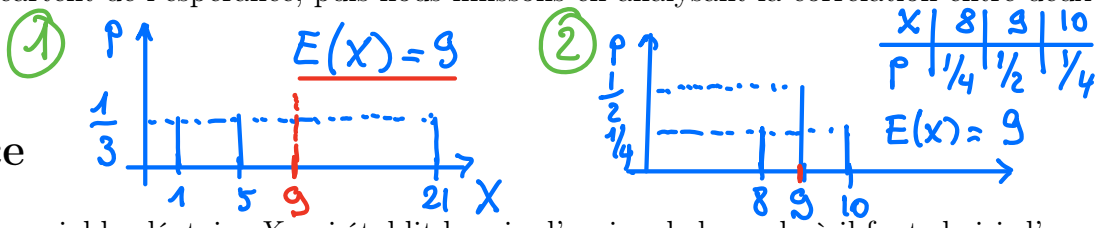
V. Variance et corrélation

La semaine passée, nous avons travaillé avec des variables aléatoires et défini l'espérance. Celle-ci donne le résultat auquel on peut s'attendre en moyenne. Pour terminer l'étude des variables aléatoires, nous définissons encore la variance, une mesure qui indique à quel point les valeurs d'une variable s'écartent de l'espérance, puis nous finissons en analysant la corrélation entre deux variables.



1 Variance

Considérons la variable aléatoire X qui établit le prix d'un jeu de hasard où il faut choisir l'une de trois portes qui cachent 1 franc, 5 francs et 21 francs.
L'espérance vaut donc $\mu = E[X] = 1 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} + 21 \cdot \frac{1}{3} = 9$, mais cette mesure ne nous dit pas comment sont réparties les valeurs de X autour d'elle.
Pour cela, considérons la différence entre les valeurs que prend X et l'espérance. Comme $X - \mu$ prend des valeurs positives et négatives, c'est une variable difficile à analyser, et on préfère dans la pratique travailler avec son carré.

Définition 1.1. Soit X une variable aléatoire et $\mu = E[X]$ son espérance. La *variance* de X est le nombre $\mathrm{Var}(X) = E[(X - \mu)^2]$.

La variance du jeu ci-dessus est donc

1) $\mathrm{Var}(X) = \frac{1}{3}(1-9)^2 + \frac{1}{3}(5-9)^2 + \frac{1}{3}(21-9)^2 \simeq 74{,}67$

⚠ l'unité de la variance = l'unité de X^2 (ici des (francs)^2)

déf 1.5 → $\sigma = \sqrt{\mathrm{Var}(X)} \simeq \sqrt{74{,}67} \simeq 8{,}64$

interprétation → $E(X) - \sigma = 9 - 8{,}64 = 0{,}36$; $E(X) + \sigma = 9 + 8{,}64 = 17{,}64$ } $I = [0{,}36 ; 17{,}64]$

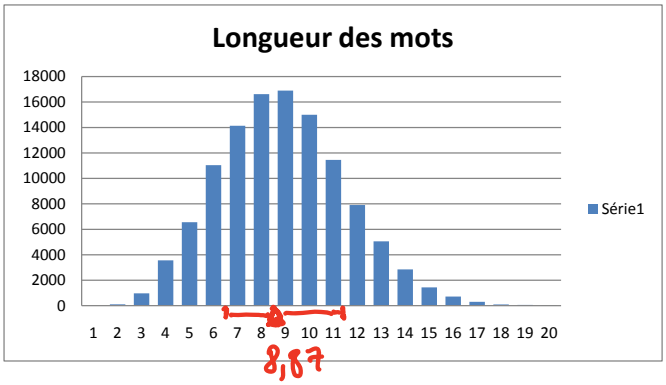
$X \in I$ avec une probabilité supérieure ou égale à 50%

2) $\mathrm{Var}(X) = \frac{1}{4}(8-9)^2 + \frac{1}{2}(9-9)^2 + \frac{1}{4}(10-9)^2 = \frac{1}{4} + 0 + \frac{1}{4} = \frac{1}{2}$

