ESSAI SUR LA STATISTIQUE

Samir Aït-Amrane

RÉSUMÉ : Dans cet article, nous allons expliquer quelques notions de base de la statistique, d'abord dans le cas d'une variable, ensuite dans le cas de deux variables, en organisant les idées et en faisant un parallèle entre certaines formules statistiques et probabilistes qui se ressemblent. Nous disons également un petit mot sur l'économétrie, les séries temporelles et les processus stochastiques et nous proposons quelques références bibliographiques où ces notions sont bien expliquées.

INTRODUCTION

Dans cet article, nous allons expliquer quelques notions de base de la statistique, en faisant un parallèle entre certaines formules que l'on rencontre en statistique et des formules semblables que l'on rencontre en théorie des probabilités, en organisant les idées et en indiquant au fur et à mesure quelques références bibliographiques où ces notions sont bien expliquées.

Dans la première section, nous rappelons les formules qui concernent une variable ainsi que certaines notions comme l'échantillonnage, l'estimation, la loi des grands nombres et le théorème central limite. Dans l'exemple 4, qui achève cette première section, nous nous intéressons à l'estimateur naturel de la moyenne qui possède toutes les meilleures propriétés : il est sans biais, consistant, il est aussi l'estimateur des moindres carrés ordinaires et le plus efficace parmi une certaine catégorie d'estimateurs. En passant, nous essayons d'expliquer en quelques mots pourquoi une bonne moitié de mathématiciens trouvent que les vérités mathématiques on ne les invente pas mais on les découvre.

Dans la deuxième section, nous rappelons les formules qui concernent deux variables ainsi que quelques notions autour de la régression linéaire, en disant un petit mot sur l'économétrie puisque nous nous intéressons à des variables économiques dans les exemples traités. Nous verrons par exemple que les économètres ont des tours de passe passe pour transformer une variable aléatoire en une variable déterministe et inversement (exemples 5 et 6). Enfin, nous disons un petit mot sur les séries temporelles et les processus stochastiques et nous terminons par un exemple sur le modèle d'équilibre des actifs financiers (MEDAF) en faisant le parallèle entre les formules statistiques et probabilistes qui se ressemblent.

1. STATISTIQUE UNIVARIÉE

La statistique est une discipline qui consiste à étudier n'importe quel caractère, ou variable, associé aux éléments d'une population. La **statistique descriptive** consiste à décrire et à résumer les données obtenues à l'aide de tableaux et de graphiques, selon le type de données et la taille de la population, ou encore en mesurant les tendances centrales et de dispersion (voir [Sim] chap. 1 et [Ive] chap. 3).

Considérons une population à n individus et soient $x_1,\cdots,x_n$ les valeurs prises par une variable quantitative. La **moyenne** arithmétique $\bar{x}$ de la **série statistique** $x_1,\cdots,x_n$ est définie par :

$$\bar{x} = \frac{x_1+\cdots+x_n}{n} \tag{1}$$

Les autres mesures de tendance centrale sont décrites dans [Do] au chapitre 4 ou dans [Joh] au chapitre 2. La **variance** s_x^2 de $x_1,\cdots,x_n$ est définie par :

$$s_x^2 = \frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n} \tag{2}$$

La racine carrée de la variance s_x s'appelle **l'écart-type** de $x_1,\cdots,x_n$. Les autres mesures de dispersion sont décrites dans [Do] au chapitre 5 ou dans [Joh] au chapitre 3.

Exemple 1 : Considérons le problème de sondage avec réponses binaires, comme par exemple celui qui consiste à déterminer le pourcentage p d'une population qui répond par « oui » à une certaine question. Il s'agit d'un caractère qualitatif puisque les données observées, à savoir les réponses « oui » et « non », ne sont pas numériques. On peut transformer ces observations en des données numériques en associant à chaque individu qui répond par « oui » le nombre 1 et à chaque individu qui répond par « non » le nombre 0. On appelle cela une **transformation de données** (voir [Do] section 2.4.1). Si la population est de taille n, soient $x_1, \cdots, x_n$ les valeurs prises par la variable quantitative obtenue. Le nombre d'individus qui répondent par « oui » n'est rien d'autre que $x_1 + \cdots + x_n$, et la moyenne arithmétique p de $x_1, \cdots, x_n$:

$$p = \frac{x_1 + \cdots + x_n}{n}$$

est le pourcentage de l'ensemble des individus de la population qui répondent par « oui », et qu'on appelle la **proportion de succès**. □

Si la population est de taille importante, il peut être difficile ou impossible de sonder tous les individus de la population. Dans une telle situation, il faudrait sélectionner un échantillon de taille raisonnable qui soit le plus **représentatif** possible de la population et utiliser la proportion de succès au sein de l'échantillon afin d'estimer la proportion de succès au sein de la population. C'est l'objet de la **statistique inférentielle**, qui consiste à faire de la statistique descriptive au niveau de l'échantillon, et à généraliser les résultats obtenus à la population.

Exemple 2 : Supposons que notre problème de sondage avec réponses binaires concerne une population de taille importante N et soit p la proportion de succès que nous cherchons à estimer. Sélectionner un individu au hasard au sein de la population est une expérience aléatoire dont l'ensemble fondamental Ω est l'ensemble des individus de la population. Soit X la variable aléatoire qui associe à chaque individu qui répond par « oui » le nombre 1 et à chaque individu qui répond par « non » le nombre 0.

