Variance et Écart Type

- Variance = indicateur de dispersion

Variance

Soit X une variable aléatoire discrète avec $X(\Omega) = \{x_1, x_2, \ldots\}$, fini ou infini, admettant une espérance. La variance de X, noté var(X), est le réel

$$\mathrm{var}(X) = E[(X - E[X])^2] = \sum_i (x_i - E[X])^2 P(X = x_i) = E[X^2] - E[X]^2$$

à condition que cette série converge

Écart type

Soit X une variable aléatoire discrète admettant une variance. L'écart type de X, noté σ_X, est le réel $\sigma_X = \sqrt{\mathrm{var(X)}}$

- Écart type (ou variance) d'une variable aléatoire discrète X = mesure (très grossière) de la dispersion des valeurs que peut prendre X autour de sa moyenne (espérance)

$\Longrightarrow$ plus l'écart type est petit, plus il y a des chances que X soit proche de son espérance

Exemple

Lancer de deux dés successivement

- $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$
- $X : \Omega \to \{2, 3, \ldots, 12\}$ avec $X((i, j)) = i + j$
- $\mathrm{var}(X) = E[X^2] - E[X]^2$
- On a calculé $E[X] = 7$
- X^2 est une variable aléatoire (i.e., $g : \mathbb{R} \to \mathbb{R}$ avec $g(x) = x^2$)
- Théorème du transfert: $E[X^2] = \sum_{i=2}^{12} i^2 P(X = i)$
  $E[X^2] = 4.\frac{1}{36} + 9.\frac{2}{36} + 16.\frac{3}{36} + 25.\frac{4}{36} + 36.\frac{5}{36} + 49.\frac{6}{36} + 64.\frac{5}{36} + 81.\frac{4}{36} + 100.\frac{3}{36} + 121.\frac{2}{36} + 144.\frac{1}{36} = \frac{1974}{36} \approx 54.83$
- $\mathrm{var}(X) = \frac{1974}{36} - 7^2 = \frac{210}{36} = \frac{35}{6} \approx 5.83$
- $\sigma_X = \sqrt{\frac{35}{6}} \approx 2.415$

Variance - Propriété

Proposition

Si X est une varible aléatoire discrète finie ayant une variance nulle, alors X est une variable aléatoire constante

Proposition

Si X est une varible aléatoire admettant une variance, alors pour tous $a, b \in \mathbb{R}$ on a $\mathrm{var}(aX + y) = a^2\mathrm{var}(X)$

Proposition

Si X et Y sont deux variables aléatoires indépendantes, alors $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$

Proposition

Si $X_1, \ldots, X_n$ sont des variables aléatoires (deux à deux) indépendantes. Alors $\mathrm{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{var}(X_i)$

Covariance

- Covariance = mesure de comment deux variables aléatoires varient ensemble

Covariance

Soient X et Y deux variables aléatoires telles que $E[X^2]$ et $E[Y^2]$ existent. La covariance de X et Y, notée $\mathrm{cov}(X, Y)$, est définie par $\mathrm{cov}(X, Y) = \sigma_{XY} = E[(X - E[X])((Y - E[Y])] = E[XY] - E[X]E[Y]$

- Loi conjointe ($f : \mathbb{R}^2 \to \mathbb{R}$ avec $f(x, y) = (x - E[X])(y - E[Y])$)
  $\mathrm{cov}(X, Y) = \sum_{(x,y)\in X(\Omega)\times Y(\Omega)} (x - E[X])(y - E[Y])P(X = x, Y = y)$
- Loi marginale

$$\sum_{(x,y)\in X(\Omega)\times Y(\Omega)} xP(X = x, Y = y) = \sum_{x\in X(\Omega)} x \sum_{y\in Y(\Omega)} P(X = x, Y = y) = \sum_{x\in X(\Omega)} xP(X = x)$$

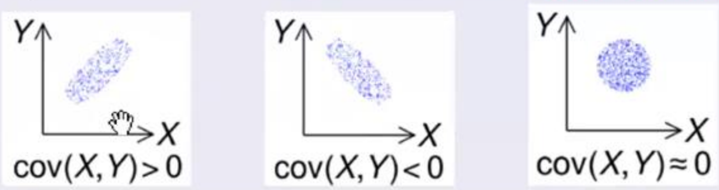
$$\Longrightarrow \mathrm{cov}(X, Y) = \left(\sum_{(x,y)\in X(\Omega)\times Y(\Omega)} xyP(X = x, Y = y) \right) - E[X]E[Y]$$

Covariance - Interprétation

Remarque

La covariance de deux variables aléatoires X et Y correspond à la direction de la relation linéaire entre ces deux variables

