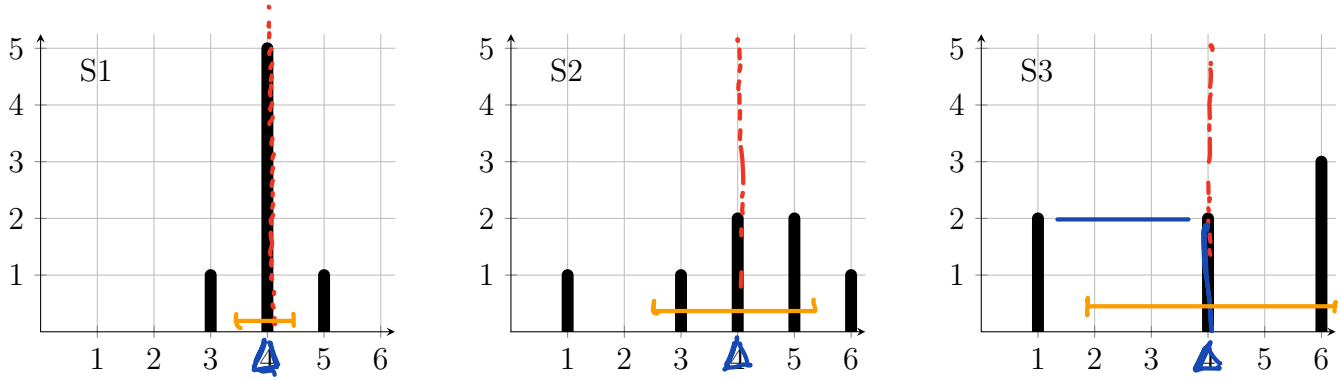
V. Variance et corrélation

La semaine passée, nous avons travaillé avec des variables aléatoires et défini l'espérance.

Pour terminer l'étude des variables aléatoires, nous définissons encore la variance, une mesure de dispersion des valeurs d'une variable autour de l'espérance, puis nous finissons en analysant la corrélation entre deux variables.

1 Variance

Exemple 1.1. Considérons les trois séries $S_1 = \{3, 4, 4, 4, 4, 4, 5\}$, $S_2 = \{1, 3, 4, 4, 5, 5, 6\}$ et $S_3 = \{1, 1, 4, 4, 6, 6, 6\}$. Elles possèdent la même moyenne et la même médiane, toutes deux égales à 4, mais il est clair que les données des séries 2 et 3 sont plus dispersées que celles de la série 1.



S'il s'agit des notes obtenues lors d'un test dans une petite classe, ses séries représentent des situations fort différentes. Il faut une **mesure de dispersion** pour compléter l'analyse statistique.

Pour cela, considérons dans chaque série les écarts entre les valeurs et la moyenne . Comme ces écarts peuvent être positifs ou négatifs, on les élève au carré.

Définition 1.2. Soit X une variable aléatoire et $\mu = E[X]$ son espérance. La *variance* de X vaut

$$\mathrm{Var}(X) = E[(X - \mu)^2].$$

L'unité de la variance diffère de l'unité de la variable aléatoire. On lui préfère l'écart-type.

Définition 1.3. Soit X une variable aléatoire. L'*écart-type* de X est le nombre

$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \sigma_X.$$

Remarque 1.4. En théorie des probabilités, l'inégalité de Bienaymé-Tchebychev montre que, quelle que soit la distribution d'une variable aléatoire ou statistique, au moins la moitié des valeurs se trouvent dans l'intervalle $[\bar{x} - \sqrt{2}\sigma ; \bar{x} + \sqrt{2}\sigma]$.
Pratiquement, en statistique, l'intervalle $[\bar{x} - \sigma ; \bar{x} + \sigma]$ contient déjà souvent la majorité des données.

Exemple 1.5. Calculons les variances puis les écart-types des séries de l'exemple 1.1 et donnons-en une interprétation selon la remarque 1.4 ci-dessus.