Si chaque individu de la population est tiré avec la même probabilité $(1/N)$, alors la probabilité de l'événement $(X = 1)$ est égale à p et la probabilité de l'événement $(X = 0)$ est égale à $1 - p$. La variable aléatoire X obéit donc à la loi de Bernoulli de paramètre p. S'il est impossible de sonder tous les individus de la population alors il est également impossible de trouver la valeur exacte du paramètre p de la population. Il faudrait alors sélectionner un échantillon représentatif de la population, calculer la proportion de succès au sein de l'échantillon et utiliser la valeur obtenue comme valeur estimative du paramètre p. □

Considérons maintenant une population quelconque de taille N et soit $\Omega = \{\omega_1, \ldots, \omega_N\}$ l'ensemble des individus de cette population. Soit X un caractère quantitatif associé à cette population et supposons les tirages équiprobables. Comme expliqué dans l'exemple 2, on peut considérer X comme une variable aléatoire définie sur Ω qui associe à chaque individu ω_i la valeur observée $X(\omega_i)$ de ce caractère pour $i = 1, \ldots, N$. L'espérance μ et la variance σ^2 de la variable aléatoire X (voir [Rom] sections 1.6 et 1.7) :

$$\mu = E(X) = \sum_{x \in \mathrm{Im}(X)} xP(X = x) = \frac{1}{N}\sum_{i=1}^{N} X(\omega_i) \qquad (1')$$

et

$$\sigma^2 = Var(X) = E((X - \mu)^2) = \frac{1}{N}\sum_{i=1}^{N} (X(\omega_i) - \mu)^2 \qquad (2')$$

qui sont souvent les **paramètres** de la population que l'on cherche à estimer pour une variable, coïncident avec la moyenne arithmétique et la variance de la série statistique $X(\omega_1), \ldots, X(\omega_N)$ respectivement. La méthode d'échantillonnage la plus couramment utilisée est celle d'**échantillonnage aléatoire simple** (voir [Do] section 10.3.1 et [DDV] chap. 10), et consiste à sélectionner des individus au sein de la population de façon aléatoire et équiprobable.

En prélevant un échantillon de taille n de la population, soient $x_1,\cdots,x_n$ les valeurs observées sur cet échantillon, qui sont aléatoires puisqu'elles varient d'un échantillon à l'autre. Soit alors X_i la variable aléatoire qui associe au $i^{ème}$ individu tiré la valeur x_i prise par le caractère X, pour $i=1,\ldots,n$. Si l'échantillonnage est avec remise alors les tirages sont indépendants, et les variables aléatoires $X_1,\cdots,X_n$ sont indépendantes et de même loi que la variable parente X. Par extension, on appelle **échantillon** un n-uple de variables aléatoires $(X_1,\cdots,X_n)$ indépendantes et de même loi, et on appelle **réalisation** de l'échantillon $(X_1,\cdots,X_n)$ tout n-uple $(x_1,\cdots,x_n)$ de valeurs observées.

Il n'est pas intéressant de détenir plusieurs fois un même individu dans un échantillon, c'est pour cela qu'en pratique l'échantillonnage est réalisé sans remise. Si l'échantillonnage est sans remise, les variables aléatoires obtenues $X_1,\cdots,X_n$ ne sont plus indépendantes ni équidistribuées. Mais si la taille de la population N est suffisamment grande par rapport à la taille de l'échantillon n, alors les variables $X_1,\cdots,X_n$ sont « approximativement » indépendantes et équidistribuées même si l'échantillonnage est sans remise (voir [R1] section 6.6 et [WW] section 6.5). L'intérêt d'avoir des variables aléatoires $X_1,\cdots,X_n$ qui soient indépendantes et équidistribuées est de pouvoir appliquer certains théorèmes comme la loi des grands nombres et le théorème central limite.

Soient $X_1,\cdots,X_n$ des variables aléatoires indépendantes et de même loi (un échantillon), qui ont donc même espérance μ que l'on suppose finie. La **loi des grands nombres** dit que la moyenne empirique de $X_1,\cdots,X_n$:

$$M_n = \frac{X_1+\cdots+X_n}{n}$$

converge en probabilité, mais aussi presque sûrement, vers μ quand n tend vers l'infini (voir [BT], [GW] ou [R2]). Pour les modes de convergence voir par exemple [GS] chap. 7.

Si nous supposons de plus que la variance commune σ^2 de $X_1,\cdots,X_n$ est finie, alors le **théorème central limite** dit que si n est assez grand, la somme $S_n = X_1+\cdots+X_n$ obéit à une loi normale. Comme S_n a pour espérance $n\mu$ et pour variance $n\sigma^2$, S_n obéit donc à la loi normale $N(n\mu,n\sigma^2)$. Plus précisément, le théorème dit que $(S_n-n\mu)/\sqrt{n\sigma^2}$ converge en loi vers la loi normale centrée réduite $N(0,1)$ (voir [BT], [R1], [GW] ou [R2]).

Exemple 3 : Revenons à notre exemple de sondage avec réponses binaires d'une population de taille importante N dont nous cherchons à estimer la proportion de succès p qui n'est rien d'autre que l'espérance de la variable X décrite dans l'exemple 2 ci-dessus. Si $x_1,\cdots,x_n$ désignent les valeurs observées sur un échantillon de taille n on peut alors utiliser la proportion de succès :

$$m = \frac{x_1+\cdots+x_n}{n}$$

au sein de cet échantillon pour estimer le paramètre p. Si l'on suppose que $(x_1,\cdots,x_n)$ est une réalisation de l'échantillon $(X_1,\cdots,X_n)$ où $X_1,\cdots,X_n$ sont des variables aléatoires indépendantes et de même loi que X, il est clair que m ne représente qu'une valeur parmi d'autres prise par la variable aléatoire :

$$M = \frac{X_1+\cdots+X_n}{n}$$

La loi des grands nombres indique que plus n est grand, plus M est « proche » de p. On appelle M un estimateur de p et m une estimation de p. □

