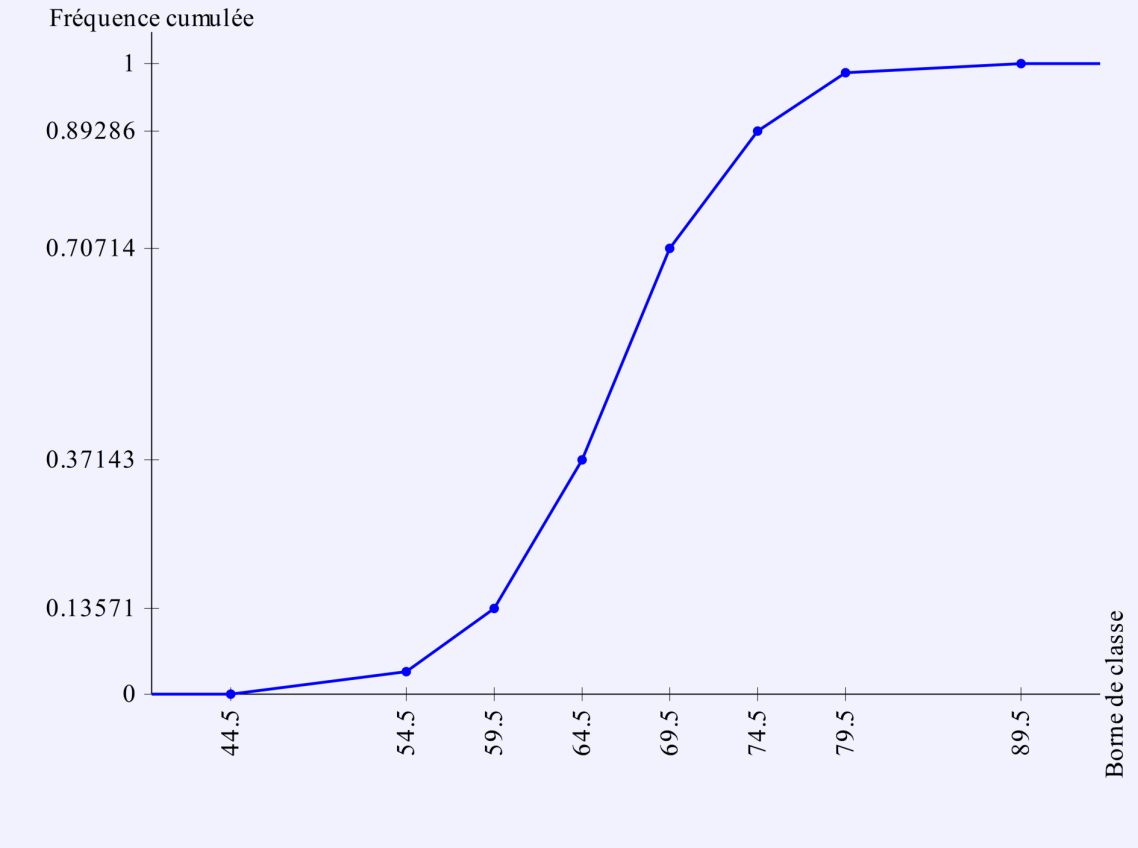
Borne	Fréquence cumulée
b_0	$F_0 = 0$
b_1	$F_1 = \frac{N_1}{n}$
b_2	$F_2 = \frac{N_2}{n}$
b_3	$F_3 = \frac{N_3}{n}$
b_4	$F_4 = \frac{N_4}{n}$
b_5	$F_5 = \frac{N_5}{n}$
...	...
	1

La valeur de la dernière fréquence cumulée est 1.

On remarquera que ce tableau comporte une ligne de plus que celui des fréquences. On dira donc «la fréquence d'une classe» et «la fréquence cumulée jusqu'à la borne».

Fonction de distribution des fréquences

En abscisses, on porte les bornes des classes.
En ordonnées, on porte les fréquences cumulées.



Classe modale

La classe modale est celle dont la *densité* (et non la fréquence) est la plus élevée. Pour tenir compte de la possibilité d'avoir plusieurs classes ex-aequo, il vaut mieux parler de l'ensemble des classes modales.

Moyenne

Pour calculer la moyenne, on fait appel aux centres des classes

Centre de la classe	Effectif
$c_1 = \frac{b_0 + b_1}{2}$	n_1
$c_2 = \frac{b_1 + b_2}{2}$	n_2
$c_3 = \frac{b_2 + b_3}{2}$	n_3
...	...
Total	$n = n_1 + n_2 + n_3 + ...$

La moyenne se calcule alors comme suit

$$m = \frac{c_1 \cdot n_1 + c_2 \cdot n_2 + c_3 \cdot n_3 + ...}{n}$$

ce qui équivaut à

$$m = c_1 \cdot f_1 + c_2 \cdot f_2 + c_3 \cdot f_3 + ...$$

Médiane

Le deuxième quartile est appelé *médiane* et correspond à une fréquence cumulée de 0.5

On commence par repérer, dans le tableau des fréquences cumulées, l'intervalle dans lequel se trouve la fréquence cumulée 0.5. Restreinte à cet intervalle $[a; b[$, la fonction de distribution F est affine et monotone croissante.

