Capes de Mathématiques
Statistique descriptive élémentaire
Formulaire

Jeanne Fine

Septembre 2006

Un cours de statistique descriptive élémentaire plus convivial est disponible en ligne sur le site : http ://www2.toulouse.iufm.fr/mathematiques/

Si vous repérez des erreurs, merci de me les signaler (jeanne.fine@toulouse.iufm.fr).

Table des matières

1 Traitement statistique d'une variable catégorielle **2**
1.1 Distributions d'effectifs et de fréquences ; diagrammes en barres ou en secteurs 3

2 Traitement statistique de deux variables catégorielles **3**
2.1 Distributions conjointe et marginales des effectifs et des fréquences 4
2.2 Distributions conditionnelles (profils lignes et profils colonnes) 5
2.3 Résumé numérique, écart à l'indépendance : le Khi2 6

3 Traitement statistique d'une variable réelle **7**
3.1 Distributions d'effectifs et de fréquences ; représentations graphiques 8
3.1.1 Cas d'une variable réelle discrète ; diagramme en bâtons 8
3.1.2 Cas d'une variable réelle continue ; histogramme 9
3.2 Effectifs cumulés et fréquences cumulées ; représentations graphiques 10
3.2.1 Cas d'une variable réelle discrète : fonction en escaliers 10
3.2.2 Cas d'une variable réelle continue : fonction continue affine par morceaux 10

3.3 Caractéristiques de position et de dispersion 11
3.3.1 Mode et étendue 11
3.3.2 Médiane, quartiles, écart interquartiles, déciles, diagramme en boîte 11
3.3.3 Quartiles et écart inter-quartiles 13
3.3.4 Moyenne, variance et écart-type 16

4 Traitement statistique d'une variable catégorielle et d'une variable réelle 18
4.1 Décomposition de la moyenne et de la variance sur une partition 19
4.2 Résumé numérique : rapport de corrélation 20

5 Traitement statistique de deux variables réelles 20
5.1 Distributions d'effectifs et de fréquences 20
5.2 Représentation graphique : le graphe plan 21
5.3 Résumés numériques : la covariance, le coefficient de corrélation linéaire 21
5.4 Droites de régression linéaire ; prévisions 23
5.4.1 Droite de régression linéaire de Y en X 23
5.4.2 Droite de régression linéaire de X en Y 25
5.4.3 Application à la détermination de la tendance linéaire d'une série chronologique 25

1 Traitement statistique d'une variable catégorielle

Soit $\Omega = \{\omega_1, ..., \omega_n\}$ une population sur laquelle est définie une variable catégorielle A et $A(\Omega) = \{a_1, ..., a_I\}$ l'ensemble des valeurs ou *modalités* de la variable.

Pour $i \in [1, I]$, on note $[A = a_i]$ l'ensemble $A^{-1}(\{a_i\}) = \{\omega \in \Omega \ ; \ A(\omega) = a_i\}$, c'est-à-dire, la *catégorie* d'individus associée à la modalité a_i.

L'ensemble $\{[A = a_i] \ ; \ i \in [1, I]\}$ est la *partition de* Ω *engendrée par* A.

Pour $i \in [1, I]$, on pose $n_i = card([A = a_i])$, n_i est l'*effectif* associé à la modalité a_i et on a $n = \sum_{i=1}^{I} n_i$, enfin, on pose $f_i = n_i/n$, f_i est la *fréquence* associée à la modalité a_i et on a $1 = \sum_{i=1}^{I} f_i$.

1.1 Distributions d'effectifs et de fréquences ; diagrammes en barres ou en secteurs

La *distribution d'effectifs* de la variable catégorielle A est l'ensemble

$$\{(a_i, n_i) \ ; \ i \in [1, I]\}$$

et la *distribution de fréquences* de la variable catégorielle A est l'ensemble

$$\{(a_i, f_i) \ ; \ i \in [1, I]\} \, ;$$

ces distributions sont souvent présentées sous forme de tableau :

$Valeurs$	$Effectifs$	$Fréquences$
a_1	n_1	f_1
...	...	...
a_i	n_i	f_i
...	...	...
a_I	n_I	f_I
$\sum$	n	1

Les distributions d'effectifs et de fréquences peuvent être représentées graphiquement par un *diagramme en barres* : on place les valeurs de la variable sur un axe horizontal et on élève au dessus d'elles des barres dont les hauteurs sont proportionnelles aux effectifs (et donc aussi aux fréquences).

Ces distributions peuvent également être représentées par un *diagramme en secteurs*. On représente alors un disque dont les secteurs angulaires représentant les catégories ont des mesures angulaires proportionnelles aux effectifs et aux fréquences.

2 Traitement statistique de deux variables catégorielles

Soit Ω une population de taille n sur laquelle sont définies deux variables catégorielles A et B à respectivement I et J modalités.
On pose $A(\Omega) = \{a_1, ..., a_i, ..., a_I\}$ et $B(\Omega) = \{b_1, ..., b_j, ..., b_J\}$.

Nous avons vu qu'à une variable catégorielle pouvait être associée la partition de la population engendrée par cette variable : $\{[A = a_i] \ ; \ i \in [1, I]\}$

pour la variable A et $\{[B = b_j]\ ;\ j \in [1, J]\}$ pour la variable B.

On en déduit une nouvelle partition de Ω formée de $I \times J$ ensembles, appelée *partition croisée* : $\{[A = a_i] \cap [B = b_j]\ ;\ i \in [1, I],\ j \in [1, J]\}$, associée au couple de variables catégorielles (A, B).

2.1 Distributions conjointe et marginales des effectifs et des fréquences

Pour $i \in [1, I]$ et $j \in [1, J]$, on note n_{ij} le nombre d'individus qui prennent la modalité a_i de A et la modalité b_j de B et on pose :

$$n_{i.} = \sum_{j=1}^{J} n_{ij} \quad \text{et} \quad n_{.j} = \sum_{i=1}^{I} n_{ij}.$$

Le nombre total n d'individus vérifie alors :

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j}.$$

On pose $f_{ij} = n_{ij}/n$, $f_{i.} = \sum_{j=1}^{J} f_{ij} = n_{i.}/n$, $f_{.j} = \sum_{i=1}^{I} f_{ij} = n_{.j}/n$.
On a alors :

$$1 = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} = \sum_{i=1}^{I} f_{i.} = \sum_{j=1}^{J} f_{.j}.$$

La *distribution conjointe des effectifs*, correspondant à la partition croisée, au centre, et les deux *distributions marginales des effectifs*, correspondant à chacune des deux variables, dans les marges, sont alors présentées dans le tableau suivant, appelé aussi *tableau de contingence*.

$A \setminus B$	b_1 ... b_j ... b_J	*Ensemble*
a_1		
...		
a_i	n_{ij}	$n_{i.}$
...		
a_I		
Ensemble	$n_{.j}$	n

De façon analogue, la *distribution conjointe des fréquences* et les deux *distributions marginales des fréquences* sont présentées dans le tableau suivant.