$\sigma(X) = \sqrt{0{,}5} \simeq 0{,}7$ $\quad I = [8{,}3 ; 9{,}7]$

Exemple 1.2. Une étude de John Henrick, 2008. (Statistique)
Supposons que l'on choisisse au hasard l'un des 114799 mots du dictionnaire anglais "Chambers Twentieth Century Dictionary". La variable aléatoire X compte alors la longueur de ce mot. Voici un tableau récapitulatif qui indique le nombre de mots pour chaque longueur entre 1 et 20 :

1	2	3	4	5	6	7	8	9	10
0	102	966	3559	6562	11041	14131	16616	16899	15002
11	12	13	14	15	16	17	18	19	20
11458	7910	5059	2855	1434	720	308	98	58	21



L'espérance de cette variable aléatoire mesure donc la longueur moyenne des mots de la langue anglaise. Nous avons

$$E[X] = 1 \cdot 0 + 2 \cdot \frac{102}{114799} + 3 \cdot \frac{966}{114799} + \cdots + 20 \cdot \frac{21}{114799}$$
$$\cong 8,87$$

La longueur moyenne vaut presque 9, ce qu'on aurait presque pu deviner en observant l'histogramme en colonnes. Qui l'eût cru ? Quelle est la variance ? L'histogramme semble indiquer que la majorité des mots ont une longueur proche de 9, si bien que notre intuition nous dit que la variance sera plutôt faible.

$$\mathrm{Var}(X) = (6,87)^2 \cdot \frac{102}{114799} + (5,87)^2 \cdot \frac{966}{114799} + \cdots + (11,13)^2 \cdot \frac{21}{114799}$$
$$\cong 7,07$$
$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{7,07} \simeq 2,66$$
$$8,87 - 2,66 \cong 6,2$$
$$8,87 + 2,66 \simeq 11,53$$

ce qui est effectivement assez faible s'agissant d'une mesure concernant une variable aléatoire qui prend des valeurs entre 0 et 21...

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.

définition formule de König

Proposition 1.3. *Soit X une variable aléatoire. Alors*

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

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

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

Exemple 1.4. En reprenant l'exemple ??, on peut calculer la variance à l'aide de la formule :

$$\text{Var}(X) = E[X^2] - (E[X])^2$$
$$= 2^2 \cdot \frac{102}{114'799} + 3^2 \cdot \frac{966}{114'799} + \dots + 20^2 \cdot \frac{21}{114'799} - 8{,}87^2 = 7{,}07$$

Définition 1.5. Soit X une variable aléatoire. L'*écart-type* de X est le nombre

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

2 Covariance

Notre but ultime dans l'étude des variables aléatoires est de pouvoir 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 intimement liées, ou non.

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

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

Remarque : $\text{Cov}(X,X) = E\left[(X-E[X])^2\right] = \text{Var}(X)$

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 (on utilise un jeu de 36 cartes).

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

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

Calculons $(X-E[X])(Y-E[Y])$ pour chaque combinaison (x,y) :

- Carte As de trèfle $\to (1;1)$ $\quad p_{11} = (1-\frac{1}{9})(1-\frac{1}{4}) = \frac{2}{3}$
- Carte As pas trèfle $\to (1;0)$ $\quad p_{10} = (1-\frac{1}{9})(0-\frac{1}{4}) = -\frac{2}{9}$
- Carte Pas As, trèfle $\to (0;1)$ $\quad p_{01} = (0-\frac{1}{9})(1-\frac{1}{4}) = -\frac{1}{12}$
- Carte ni As, ni trèfle $\to (0;0)$ $\quad p_{00} = (0-\frac{1}{9})(0-\frac{1}{4}) = \frac{1}{36}$

$$\mathrm{Cov}(X;Y) = \frac{1}{36}\cdot\frac{2}{3} + \frac{3}{36}\cdot\left(-\frac{2}{9}\right) + \frac{8}{36}\left(-\frac{1}{12}\right) + \frac{24}{36}\cdot\frac{1}{36}$$
$$= \ldots = 0$$

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 où 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* $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$.

