

## Rappels utiles

### Théorème Spectral

On peut décomposer une matrice symétrique  $A$  en  $A = Q\Lambda Q^T$  où  $Q$  est une matrice orthogonale (rotation) et  $\Lambda$  est une matrice diagonale (scalation).

### Série de Taylor

Expontielle :

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

### Valeur d'une somme

$$\sum_{x=0}^{\infty} q^x = \frac{1}{1-q} \text{ si } |q| < 1$$

### Indicator Function

$$I(\text{some expression}) = \begin{cases} 1 & \text{if the expression is true} \\ 0 & \text{otherwise} \end{cases}$$

### Analyse dimensionnelle

Si on intègre  $f_X$  on trouve une probabilité, donc par exemple une CDF.

Si on intègre  $f_{XY}$  une fois, on trouve une autre fonction de densité de probabilité (de  $X$  ou de  $Y$ ), qu'on peut intégrer pour trouver une probabilité.<sup>21</sup>

### Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

## Distributions

- **probability mass function** pour les distributions discrètes (binomiale, poisson, etc), **probability density function** pour les distributions continues (exponentielle, normale, etc).
- “distribution function” ≡ “cumulative distribution function”. Donc quand on nous demande la distribution function d’une variable c’est la fonction qui  $\forall t$  donne  $P(X \leq t)$ .
- Quand on demande la PDF souvent c’est plus simple de trouver la CDF puis de dériver.

### P.D.F $\Leftrightarrow$ CDF

On a la P.D.F  $f(x)$  et on veut la C.D.F  $G(y)$ , avec  $Y = \frac{1}{X}$ .

D’abord on définit nos fonctions pour passer de  $x$  à  $y$  :

$$r(x) = \frac{1}{x} \text{ et } s(y) = \frac{1}{y}$$

$$G(y) = P(Y \leq y) = P\left(\frac{1}{X} \leq y\right) = P\left(X \geq \frac{1}{y}\right) = 1 - P\left(X < \frac{1}{y}\right)$$

$$G(y) = 1 - F\left(\frac{1}{y}\right)$$

$$\frac{dG(y)}{dy} = \frac{d\left(1 - F\left(\frac{1}{y}\right)\right)}{dy}$$

$$g(y) = -\frac{dF}{dy}\left(\frac{1}{y}\right) \cdot \left| -\frac{1}{y^2} \right| \text{ (on s'intéresse à la croissance, on enlève le signe -)}$$

$$g(y) = -f\left(\frac{1}{y}\right) \cdot \frac{1}{y^2}$$

Et ensuite pour trouver  $G(y)$  on intègre.

## Expected Value

Continue :

$$\int_{-\infty}^{+\infty} f_D(x)xdx$$

Attention, c'est la P.D.F. qu'on intègre, parfois il faut dériver la C.D.F.

## Variance

$$\text{var}(X) = E(X^2) - E(X)^2$$

donc, quand continue :

$$\text{var}(X) = \int_{-\infty}^{+\infty} f_D(x)x^2dx - E(X)^2$$

Standard deviation :

$$\sigma = \sqrt{\text{var}(X)}$$

if  $X_1$  et  $X_2$  independent:

$$\text{var}(X_1 + aX_2) = \text{var}(X_1) + a^2\text{var}(X_2)$$

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{ cov}(X, Y)$$

## Covariance

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

if  $X, Y$  are independent then the covariance is zero (the converse is false!).

Linearité de la covariance :

$$\text{cov}(X + Y, Z + W) = \text{cov}(X, Z) + \text{cov}(X, W) + \text{cov}(Y, Z) + \text{cov}(Y, W)$$

Nous permet de réécrire la variance de la somme de variables aléatoires :

$$\text{var}(a + bX + cY) = b^2\text{var}(X) + 2bc \text{ cov}(X, Y) + c^2 \text{ var}(Y)$$

## Covariance matrix

Pour un vecteur de variables aléatoires  $(X_1, \dots, X_p)$

$$\text{var}(X) = \Omega = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_p) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \dots & \text{cov}(X_2, X_p) \\ \dots & \dots & \dots & \dots \\ \text{cov}(X_p, X_1) & \text{cov}(X_p, X_2) & \dots & \text{var}(X_p) \end{pmatrix}$$

Sachant que  $\text{cov}(X_i, X_j) = (\text{notamment}) \text{ corr}(X_i, X_j)\sigma_i\sigma_j$

Pour un vecteur  $(X_1, X_2)$  de correlation  $p$  et de variance  $\sigma_1, \sigma_2$  :

$$\Omega = \begin{pmatrix} \sigma_1^2 & p\sigma_1\sigma_2 \\ p\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

## Correlation

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\{\text{var}(X)\text{var}(Y)\}^{\frac{1}{2}}}$$

$$\text{corr}(X, Y) = \frac{E(XY) - E(X)E(Y)}{\{\text{var}(X)\text{var}(Y)\}^{\frac{1}{2}}}$$

toujours entre  $-1$  et  $1$ .

une corrélation de  $0$  ne signifie pas que les variables sont indépendantes (il peut y avoir d'autres types de corrélation).