Plus généralement, soit θ un paramètre d'une population quelconque que l'on cherche à estimer, associé à un caractère quantitatif X, et soit $(X_1,\cdots,X_n)$ un échantillon aléatoire de variable parente X. On appelle **estimateur** de θ toute fonction de l'échantillon $(X_1,\cdots,X_n)$:

$$T = f(X_1,\cdots,X_n)$$

Un estimateur est donc une variable aléatoire et on voit qu'il y'a une infinité d'estimateurs d'un paramètre donné, sauf que très peu d'entre eux sont intéressants. Si $(x_1,\cdots,x_n)$ est une réalisation de $(X_1,\cdots,X_n)$ on appelle $f(x_1,\cdots,x_n)$ une **estimation** de θ. On parle alors d'**estimation ponctuelle** puisque nous estimons le paramètre θ à partir d'une seule réalisation $(x_1,\cdots,x_n)$ de l'échantillon $(X_1,\cdots,X_n)$ (voir [Do] chap. 10).

Un estimateur T d'un paramètre θ est dit **sans biais** si l'espérance de T est égale à θ : $E(T)=\theta$. Si T n'est pas sans biais, on appelle $E(T)-\theta$ le **biais** de l'estimateur T. On dit que T est **consistant** si $T = f(X_1,\cdots,X_n)$ converge en probabilité vers θ quand n tend vers l'infini. Si T_1 et T_2 sont deux estimateurs de θ, on dit que T_1 est plus **efficace** que T_2 si :

$$Var(T_1) < Var(T_2)$$

En effet, plus la variance d'un estimateur sans biais T est faible, moins les valeurs prises par la variable aléatoire T sont dispersées autour de leur moyenne θ, et plus ces valeurs sont proches de θ. Or les valeurs prises par $T = f(X_1,\cdots,X_n)$ ne sont rien d'autres que les valeurs $f(x_1,\cdots,x_n)$ lorsque $(x_1,\cdots,x_n)$ décrit l'ensemble des réalisations de l'échantillon $(X_1,\cdots,X_n)$. Un estimateur $T = f(X_1,\cdots,X_n)$ sans biais et à variance faible permet donc de fournir des estimations $f(x_1,\cdots,x_n)$ proches du paramètre à estimer, c'est dans ce sens que l'estimateur T est efficace. Dans l'exemple 3 ci-dessus, la variance de M est : $Var(M)=\sigma^2/n$ où σ^2 désigne la variance de la variable X, on voit donc que plus n est grand plus l'estimateur M est efficace.

Exemple 4 : Soit X un caractère quantitatif associé à une population quelconque, μ l'espérance de X et $(X_1,\cdots,X_n)$ un échantillon de variable parente X. Il est clair que l'estimateur :

$$M = \frac{X_1+\cdots+X_n}{n}$$

de μ est sans biais et consistant en vertu de la loi des grands nombres. De plus, M est l'estimateur le plus efficace parmi tous les estimateurs sans biais qui sont fonction linéaire de $X_1,\cdots,X_n$ (voir [GHJ] section 3.4.4). On peut aussi montrer que M est l'estimateur des **moindres carrés ordinaires (MCO)** de μ. En effet, si $(x_1,\cdots,x_n)$ désigne une réalisation de $(X_1,\cdots,X_n)$, la valeur de m qui minimise la somme des carrés $\sum_{i=1}^{n}(x_i-m)^2$ est précisément :

$$m = \frac{x_1+\cdots+x_n}{n}$$

C'est parce que la nature est bien faite que cet estimateur le plus naturel est le meilleur dans tous les sens du terme. En effet, les mathématiques sont bâties sur des axiomes qu'on découvre dans la nature par intuition, en respectant des règles de logique qu'on découvre dans notre nature puisqu'elles ne sont qu'une formalisation du raisonnement humain. C'est pour cela qu'en mathématiques on ne peut jamais décider si une affirmation va être vraie ou fausse mais c'est toujours quelque chose qu'on découvre. La preuve est qu'aux 4 coins du monde tous les mathématiciens sont naturellement d'accord sur ce qui est vrai et ce qui est faux en mathématiques. Il est donc clair que les vérités mathématiques on ne les invente pas mais on les découvre.

Le théorème central limite entraîne que la suite de variables aléatoires :

$$Z = \frac{M - \mu}{\sigma/\sqrt{n}} = \frac{\dfrac{X_1 + \cdots + X_n}{n} - \mu}{\sigma/\sqrt{n}}$$

converge en loi vers la loi normale centrée réduite $N(0,1)$, ce qui signifie que pour n assez grand on peut supposer que la variable Z obéit à la loi $N(0,1)$. Dans ce cas, la table de Gauss donne la probabilité suivante :

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

ce qui entraîne :

$$P(M \in [\mu - e_n ; \mu + e_n]) = 0.95$$

où l'on a noté $e_n = 1.96\sigma/\sqrt{n}$ (on suppose la valeur de σ connue). Cela signifie que 95% des réalisations $(x_1, \cdots, x_n)$ de l'échantillon $(X_1, \cdots, X_n)$ vont avoir une moyenne arithmétique :

$$m = \frac{x_1 + \cdots + x_n}{n}$$

qui appartient à l'intervalle $[\mu - e_n ; \mu + e_n]$, ce qui équivaut à dire que μ appartient à $[m - e_n ; m + e_n]$ pour 95% des réalisations $(x_1, \cdots, x_n)$. Pour une réalisation donnée $(x_1, \cdots, x_n)$ de moyenne arithmétique m, on appelle $[m - e_n ; m + e_n]$ l'**intervalle de confiance** de niveau de confiance 95% pour l'estimation de μ. □