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

$$\mathrm{Cov}(X,Y) = E\left[(X-E[X])(Y-E[Y])\right]$$
$$= E\left[XY - XE[Y] - E[X]\cdot Y + E[X]\cdot E[Y]\right]$$
$$= E[X\cdot Y] - E\left[XE[Y]\right] - E\left[E[X]\cdot Y\right] + E\left[E[X]\cdot E[Y]\right]$$
$$= E[X.Y] - E[X]E[Y] - E[X]\cdot E[Y] + E[X]\cdot E[Y]$$
$$= E[X.Y] - E[X]\cdot E[Y]$$

□

Exemple 2.4. En reprenant l'exemple 2.2., on peut calculer la covariance à l'aide de la formule :

$$\mathrm{Cov}(X;Y) = \underbrace{1\cdot 1\cdot\frac{1}{36} + 1\cdot 0\cdot\frac{3}{36} + 0\cdot 1\cdot\frac{8}{36} + 0\cdot 0\,\frac{24}{36}}_{E(X\cdot Y)} - \underbrace{\frac{1}{9}\cdot\frac{1}{4}}_{E[X]\cdot E[Y]} = \frac{1}{36} - \frac{1}{36} = 0$$

On définit pour des variables aléatoires la notion d'indépendance que nous avons déjà vue pour des événements. En gros, deux variables sont indépendantes si tous les événements décrits par elles 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

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

Pour calculer la covariance de deux variables indépendantes, nous devons savoir calculer l'espérance de leur produit.

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

Démonstration. Nous devons montrer que $E[XY] = E[X]E[Y]$ et donc calculer l'espérance du produit :

$$E[X\cdot Y] = \sum_c c\cdot P(X\cdot Y = c) = \sum_{c\neq 0} c\, P\Big\{\bigcup_{a\neq 0}\{X=a;\, Y=\tfrac{c}{a}\}\Big\}$$

$$= \sum_{c\neq 0} c \sum_{a\neq 0} P\{X=a;\, Y=\tfrac{c}{a}\}$$

$\amalg \to \Sigma$; X et Y indép $\to P\{X=a\}\cdot P\{Y=\tfrac{c}{a}\}$

$$= \sum_{a\neq 0} a\, P\{X=a\} \sum_{c\neq 0} \frac{c}{a}\cdot P\{Y=\tfrac{c}{a}\}$$

en posant $b = \frac{c}{a}$

$$= E[X]\cdot E[Y]$$

□

Tu verras en exercice que la réciproque est fausse : La covariance peut être nulle sans pour autant que les variables soient indépendantes. Notre but suivant est d'étudier la corrélation entre deux variables.

Exemple 2.7. Dans le tableau ci-dessous, X désigne le taux en % d'alphabétisation des femmes et Y le taux en ‰ de mortalité infantile. 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

La moyenne du taux d'alphabétisation est $E[X] = 62,125$% alors que la moyenne du taux de mortalité infantile est $E[Y] = 57,1$‰
Les variances respectives sont $\mathrm{Var}(X) = E[X^2] - E[X]^2 = 4768,1475 - (62,125)^2 = 908,631875$ et $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = 5056,6625 - (57,1)^2 = 1796,2525$.

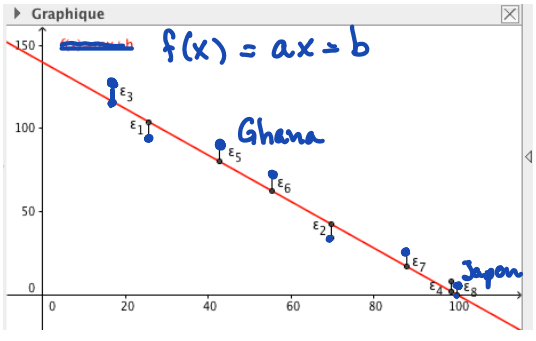
Passons maintenant à la covariance. On applique la formule que nous avons démontrée ci-dessus :

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 2290,85875 - 62,125 \cdot 57,1 = -1256,47875 \neq 0$$

Ce nombre n'est pas nul et suggère que les variables X et Y ne sont pas indépendantes ! Mais la valeur de la covariance est grande et ne nous permet pas de définir si le lien est fort ou non. Nous avons besoin d'un autre outil pour mesurer à quel point X et Y sont corrélées...

