Cours 8 : Séries statistiques à une et deux dimensions

Cette leçon introduit les outils statistiques de base qui seront utilisés dans les deux prochaines leçons consacrées à la régression linéaire. Ces outils servent à étudier des séries statistiques, en les localisant (par leur moyenne), en quantifiant leur dispersion (par leur variance ou leur écart-type) et, lorsqu'il y en a plusieurs, en comparant leurs variations (par leur covariance et leur corrélation).

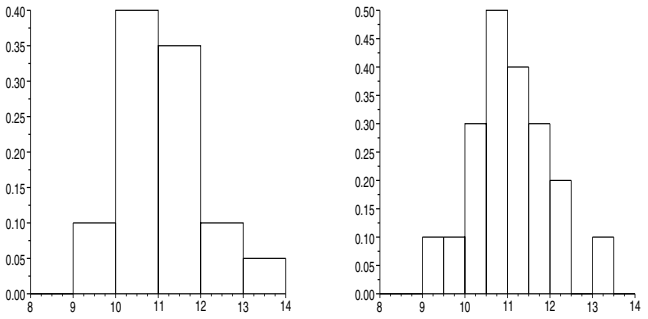
1 Séries statistiques à une dimension

Une *série statistique* est simplement une suite de mesures comme par exemple la suite, mesurée en centimètres des tailles de 20 plants que l'on étudie (présentées par ordre croissant) :

$$9,3 \quad 9,7 \quad 10,1 \quad 10,2 \quad 10,4 \quad 10,6 \quad 10,7 \quad 10,7 \quad 10,9 \quad 11$$

$$11,1 \quad 11,1 \quad 11,3 \quad 11,3 \quad 11,6 \quad 11,7 \quad 11,9 \quad 12,3 \quad 12,4 \quad 13,4.$$

Par la suite on désignera par x_i une telle suite, l'indice i prenant les valeurs entières de 1 à n (ici $n = 20$, $x_1 = 9,3$, $x_2 = 9,7$, ...). Pour comprendre une telle série la première idée est de la représenter à travers un histogramme. A noter cependant qu'il n'y a pas une façon unique de le faire. Par exemple, pour le premier histogramme ci-dessous, on a regroupé les mesures comprises entre 9 et 10 (ici 2 mesures), puis celles comprises entre 10 et 11 (ici, 8 mesures), et ainsi de suite. Dans le second histogramme, qui correspond aux même données, les intervalles de classes ne sont plus d'un centimètre mais d'un demi centimètre.



Mais quelque soit la façon dont on procède pour tracer l'histogramme, on fait en sorte que la surface qu'il occupe (somme des surface des barres) soit égale à 1. Ainsi, pour une série de 20 termes, si la largeur de classes est 1 (comme dans le premier histogramme), la hauteur des barres sera égale à l'effectif de la classe multiplié par $\frac{1}{20} = 0,05$; mais si la largeur des classes est $0,5$ (deuxième histogramme), alors il faudra multiplier par deux l'unité en hauteur en passant de $\frac{1}{20}$ à $\frac{1}{10}$.

Moyenne :

L'histogramme fournit un bon résumé graphique des données, toujours utile pour commencer l'analyse d'une série, mais le plus souvent il faudra l'acompagner d'autres résumés, quantitatifs cette fois, dont les plus utilisés sont la moyenne, la variance et l'écart-type.

Définition : La *moyenne* est simplement donnée par

$$\mu = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La moyenne de la série des 20 mesures de notre exemple est égale à $11,8$. C'est une grandeur que l'on appelle un *paramètre de position* parce qu'elle indique autour de quelle valeur (ici 11,8) se situent les mesures de la série.

Mais si l'on s'en tenait à la moyenne on ne saurait rien sur l'*étendue* de la série : deux séries de même moyenne peuvent être très différentes, l'une regroupée autour de cette moyenne et l'autre au contraire très dispercée. On a donc besoin d'un autre paramètre qui, lui, quantifie cette étendue.