- $\mathrm{cov}(X, Y) > 0$: X et Y évoluent dans le même sens (positivement corrélées)
- $\mathrm{cov}(X, Y) < 0$: X et Y évoluent dans le sens contraire (négativement corrélées)
- $\mathrm{cov}(X, Y) \approx 0$: X et Y ne sont liées entre elles (non-corrélée)



- $\mathrm{cov}(X, Y)$ qualifie la relation des variables aléatoires X et Y
- l'unité de $\mathrm{cov}(X, Y)$ est égale au produit des unités de X et Y
- $\mathrm{cov}(X, Y)$ ne quantifie pas la force de la relation

Covariance - Exemple #1

- Soit X une variable aléatoire telle que $P(X = x) = \frac{1}{5}$ pour $x \in \{-2, -1, 0, 1, 2\}$
- Soit $Y = X^2$
- $\mathrm{cov}(X, Y)$?

Y \ X	-2	-1	0	1	2	$P(Y = y)$
0	0	0	$\frac{1}{5}$	0	0	$\frac{1}{5}$
1	0	$\frac{1}{5}$	0	$\frac{1}{5}$	0	$\frac{2}{5}$
4	$\frac{1}{5}$	0	0	0	$\frac{1}{5}$	$\frac{2}{5}$
$P(X = x)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	

  - $E[X] = 0$ et $E[Y] = 2$
  - $\mathrm{cov}(X, Y) = \frac{1}{5}(-8 - 1 + 1 + 8) - 0 = 0$
- X et Y indépendants ?
  - $P(X = -2, Y = 0) = 0 \neq P(X = -2)P(Y = 0) = \frac{1}{5} \cdot \frac{1}{5}$
  - X et Y sont dépendants : si la valeur de X est connue, celle de Y est connue avec certitude
  - X et Y ont une relation quadratique (pas linéaire)

Covariance - Propriétés

Proposition

(i) $\text{cov}(X, c) = E[(X - E[X])((c - c)] = 0$
(ii) $\text{cov}(X, Y) = \text{cov}(Y, X)$
(iii) $\text{cov}(aX_1 + bX_2, Y) = a.\text{cov}(X_1, Y) + b.\text{cov}(X_2, Y)$ pour $a, b \in \mathbb{R}$
(iv) $\text{cov}(X, aY_1 + bY_2) = a.\text{cov}(X, Y_1) + b.\text{cov}(X, Y_2)$ pour $a, b \in \mathbb{R}$
(v) $\text{cov}(aX, bY) = ab.\text{cov}(X, Y)$ pour $a, b \in \mathbb{R}$
(vi) $\text{cov}(X, X) = E[X^2] - E[X]^2 = \text{var}(X)$
(vii) Si X et Y sont indépendantes alors $\text{cov}(X, Y) = 0$ (pas le contraire)
(viii) $|\text{cov}(X, Y)| \leq \sqrt{var(X)} + \sqrt{var(Y)}$ (inégalité de Cauchy-Schwarz)

Proposition

Si $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ sont des variables aléatoires alors

- $\text{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{cov}(X_i, X_j)$
- $\text{cov}(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{cov}(X_i, Y_j)$ pour $a_1, \ldots, a_n, b_1, \ldots, b_m \in \mathbb{R}$

Covariance - Exemple #2

Soient n boules, numérotées de 1 à n, aléatoirement mises dans n boîtes, numérotées de 1 à n, à raison de une boule par boîte

- $\Omega = \{\text{permutations}\}$
- Soit $S_n =$ le nombre de boules mises dans les boîtes de même numéro
- $S_n(\Omega) = \{0, 1, \ldots, n\}$
- $E[S_n]$?
  - Soient $X_j = 1$ si boule j dans boîte j, 0 sinon ($j = 1, \ldots, n$)
  - $S_n = X_1 + \cdots + X_n$
  - $P(X_j = 1) = \frac{(n-1)!}{n!} = \frac{1}{n}$ et $E[X_j] = \frac{1}{n}$ pour ($j = 1, \ldots, n$)
  - $E[S_n] = E[X_1] + \cdots + E[X_n] = n.\frac{1}{n} = 1$
- $\text{var}(S_n)$?
  - $X_j^2 = X_j$ ($j = 1, \ldots, n$)
  - $\text{var}(X_j) = E[X_j^2] - E[X_j]^2 = E[X_j] - E[X_j]^2 = \frac{1}{n} - \frac{1}{n^2} = \frac{n-1}{n^2}$ ($j = 1, \ldots, n$)
  - $P(X_j X_k = 1) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$ ($1 \leq j < k \leq n$)
  - $\text{cov}(X_j, X_k) = E[X_j X_k] - E[X_j]E[X_k] = \frac{1}{n(n-1)} - 1 = \frac{1}{n^2(n-1)}$ ($1 \leq j < k \leq n$)
  - $\text{var}(S_n) = \sum_{j=1}^{n} \text{var}(X_j) + 2 \sum_{1 \leq j < k \leq n} \text{cov}(X_j, X_k) = n\frac{n-1}{n^2} + 2\binom{n}{2}\frac{1}{n^2(n-1)} = 1$