Variable aléatoire	Fréquence cumulée
a	$F(a)$
médiane	0.5
b	$F(b)$

Sachant que $F(a) \leq 0.5 < F(b)$, on détermine le nombre médiane tel que $a \leq$ médiane $< b$ au moyen de la *formule d'interpolation* :

$$\text{médiane} = a + \frac{b - a}{F(b) - F(a)}(0.5 - F(a))$$

Des exemples de calculs sont donnés dans les *corrigés des exercices* portant sur les variables aléatoires continues.

Interprétation : les données inférieures à la médiane constituent la moitié de l'effectif.

Étendue

L'étendue est l'écart entre les bornes extrêmes.

Variance

La variance est la moyenne des carrés des écarts à la moyenne :

$$V = \frac{(c_1 - m)^2 \cdot n_1 + (c_2 - m)^2 \cdot n_2 + (c_3 - m)^2 \cdot n_3 + ...}{n}$$

$$V = (c_1 - m)^2 \cdot f_1 + (c_2 - m)^2 \cdot f_2 + (c_3 - m)^2 \cdot f_3 + ...$$

Il s'agit ici de la variance empirique non corrigée (à distinguer de l'estimateur de la variance théorique obtenu en multipliant par $\frac{n}{n-1}$).

Écart-type

L'écart-type est égal à la racine carrée de la variance :

$$s = \sqrt{V}$$

Il s'agit ici de l'écart-type non corrigé (à distinguer de l'estimateur de l'écart-type théorique obtenu en multipliant par $\sqrt{\frac{n}{n-1}}$).

Premier quartile Q_1

Le premier quartile correspond à une fréquence cumulée de 0.25

On commence par repérer, dans le tableau des fréquences cumulées, l'intervalle dans lequel se trouve la fréquence cumulée 0.25. Restreinte à cet intervalle $[a; b[$, la fonction de distribution F est affine et monotone croissante.

Variable aléatoire	Fréquence cumulée
a	$F(a)$
Q_1	0.25
b	$F(b)$

Sachant que $F(a) \leq 0.25 < F(b)$, on détermine Q_1 tel que $a \leq Q_1 < b$ au moyen de la *formule d'interpolation* :

$$Q_1 = a + \frac{b - a}{F(b) - F(a)}(0.25 - F(a))$$

Interprétation : les données inférieures au premier quartile constituent le quart de l'effectif.

Troisième quartile Q_3

Le troisième quartile correspond à une fréquence cumulée de 0.75

On commence par repérer, dans le tableau des fréquences cumulées, l'intervalle dans lequel se trouve la fréquence cumulée 0.75. Restreinte à cet intervalle $[a; b[$, la fonction de distribution F est affine et monotone croissante.

Variable aléatoire	Fréquence cumulée
a	$F(a)$
Q_3	0.75
b	$F(b)$

Sachant que $F(a) \leq 0.75 < F(b)$, on détermine Q_3 tel que $a \leq Q_3 < b$ au moyen de la *formule d'interpolation* :

$$Q_3 = a + \frac{b-a}{F(b) - F(a)}(0.75 - F(a))$$

Interprétation : les données inférieures au troisième quartile constituent les trois quarts de l'effectif.

Intervalle interquartile

L'intervalle interquartile est l'écart entre les premier et troisième quartiles :

$$Q_3 - Q_1$$

Interprétation : dans l'intervalle $[Q_1, Q_3[$ se situe la moitié de l'effectif.

Inégalité de Bienaymé-Tchebychev

L'inégalité de Bienaymé-Tchebychev donne un minorant de la probabilité d'un intervalle centré sur la moyenne :

$$P([\mu - k\sigma, \mu + k\sigma[) \geq 1 - \frac{1}{k^2}$$

Avec les valeurs empiriques $\mu = m$, $\sigma = s$, et pour $k = \sqrt{\frac{1}{1-t}}$, on a

$$P([m - k \cdot s, m + k \cdot s[) \geq t$$

L'inégalité de Bienaymé-Tchebychev n'est pas une estimation. Par exemple, pour $k = 2.23607$, elle donne 0.8 comme minorant alors que, si la distribution est normale, la probabilité de l'intervalle est 0.974653. Cependant, elle a l'avantage de s'appliquer à toutes les distributions, qu'elles soient normales ou non.

Variable centrée réduite

La variable aléatoire est centrée en lui soustrayant la moyenne, puis réduite en divisant par l'écart-type

$$z_0 = \frac{b_0 - m}{s}$$
$$z_1 = \frac{b_1 - m}{s}$$
$$z_2 = \frac{b_2 - m}{s}$$
$$\ldots$$

Les fréquences demeurant inchangées, on part des données ainsi modifiées :

Classe	Fréquence
$[z_0; z_1[$	f_1
$[z_1; z_2[$	f_2
$[z_2; z_3[$	f_3
...	...

Par construction, la moyenne de la variable centrée réduite est nulle et l'écart-type vaut 1.

Comparaison avec la densité normale

La densité de la loi normale de Gauss-Laplace, dans le cas où la moyenne est 0 et l'écart-type est 1, a pour expression :

$$f(x) = \frac{\exp(-\frac{x^2}{2})}{\sqrt{2\pi}}$$

Pour effectuer une comparaison visuelle, le graphique de la densité normale est superposé à l'histogramme de la variable centrée réduite.

Marcel Délèze

Lien vers la page mère : *Statistique descriptive*

www.deleze.name/marcel/sec2/stat-descr/index.html