Variance et écart type d'une série

Considérons un exemple. Soit la série

$$12,1 \quad 16,3 \quad 13,2 \quad 13,5 \quad 14,9$$

dont la moyenne est 14. La première idée est de calculer pour chaque terme de la série leur é*cart à la moyenne* c'est-à-dire ici leur écart à 14 :

$$-1,9 \quad 2,3 \quad -0,8 \quad -0,5 \quad 0,9$$

On trouve trois écarts négatifs correspondant à des mesures inférieures à la moyenne et deux écarts positifs correspondant à des mesures supérieures à la moyenne. Comme mesure de l'étendue, on peut avoir l'idée de faire la moyenne de ces écarts

$$\frac{-1,9+2,3-0,8-0,5+0,9}{5} = \frac{0}{5} = 0.$$

Mais on voit que cette moyenne n'est pas une bonne mesure de l'étendue et en fait, un petit calcul permet de vérifier que cette moyenne sera toujours nulle. Pour contourner cette difficulté, on peut avoir l'idée de faire la moyenne des écarts, mais en les comptant tous positivement cette fois $\frac{1,9+2,3+0,8+0,5+0,9}{5} = 1,28$. On obtient une quantité que l'on appelle la *déviation moyenne* qui est bien une mesure de l'étendue mais qui n'est pas la quantité la plus souvent utilisée pour cela. Celle que l'on utilise en général (car elle s'avère plus maniable dans les calculs) est la variance, ici

$$\frac{(1,9)^2+(2,3)^2+(0,8)^2+(0,5)^2+(0,9)^2}{5} = \frac{10,6}{5} = 2,12.$$

Définition : La *variance* d'une série (x_i) est la moyenne des carrés des écarts à la moyenne, c'est-à-dire

$$\text{Var}\,(x) = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

A noter que la variance de données exprimées en *cm* sera exprimée en cm^2. Si l'on veut mesurer l'étendue de la série en utilisant les mêmes unités que la série elle-même, il faut prendre la racine carré de la variance.

Définition : L'é*cart type* d'une série (x_i) (an anglais, *standard deviation*) est la racine carrée de sa variance, c'est-à-dire

$$\sigma(x) = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Très souvent, les histogrammes de séries statistiques ont la forme d'une cloche semblable à la *cloche de Gauss*, graphe de la fonction $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. C'est le cas notamment lorsque le nombre n de termes de la série est grand. Pour cette courbe de référence, dont le sommet est situé en $x = \mu$, on peut montrer que 65% de la surface totale (qui vaut 1) est situé au dessus de l'intervalle $[\mu-\sigma, \mu+\sigma]$. De la même façon, on peut montrer que 95% est situé au dessus de l'intervalle $[\mu-2\sigma, \mu+2\sigma]$. On peut donc retenir que, très grossièrement, un histogramme en forme de cloche est "centré" sur sa moyenne μ et son étendue est presque confondue avec l'intervalle $[\mu-2\sigma, \mu+2\sigma]$ et situé aux $\frac{2}{3}$ au dessus de l'intervalle $[\mu-\sigma, \mu+\sigma]$.

Propriétés de la moyenne et de la variance :

Si l'on modifie une série (x_i) en (ax_i+b), par exemple en changeant les unités ou l'origine des mesures, alors il est facile de vérifier que la moyenne, la variance et l'écart-type de la nouvelle série s'exprime en fonction de celles de la série originale par les formules suivantes :

– $\mu(ax+b) = a\mu(x) + b$
– $\text{Var}\,(ax+b) = a^2\text{Var}\,(x)$
– $\sigma(ax+b) = a\sigma(x)$ si $a > 0$