Série 3

$$\mathrm{Var}(S_3) = \frac{2}{7}(1-4)^2 + \frac{2}{7}(4-4)^2 + \frac{3}{7}(6-4)^2 = \frac{18+0+12}{7} = \frac{30}{7}$$

d'où $\sigma = \sqrt{\frac{30}{7}} \simeq 2{,}07$ $\qquad \mu - \sigma = 4 - 2{,}07 = 1{,}93$ $\qquad \mu + \sigma = 6{,}07$

$\Rightarrow$ Plus de la moitié des données $\in [1{,}93 ; 6{,}07]$

Série 1 : $\mathrm{Var}(S_1) = \frac{2}{7}$, $\sigma = \sqrt{\frac{2}{7}} \simeq 0{,}53 \rightarrow I = [3{,}5 ; 4{,}5]$

A des fins calculatoires, il est souvent plus commode d'utiliser la formule donnée dans le théorème suivant. Il dit que la variance mesure la différence entre le carré de l'espérance et l'espérance au carré.

Proposition 1.6. Formule de König *Soit X une variable aléatoire. Alors*

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 .$$

Démonstration. Si $E[X] = \mu$, la définition de la variance utilise le carré $(X-\mu)^2 = X^2 - 2\mu X + \mu^2$.

Par linéarité de l'espérance, on a

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right] = E\left[X^2 - 2\mu X + \mu^2\right]$$

$$= E[X^2] - 2\mu \underbrace{E[X]}_{=\mu} + \mu^2$$

$$= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2$$

$$= E[X^2] - (E[X])^2$$

□

Exemple 1.7. Calculons la variance des séries de l'exemple 1.1 avec la formule de König :

$$\text{Var}(S_2) = \frac{1^2+3^2+2\cdot 4^2+2\cdot 5^2+6^2}{7} - 4^2$$

$$= \frac{1+9+32+50+36}{7} - \frac{112}{7} = \frac{128-112}{7} = \frac{16}{7}$$

$$\Rightarrow \quad \sigma = \sqrt{\frac{16}{7}} = 1{,}51$$

2 Covariance

Dans l'étude des variables aléatoires, il est parfois souhaité de décider à quel point deux variables sont corrélées. Lorsqu'elles sont indépendantes, elles ne le sont pas du tout, mais dans le cas contraire, il se peut qu'elles soient plus ou moins fortement liées.

Définition 2.1. Soient X et Y deux variables aléatoires définies sur le même ensemble fondamental. La *covariance* est définie par

$$\text{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])].$$

N.B : $\text{Cov}(X,X) = \text{Var}(X)$

Exemple 2.2. Soient X et Y deux variables aléatoires valant 1 ou zéro, selon que l'on tire un as ou non dans le cas de X, et une carte de trèfle ou non dans le cas de Y dans un jeu de 36 cartes.

$E[X] = 1\cdot P\{X=1\} + 0\cdot P\{X=0\} = P\{X=1\} = \frac{4}{36} = \frac{1}{9}$

De même $E[Y] = P\{Y=1\} = \frac{9}{36} = \frac{1}{4}$

Calculons $(X-E[X])(Y-E[Y])$ et $P\{(x;y)\}$ pour chaque couple (X,Y)

$(1;1)$ → as trèfle $P\{(1;1)\} = \frac{1}{36}$ et $\left(1-\frac{1}{9}\right)\left(1-\frac{1}{4}\right) = \frac{2}{3}$

$(1;0)$ → 3 autres as $P\{(1;0)\} = \frac{3}{36} = \frac{1}{12}$ et $\left(1-\frac{1}{9}\right)\left(0-\frac{1}{4}\right) = -\frac{2}{9}$

$(0;1)$ → trèfle pas as $P\{(0;1)\} = \frac{8}{36} = \frac{2}{9}$ et $\left(0-\frac{1}{9}\right)\left(1-\frac{1}{4}\right) = -\frac{1}{12}$