Matrice de Covariance

Matrice de covariance

Soient $X_1, \dots, X_n$ n variables alétoires. La $n \times n$ matrice

$$\Sigma = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \dots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \dots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \dots & \mathrm{var}(X_n) \end{bmatrix}$$

est appelée matrice de covariance de $(X_1, \dots, x_n)$

$$\mathbf{z}^T \Sigma \mathbf{z} = \sum_{1 \le i,j \le n} z_i \mathrm{cov}(X_i, X_j) z_j = \sum_{1 \le i,j \le n} \mathrm{cov}(z_i X_i, z_j X_j) = \mathrm{var}(X_1, \dots, X_n)$$

= somme (variance) ??

Propriétés

1. La matrice de covariance est symétrique
2. La matrice de covariance est semi-définie positive (i.e., $\mathbf{z}^T \Sigma \mathbf{z} \ge 0$ pour tour vecteur non-nul $\mathbf{z}$ ou les valeurs propres de Σ sont positives ou nulles)
3. La matrice de covariance est inversible

Matrice de Covariance - Exemple

- Soient X_1 et X_2 deux variables aléatoires
- Matrice de covariance $\Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$
- $\mathrm{var}(3X_1 + 4X_2)$?
- $\mathrm{var}(3X_1 + 4X_2) = \begin{bmatrix} 3 & 4 \end{bmatrix} \Sigma \begin{bmatrix} 3 \\ 4 \end{bmatrix}$
- $\mathrm{var}(3X_1 + 4X_2) = \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix}$
- $\mathrm{var}(3X_1 + 4X_2) = 92$

Coefficient de Corrélation

- les valeurs de covariance ne sont pas normalisées (entre $-\infty$ et $+\infty$)

$\Rightarrow$
1. difficile de comparer des covariances
2. difficile de déterminer la force de la relation linéaire entre deux variables

- Coefficient de corrélation = covariance normalisée

$\Rightarrow$
1. plus de dépendance aux unités
2. mesure la force de la relation linéaire entre deux variables

Coefficient de Corrélation

Le coefficient de corrélation $\rho(X, Y)$ de deux variables aéatoires X et Y, de variances non-nulles, est égale à la covariance de X et Y divisée par le produit de leur écart-type, i.e., $\rho(X, Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$

Propriétés

(i) $-1 \leq \rho(X, Y) \leq 1$
(ii) $\rho(X, Y) = 1$ si et seulement si $Y = aX + b$ avec $a > 0$
(iii) $\rho(X, Y) = -1$ si et seulement si $Y = aX + b$ avec $a < 0$
(iv) Si X et Y sont indépendantes alors $\rho(X, Y) = 0$ (pas le contraire)

Coefficient de Corrélation - Exemple

- Lancer d'une pièce trois fois de rang
- X = nombre de fois où Face est obtenu durant les deux premiers lancers
- Y = nombre de fois où Face est obtenu durant les deux derniers lancers
- cov(X, Y)?
Y \ X	0	1	2	$P(Y = y)$
0	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{1}{4}$
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$
2	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	
  - $E[X] = E[Y] = 1$
  - $E[XY] = \sum_{x,y \in \cancel{\{0,1,2\}}\ \{0,1,2\}} xyP(X = x, Y = y) = 1.\frac{1}{4} + 2.\frac{1}{8} + 2.\frac{1}{8} + 4.\frac{1}{8} = \frac{5}{4}$
  - cov$(X, Y) = E[XY] - E[X]E[Y] = \frac{5}{4} - 1 = \frac{1}{4}$
- $\rho(X, Y)$?
  - var$(X) = E[X^2] - E[X]^2 = \left(1^2.\frac{1}{2} + 2^2.\frac{1}{4}\right) - 1^2 = \frac{1}{2}, \sigma_Y = \frac{1}{\sqrt{2}}$
  - var$(Y) = \frac{1}{2}, \sigma_Y = \frac{1}{\sqrt{2}}$
  - $\rho(X, Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_y} = \frac{\frac{1}{4}}{\frac{1}{\sqrt{2}}.\frac{1}{\sqrt{2}}} = \frac{1}{2}$
  - Y augmente quand X augmente, Y diminue quand X diminue (car 2$^{\text{ème}}$ lancer inclu dans X et Y)