Des propriétés précédentes on peut aussi déduire une nouvelle formule de la variance qu'on appelle la *formule de Huygens* et qui est très utile en pratique :

$$\text{Var}\,(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2.$$

En effet, la variance est par définition égale à $\text{Var}\,(x) = \mu((x-\mu(x))^2)$. Or

$$\mu((x-\mu)^2) = \mu(x^2 - 2\mu(x)x + \mu(x)^2) = \mu(x^2) - 2\mu(x)\mu(x) + \mu(\mu(x)^2) = \mu(x^2) - 2(\mu(x))^2 + (\mu(x)^2).$$

Donc pour calculer la variance d'une série (lorsqu'on ne le fait pas directement avec un logiciel), il est souvent commode de présenter les calculs dans un tableau du type suivant,

x_i	$12,1$	$16,3$	$13,2$	$13,5$	$14,9$	14
x_i^2	$146,41$	$265,69$	$174,24$	$182,25$	$222,01$	198,12

la dernière colonne étant la moyenne des termes de la ligne. Pour le calcul de la variance, on applique alors la formule de huygens $\text{Var}\,(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2$ et donc, en utilisant les deux nombres de la dernière colonne, on obtient ici $\text{Var}\,(x) = 198,12 - (14)^2 = 2,12$.

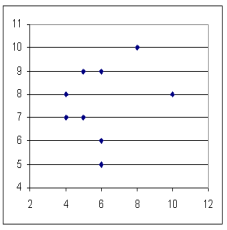
2 Séries statistiques à deux dimensions

Lorsque l'on dispose de deux séries, x_i et y_i de même nombre de termes, on peut, biensur, les étudier séparemment, tracer pour chacune d'elles un histogramme, et aussi calculer, pour chacune d'elle, la moyenne, $\mu(x)$ et $\mu(y)$, la variance, $\text{Var}\,(x)$ et $\text{Var}\,(y)$, et l'écart-type, $\sigma(x)$ et $\sigma(y)$. Mais s'il s'agit par exemple de deux mesures prises sur les mêmes individus, taille et nombre de feuilles d'une plante dont on étudie la croissance, ou concentration de deux substances dans un liquide dont on étudie l'évolution au cours du temps, on peut souhaiter étudier non seulement les deux séries séparémment mais aussi leur interaction : les deux quantités augmentent-elles de concert ou au contraire l'une diminue-t-elle lorsque l'autre augmente et cela se fait-il dans les même proportions ou non ?

A ces questions, il y a une réponse graphique, le nuage de points associé, et une réponse quantitative, le calcul de la covariance et du coefficient de corrélation linéaire.

La réponse graphique consiste à tracer le *nuage de points* (x_i, y_i) dans le plan (x, y), que l'on appelle encore *diagramme de dispersion*. Chaque point a comme abscisse le ième terme de la première série et comme ordonnée le ième terme de la deuxième série.

Considérons un exemple. On a fait pousser un premier lot de 10 plants de haricots dans des conditions que nous appellons *Conditions A* et 10 autres plants identiques dans des conditions que nous appellons *Conditions B*. Pour chaque plant, on a compté le nombre de feuilles après 30 jours et on a obtenu les données suivantes :



Conditions A	4	6	5	6	8	4	6	5	10	5
Conditions B	7	5	9	6	10	8	9	7	8	7

Le nuage de point correspondant est localisé autour d'un *point moyen* qu'on appelle le *centre de gravité* du nuage, dans cet exemple le point de coordonnées $(5,9\ \ 7,6)$.

Définition : Le centre de gravité du nuage représentant les deux séries x_i et y_i de moyennes $\mu(x)$ et $\mu(y)$ rspectivement est le point $G = (\mu(x), \mu(y))$ dont les coordonnées sont les moyennes des deux séries.

On peut aussi relier l'étendue du nuage aux deux écarts-type $\sigma(x)$ et $\sigma(y)$ des séries x et y, ici, $\sigma(x) = 0,59$ et $\sigma(y) = 0,48$. Le premier est lié à l'étendue horizontale (dans le sens de l'axe des x), et le second à l'étendue verticale (dans le sens de l'axe des y).