$(0;0)$ → ni trèfle, ni as $P\{(0;0)\} = \frac{24}{36} = \frac{2}{3}$ et $\left(0-\frac{1}{9}\right)\left(0-\frac{1}{4}\right) = \frac{1}{36}$

$$\Rightarrow \text{Cov}(X;Y) = \frac{1}{36}\cdot\frac{2}{3} + \frac{1}{12}\left(-\frac{2}{9}\right) + \frac{2}{9}\left(-\frac{1}{12}\right) + \frac{2}{3}\cdot\frac{1}{36} = \frac{1}{54} - \frac{1}{54} - \frac{1}{54} + \frac{1}{54} = 0$$

Avant de continuer, observons deux choses. La première est que cette manière de calculer est quelque peu laborieuse, et nous allons développer d'autres méthodes de calcul. La seconde est que la covariance est nulle dans ce cas car les variables sont indépendantes, et nous verrons qu'effectivement la covariance permet de mesurer la corrélation entre X et Y.

Proposition 2.3. *On a* $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$. (Idem formule de König)

Démonstration. On développe le produit $(X - E[X])(Y - E[Y])$ et on applique la linéarité de l'espérance pour obtenir

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[X\cdot Y] - E[X\cdot E[Y]] - E[E[X]\cdot Y] + E[E[X]\cdot E[Y]] \\ &= E[XY] - E[Y]\cdot E[X] - E[X]E[Y] + E[X]\cdot E[Y] \\ &= E[XY] - E[Y]\cdot E[X]\end{aligned}$$

□

Exemple 2.4. Calculons la covariance de l'exemple 2.2 à l'aide de la formule :

$$\begin{aligned}\mathrm{Cov}(X,Y) &= 1\cdot 1\cdot\tfrac{1}{36} + 1\cdot 0\cdot \ldots + 0\cdot 1 \ldots + 0\cdot 0\cdot\ldots - \tfrac{1}{4}\cdot\tfrac{1}{9} \\ &= \tfrac{1}{36} - \tfrac{1}{36} = 0\end{aligned}$$

La notion d'indépendance que nous avons vue pour les événements se définit aussi pour variables aléatoires : deux variables sont indépendantes si tous les événements qu'elles décrivent le sont.

Définition 2.5. Deux variables aléatoires X et Y sont *indépendantes* si pour tout choix de sous-ensembles $A, B \subset \mathbb{R}$ on a

Rappel : A,B indépendants $\Leftrightarrow P(A\cap B) = P(A)\cdot P(B)$

$$P\{X\in A, Y\in B\} = P\{X\in A\}\cdot P\{Y\in B\}$$

Proposition 2.6. *Si* X *et* Y *sont indépendantes, alors* $\mathrm{Cov}(X,Y) = 0$.

Démonstration. Avec la proposition 2.3, il suffit de montrer que $E[XY] = E[X]E[Y]$.

$$\begin{aligned}E[X\cdot Y] &= \sum_a\sum_b a\cdot b\cdot \underbrace{P\{X=a \text{ et } Y=b\}}_{= P\{X=a\}\cdot P\{Y=b\} \text{ car } X \text{ et } Y \text{ sont indépendantes.}} \\ &= \sum_a\sum_b a\cdot b\cdot P\{X=a\}\cdot P\{Y=b\} \\ &= \underbrace{\sum_a a\cdot P\{X=a\}}_{E[X]}\cdot\underbrace{\sum_b b\cdot P\{Y=b\}}_{E[Y]} \\ &= E[X]\cdot E[Y]\end{aligned}$$

□

Vous verrez en exercice que la réciproque est fausse : la covariance peut être nulle sans pour autant que les variables soient indépendantes.