2. STATISTIQUE BIVARIÉE

Si X et Y sont deux variables aléatoires définies sur le même espace de probabilité (Ω, P), la **covariance** entre X et Y est définie par :

$$Cov(X,Y) = E((X - E(X))(Y - E(Y)))$$

Rappelons que si X et Y sont indépendantes alors leur covariance est nulle, mais la réciproque n'est pas toujours vraie. Si les variables aléatoires X et Y sont discrètes, on a :

$$Cov(X,Y) = \sum_{x \in \mathrm{Im}\, X} \sum_{y \in \mathrm{Im}\, Y} (x - E(X))(y - E(Y)) P(X = x, Y = y) \qquad (3)$$

Le **coefficient de corrélation** ρ entre X et Y est défini par :

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad (4)$$

où σ_X (resp. σ_Y) désigne l'écart-type de la variable X (resp. Y) et on suppose bien sûr $\sigma_X \sigma_Y \neq 0$. L'expression (3) indique que si les variables X et Y varient dans le même sens, leur covariance, ainsi que leur coefficient de corrélation (qui est du même signe que la covariance), ont tendance à être positifs. De même, quand les variables varient en sens opposé, leur covariance et leur coefficient de corrélation ont tendance à être négatifs (voir [WW] section 5.3 ou [BT] section 4.2). Le coefficient de corrélation vérifie toujours :

$$-1 \le \rho(X,Y) \le 1.$$

Les cas extrêmes $\rho(X,Y) = \pm 1$ correspondent à une parfaite **liaison linéaire** entre X et Y. On a $\rho(X,Y) = 1$ si et seulement si l'une des variables est une fonction affine croissante de l'autre, c'est-à-dire, par exemple, qu'il existe deux nombres réels α et β avec $\beta > 0$ tels que $Y = \alpha + \beta X$ (presque sûrement) et on dit que X et Y sont parfaitement corrélées positivement. On a également $\rho(X,Y) = -1$ si et seulement si l'une des variables est une fonction affine décroissante de l'autre, c'est-à-dire qu'il existe deux nombres réels α et β avec $\beta < 0$ tels que $Y = \alpha + \beta X$ (presque sûrement) et on dit que X et Y sont parfaitement corrélées négativement (voir [GW] section 7.3 et [Rom] section 1.8). Dans les deux cas nous avons $\beta = \rho \dfrac{\sigma_Y}{\sigma_X}$ (voir [R2] section 7.3) où $\rho = \rho(X,Y)$ de sorte que la relation entre X et Y devient :

$$Y = \alpha + \rho \frac{\sigma_Y}{\sigma_X} X \tag{5}$$

Soient maintenant X et Y deux caractères quantitatifs associés à une même population, que l'on peut donc considérer comme deux variables aléatoires comme on l'a expliqué au paragraphe précédent, et soit $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ l'ensemble des valeurs prises par le couple (X,Y) sur un échantillon de taille n tiré de cette population. La **covariance** de la **série statistique bivariée** $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ est par définition égale à :

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{3'}$$

où $\bar{x}$ (resp. $\bar{y}$) désigne la moyenne arithmétique de la série $x_1, \cdots, x_n$ (resp. $y_1, \cdots, y_n$). Par exemple, si $s_{xy} = 0$ alors $s_{x+y}^2 = s_x^2 + s_y^2$. En effet :

$$\begin{aligned} s_{x+y}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i + y_i - \bar{x} - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + 2(x_i - \bar{x})(y_i - \bar{y}) \right] \\ &= s_x^2 + s_y^2 + 2s_{xy} \\ &= s_x^2 + s_y^2 \end{aligned}$$

Le **coefficient de corrélation** r de $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ est défini par :

$$r = \frac{s_{xy}}{s_x s_y} \tag{4'}$$

où s_x (resp. s_y) désigne l'écart-type de la série $x_1, \cdots, x_n$ (resp. $y_1, \cdots, y_n$) et on suppose bien sûr $s_x s_y \neq 0$. Ce coefficient de corrélation r mesuré sur un échantillon sert à estimer le coefficient de corrélation $\rho(X,Y)$ du couple (X,Y) au niveau de la population (voir [WW] section 14.1), de la même manière que la proportion de succès au sein d'un échantillon sert à estimer la proportion de succès au sein de la population (voir l'exemple 2 ci-dessus).

L'expression qui définit la covariance indique que si les variables X et Y varient dans le même sens, leur covariance, ainsi que leur coefficient de corrélation (qui est du même signe que la covariance), ont tendance à être positifs. De même, quand les variables varient en sens opposé, la covariance et le coefficient de corrélation ont tendance à être négatifs (voir [DDV] section 4.3.3 ou [R1] section 2.6).