Covariance de deux séries

Mais si tous les points du nuage étaient pratiquement situés sur une même droite (nuage de forme alongée), le nuage pourrait être très étendu, à la fois en x et à la fois en y sans pour autant être étendu en surface dans le plan. Pour mesurer "l'étendue en surface"du nuage, on calcule la covariance des deux séries.

Définition : La *covariance* de deux séries (x_i) et (y_i) est la moyenne des produits des écarts à la moyenne, c'est-à-dire

$$\mathrm{Cov}\,(x,y) = \frac{(x_1-\mu(x))(y_1-\mu(y)) + \ldots + (x_n-\mu(x))(y_n-\mu(y))}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu(x))(y_i-\mu(y)).$$

A noter que si $y = x$, $\mathrm{Cov}\,(x,x) = \mathrm{Var}\,(x)$

Comme pour la variance, on a pour la covariance une *formule de Huygens* qui permets souvent de faire les calculs plus facilement. On vérifie en effet (le faire en exercice) que la covariance est aussi égale à la moyenne des produits moins le produit des moyennes :

$$\mathrm{Cov}\,(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu(x)\mu(y).$$

Donc pour calculer la covariance de deux séries (lorsqu'on ne le fait pas directement avec un logiciel), il est souvent commode de présenter les calculs dans un tableau du type suivant

x_i	4	6	5	6	8	4	6	5	10	5	5,9
y_i	7	5	9	6	10	8	9	7	8	7	7,6
$x_i y_i$	28	30	45	36	80	32	54	35	80	35	45,5

la dernière colonne donnant les moyennes des termes de la ligne. En appliquant alors la formule de Huygens, $\mathrm{Cov}\,(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu(x)\mu(y)$, il suffit d'utiliser les trois nombres de la dernière colonne pour calculer la covariance. Dans l'exemple, $\mathrm{Cov}\,(x,y) = 45,5 - (5,9)(7,6) = 0,66$.

Le plus souvent, ce calcul est fait par le logiciel que l'on utilise. Ensuite le plus important de savoir interpréter la valeur trouvée. Le signe de la covariance renseigne sur le fait que les deux séries varient *dans le même sens ou en sens opposé :* une covariance positive indique que l'une augmente quand l'autre augmente ou diminiue quand l'autre diminue.

Par contre la valeur absolue de la covariance (le fait qu'elle ait une valeur petite ou grande) fournit peu d'informations pertinentes car elle dépend des unités dans lesquelles ont été exprimées les deux séries (x_i) et (y_i). Pour obtenir un équivalent de la covariance qui soit indépendant des unités dans lesquelles ont été exprimé les deux séries de mesures il convient de calculer le coefficient de correlations linéaire. On peut montrer que sa valeur est toujours comprise entre -1 et $+1$.

Définition : Le *coefficient de corrélation linéaire* de deux séries (x_i) et (y_i) est la quantité

$$\rho(x,y) = \frac{\mathrm{Cov}\,(x,y)}{\sqrt{\mathrm{Var}\,(x)\mathrm{Var}\,(y)}}.$$

On dit que les deux séries sont *faiblement corrélées* lorsque $\rho(x,y)$ est proche de zéro, et dans ce cas le nuage a une forme très dispercée. Au contraire si $\rho(x,y)$ est proche de -1 ou $+1$, les deux séries sont dites *fortement corrélées* (positivement si $\rho(x,y) > 0$ et négativement si $\rho(x,y) < 0$) et dans ce cas le nuage est regroupé le long d'une droite (de pente positive si $\rho(x,y) \simeq +1$ et de pente négative si $\rho(x,y) \simeq -1$). Mais nous reviendrons sur ce coefficient de corrélation linéaire car il joue un grand rôle dans la méthode de la régression linéaire qui fait l'objet des deux prochaines leçons.