$A \backslash B$	$b_1 \ldots b_j \ldots b_J$	*Ensemble*
a_1		
...		
a_i	f_{ij}	$f_{i.}$
...		
a_I		
Ensemble	$f_{.j}$	1

2.2 Distributions conditionnelles (profils lignes et profils colonnes)

La distribution des fréquences de la variable B conditionnelle au sous-ensemble $[A = a_i]$ correspond à la *ième* ligne du tableau suivant. Nous avons donc I distributions conditionnelles de B (appelées aussi *profils lignes*) que l'on peut comparer à la distribution marginale de B, dernière ligne du tableau.

$A \setminus B$	$b_1 \ldots b_j \ldots b_J$	*Ensemble*
a_1		1
...		
a_i	$n_{ij}/n_{i.}$	1
...		
a_I		1
Ensemble	$n_{.j}/n$	1

De façon analogue, *la distribution des fréquences de la variable A conditionnelle au sous-ensemble* $[B = b_j]$ correspond à la *jème* colonne du tableau suivant. Nous avons donc J distributions conditionnelles de A (appelées aussi *profils colonnes*) que l'on peut comparer à la distribution marginale de A, dernière colonne du tableau.

$A \searrow B$	$b_1 \ldots b_j \ldots b_J$	*Ensemble*
a_1		
...		
a_i	$n_{ij}/n_{.j}$	$n_{i.}/n$
...		
a_I		
Ensemble	1 ... 1 ... 1	1

2.3 Résumé numérique, écart à l'indépendance : le Khi2

Nous traduirons l'absence de liaison (ou *indépendance*) entre les deux variables catégorielles A et B par l'une des propriétés équivalentes suivantes :
1) égalité des profils lignes $\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$ pour tout i et j,
2) égalité des profils colonnes $\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$ pour tout i et j,
3) $n_{ij} = \frac{n_{i.}n_{.j}}{n}$ pour tout i et j.

Le *tableau théorique d'indépendance*, construit sur les distributions marginales des effectifs de A et de B est alors le suivant :

$A \setminus B$	b_1 ... b_j ... b_J	*Ensemble*
a_1		
...		
a_i	$\frac{n_{i.}n_{.j}}{n}$	$n_{i.}$
...		
a_I		
Ensemble	$n_{.j}$	n

On peut à présent construire le *tableau des écarts entre les effectifs observés et les effectifs théoriques d'indépendance*. On obtient alors le tableau suivant :

$A \setminus B$	b_1 ... b_j ... b_J	*Ensemble*
a_1		
...		
a_i	$n_{ij} - \frac{n_{i.}n_{.j}}{n}$	0
...		
a_I		
Ensemble	0	0

La somme des éléments de chaque ligne (resp. de chaque colonne) est nulle. On a donc des écarts positifs et des écarts négatifs. Un écart positif indique une *sur-représentation* du couple de modalités correspondant par rapport à l'indépendance, un écart négatif indique une *sous-représentation*.

Dans le cadre de la statistique inférentielle, l'écart entre le tableau de contingence des effectifs observés et le tableau d'effectifs théoriques est mesuré à l'aide de l'indice suivant, appelé *indice du* χ^2 *ou Khi2* (le "2" vient du

fait qu'il s'agit du carré d'une distance entre les deux tableaux) :

$$\begin{aligned} Khi2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\left(\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right) \\ &= n\left(\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{f_{ij}^2}{f_{i.}f_{.j}} - 1\right) = n\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \end{aligned}$$

La première égalité correspond à la définition, les deux autres s'en déduisent aisément. Les deux dernières égalités montrent que le Khi2 est proportionnel à la taille de la population (ou de l'échantillon).

Si on pose :

$$ctr_{ij} = \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \,/\, Khi2,$$

alors le *tableau de contribution au* $Khi2$ est défini par :

$A \setminus B$	b_1 ... b_j ... b_J
a_1	
...	
a_i	ctr_{ij}
...	
a_I	

La somme pour $(i,j) \in \{1,...,I\} \times \{1,...,J\}$ des ctr_{ij} est alors égale à 1 et il est possible de repérer les couples de modalités contribuant le plus à l'écart à l'indépendance, le signe de l'écart étant donné par le tableau des écarts.

3 Traitement statistique d'une variable réelle

Soit $\Omega = \{\omega_1, ..., \omega_n\}$ une population sur laquelle est définie une variable réelle X. Pour alléger les écritures, on pose $X_k = X(\omega_k)$ pour $k \in [1, n]$.

Si l'ensemble des valeurs possibles de la variable est un ensemble fini ou dénombrable, la variable est dite *discrète* ; si la variable X peut prendre a priori toute valeur d'un intervalle de $\mathbb{R}$, les valeurs de la variable sont alors regroupées en "classes" et on fait l'hypothèse que, dans chaque classe, les effectifs (et donc aussi les fréquences) sont uniformément distribués. L'approximation ainsi construite est une variable dite *continue* (que l'on notera encore X).

3.1 Distributions d'effectifs et de fréquences ; représentations graphiques

3.1.1 Cas d'une variable réelle discrète ; diagramme en bâtons

Soit X une variable réelle discrète et $X(\Omega) = \{x_i \, ; \, i \in [1, I]\}$ l'ensemble des valeurs distinctes de la variable rangées dans l'ordre croissant.

Pour $i \in [1, I]$, on pose $[X = x_i] = X^{-1}(\{x_i\}) = \{\omega \in \Omega \, ; \, X(\omega) = x_i\}$.

L'ensemble $\{[X = x_i] \ ; \ i \in [1, I]\}$ est la partition de Ω engendrée par X.

Pour $i \in [1, I]$, on pose $n_i = card([X = x_i])$, on a alors $n = \sum_{i=1}^{I} n_i$ et on pose $f_i = n_i/n$, on a alors $1 = \sum_{i=1}^{I} f_i$.

La *distribution d'effectifs* de la variable réelle discrète X est l'ensemble

$$\{(x_i, n_i) \ ; \ i \in [1, I]\}$$

et sa *distribution de fréquences* est l'ensemble

$$\{(x_i, f_i) \ ; \ i \in [1, I]\} \, ;$$

ces distributions sont souvent présentées sous forme de tableau :

$Valeurs$	$Effectifs$	$Fréquences$
x_1	n_1	f_1
...	...	...
x_i	n_i	f_i
...	...	...
x_I	n_I	f_I
$\sum$	n	1

Les distributions d'effectifs et de fréquences d'une variable réelle discrète sont représentées graphiquement par un *diagramme en bâtons* : on place les valeurs de la variable sur un axe horizontal représentant la droite réelle (c'est-à-dire, respectant l'ordre et la distance des nombres réels) et on élève au dessus de ces valeurs des bâtons dont les hauteurs sont proportionnelles aux effectifs (et donc aussi aux fréquences).