3 Régression linéaire des moindres carrés

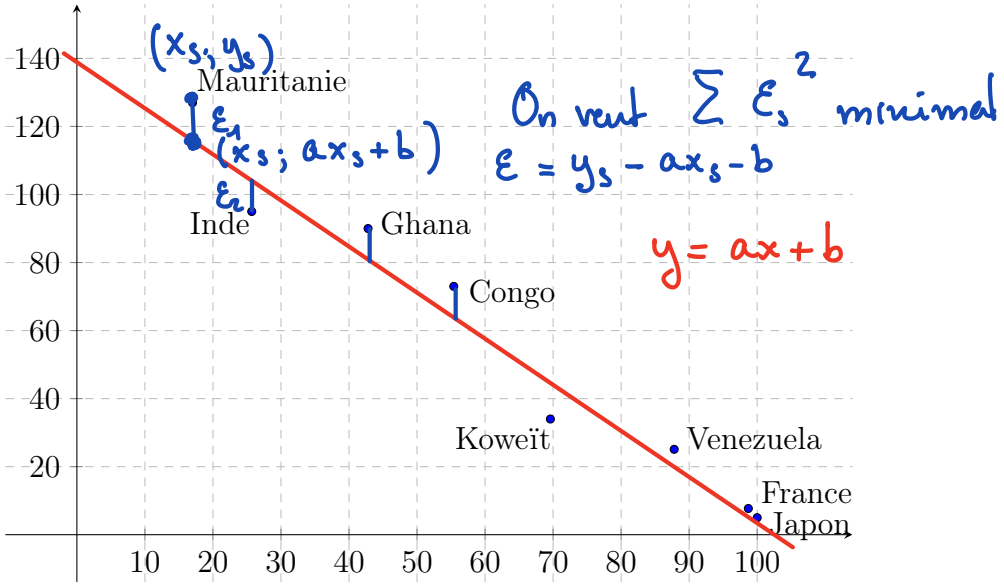
Exemple 3.1. Dans le tableau ci-dessous, X désigne le taux en % d'alphabétisation des femmes et Y le taux en ‰ de mortalité infantile dans les années 1980. On peut en effet supposer que le taux d'alphabétisation des femmes a de l'impact sur le taux de mortalité infantile.

Pays	**Inde**	**Koweït**	**Mauritanie**	**France**	**Ghana**	**Congo**	**Venezuela**	**Japon**
X [%]	25.7	69.6	17	98.7	42.8	55.4	87.8	100
Y [‰]	95	34	127	7.7	90	73	25.1	5

Dans cet exemple, l'ensemble fondamental S est l'ensemble des pays étudiés.

Pour tout élément $s \in S$, nous disposons de deux valeurs $x_s = X(s)$ et $y_s = Y(s)$.
Si nous plaçons toutes les paires (x_s, y_s) dans le plan $\mathbb{R}^2$, nous obtenons un nuage de points.
Le but est de tracer une droite qui approche "le mieux possible" ce nuage.

Taux d'alphabétisation des femmes et taux de mortalité infantile.



La forme du nuage de points suggère que Y est liée à X par une relation affine. Mais laquelle ? Nous allons déterminer la droite de régression obtenue avec la méthode des moindres carrés.

Historiquement, le premier texte paru, faisant mention de la méthode des moindres carrés, est dû à Adrien-Marie Legendre (1752-1833), dans un article sur ses « nouvelles méthodes pour la détermination des orbites des comètes » publié en 1805. Un an plus tard, Carl Friedrich Gauss (1777-1855) fait aussi allusion à cette méthode.