Le coefficient de corrélation r est toujours compris entre -1 et 1 et nous avons $r=1$ si et seulement si les points de coordonnées $(x_1,y_1),\ldots,(x_n,y_n)$ sont parfaitement alignés sur une droite ascendante, ce qui revient à dire qu'il existe deux nombres réels a et b avec $b>0$ tels que $y_i=a+bx_i$ pour tout $i=1,\ldots,n$. De même, $r=-1$ si et seulement si les points de coordonnées $(x_1,y_1),\ldots,(x_n,y_n)$ sont parfaitement alignés sur une droite descendante, ce qui revient à dire qu'il existe deux nombres réels a et b avec $b<0$ tels que $y_i=a+bx_i$ pour tout $i=1,\ldots,n$ (voir [R1] section 2.6). Dans les deux cas nous avons $b=r\dfrac{s_y}{s_x}$ de sorte que les points de coordonnées $(x_1,y_1),\ldots,(x_n,y_n)$ appartiennent à la droite d'équation :

$$y=a+r\frac{s_y}{s_x}x \qquad (5')$$

C'est pour cela qu'on appelle r le **coefficient de corrélation linéaire**, plus les points de coordonnées $(x_1,y_1),\ldots,(x_n,y_n)$ se concentrent autour d'une droite de pente positive (resp. négative), plus la valeur de r est proche de 1 (resp. de -1). La réciproque n'est pas toujours vraie, si r est proche de 1 ou de -1 cela n'implique pas toujours que les points soient à peu près alignés (voir [DDV] section 4.5 ou [Joh] chap. 3).

En économétrie, on s'intéresse beaucoup à ce type de liaison linéaire entre deux variables économiques X et Y. Une relation telle que $Y=\alpha+\beta X$ indique qu'une variation d'une unité de la variable X permet de **prédire** une variation de β unités de la variable Y. Ce sont les économistes qui décident si une variable économique X a un **effet causal** sur une variable économique Y (voir [Woo] chap. 1 et [Ive] chap. 8), et si on peut exprimer Y comme fonction de X : $Y=f(X)$.

Les variables X et Y étant aléatoires, une relation exacte du type $Y=f(X)$ ne peut pas exister. C'est pour cela que les économètres rajoutent un terme d'erreur ε, de sorte à avoir $Y=f(X)+\varepsilon$, le terme d'erreur ε étant une variable aléatoire qui correspond à l'écart entre Y et son approximation par la fonction $f(X)$ qu'on appelle la **fonction de régression.** On essaye bien sûr de trouver une fonction f telle que $f(X)$ soit aussi proche que possible de Y. Au sens des **moindres carrés** la fonction $f(X)=E[\mathrm{Y}\mid X]$ est la plus proche de Y, où $E[\mathrm{Y}\mid X]$ désigne l'espérance conditionnelle de Y sachant X (voir [R2] section 7.5). Cela signifie que pour toute fonction g nous avons l'inégalité (voir [R2] section 7.6 ou [BT] section 4.6) :

$$E\left[(Y-E[Y\mid \mathrm{X}])^2\right]\le E\left[(Y-g(X))^2\right]$$

La variable $\varepsilon=Y-f(X)$ avec $f(X)=E[\mathrm{Y}\mid X]$ vérifie (voir [BT] section 4.6) les relations suivantes :

i) $E[\varepsilon]=0$ et $E[\varepsilon\mid \mathrm{X}=\mathrm{x}]=0$ pour tout x
ii) $Cov(f(X),\varepsilon)=0$
iii) $Var(Y)=Var(f(X))+Var(\varepsilon)$

Si la fonction f est affine on obtient le modèle de **régression linéaire simple** :

$$Y=\alpha+\beta X+\varepsilon \qquad (\mathrm{RL})$$

qui permet d'expliquer la variable Y (**variable expliquée**) à partir de la variable X (**variable explicative**). Avec la condition $E[\varepsilon\mid \mathrm{X}]=0$ (ce qui équivaut à $E[\varepsilon\mid \mathrm{X}=\mathrm{x}]=0$ pour tout x), l'équation (RL) est équivalente à $E[Y\mid \mathrm{X}]=\alpha+\beta X$. En posant $Z=\alpha+\beta X$ les trois relations ci-dessus deviennent :

$$E[\varepsilon]=0, \qquad Cov(Z,\varepsilon)=0 \qquad \text{et} \qquad Var(Y)=Var(Z)+Var(\varepsilon) \tag{6}$$

En théorie, les valeurs de α et β qui permettent un meilleur rapprochement au sens des moindres carrés des variables $\alpha+\beta X$ et Y, c'est-à-dire tels que $E\left[(Y-(\alpha+\beta X))^2\right]$ soit minimum, sont donnés par :

$$\beta=\rho\frac{\sigma_Y}{\sigma_X} \qquad \text{et} \qquad \alpha=E[Y]-\beta E[X] \tag{7}$$

où ρ désigne le coefficient de corrélation entre X et Y (voir [Rom] section 1.8 ou [R2] section 7.6). Cependant, l'équation (RL) est supposée valable au niveau de la population et les paramètres α et β sont inconnus et doivent être là encore estimés à partir d'un échantillon. Soit alors une réalisation $(x_1,y_1),\ldots,(x_n,y_n)$ de l'échantillon $(X_1,Y_1),\ldots,(X_n,Y_n)$ et cherchons le meilleur ajustement linéaire de ce nuage de points, c'est-à-dire deux nombres a et b vérifiant :

$$y_i=a+bx_i+e_i \qquad \text{pour } i=1,\ldots,n$$

de sorte que chaque valeur observée y_i soit la plus proche possible de la valeur ajustée $z_i=a+bx_i$, ce qui revient à minimiser les erreurs d'ajustement $e_i=y_i-z_i$. La méthode des **moindres carrés ordinaires (MCO)** consiste à trouver les valeurs de a et b qui minimisent la somme des carrés de ces résidus $\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2$ et en résolvant on trouve :

$$b=\frac{s_{xy}}{s_x^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \text{et} \qquad a=\bar{y}-b\bar{x}$$