3.1.2 Cas d'une variable réelle continue ; histogramme

Soit X une variable réelle et $[\{[x_{i-1}, x_i[\ ;\ i \in [1, I]\}$ un *recouvrement* de $X(\Omega)$ (i.e. $X(\Omega) \subset \cup_{i=1}^{I} [x_{i-1}, x_i[$) en intervalles de $\mathbb{R}$ non vides (i.e. $x_i > x_{i-1}$) deux à deux disjoints, appelés *classes*. On a alors : $X(\Omega) \subset [x_0, x_I[$.

Pour $i \in [1, I]$, on pose :

$$[x_{i-1} \leq X < x_i] = X^{-1}([x_{i-1}, x_i[) = \{\omega \in \Omega\ ;\ x_{i-1} \leq X(\omega) < x_i\}.$$

Pour $i \in [1, I]$, on pose $n_i = card([x_{i-1} \leq X < x_i])$ et $f_i = n_i/n$; on a alors $n = \sum_{i=1}^{I} n_i$ et $1 = \sum_{i=1}^{I} f_i$.

On suppose que, dans chaque classe, les effectifs (et donc aussi les fréquences) sont uniformément distribués.

On note encore X la variable continue ainsi construite (utilisée comme approximation de la variable initiale).

On pose, pour $i \in [1, I]$, $a_i = x_i - x_{i-1}$ appelé *amplitude* de la classe i et $h_i = f_i/a_i$. La *densité de fréquences* de la variable réelle continue X est la fonction h définie par :

$$\forall x \in \mathbb{R},\ h(x) = \sum_{i=1}^{I} h_i \mathbf{1}_{[x_{i-1}, x_i[}(x).$$

La *distribution d'effectifs* de la variable réelle continue X est l'ensemble

$$\{([x_{i-1}, x_i[, n_i)\ ;\ i \in [1, I]\}$$

et sa *distribution de fréquences* est l'ensemble

$$\{([x_{i-1}, x_i[, f_i)\ ;\ i \in [1, I]\}\,;$$

ces distributions sont souvent présentées sous forme de tableau :

$Classes$	$Effectifs$	$Fréquences$
$[x_0, x_1[$	n_1	f_1
...	...	...
$[x_{i-1}, x_i[$	n_i	f_i
...	...	...
$[x_{I-1}, x_I[$	n_I	f_I
$\sum$	n	1

Les distributions d'effectifs et de fréquences sont représentées graphiquement par un *histogramme* : on place les limites des classes de la variable sur un axe horizontal représentant la droite réelle et on élève au dessus des intervalles représentant les classes des rectangles dont les aires sont proportionnelles aux effectifs et aux fréquences. Les fréquences sont donc représentées par la surface située entre le graphe de la densité de fréquences et l'axe horizontal.

L'unité d'aire utilisée pour le graphique doit être reportée dans un coin du graphique pour éviter toute confusion sur la lecture des effectifs et des fréquences.

3.2 Effectifs cumulés et fréquences cumulées ; représentations graphiques

Que la variable réelle soit discrète ou continue, la fonction des effectifs cumulés (resp. la fonction des fréquences cumulées appelée aussi *fonction de répartition*) est définie sur $\mathbb{R}$ par :

$$N(x) = card([X \leq x]), \text{ (resp. } F(x) = N(x)/n = fréq([X \leq x])).$$

C'est une fonction croissante à valeurs dans $[0, n]$ (resp. $[0, 1]$).

On reprend les notations précédemment introduites et on pose, pour i appartenant à $[1, I]$, $N_i = \sum_{j=1}^{i} n_j$ et $F_i = \sum_{j=1}^{i} f_j$.

3.2.1 Cas d'une variable réelle discrète : fonction en escaliers

Dans le cas d'une *variable réelle discrète*, en posant $x_{I+1} = +\infty$, on montre que l'on a :

$N(x) = \sum_{i=1}^{I} N_i \mathbf{1}_{[x_i, x_{i+1}[}(x)$ (resp. $F(x) = \sum_{i=1}^{I} F_i \mathbf{1}_{[x_i, x_{i+1}[}(x)$)

c'est-à-dire une fonction en escaliers, continue à droite (et continue à gauche sauf aux points d'abscisses x_i, $i \in [1, I]$).

3.2.2 Cas d'une variable réelle continue : fonction continue affine par morceaux

Dans le cas d'une *variable réelle continue*, en posant $N_0 = F_0 = 0$, on montre que l'on a :

$N(x) = \sum_{i=1}^{I} \left[\frac{x - x_{i-1}}{x_i - x_{i-1}} (N_i - N_{i-1}) + N_{i-1} \right] \mathbf{1}_{[x_{i-1}, x_i[}(x) + n \mathbf{1}_{[x_I, +\infty[}(x)$

(resp. $F(x) = \sum_{i=1}^{I} \left[\frac{x - x_{i-1}}{x_i - x_{i-1}} (F_i - F_{i-1}) + F_{i-1} \right] \mathbf{1}_{[x_{i-1}, x_i[}(x) + \mathbf{1}_{[x_I, +\infty[}(x)$)

c'est-à-dire une fonction continue affine par morceaux vérifiant $N(\mathbb{R}) = [0, n]$

(resp. $F(\mathbb{R}) = [0, 1]$).

On peut remarquer que l'on a :

$$\frac{F_i - F_{i-1}}{x_i - x_{i-1}} = \frac{f_i}{a_i} = h_i \text{ et } \forall x \in R,\ F(x) = \int_{-\infty}^{x} h(t)dt,$$

h désignant la densité de fréquences.

3.3 Caractéristiques de position et de dispersion

Pour des variables réelles il est possible de proposer des résumés *numériques.*

Pour résumer une distribution d'effectifs ou de fréquences, on proposera une caractéristique de *tendance centrale* ou de *position* autour de laquelle se distribuent les données et une caractéristique mesurant la *dispersion* des données.

3.3.1 Mode et étendue

Dans le cas d'*une variable réelle discrète*, le *mode* ou *valeur modale* est la valeur de la variable ayant le plus fort effectif (ou la plus forte fréquence). L'*étendue* est l'écart entre la plus grande et la plus petite valeur ($= x_I - x_1$ en utilisant les notations introduites précédemment).

Dans le cas d'une *variable réelle continue,* la *classe modale* est celle dont la densité de fréquences est la plus élevée. L'*étendue* est l'écart $x_I - x_0$ en utilisant les notations introduites précédemment.

Ces deux caractéristiques, mode et étendue, de position et de dispersion respectivement, sont très sommaires et peu utilisées.

3.3.2 Médiane, quartiles, écart interquartiles, déciles, diagramme en boîte

Médiane

Une valeur m de la variable réelle X est *médiane* si au moins 50% de la population prend une valeur inférieure ou égale à m et au moins 50% de la population une valeur supérieure ou égale à m, c'est-à-dire, m est médiane si $freq(X \leq m) \geq 0.5$ et $freq(X \geq m) \geq 0.5$ ou encore si : $freq(X < m) \leq 0.5 \leq freq(X \leq m)$

En utilisant la fonction de répartition F, on en déduit que l'ensemble $M(X)$ des médianes de X est :

$$M(X) = \{m \in \mathbb{R} \; ; \; \lim_{\substack{x \to m \\ <}} F(x) \leq 0.5 \leq F(m)\}$$

On peut préciser la définition selon que la variable est discrète ou continue.