3 Regression linéaire et corrélation

Rappelons que nous étudions deux variables X et Y dont nous disposons des valeurs sur l'ensemble fondamental. Autrement dit, 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 du jeu est de tracer une droite qui "approxime" le mieux possible ce nuage.



Chaque point correspond à un pays

On cherche la droite qui minimise $\sum \varepsilon_i^2 = \sum \left(y_i - (ax_i + b)\right)^2$

On cherche donc une droite d'équation $y = ax + b$ telle que la valeur $ax_s + b$ soit proche de 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 veut minimiser la valeur absolue de cette différence, ou mieux encore son carré.
Comment donc trouver a et b pour que le nombre

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

soit minimal ? Regardons d'abord la dépendance en fonction de b. On développe cette somme de carrés :

$$S(b) = \sum_s \big((y_s - ax_s) - b\big)^2 = \sum_s \big((y_s-ax_s)^2 - 2(y_s-ax_s)b + b^2\big)$$
$$= \sum_s (y_s-ax_s)^2 - 2b\sum_s (y_s-ax_s) + \sum_s b^2$$
$$\overset{n=|S|}{=} n\,b^2 - 2\sum_s (y_s-ax_s)\cdot b + \sum_s (y_s-ax_s)^2$$
$$= n\left(b - \tfrac{1}{n}\sum_s (y_s-ax_s)\right)^2 + \dots$$

$S(b)$ est minimal lorsque $b = \frac{1}{n}\sum_s (y_s - ax_s)$

$$\iff b = \tfrac{1}{n}\sum_s y_s - a\,\tfrac{1}{n}\sum x_s$$
$$\iff b = \bar{y} - a\bar{x} \iff \bar{y} = a\bar{x} + b$$

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 trouvée dans l'expression de départ $\sum_s (y_s - ax_s - b)^2$. La somme à minimiser est maintenant

$$S(a) = \sum_s \big(y_s - ax_s \underbrace{-\bar{y} + a\bar{x}}_{=b}\big)^2 = \sum_s \big((y_s-\bar{y}) - a(x_s-\bar{x})\big)^2$$
$$= \sum_s \Big[(y_s-\bar{y})^2 - 2a(x_s-\bar{x})(y_s-\bar{y}) + a^2(x_s-\bar{x})^2\Big]$$
$$= \sum_s (x_s-\bar{x})^2\cdot a^2 - 2\sum (x_s-\bar{x})(y_s-\bar{y})\cdot a + \dots$$
$$= n\,\mathrm{Var}(X)\cdot a^2 - 2n\,\mathrm{Cov}(X,Y)\cdot a + \dots$$
$$= n\left(a\sqrt{\mathrm{Var}(X)} - \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}\,X}}\right)^2 + \dots$$

$S(a)$ est minimal lorsque $a = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$

Nous avons démontré le résultat suivant :

Théorème 3.1. *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 $(y_s - x_s)^2$ est*

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

Exemple 3.2. Taux d'alphabétisation et de mortalité
Dans l'exemple du taux d'alphabétisation et du taux de mortalité, nous obtenons

$$Z = -1,38(X - 62,125) + 57,1 = -1,38X + 143,008$$

Si on calcule l'erreur moyenne faite avec cette approximation, on obtient la valeur

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

Ceci motive l'introduction du coefficient de corrélation.

Définition 3.3. 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 sont non nulles par

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

Dans l'exemple du taux d'alphabétisation et du taux de mortalité, nous obtenons

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{-1256,47875}{\sqrt{908,631875 \cdot 1796,2525}} \cong -0,9835.$$

Ce nombre est très proche de 1 en valeur absolue, ce qui indique une corrélation forte des deux variables.

Remarque 3.4. a) Le coefficient de corrélation linéaire prend des valeurs comprises entre -1 et 1.

b) 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.

c) 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.