(voir [Do] section 15.3 ou [DDV] section 4.4.1), et on suppose bien sûr que les x_i ne sont pas tous égaux. Un calcul facile montre que $\bar{e}=0$ et $\bar{z}=\bar{y}$, ce qui entraîne :

$$\begin{aligned}
s_{ze} &= \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})(e_i-\bar{e}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(a+bx_i-\bar{y})(y_i-a-bx_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}b(x_i-\bar{x})[(y_i-\bar{y})-b(x_i-\bar{x})] \\
&= bs_{xy}-b^2s_x^2 \\
&= 0
\end{aligned}$$

La covariance entre les séries (z_i) et (e_i) étant nulle, cela entraîne que la variance de leur somme est égale à la somme de leurs variances $s_y^2=s_z^2+s_e^2$. D'où les relations :

$$\bar{e}=0, \qquad s_{ze}=0 \qquad \text{et} \qquad s_y^2=s_z^2+s_e^2 \tag{6'}$$

Si r désigne le coefficient de corrélation entre les séries (x_i) et (y_i) on obtient :

$$b = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x s_y} \times \frac{s_y}{s_x} = r\frac{s_y}{s_x}$$

D'où les relations :

$$b = r\frac{s_y}{s_x} \qquad \text{et} \qquad a = \bar{y} - b\bar{x} \tag{7'}$$

Le but du jeu étant d'expliquer la variation de la variable Y à partir de la variation de la variable X, et comme $a + bX$ varie systématiquement avec X, on appelle s_z^2 la **composante systématique** ou la **variation expliquée** puisqu'elle représente la partie de la variation de Y qui est expliquée par la variation de X au sein de l'échantillon. On appelle s_e^2 la **composante non systématique** ou la **variation non expliquée**, et on appelle s_y^2 la **variation totale**.

Un calcul facile montre que $s_z^2 = b^2 s_x^2 = r^2 s_y^2$. On appelle r^2 le **coefficient de détermination**, il permet de déterminer le pourcentage de la variation de Y qui est expliquée par la variation de X au sein de l'échantillon puisque r^2 s'obtient en divisant s_z^2 par s_y^2. D'ailleurs, dans certains livres de statistique, on définit d'abord le coefficient de détermination comme étant égal à s_z^2 / s_y^2, ensuite on définit le coefficient de corrélation r comme étant égal à la racine carrée du coefficient de détermination, et ayant le même signe que le coefficient de régression b (voir [Zu] chap. 11). Pour en savoir plus sur le coefficient de détermination voir [Ken] chapitres 2,5 et 6.

Revenons à notre modèle (RL) et à notre échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ et pour $i = 1, \ldots, n$ posons $\varepsilon_i = Y_i - \alpha - \beta X_i$, ce qui entraîne $Y_i = \alpha + \beta X_i + \varepsilon_i$.

Exemple 5 : Supposons que Y représente le rendement des cultures de blé et X représente la quantité d'engrais utilisée (voir [WW] chap. 11 & 12). La particularité de cet exemple est que la variable X n'est pas aléatoire puisque la quantité d'engrais x_i utilisée pour la $i^{ème}$ parcelle est prédéterminée par le cultivateur, la variable X_i n'est donc pas aléatoire mais **déterministe** et prend la valeur x_i qu'on lui a fixée. Dans ce cas, les variables $X_1, \cdots, X_n$ ne sont pas équidistribuées comme dans le cas d'un échantillon aléatoire puisque chaque variable X_i est constante égale à x_i ce qui entraîne, par exemple, que les X_i n'ont pas la même espérance.

Même si les variables $Y_1, \ldots, Y_n$ sont aléatoires et indépendantes, elles ne peuvent pas être équidistribuées non plus puisqu'elles ont aussi des espérances différentes $E[Y_i] = \alpha + \beta x_i$ (voir [DKL] section 17.4). Dans le cas où la variable X est supposée déterministe, il existe des protocoles expérimentaux modernes qui permettent d'attribuer des valeurs à X en suivant une répartition aléatoire générée par ordinateur, la variable X devient aléatoire et les couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ deviennent indépendants et équidistribués (voir [SW] section 1.4.2). □

Dans le modèle de régression linéaire on traite séparément les cas X aléatoire et X déterministe. Dans le cas X aléatoire on impose, en effet, l'hypothèse que les couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ forment un échantillon aléatoire, c'est-à-dire qu'ils soient indépendants et équidistribués (voir [HGL] chap. 10 ou [Doug] chap. 8). Dans les livres d'économétrie on commence souvent par traiter le cas X déterministe qui permet une analyse plus simple des propriétés des estimateurs MCO :

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{et} \qquad \hat{a} = \bar{Y} - \bar{x}\hat{b},$$

de α et β et les hypothèses sur le modèle sont déterminées de sorte à obtenir des estimateurs MCO **performants**, c'est-à-dire sans biais et de variance minimale (voir [Guj] section 3.4). Rappelons les 7 hypothèses du modèle telles qu'elles sont présentées dans [HBF] section 2.2.3 et qui permettent une parfaite simplification de l'analyse des estimateurs MCO :

1) La variable X est déterministe et les valeurs $x_1, \cdots, x_n$ ne sont pas toutes égales

2) Les variables $\varepsilon_1, \ldots, \varepsilon_n$ sont aléatoires et $E(\varepsilon_i) = 0$ pour $i = 1, \ldots, n$

3) Homoscédasticité : $E(\varepsilon_i^2) = Var(\varepsilon_i) = \sigma^2 < +\infty$ pour $i = 1, \ldots, n$

4) Absence de corrélation : $E(\varepsilon_i \varepsilon_j) = 0$ pour tout $i \neq j$

5) Paramètres constants : les paramètres α, β, σ sont inconnus mais constants avec $\sigma > 0$