Cas d'une variable réelle discrète

Si l'effectif total n est impair, $n = 2k + 1$, k entier positif, la médiane est unique et égale à la valeur prise par le $(k+1)$ème individu lorsque les valeurs sont ordonnées dans l'ordre croissant.

Si l'effectif total n est pair, $n = 2k$, k entier positif, on note a la valeur prise par le $k$ème individu et b celle prise par le $(k+1)$ème individu, alors tout réel de l'intervalle $[a, b]$ est médiane. Pour avoir unicité, on convient alors de prendre pour médiane la valeur $(a + b)/2$.

En utilisant la fonction de répartition, on peut envisager deux cas.

Cas $F^{-1}(\{0.5\}) = \emptyset$, c'est-à-dire cas n impair ou n pair avec $a = b$
On a alors :
$\exists i_0 \in [1, I]$, $(F(x_{i_0}) > 0.5)$ et $(\forall x \in \mathbb{R})$, $(x < x_{i_0}) \Longrightarrow (F(x) < 0.5)$.
On peut en déduire : $M(X) = \{x_{i_0}\}$ $(= \inf\{x \in \mathbb{R} \; / \; F(x) \geq 0.5\})$.

Cas $F^{-1}(\{0.5\}) \neq \emptyset$, c'est-à-dire cas n pair avec $a < b$
On a alors :
$card(I) > 1$ et $\exists i_0 \in [1, I - 1]$ tel que $F^{-1}(\{0.5\}) = [x_{i_0}, x_{i_0+1}[$.
On peut en déduire : $M(X) = [x_{i_0}, \; x_{i_0+1}]$ $(= [a, b])$.

Cas d'une variable réelle continue
La fonction de répartition F étant continue et croissante sur $\mathbb{R}$ et d'image $[0, 1]$, on en déduit :

$$M(X) = F^{-1}(\{0.5\})$$

et que cet ensemble est non vide.

La fonction F étant affine par morceaux, on peut envisager deux cas.

Cas où $F^{-1}(\{0.5\})$ est un singleton.

Il existe $i_0 \in [1, I]$ tel que $F(x_{i_0-1}) \leq 0.5 < F(x_{i_0})$, et l'application de $[x_{i_0-1}, x_{i_0}]$ vers $[F(x_{i_0-1}), F(x_{i_0})]$ qui à x associe $F(x)$ est bijective.

Il existe donc un unique réel m de $[x_{i_0-1}, x_{i_0}]$ tel que $F(m) = 0.5$, c'est l'unique médiane de X.

On déduit la valeur de m de l'équation :

$$\frac{m - x_{i_0-1}}{x_{i_0} - x_{i_0-1}} = \frac{0.5 - F(x_{i_0-1})}{F(x_{i_0}) - F(x_{i_0-1})}$$

(détermination de la médiane par interpolation linéaire).

Cas où $F^{-1}(\{0.5\})$ est un intervalle fermé non réduit à un singleton.

On a alors $card(I) > 2$ et il existe i_0 et i_1 de $[1, I]$ tels que $1 \leq i_0 < i_1 \leq I - 1$, $F(x_{i_0-1}) < 0.5$, $F(x_i) = 0.5$ pour $i \in [i_0, i_1]$ et $F(x_{i_1+1}) > 0.5$. alors $M(X) = F^{-1}(\{0.5\}) = [x_{i_0}, x_{i_1}]$. Une telle situation signifie qu'aucun individu ne prend une valeur comprise entre x_{i_0} et x_{i_1}.

Pour avoir unicité, on peut convenir de prendre pour médiane la valeur centrale de l'intervalle.

3.3.3 Quartiles et écart inter-quartiles

Les *quartiles* partagent la population en quatre sous-populations d'effectifs égaux. Plus précisément :

q_1 est *premier quartile* si au moins 25% de la population prend une valeur inférieure ou égale à q_1 et au moins 75% une valeur supérieure ou égale à q_1.

q_2 est *deuxième quartile* si au moins 50% de la population prend une valeur inférieure ou égale à q_2 et au moins 50% une valeur supérieure ou égale à q_2 .

q_3 est *troisième quartile* si au moins 75% de la population prend une valeur inférieure ou égale à q_3 et au moins 25% une valeur supérieure ou égale à q_3.

En utilisant la fonction de répartition, l'ensemble $Q_1(X)$ des premiers quartiles est :

$$Q_1(X) = \{q_1 \in \mathbb{R} \; ; \; \lim_{\substack{x \to q_1 \\ <}} F(x) \leq 0.25 \leq F(q_1)\},$$

l'ensemble $Q_3(X)$ des troisièmes quartiles est :

$$Q_3(X) = \{q_3 \in \mathbb{R} \; ; \; \lim_{\substack{x \to q_3 \\ <}} F(x) \leq 0.75 \leq F(q_3)\},$$

enfin, l'ensemble $Q_2(X)$ des deuxièmes quartiles n'est autre que l'ensemble $M(X)$ des médianes.

Cas d'une variable discrète

Si la taille de la population n'est pas un multiple de 4, $n = 4k+l$, k entier positif, $1 \leq l \leq 3$, alors le premier (resp. le troisième) quartile est unique et correspond à la valeur prise par le $(k+1)$*ème* individu lorsque les valeurs de la variable sont rangées dans l'ordre croissant (resp. décroissant).

Si la taille de la population est un multiple de 4, $n = 4k$, k entier positif, soit a et b (resp. b' et a') les valeurs prises par les k*ème* et $(k+1)$*ème* individus lorsque les valeurs de la variable sont rangées dans l'ordre croissant (resp. décroissant), alors tout réel de $[a, b]$ (resp. $[a', b']$) est premier (resp. troisième) quartile. Pour avoir unicité, on convient de prendre pour premier (resp. troisième) quartile la valeur $(a+b)/2$ (resp. $(a'+b')/2$).

Cas d'une variable continue

La fonction de répartition F étant une fonction continue, on en déduit : $Q_1(X) = F^{-1}(\{0.25\})$ (resp. $Q_3(X) = F^{-1}(\{0.75\})$).

Si $Q_1(X)$ (resp. $Q_3(X)$) est un singleton, le premier (resp. troisième) quartile peut être déterminé par interpolation linéaire comme indiqué ci-dessus pour la médiane.

Sinon, il s'agit d'un intervalle fermé correspondant à une classe vide (ou plusieurs). Pour avoir unicité, on prend pour premier (resp. troisième) quartile le centre de l'intervalle.

Lorsque q_1 et q_3 sont définis de manière unique, *l'intervalle interquartiles* est l'intervalle $[q_1, q_3\,]$ et *l'écart interquartiles* est le réel $q_3 - q_1$.

Déciles, écart interdéciles

Les *déciles* partagent la population en dix sous-populations d'effectifs égaux.