On cherche donc une droite d'équation $y = ax + b$ telle que les valeurs $ax_s + b$ soient proches des valeurs y_s. Concrètement, on aimerait minimiser la distance entre ces deux valeurs, pour tous les s. Puisque la différence $\varepsilon_s = y_s - ax_s - b$ peut être positive ou négative, on va minimiser la somme des carrés de ces différence ε_s.
Calculons donc a et b pour minimiser l'expression

$$S = \sum_s (y_s - ax_s - b)^2$$

Développons d'abord cette somme de carrés comme une fonction de b :

$$S(b) = \sum_s ((y_s - ax_s) - b)^2 = \sum_s \left((y_s - ax_s)^2 - 2(y_s - ax_s)\, b + b^2 \right)$$

($A = y_s - ax_s$, $B = b$, A^2, $-2AB$, $+B^2$)

$$= \sum_s (y_s - ax_s)^2 - 2b \sum_s (y_s - ax_s) + \sum_s b^2$$

$n = |S|$

$$= nb^2 - 2b \frac{n}{n} \sum_s (y_s - ax_s) + \sum_s (y_s - ax_s)^2$$

$$= n\left(b^2 - 2b \frac{1}{n} \sum (y_s - ax_s) \right) + \sum_s (y_s - ax_s)^2$$

$$= n \left(b - \frac{1}{n} \sum (y_s - ax_s) \right)^2 + \dots$$

↑ ne dépend pas de b

$S(b)$ est minimale si $b = \frac{1}{n} \sum (y_s - ax_s) = \frac{1}{n} \sum y_s - a \cdot \frac{1}{n} \sum x_s = \mu_y - a\mu_x$

$\Rightarrow b = \mu_y - a\mu_x \Leftrightarrow \mu_y = a\mu_x + b$

En d'autre terme, la droite de régression que nous cherchons passe par le point "moyen" (μ_X, μ_Y) du nuage de points. Il reste donc à déterminer a.
Pour déterminer a, nous remplaçons b par la valeur $\mu_y - a\mu_x$ que nous venons de trouver dans l'expression $\sum_s (y_s - ax_s - b)^2$ de départ. La somme à minimiser est maintenant

$$\begin{aligned} S(a) &= \sum_s (y_s - ax_s - (\mu_y - a\mu_x))^2 = \sum_s ((y_s - \mu_y) - a(x_s - \mu_x))^2 \\ &= \sum_s (y_s - \mu_y)^2 - 2a \sum_s (y_s - \mu_y)(x_s - \mu_x) + a^2 \sum_s (x_s - \mu_x)^2 \\ &= n \operatorname{Var}(Y) - 2na \operatorname{Cov}(X,Y) + na^2 \operatorname{Var}(X) \\ &= n \left(a\sqrt{\operatorname{Var}(X)} - \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)} \right)^2 + \dots \end{aligned}$$

Comme précédemment, on conclut que $S(a)$ est minimale lorsque $a = \dfrac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}$

Nous avons démontré le résultat suivant :

Théorème 3.2. *Soit S l'ensemble fondamental de deux variables aléatoires X et Y. Alors l'approximation affine de Y en fonction de X qui minimise la somme des carrés $(y_s - x_s)^2$ pour autant que la variances de X soit non nulle est*

$$Z = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - \mu_X) + \mu_Y.$$

Exemple 3.3. Dans l'exemple 3.1 des taux d'alphabétisation et de mortalité, nous calculons $E[X] = 62,125$, $E[Y] = 57,1$ puis $\mathrm{Var}(X) = E[X^2] - E[X]^2 = 4768,15 - (62,125)^2 = 908,63$.

Et pour la covariance, $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 2290,86 - 62,125 \cdot 57,1 = -1256,48$.

D'où le modèle affine $Z = \dfrac{-1256,48}{908,63}(X - 62,125) + 57,1$ c'est-à-dire $Z = -1,38\,X + 143,01$.

4 Coefficient de détermination et coefficient de corrélation

La droite des moindres carrés peut être déterminée quelle que soit la forme du nuage de points pour autant que $\mathrm{Var}(X) \neq 0$, mais ce n'est pas forcément un modèle judicieux de relation entre les variables X et Y. Il faut donc disposer d'outils de mesure de la qualité de la relation obtenue. Vous montrerez en exercice qu'avec la droite des moindres carrés, la moyenne des carrés des écarts vaut

$$\frac{1}{n}\sum_s \varepsilon_s^2 \;=\; \mathrm{Var}(Y)\left(1 - \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}\right)$$

Ce résultat motive l'introduction du coefficient de détermination et du coefficient de corrélation comme indicateurs de pertinence du modèle des moindres carrés.