6) Modèle linéaire : $Y_i = \alpha + \beta x_i + \varepsilon_i$ pour $i = 1, \ldots, n$

7) les variables $\varepsilon_1, \ldots, \varepsilon_n$ suivent conjointement une distribution normale

Les hypothèses 1,2,5,6 entraînent que les estimateurs $\hat{a}$ et $\hat{b}$ de α et β respectivement sont sans biais (voir [HBF] section 2.2.4 ou [Doug] section 2.5). Les hypothèses 1 à 6 entraînent que $\hat{a}$ et $\hat{b}$, qui sont linéaires en $Y_1, \ldots, Y_n$, sont les estimateurs les plus efficaces (de plus faibles variances) parmi tous les estimateurs sans biais et linéaires en $Y_1, \ldots, Y_n$ de α et β respectivement, c'est le **théorème de Gauss-Markov** (voir [HBF] section 2.2.5).

Exemple 6 : Supposons que X représente le revenu et Y la consommation alimentaire hebdomadaires (en euros) d'une population de ménages. Si on réalise un échantillonnage aléatoire simple, on obtient un échantillon aléatoire $(X_1, Y_1), \ldots, (X_n, Y_n)$ et la variable X n'est donc pas déterministe. On peut aussi réaliser un **échantillonnage stratifié** (voir [Do] section 10.3.2 ou [HJ] section 13.4) qui consiste à découper la population en strates de sorte à avoir le même niveau de revenu X dans chaque strate, et ensuite effectuer un échantillonnage aléatoire simple dans chacune de ces strates. On dit alors que les valeurs de X sont **fixées dans un échantillonnage répété** (voir [Guj] section 3.2, [HGL] chap. 2, [Woo] section 2.5), la variable X est considérée comme étant déterministe et les estimateurs MCO sont performants. □

Dans les exemples 5 et 6 ci-dessus les données sont observées sur une seule période de temps pour plusieurs individus d'une population. On appelle ce type de données des données en **coupes instantanées** ou en **coupes transversales**. On peut aussi s'intéresser à la progression dans le temps de certaines variables économiques comme le PIB ou le taux de chômage dans un pays donné ou encore le cours d'une action. L'ensemble des données d'une variable X, observées sur un seul individu et à différents moments du temps, disons x_t pour $t = t_1, \ldots, t_n$, s'appelle une **série temporelle** ou une **série chronologique** (voir [Joh], [Ham], [Woo], [Wei]). L'observation x_t étant inconnue avant la date t, x_t est donc une réalisation d'une variable aléatoire X_t. L'échantillon de données temporelles $(x_{t_1}, \ldots, x_{t_n})$ est une réalisation du processus stochastique $(X_{t_1}, \ldots, X_{t_n})$ (voir [BD] section 1.2), un **processus stochastique** étant par définition une famille de variables aléatoires $(X_t)_{t \in T}$ indexées par le temps, où T est un sous-ensemble dénombrable ou non de l'ensemble des nombres réels (voir [Gal], [BZ], [Law], [KMT], [Shr]). Dans la littérature, l'expression « série temporelle » peut également signifier « processus stochastique ».

Exemple 7 : Soit R le taux de rentabilité (boursière) d'un actif financier et R_M le taux de rentabilité du portefeuille de marché contenant tous les actifs. Le **bêta** de l'actif est le coefficient de régression β de $R - R_f$ sur $R_M - R_f$, où R_f désigne le taux d'intérêt sans risque :

$$R - R_f = \alpha + \beta\left(R_M - R_f\right) + \varepsilon$$

Les coefficients α et β peuvent être estimés à l'aide des formules (7') à partir d'un échantillon de données temporelles $\left(R_t, R_{M,t}\right)$ pour $t = t_1, \ldots, t_n$ et on a alors (voir [Ziv] section 6.3 exemple 35, [RHF] chap. 6 et [Bro]) :

$$R_t - R_{f,t} = \alpha + \beta\left(R_{M,t} - R_{f,t}\right) + \varepsilon_t$$

où R_t (resp. $R_{M,t}$, resp. $R_{f,t}$) désigne le **taux de rentabilité** de l'actif (resp. le taux de rentabilité du portefeuille de marché, resp. le taux d'intérêt sans risque) entre les dates $t-1$ et t (voir [Wil] section 3.5).

Dans le cas des séries temporelles, les observations temporellement proches les unes des autres sont souvent corrélées. C'est notamment le cas des termes d'erreurs $\varepsilon_{t_1}, \ldots, \varepsilon_{t_n}$ et la condition 4) ci-dessus d'absence de corrélation des termes d'erreurs dans le modèle de régression est donc violée. D'ailleurs, l'estimateur MCO $\hat{b}$ n'est pas de variance minimale parmi les estimateurs linéaires sans biais de β (voir [Guj] section 12.2).

Terminons par la célèbre relation du **MEDAF** (Modèle d'équilibre des actifs financiers) :

$$E(R) - R_f = \beta\left(E(R_M) - R_f\right)$$

où $E(R)$ désigne l'espérance mathématique du taux de rentabilité futur R de l'actif, qui est donc une variable aléatoire, dont la distribution de probabilité est estimée à partir de la distribution de fréquence d'un échantillon de taux de rentabilité passés de l'actif, et peut être calculé aussi bien par la formule (1) que par la formule (1') qui sont identiques dans ce cas. Les formules (2) et (2') qui permettent de calculer la variance de R coïncident également dans ce cas particulier, et on remplacera bien sûr N par n dans les formules (1') et (2') puisque nous considérons un échantillon de taille n. Il y'a peut être aussi des cas particuliers où les covariances (3) et (3') coïncident, sachant qu'il y'a plus de termes dans la somme de la formule (3) que dans la somme de la formule (3') et il est clair que les coefficients de corrélation (4) et (4') sont définis par des formules analogues.