Plus précisément, on a pour les premier et neuvième déciles, la définition se généralisant aisément aux autres déciles :

q_1 est *premier décile* si au moins 10% de la population prend une valeur inférieure ou égale à d_1 et au moins 90% une valeur supérieure ou égale à d_1.

q_9 est *neuvième décile* si au moins 90% de la population prend une valeur inférieure ou égale à d_9 et au moins 10% une valeur supérieure ou égale à d_9.

Cas d'une variable discrète

Si la taille de la population n'est pas un multiple de 10, $n = 10k + l$, k entier positif, $1 \leq l \leq 9$, alors le premier (resp. le neuvième) décile est unique et correspond à la valeur prise par le $(k+1)$*ème* individu lorsque les valeurs de la variable sont rangées dans l'ordre croissant (resp. décroissant).

Si la taille de la population est un multiple de 10, $n = 10k$, k entier positif, soit a et b (resp. b' et a') les valeurs prises par les k*ème* et $(k+1)$*ème* individus lorsque les valeurs de la variable sont rangées dans l'ordre croissant (resp. décroissant), alors tout réel de $[a, b]$ (resp. $[a', b']$) est premier (resp. neuvième) décile. Pour avoir unicité, on convient de prendre pour premier (resp. neuvième) décile la valeur $(a + b)/2$ (resp. $(a' + b')/2$).

Cas d'une variable continue

L'ensemble des premiers (resp. neuvièmes) déciles est $D_1(X) = F^{-1}(\{0.1\})$ (resp. $D_9(X) = F^{-1}(\{0.9\})$).

On procède alors comme pour les quartiles selon que l'ensemble est un singleton ou un intervalle fermé.

Lorsque d_1 et d_9 sont définis de manière unique, *l'intervalle interdéciles* est l'intervalle $[d_1, d_9\,]$ et *l'écart interdéciles* est le réel $d_9 - d_1$.

Diagramme en boîte

Une représentation graphique s'appuyant sur les quartiles résume l'histogramme et permet de comparer plusieurs distributions. Il s'agit du diagramme "*boîte et moustaches*" ("box-plot" en anglais).

La boîte, rectangulaire de largeur arbitraire, est limitée en longueur par le premier et troisième quartile ; à l'intérieur de la boîte est indiquée la médiane et, de part et d'autre de la boîte, des segments (les "moustaches") représentent les valeurs extérieures à l'intervalle interquartiles, les extrémités de ces segments indiquent les valeurs extrêmes de la variable. (Les extrémités des segments peuvent aussi correspondre aux premier et neuvième déciles.)

Propriétés

Soit X une variable réelle, dont on note m, q_1, q_3, d_1 et d_9 la médiane, les premier et troisième quartiles, les premier et neuvième déciles (définis éventuellement de manière unique en utilisant la convention présentée ci-dessus) et donc $(q_3 - q_1)$ et $(d_9 - d_1)$ les écarts interquartiles et interdéciles.

Soit $Y = aX + b$, $a \neq 0$, alors la médiane, les premier et troisième quartiles, les premier et neuvième déciles de Y sont respectivement $am + b$, $aq_1 + b$, $aq_3 + b$, $ad_1 + b$ et $ad_9 + b$ si $a > 0$, $am + b$, $aq_3 + b$, $aq_1 + b$, $ad_9 + b$ et $ad_1 + b$ si $a < 0$. On en déduit que les écarts interquartiles et interdéciles de Y sont respectivement $|a|(q_3 - q_1)$ et $|a|(d_9 - d_1)$. La preuve est évidente puisque seul l'ordre des valeurs intervient dans les définitions.

Soit $x_1 \leq x_2 \leq ... \leq x_n$ les valeurs d'une variable réelle X définie sur une population de taille n, rangées dans l'ordre croissant. L'ensemble des réels x qui minimisent $f(x) = \sum_{j=1}^{n} |x_j - x|$ est égal à l'ensemble des médianes.

En effet, on a :

$f(x) = \sum_{j=1}^{n} x_j - nx$, si $x \leq x_1$

$f(x) = -\sum_{j=1}^{k} x_j + \sum_{j=k+1}^{n} x_j - (n-2i)x$, si $x_i \leq x \leq x_{i+1}$ pour $1 \leq i \leq n-1$

$f(x) = -\sum_{j=1}^{n} x_j + nx$, si $x \geq x_n$.

Si n est impair, $n = 2k + 1$, la fonction est strictement décroissante pour $x \leq x_{k+1}$ et strictement croissante pour $x \geq x_{k+1}$, le minimum est donc x_{k+1}.

Si n est pair, $n = 2k$, la fonction est strictement décroissante pour $x \leq x_k$, elle est constante pour $x_k \leq x \leq x_{k+1}$, elle est strictement croissante pour $x \geq x_{k+1}$, toute valeur de $[x_k, x_{k+1}]$ minimise $f(x)$.

L'ensemble des réels qui minimisent $\sum_{i=1}^{n} |x_i - x|$ est l'ensemble des médianes de la série statistique.

3.3.4 Moyenne, variance et écart-type

Dans le cas d'une *variable réelle discrète*, la *moyenne (arithmétique)* $\overline{X}$ de la variable X est définie par :

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\sum_{i=1}^{I} n_i x_i = \sum_{i=1}^{I} f_i x_i$$

La première expression est utilisée lorsque l'on travaille sur le tableau initial des données "individus x variables", la seconde et la troisième lorsqu'on

utilise la distribution des effectifs et la distribution des fréquences de la variable X. On remarquera que la moyenne est exprimée dans la même unité de mesure que la variable X.

La *variance* $\text{var}(X)$ de la variable X est la moyenne des carrés des écarts à la moyenne, c'est-à-dire :

$$\text{var}(X) = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{I} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{I} f_i(x_i - \overline{X})^2$$

Il est aisé de vérifier, en développant les carrés, que la variance est aussi égale à la moyenne des carrés moins le carré de la moyenne, c'est-à-dire :

$$\text{var}(X) = \frac{1}{n}\sum_{k=1}^{n} {X_k}^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{I} n_i x_i^2 - \overline{X}^2 = \sum_{i=1}^{I} f_i x_i^2 - \overline{X}^2$$

On remarquera que la variance de X est mesurée dans l'unité de mesure de la variable élevée au carré. On définit alors l'*écart-type* de X, noté $s(X)$, comme la racine carrée de la variance de X :

$$s(X) = \sqrt{\text{var}(X)}$$

L'écart-type de X est mesuré dans la même unité de mesure que X.

Dans le cas d'une *variable continue*, l'hypothèse de distribution uniforme dans chaque classe implique que la moyenne des valeurs prises par les individus d'une classe est la valeur centrale de la classe.
Aussi, les calculs de moyenne, variance et écart-type se font sur les centres de classes c_i ($c_i = (x_{i-1} + x_i)/2$), donc sur les distributions d'effectifs et de fréquences $\{(c_i, n_i)\ ;\ i \in [1, I]\}$ et $\{(c_i, f_i)\ ;\ i \in [1, I]\}$ respectivement.