Loi Uniforme

Loi uniforme (cas discret)

Une variable aléatoire discrète X suit une loi uniforme sur $\{1, \dots, n\}$, notée $\mathcal{U}(n)$, si $P(X = i) = \frac{1}{n}$ pour tout $i \in \{1, \dots, n\}$

Loi uniforme (cas discret) - Espérance et variance

(i) $E[X] = \frac{n+1}{2}$

(ii) $\text{var}(X) = \frac{n^2-1}{12}$

Exemple:

- Lancer d'un dé à six faces ($\Omega = \{1, \dots, 6\}$)
- X = numéro de la face supérieure ($X(\Omega) = \{1, \dots, 6\}$)
- $X \sim \mathcal{U}(6)$
- $E[X] = \frac{6+1}{2} = \frac{7}{2}$
- $\text{var}(X) = \frac{6^2-1}{12} = \frac{35}{12}$

Loi de Bernoulli

Loi de Bernoulli

Une variable aléatoire X suit une loi de Bernoulli de paramètre p, notée $\mathcal{B}(1, p)$, si $X(\Omega) = \{0, 1\}$ et $P(X = 1) = p$

Loi uniforme (cas continu) - Espérance et variance

(i) $E[X] = p$

(ii) $\text{var}(X) = pq$ avec $q = 1 - p$

Exemple:

- Une carte est tirée dans un jeu de 52 cartes
- Gain de 1 euro si un as est tiré; gain nul si une autre carte est tirée
- X = gain
- $\Omega = \{$une carte parmi les 52$\}$ - équiprobabilité
- $X(\Omega) = \{0, 1\}$
- $P(X = 1) = \frac{4}{52} = \frac{1}{13}$
- $X \sim \mathcal{B}(1, \frac{1}{13})$
- $E[X] = \frac{1}{13}$
- $\text{var}(X) = \frac{1}{13}(1 - \frac{1}{13}) = \frac{12}{169}$

Loi binomiale

Loi binomiale

Une variable aléatoire X suit une loi binomiale de paramètre (n, p), où $n \in \mathbb{Z}_+^*$ et $0 \leq p \leq 1$, notée $\mathcal{B}(n, p)$, si $X(\Omega) = \{0, 1, \dots, n\}$ et $P[X = i] = \binom{n}{i} p^i q^{n-i}$ où $q = 1 - p$

Loi binomiale - Espérance et variance

(i) $E[X] = np$

(ii) $\mathrm{var}(X) = npq$

Exemple:

- Une urne contient a boules blanches et b boules rouges. On tire n boules une à une succèssivement avec remise.
- X = nombre de boules blanches tirées
- $\Omega = \{(x_1, \dots, x_n) : x_i \in \{\text{blanche}, \text{rouge}\}, i = 1, \dots, n\}$
- $P(X = i) = \binom{n}{i} \left(\frac{a}{a+b}\right)^i \left(\frac{b}{a+b}\right)^{n-i}, i = 1, \dots, n$
- $p = \frac{a}{a+b}$ et $X \sim \mathcal{B}(n, \frac{a}{a+b})$
- $E[X] = \frac{na}{a+b}$
- $\mathrm{var}(X) = n\frac{a}{a+b}(1 - \frac{a}{a+b}) = \frac{nab}{(a+b)^2}$

Loi binomiale et Loi de Bernoulli

Propriété

Soient n variables de Bernoulli mutuellement indépendantes et de de même paramètre $X_i \sim \mathcal{B}(1, p)$, alors $X_1 + \dots + X_n \sim \mathcal{B}(n, p)$

- Schéma binomial : tirages successifs avec remise (obtention de i succès parmi n épreuves de Bernoulli identiques et mutuellement indépendantes)
- $E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np$
- $\mathrm{var}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathrm{var}[X_i] = \sum_{i=1}^{n} pq = npq$

Proposition

Soient $X_1, \dots, X_k$ des variables aléatoires mutuellement indépendantes telles que $X_i \sim \mathcal{B}(n_i, p)$, $i = 1, \dots, k$, alors $X_1 + \dots + X_k \sim \mathcal{B}(n_1 + \dots + n_k, p)$

Slide 40 bug… Loi Géométrique

Loi de Pascal

- Loi de Pascal : on s'arrête au r$^{\text{ème}}$ succès

Loi de Pascal

Une variable aléatoire X suit une loi de Pascal de paramètre (r, p), où $0 < p \leq 1$, notée $\mathcal{G}(r, p)$, si $X(\Omega) = \{r, r+1, \dots\}$ et $P(X = i) = \binom{i-1}{r-1} p^r q^{i-r}$ où $q = 1 - p$