Définition 4.1. Soient X et Y deux variables aléatoires. Le *coefficient de détermination* $\rho^2(X,Y)$ ou plus simplement ρ^2 est défini pour autant que les variances de X et Y soient non nulles par

$$\rho^2 = \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}.$$

Par définition, ρ^2 est compris entre 0 et 1 et peut être interprété comme un pourcentage.
Si $\rho^2 = 1$, alors $\sum \varepsilon_s^2 = 0$, ce qui signifie que la relation linéaire établie entre X et Y explique le 100% de la variance de Y.
Si $\rho^2 < 1$, alors $\sum \varepsilon_s^2 > 0$ et les points du nuage ne sont pas parfaitement alignés. Seule une proportion de ρ^2 de la variance de Y s'explique par la relation établie entre X et Y alors qu'une proportion de $(1 - \rho^2)$ de la variance de Y provient d'autres facteurs d'influence que X.

Pour une **interprétation correcte** du coefficient de détermination ρ^2, on peut utiliser la formulation suivante :

Selon le modèle de régression, *nom de la variable X* explique $\rho^2 \cdot 100\%$ de la variance de *nom de la variable Y* ; $(1-\rho^2) \cdot 100\%$ de cette variance est imputable à d'autres facteurs.

Exemple 4.2. Pour le modèle de régression de l'exemple 3.3, le coefficient de détermination, après calcul de $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = 5056,66 - (57,1)^2 = 1796,25$, est obtenu comme suit :

$$\rho^2 = \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)} = \frac{(-1256,48)^2}{908,63 \cdot 1796,25} \cong 0,9678 \cong 97\%.$$

Interprétation : le taux d'analphabétisme des femmes explique 97% de la variance du taux de mortalité infantile. En conséquence, seulement 3% de cette variance s'explique par d'autres facteurs comme le taux de vaccination par exemple.

Définition 4.3. Soient X et Y deux variables aléatoires. Le *coefficient de corrélation* $\rho(X,Y)$ est défini pour autant que les variances de X et Y soient non nulles par

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Exemple 4.4. Dans l'exemple 3.1 des taux d'alphabétisation et de mortalité, nous obtenons

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{-1256,48}{\sqrt{908,63 \cdot 1796,25}} \cong -0,984.$$

Ce nombre est négatif car le taux de mortalité infantile décroît lorsque le taux d'alphabétisation des femmes croît. Par ailleurs, il est très proche de 1 en valeur absolue, ce qui indique une corrélation forte des deux variables.

Remarque 4.5.

a) Le coefficient de corrélation linéaire prend des valeurs comprises entre -1 et 1.

b) Si $\rho = 1$ ou -1, les points sont parfaitement alignés.

c) On considère généralement que la corrélation linéaire est
   - **forte** si $|\rho| \geq 0.9$; la régression linéaire exprime parfaitement le lien entre les données ;
   - **moyenne** si $0.6 \leq |\rho| < 0.9$; le modèle linéaire peut être considéré comme acceptable ;
   - **faible** si $0.2 \leq |\rho| < 0.6$; le modèle linéaire doit être remis en cause ;
   - **nulle** si $|\rho| \leq 0.2$; dans ce cas le modèle linéaire doit être rejeté. On dit alors que les variables X et Y sont **non-corrélées** linéairement.

d) Si $\rho > 0$, X et Y sont **corrélées positivement** ; la droite de régression a une pente positive. Si $\rho < 0$, X et Y sont **corrélées négativement** ; la droite de régression a une pente négative.

e) le coefficient de détermination est égal au carré du coefficient de corrélation.