Finalement pour les calculs de moyenne et variance, on remplace la variable réelle initiale pouvant prendre toute valeur d'un intervalle donné par la variable réelle discrète C dont les distributions d'effectifs et de fréquences sont $\{(c_i, n_i)\ /\ i \in [1, I]\}$ et $\{(c_i, f_i)\ /\ i \in [1, I]\}$ respectivement. Aussi, il est courant de dire que le regroupement des valeurs de la variable en classes revient à *discrétiser* la variable initiale (potentiellement continue).

Propriétés de la moyenne, de la variance et de l'écart-type

Si on pose $Z = aX + b$, avec a et b réels, alors on a :

$$\overline{Z} = a\overline{X} + b, \quad \text{var}(Z) = a^2\text{var}(X) \quad \text{et} \quad s(Z) = |a|\, s(X).$$

On a de plus : X constante $\Leftrightarrow \text{var}(X) = 0 \Leftrightarrow s(X) = 0$.

Soit $X_1, ..., X_n$ les valeurs prises par une variable réelle X sur les n individus d'une population, alors l'ensemble des réels x qui minimisent : $f(x) = \sum_{k=1}^{n}(X_k - x)^2$ est réduit au singleton $\{\overline{X}\}$

En effet, il est aisé de vérifier l'égalité, connue sous le nom de formule de Huyghens :

$\sum_{k=1}^{n}(X_k - x)^2 = \sum_{k=1}^{n}(X_k - \overline{X})^2 + (\overline{X} - x)^2$

On en déduit : $\forall x \in \mathbb{R},\ f(x) \geq \sum_{k=1}^{n}(X_k - \overline{X})^2 = n\,\text{var}(X)$.

Le minimum de $\sum_{k=1}^{n}(X_k - x)^2$ est donc obtenu pour $x = \overline{X}$ et le minimum est égal à n fois la variance de X.

Variables centrées, réduites, centrées et réduites

Une variable dont la moyenne est nulle est dite *variable centrée.* Une variable dont l'écart-type est égal à 1 est dite *variable réduite.* Soit X une variable non constante, $\overline{X}$ sa moyenne, $s(X)$ son écart-type, alors la variable $X - \overline{X}$ est une variable centrée, appelée *variable X centrée*, la variable $X/s(X)$ est une variable réduite, appelée *variable X réduite* et la variable $(X - \overline{X})/s(X)$ est une variable centrée et réduite, appelée *variable X centrée réduite.*

4 Traitement statistique d'une variable catégorielle et d'une variable réelle

Soit $\Omega = \{\omega_1, ..., \omega_n\}$ une population de n individus, A une variable catégorielle à I modalités définie sur Ω, $(A_i)_{i=1,...,I}$ la partition de Ω engendrée par cette variable, et X une variable réelle définie sur Ω.

On rappelle que la moyenne et variance de X sont définies par :

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\sum_{\omega\in\Omega} X(\omega)$$

et

$$\text{var}(X) = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n}\sum_{\omega\in\Omega}(X(\omega) - \overline{X})^2.$$

4.1 Décomposition de la moyenne et de la variance sur une partition

Pour tout i de $[1, I]$, soit n_i l'effectif de A_i, $\overline{X}_i$ et var_i la moyenne et variance de X sur A_i, c'est-à-dire :

$$\overline{X}_i = \frac{1}{n_i} \sum_{\omega \in A_i} X(\omega) \quad \text{et } \text{var}_i = \frac{1}{n_i} \sum_{\omega \in A_i} (X(\omega) - \overline{X}_i)^2.$$

Propriété :

La moyenne $\overline{X}$ est égale à la moyenne des moyennes (pondérées par les effectifs des catégories), c'est-à-dire :

$$\overline{X} = \sum_{i=1}^{I} \frac{n_i}{n} \overline{X}_i.$$

En effet, on a :
$\overline{X} = \frac{1}{n} \sum_{\omega \in \Omega} X(\omega) = \frac{1}{n} \sum_{i=1}^{I} \sum_{\omega \in A_i} X(\omega) = \frac{1}{n} \sum_{i=1}^{I} n_i \overline{X}_i = \sum_{i=1}^{I} \frac{n_i}{n} \overline{X}_i.$

Définitions :

La *variance intra* (ou intra-catégories, ou intra-groupes, ou intra-classes) est la moyenne des variances (pondérées par les effectifs des catégories) et la *variance inter* (ou inter-catégories, ou inter-groupes, ou inter-classes) est la variance des moyennes (pondérées par les effectifs des catégories), soit :

$$\text{var}_{intra} = \sum_{i=1}^{I} \frac{n_i}{n} \text{var}_i \quad \text{et } \text{var}_{inter} = \sum_{i=1}^{I} \frac{n_i}{n} (\overline{X}_i - \overline{X})^2.$$

Propriété :

La variance $\text{var}(X)$ est égale à la somme de la variance inter et de la variance intra, c'est-à-dire :

$$\text{var}(X) = \text{var}_{intra} + \text{var}_{inter}.$$

En effet, on a :
$\text{var}(X) = \frac{1}{n} \sum_{\omega \in \Omega} (X(\omega) - \overline{X})^2 = \frac{1}{n} \sum_{i=1}^{I} \sum_{\omega \in A_i} (X(\omega) - \overline{X}_i + \overline{X}_i - \overline{X})^2$
$= \frac{1}{n} \sum_{i=1}^{I} \sum_{\omega \in A_i} [(X(\omega) - \overline{X}_i)^2 + 2(X(\omega) - \overline{X}_i)(\overline{X}_i - \overline{X}) + (\overline{X}_i - \overline{X})^2]$
$= \frac{1}{n} \sum_{i=1}^{I} \sum_{\omega \in A_i} (X(\omega) - \overline{X}_i)^2 + \frac{2}{n} \sum_{i=1}^{I} \sum_{\omega \in A_i} (X(\omega) - \overline{X}_i)(\overline{X}_i - \overline{X}) +$
$\frac{1}{n} \sum_{i=1}^{I} \sum_{\omega \in A_i} (\overline{X}_i - \overline{X})^2.$

Comme on a : $\sum_{\omega \in A_i} (X(\omega) - \overline{X}_i) = 0$, le deuxième terme est nul et on en déduit :

$$\text{var}(X) = \frac{1}{n} \sum_{i=1}^{I} n_i \text{var}_i + \frac{1}{n} \sum_{i=1}^{I} n_i (\overline{X}_i - \overline{X})^2 = \text{var}_{intra} + \text{var}_{inter}.$$

4.2 Résumé numérique : rapport de corrélation

La variable catégorielle A et la variable réelle X sont liées si, dans chaque catégorie de la partition associée à A, les individus sont homogènes vis à vis de la variable X (variances intra-catégories faibles).

On pose :

$$\eta = \sqrt{\frac{\text{var}_{inter}}{\text{var}(X)}}$$

Cet indice (prononcer eta), appelé *rapport de corrélation* entre A et X, est compris entre 0 et 1.

Il est égal à 1 si $\text{var}_{inter} = \text{var}(X)$, c'est-à-dire si var_{intra} est nul, c'est-à-dire si $\forall i \in [1, I]$, $\text{var}_i = 0$ (dans chaque catégorie, tous les individus prennent la même valeur).