Loi de Pascal - Espérance et variance

$E[X] = r\frac{1}{p}$ et $\text{var}(X) = r\frac{q}{p^2}$

- Schéma de Pascal : temps d'attente du premier succès lors d'une répétition indépendante d'une expérience de Bernoulli
- $X_i = 1$ obtenir un succès au i$^{\text{ème}}$ essai, $X_i = 0$ sinon
- $X_i \sim \mathcal{B}(1, p)$
- suite $(X_i)_{i \in \mathbb{Z}_+^*}$ est mutuellement indépendante
- $(X = k) = \{(r-1)$ succès en $(k-1)$ expériences de Bernoulli et $X_k = 1\}$

$$\begin{aligned} P(X = k) &= P(X_1 + \cdots + X_{k-1} = r - 1 \cap X_k = 1) \\ &= P(X_1 + \cdots + X_{k-1} = r - 1) P(X_k = 1) = \binom{k-1}{r-1} p^{r-1} q^{k-r} p \end{aligned}$$

Loi de Poisson

Loi de Poisson

Une variable aléatoire X suit une loi de Poisson de paramètre λ, où $\lambda > 0$, notée $\mathcal{P}(\lambda)$, si $X(\Omega) = \mathbb{Z}_+$ et $P[X = i] = e^{-\lambda}\frac{\lambda^i}{i!}$

Loi de Poisson - Espérance et variance

$E[X] = \lambda$ et $\text{var}(X) = \lambda$

Utilisation de la loi de Poisson :

- nombre de tâches arrivant sur un serveur pendant une minute
- nombre de globules rouges par ml de sang
- nombre d'accidents de travail dans une entreprise pendant une année

Proposition

Soient $X_1, \ldots, X_k$ des variables aléatoires mutuellement indépendantes telles que $X_i \sim \mathcal{P}(\lambda_i)$, $i = 1, \ldots, k$, alors $X_1 + \cdots + X_k \sim \mathcal{P}(\lambda_1 + \cdots + \lambda_k)$

Loi de Poisson - Exemple

Un système électronique a un cycle d'opération périodique de 0.01 seconde. Dans chaque cycle, un évènement occure avec une probabilité 0.001. Quelle est la probabilité d'observer moins de 15 évènements dans un intervalle de 100 secondes ?

- $X = \{$nombre d'évènements observés pendant 100 secondes$\}$
- 100 secondes = 10000 cycles observés
- $X_i = \{$nombre d'évènements observés pendant le i$^{\text{ème}}$ cycle$\}$, $i = 1, \ldots, 10000$
- $X_i \sim \mathcal{P}(0.001)$, $i = 1, \ldots, 10000$
- $X = X_1 + \cdots + X_{10000} \sim \mathcal{P}(10000 * 0.001) = \mathcal{P}(10)$ (indépendance mutuelle)
- $P(X < 15) = P(X \leq 14) = \sum_{k=0}^{14} P(X = k) = \sum_{k=0}^{14} e^{-10}\frac{10^k}{k!} \approx 0.9165$

Loi de Poisson et Loi Binomiale

Approximation

La loi de Poisson de paramètre np est une "approximation" de la loi binomiale de paramètres (n, p) quand np est petit (e.g., $np \leq 10$) et n est grand (e.g., $n \geq 50$)

- Loi de Poisson est généralement utilisée quand il y a
  - un large nombre n d'essais
  - une petite probabilité p qu'un évènement va se réaliser pendant un essai
  - np est modéré en magnitude (i.e., pas trop grand)
- Exemple: Si 5% de la population est gauchère, quelle est la probabilité qu'un échatillon aléatoire de 100 personnes contienne au moins 2 gauchers
- $\Omega = \{100 \text{ personnes}\}$
- $X = \text{nombre de gauchers}$, $X(\Omega) = \{0, 1, \ldots, 100\}$
- $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 1)$
- $X \sim \mathcal{B}(100, \frac{5}{100}) : P(X \geq 2) = 1 - \binom{100}{0}\left(\frac{95}{100}\right)^{100} - \binom{100}{1}\left(\frac{95}{100}\right)^{99}\frac{5}{100} \approx 0.96292$
- $X \sim \mathcal{P}(100 * \frac{5}{100}) : P(X \geq 2) = 1 - e^{-5}\frac{5^0}{0!} - e^{-5}\frac{5^1}{1!} \approx 0.95957$