Par ailleurs, il est bien connu que l'estimateur naturel de la variance de la population σ^2 obtenu en remplaçant x_i par X_i et $\bar{x}$ par $\overline{X}$ dans la formule (2) est biaisé. Il faut, en effet, diviser par $n-1$ au lieu de diviser par n pour obtenir un estimateur non biaisé. Cependant, il y'a une parfaite similitude entre les relations (5) avec (5'), (6) avec (6') et (7) avec (7') qui sont pourtant bâties sur la variance et la covariance, ce qui est plutôt rassurant. Faire de la recherche en mathématiques est certes un voyage semé d'embûches, mais on découvre à chaque fois des vérités qui montrent que le monde mathématique est aussi parfait que l'univers.

Comme on l'a vu dans cet article, c'est grâce à la statistique qu'on arrive à estimer les paramètres d'une population et les probabilités d'événements futurs. La démonstration de la relation du MEDAF s'appuie sur plusieurs hypothèses (voir [Hul] annexe du chapitre 3), une démonstration est proposée dans [Rom] chap. 2 théorème 7. □

RÉFÉRENCES BIBLIOGRAPHIQUES

[BT] D. P. Bertsekas & J. N. Tsitsiklis, Introduction to Probability, Athena Scientific, 1st Edition, 2002

[BD] P. J. Brockwell & R. A. Davis, Time Series : Theory and Methods, Springer, 2nd Edition, 1991

[Bro] C. Brooks, Introductory Econometrics for Finance, Cambridge University Press, 3rd Edition, 2014

[BZ] Z. Brzezniak & T. Zastawniak, Basic Stochastic Processes, Springer, 2000

[DDV] C. Dehon, J. J. Droesbeke & C. Vermandele, Eléments de statistique, Ellipses, 6ème édition, 2015

[DKL] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä & L. E. Meester, A Modern Introduction to Probability and Statistics, Springer, 2007

[Do] Yadolah Dodge, Premiers pas en statistiques, Springer, 2006

[Doug] C. Dougherty, Introduction to Econometrics, Oxford University Press, 3rd Edition, 2007

[Gal] R. G. Gallager, Stochastic Processes, Cambridge University Press, 2013

[GHJ] W. E. Griffiths, R. C. Hill & G. G. Judge, Learning and Practicing Econometrics, Wiley, 1993

[GS] G.R. Grimmett and D.R. Stirzaker, Probability and Random Processes, Oxford Univ. P., 3rd Edition, 2001

[GW] G. Grimmett and D. Welsh, Probability : An Introduction, Oxford University Press, 1986

[Guj] D. Gujarati, Basic Econometrics, McGraw-Hill, 4th Edition, 2002

[Ham] J. D. Hamilton, Time Series Analysis, Princeton University Press, 1994

[HBF] C. Heij, P. de Boer, P. H. Franses, T Kloek & H. K. van Dijk, Econometric Methods with Applications in Business and Economics, Oxford University Press, 1st Edition, 2004

[HGL] R. C. Hill, W. E. Griffiths & G. C. Lim, Principles of Econometrics, Wiley, 3rd edition, 2007

[HJ] P. G. Hoel & R. J. Jessen, Basic Statistics for Business and Economics, Wiley, 1971

[Hul] J. Hull, Options, Futures et autres actifs dérivés, Pearson, 9ème édition, 2014

[Ive] G. R. Iversen & M. Gergen, Statistics : The conceptual approach, Springer, 1997

[Joh] R. A. Johnson & D. W. Wichern, Business Statistics, Wiley, 1st Edition, 1996

[Ken] Peter Kennedy, A Guide to Econometrics, Wiley-Blackwell, 6th Edition, 2008

[KMT] H. Kobayashi, B. L. Mark & W. Turin, Probability, Random Processes and Statistical Analysis, Cambridge University Press, 2012

[Law] Gregory F. Lawler, Introduction to Stochastic Processes, Chapman & Hall/CRC, 2nd Edition, 2006

[RHF] S. T. Rachev, M. Hoechstoetter & F. J. Fabozzi, Probability and statistics for finance, Wiley, 2010

[Rom] S. Roman, Introduction to the Mathematics of Finance : From Risk Management to Options Pricing, Springer, 2004

[R1] S. M. Ross, Introduction to probability and statistics for engineers and scientists, Elsevier,3rd edition, 2004

[R2] S. M. Ross, A first course in probability, Pearson, 9th edition, 2014

[Shr] S. E. Shreve, Stochastic Calculus for Finance I & II, Springer, 2004

[Sim] C. Simard, Notions de Statistique, Modulo, 3ème édition, 2015

[SW] J. Stock & M. Watson, Principes d'économétrie, Pearson, 3ème édition, 2012

[Wei] W. W. S. Wei, Time series analysis, Pearson, 2nd Edition, 2005

[Wil] P. Wilmott, On Quantitative Finance, John Wiley & Sons, 2nd Edition, 2006

[WW] T. H. Wonnacott & R. J. Wonnacott, Introductory Statistics, John Wiley and Sons, 1969

[Woo] J. M. Wooldridge, Introduction à l'économétrie, De Boeck, 5ème édition, 2015

[Ziv] E. Zivot & J. Wang, Modeling Financial Time Series with S-Plus, Springer, 2nd Edition, 2006

[Zu] Fadil H. Zuwaylif, General Applied Statistics, Addison Wesley, 3rd Edition, 1979