Il est égal à 0 si $\text{var}_{inter} = 0$, c'est-à-dire si $\forall i \in [1, I]$, $\overline{X}_i = \overline{X}$ (dans chaque catégorie, les individus sont hétérogènes vis à vis de la variable X, certains prennent des faibles valeurs de X, d'autres des fortes).

5 Traitement statistique de deux variables réelles

Soit $\Omega = \{\omega_1, ..., \omega_n\}$ une population sur laquelle sont définies deux variables réelles X et Y.

Soit $(X, Y)(\Omega) = \{(X(\omega), Y(\omega))\,;\, \omega \in \Omega\}$ l'image de Ω par (X, Y).

5.1 Distributions d'effectifs et de fréquences

Si $\{(x_i, y_i)\,;\, i \in [1, I]\}$ désigne l'ensemble des couples distincts de $(X, Y)(\Omega)$, soit :

$$n_i = card\{\omega \in \Omega\,;\, X(\omega) = x_i \text{ et } Y(\omega) = y_i\} \text{ et } f_i = n_i/n.$$

La distribution d'effectifs (resp. de fréquences) de (X, Y) est

$$\{((x_i, y_i), n_i)\,;\, i \in [1, I]\} \text{ (resp. } \{((x_i, y_i), f_i)\,;\, i \in [1, I]\}).$$

5.2 Représentation graphique : le graphe plan

Dans un plan ramené à un repère orthogonal, on représente les points de coordonnées (x_i, y_i), en indiquant à côté du point l'effectif n_i correspondant lorsqu'il est différent de 1. On obtient ce qu'on appelle un *nuage de points*. On choisit les unités des axes ainsi que les supports des variables X et Y de façon à ce que le nuage de points soit bien dispersé et centré sur le graphique.

Le point moyen, souvent noté G, pour centre de gravité du nuage de points pondérés par les effectifs, a pour coordonnées $(\overline{X}, \overline{Y})$.

5.3 Résumés numériques : la covariance, le coefficient de corrélation linéaire

Covariance

La *covariance des variables X et Y* est définie comme la moyenne des produits des différences à la moyenne, c'est-à-dire :

$$\begin{gathered}\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y}) \\ = \frac{1}{n}\sum_{i=1}^{I} n_i(x_i - \overline{X})(y_i - \overline{Y}) = \sum_{i=1}^{I} f_i(x_i - \overline{X})(y_i - \overline{Y})\end{gathered}$$

La première expression est utilisée lorsque l'on travaille sur le tableau initial des données "individus x variables", la seconde et la troisième lorsqu'on utilise la distribution des effectifs et la distribution des fréquences du couple de variables (X, Y).

L'unité de mesure de la covariance est le produit des unités de mesure de X et de Y.

Il est aisé de vérifier, en développant le produit, que la covariance est aussi égale à la moyenne des produits moins le produit des moyennes, c'est-à-dire :

$$\begin{gathered}\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{k=1}^{n} X_k Y_k - \overline{X}\ \overline{Y} \\ = \frac{1}{n}\sum_{i=1}^{I} n_i x_i y_i - \overline{X}\ \overline{Y} = \sum_{i=1}^{I} f_i x_i y_i - \overline{X}\ \overline{Y}.\end{gathered}$$

Propriétés :
Soit Z une autre variable réelle définie sur Ω,

$$\begin{gathered}\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X) \\ \forall (a,b) \in R^2, \\ \mathrm{cov}(aX + bY, Z) = a\mathrm{cov}(X,Z) + b\mathrm{cov}(Y,Z) \text{ et} \\ \mathrm{cov}(Z, aX + bY) = a\mathrm{cov}(Z,X) + b\mathrm{cov}(Z,Y),\end{gathered}$$

$$\text{cov}(X, X) = \text{var}(X) \geq 0,$$
$$\text{var}(X) = 0 \text{ si, et seulement si, } X \text{ est constante,}$$
$$\text{cov}(X, Y) = 0 \text{ si } X \text{ ou } Y \text{ est constante.}$$

La preuve est évidente.

L'ensemble des variables réelles définies sur Ω, muni de l'addition et de la multiplication externe par un réel est un espace vectoriel, dont l'ensemble des variables centrées est un sous-espace vectoriel.

La covariance est alors une forme bilinéaire, symétrique définie positive, c'est-à-dire, un produit scalaire sur l'espace vectoriel des variables centrées.

Coefficient de corrélation linéaire

Dans le cas où les variables X et Y sont non constantes, on définit *le coefficient de corrélation linéaire de X et Y*, noté $r(X, Y)$, par le rapport de la covariance sur le produit des écart-types de X et de Y, c'est-à-dire :

$$r(X, Y) = \frac{\text{cov}(X,Y)}{s(X)s(Y)}.$$

C'est un indice sans unité de mesure.

Propriétés

$$r(X, Y) = r(Y, X)$$
$$-1 \leq r(X, Y) \leq 1$$
$$r(X, Y) = 1 \Leftrightarrow \exists (a, b) \in R_+^* \times R,\ Y = aX + b$$
$$r(X, Y) = -1 \Leftrightarrow \exists (a, b) \in R_-^* \times R,\ Y = aX + b$$

Preuve
On pose :

$$\forall a \in R,\ \text{var}(Y - aX) = a^2\text{var}(X) - 2a\text{cov}(X, Y) + \text{var}(Y)\ (= f(a))$$

Une variance étant positive, le trinôme du second degré $f(a)$ est positif pour toute valeur de a. On en déduit que le discriminant est négatif ou nul :

$$[\text{cov}(X, Y)]^2 - \text{var}(X)\text{var}(Y) \leq 0,$$

c'est-à-dire,

$$|r(X, Y)| \leq 1,$$
$$-1 \leq r(X, Y) \leq 1.$$

Par ailleurs, $f(a)$ atteint son minimum pour :

$$a = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

et, pour cette valeur de a, on a :

$$f(a) = \frac{\text{var}(X)\text{var}(Y) - [\text{cov}(X,Y)]^2}{\text{var}(X)}.$$

On en déduit :

$$\begin{gathered}\text{var}(Y - aX) = 0 \Leftrightarrow \exists b \in R,\ Y = aX + b \\ \Leftrightarrow \text{var}(X)\text{var}(Y) = [\text{cov}(X,Y)]^2 \Leftrightarrow |r(X,Y)| = 1.\end{gathered}$$

On peut préciser alors :

$$\begin{gathered}r(X,Y) = 1 \Leftrightarrow Y = aX + b,\ a > 0, \\ r(X,Y) = -1 \Leftrightarrow Y = aX + b,\ a < 0.\end{gathered}$$

5.4 Droites de régression linéaire ; prévisions

Soit X et Y deux variables réelles non constantes définies sur une population (ou un échantillon) $\Omega = \{\omega_1, ..., \omega_n\}$, X_k et Y_k les valeurs prises par X et Y sur ω_k, $k \in [\![1, n]\!]$.

5.4.1 Droite de régression linéaire de Y en X

On s'intéresse à un modèle de la forme :

$$Y = aX + b,$$

X est la variable explicative, Y la variable à expliquer, a et b les paramètres à estimer.

Il s'agit de trouver la droite, d'équation $y = ax + b$, "la plus proche" (dans un sens à déterminer) du nuage des points de coordonnées (X_k, Y_k), $k \in [\![1, n]\!]$.

Pour cela, on introduit la variable des erreurs : $E = Y - aX - b$ (c'est-à-dire, $E_k = Y_k - aX_k - b$, $k \in [\![1, n]\!]$) et on convient de déterminer les coefficients a et b qui minimisent la moyenne (ou la somme) des carrés des erreurs (critère des moindres carrés des erreurs) :

$$f(a, b) = \frac{1}{n}\sum_{k=1}^{n}(Y_k - aX_k - b)^2.$$

La solution est la suivante :

$$a = \frac{\text{cov}(X,Y)}{\text{var}(X)} \text{ et } b = \overline{Y} - a\overline{X}.$$

En effet, on a :
$f(a,b) = \frac{1}{n}\sum_{k=1}^{n}(Y_k - aX_k - b)^2 = \frac{1}{n}\sum_{k=1}^{n}(Y_k - \overline{Y} - a(X_k - \overline{X}) + \overline{Y} - a\overline{X} - b)^2$
$= \frac{1}{n}\sum_{k=1}^{n}[Y_k - \overline{Y} - a(X_k - \overline{X})]^2 + \frac{2}{n}\sum_{k=1}^{n}[Y_k - \overline{Y} - a(X_k - \overline{X})](\overline{Y} - a\overline{X} - b) + (\overline{Y} - a\overline{X} - b)^2$
$= \frac{1}{n}\sum_{k=1}^{n}[Y_k - \overline{Y} - a(X_k - \overline{X})]^2 + (\overline{Y} - a\overline{X} - b)^2,$
le second terme étant nul.

Le minimum est obtenu lorsque le deuxième terme est nul et le premier minimum, c'est-à-dire, $b = \overline{Y} - a\overline{X}$, et a minimisant :

$$g(a) = a^2\text{var}(X) - 2a\text{cov}(X,Y) + \text{var}(Y),$$

d'où la solution.

La droite d'équation solution de ce critère : $y = ax + b$, avec $a = \frac{cov(X,Y)}{var(X)}$ et $b = \overline{Y} - a\overline{X}$, est appelée droite de régression linéaire de Y en X.

On pose :

$$\widehat{Y} = \frac{\text{cov}(X,Y)}{\text{var}(X)}(X - \overline{X}) + \overline{Y} \text{ et } \widehat{E} = Y - \widehat{Y},$$

la variable $\widehat{Y}$ est le modèle ajusté et la variable $\widehat{E}$ la variable des résidus (ou des erreurs d'ajustement).

On vérifie que $\widehat{E}$ est une variable centrée dont la variance est la valeur de $f(a,b)$ au minimum, soit : $\text{var}(\widehat{E}) = \text{var}(Y)(1 - r^2)$.

Comme on a : $\text{var}(\widehat{Y}) = \text{var}(Y)r^2$, on en déduit :

$$\text{var}(Y) = \text{var}(\widehat{Y}) + \text{var}(\widehat{E}) \text{ et } \text{cov}(\widehat{Y}, \widehat{E}) = 0.$$

La part de variance de Y expliquée par le modèle : $\text{var}(\widehat{Y})/\text{var}(Y)$ est égale au carré du coefficient de corrélation linéaire de X et de Y.

On peut utiliser le modèle ajusté :

$$\widehat{Y} = \frac{\text{cov}(X,Y)}{\text{var}(X)}(X - \overline{X}) + \overline{Y}$$

pour prévoir les valeurs de Y correspondant à des valeurs de X non encore observées. On parle *d'interpolation* lorsque x est compris entre le minimum et le maximum des valeurs prises par X, *d'extrapolation* dans le cas contraire.

5.4.2 Droite de régression linéaire de X en Y

Si l'on considère à présent le modèle : $X = a'Y + b'$, on montre, en échangeant les rôles de X et Y, que la droite de régression linéaire de X en Y, dont on peut écrire l'équation sous la forme $x = a'y + b'$, a pour coefficients

$$a' = \frac{\text{cov}(X,Y)}{\text{var}(Y)} \text{ et } b' = \overline{X} - a'\overline{Y}.$$

Si l'on représente sur le même graphique les deux droites de régression linéaire d'équations respectives :

$$y = ax + b \text{ et } y = \frac{1}{a'}x - \frac{b'}{a'},$$

elles se coupent au centre de gravité, elles sont toutes les deux croissantes (si $\text{cov}(X,Y) > 0$) ou toutes les deux décroissantes (si $\text{cov}(X,Y) < 0$) ; enfin, elles sont confondues si $r = 1$ ou -1 (car $aa' = r^2$).

5.4.3 Application à la détermination de la tendance linéaire d'une série chronologique

Une série chronologique $(Y_t)_{t=1,...,n}$ est une variable réelle dont les unités statistiques sont les instants. Ces instants peuvent être considérés comme les valeurs d'une variable réelle X (en posant $X_t = t$, pour $t = 1, ..., n$).

Lorsque les différences $Y_t - Y_{t-1}$ sont relativement constantes : $Y_t - Y_{t-1} = a$, on est amené à considérer un modèle linéaire de la forme : $Y_t = at + b$ avec $b = Y_0$.

Les coefficients a et b de la droite de régression de Y par rapport au temps sont alors :

$$a = \frac{\text{cov}(X,Y)}{\text{var}(X)} \text{ et } b = \overline{Y} - a\overline{X}$$

avec

$$\overline{X} = (n+1)/2 \text{ et } \text{var}(X) = (n^2 - 1)/12.$$

En effet, pour la variable temps, on utilise les égalités :

$$\sum_{t=1}^{n} t = n(n+1)/2 \text{ et } \sum_{t=1}^{n} t^2 = n(n+1)(2n+1)/6.$$

Pour une variable Y strictement positive, lorsque les rapports Y_t/Y_{t-1} sont relativement constants : $Y_t/Y_{t-1} = a$, on est amené à considérer un modèle de la forme : $Y_t = a^t\ b$ avec $b = Y_0$, dit *modèle exponentiel.*

On fait alors un changement de variable en posant $Z_t = \ln Y_t$. On pose alors aussi $A = \ln a$ et $B = \ln b$ et on se ramène au modèle linéaire $Z_t = At + B$.

Les coefficients A et B de la droite de régression linéaire de Z par rapport au temps sont alors :

$$A = \frac{\mathrm{cov}(X, Z)}{\mathrm{var}(X)} \text{ et } B = \overline{Z} - a\overline{X}$$

avec

$$\overline{X} = (n+1)/2 \text{ et } var(X) = (n^2 - 1)/12.$$

On en déduit les coefficients a et b :

$$a = \exp A \text{ et } b = \exp B.$$