## Moments

On appelle  $E(X^r)$  le  $r$ th moment de  $X$ .

## Moment Generating Function

$$\psi(t) = E(e^{tX})$$

$$= E\left(\sum_{n=0}^{\infty} \frac{X^n t^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E(X^n)$$

(on peut sortir les  $t$  et  $n$  de l'espérance car ils ne dépendent pas de  $X$ )

$$= \int_{-\infty}^{+\infty} e^{tx} f_X(x) dx$$

Comme on sait que dériver l'espérance de  $X$  revient à prendre l'espérance de la dérivée de  $X$  (ça apparemment ça marche pas dans tous les cas mais ici oui) :

$$E(X^n) = \varphi^{(n)}(0)$$

Comme ça le  $t$  s'annule et il reste juste tous les facteurs  $X$  devant qui s'accumulent.

$$E(X) = \varphi'(0) \text{ et } E(X^2) = \varphi''(0)$$

$$\text{var}(X) = \varphi''(0) - (\varphi'(0))^2$$

On sait que si  $X$  et  $Y$  sont indépendantes alors  $E(f(X) \cdot g(Y)) = E(f(X)) \cdot E(g(Y))$  (prouver avec l'intégrale de  $xyf_{X,Y}(x,y)$ ) donc on peut souvent exprimer la MGF d'une variable aléatoire comme le produit des MGF de ses composantes.

## Pour un vecteur

Soit  $X \in \mathbb{R}^p$  un vecteur aléatoire et  $t \in \mathbb{R}^p$  :

En fait les  $t$  c'est juste des points dans l'espace par rapport auxquels on va dériver. On utilise que  $t = 0$  pour avoir la valeur.

On transpose  $t$  pour pouvoir faire le produit scalaire.

$$\begin{aligned} \psi(t) &= E(e^{t^T X}) = E(e^{t_1 X_1 + \dots + t_p X_p}) = E(e^{t_1 X_1} \dots e^{t_p X_p}) \\ &= \psi_1(t_1) \dots \psi_p(t_p) \end{aligned}$$

L'espérance de la  $i$ ème composante du vecteur  $X$  :  $\frac{\partial \psi(t)}{\partial t_i} |_{t=0} = E(X_i)$   
Le vecteur d'espérance est :  $\nabla \psi(t) |_{t=0} = E(X)$

## Cumulant Generating Function

$$K(t) = \log(\psi(t))$$

Pratique car moins de calcul que la MGF pour trouver :

$$K'(0) = E(X) = \mu \text{ et } K''(0) = \text{var}(X) = \sigma^2$$

## Central Limit Theorem

is a formal statement of how normal distributions can approximate distributions of general sums or averages of i.i.d. random variables.

The simple version of the central limit theorem that we give in this section says that whenever a random sample of size  $n$  is taken from any distribution with mean  $\mu$  and variance  $\sigma^2$ , the sample average  $X_n$  will have a distribution that is approximately normal with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .

(la variance diminue avec la taille de l'échantillon)

for sums of random variables, the central limit theorem says that the distribution of the sum will be approximately normal with mean  $n\mu$  and variance  $n\sigma^2$ .

attention quand on approxime avec des nombres petits on doit faire attention à utiliser les bonnes bornes  $P(X \leq x) \neq 1 - P(X > x)$  mais  $P(X \leq x) = 1 - P(X < x + \varepsilon)$  (ou  $\varepsilon = 1$  dans le cas d'un entier)

## Joint random variables

Conditional pdf (2 variables)

$$f_{X/Y}(x/y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) f_Y(y) dy$$

## Law of total variance

$$\text{var}(Y) = E[\text{Var}(Y|X)] + \text{Var}(E[Y|X])$$

Le premier terme a du sens : la variance de  $Y$  c'est la moyenne des variances de  $Y$  sachant que  $X$  est égal à une valeur particulière. Mais on doit aussi prendre en compte que le fait que  $Y/X$  varie beaucoup ou soit lisse influence aussi la variance de  $Y$ .

[See this thread.](#)

## Multivariate normal distributions

Le vecteur  $X$  suit une distribution normale multivariée si toute combinaison linéaire de ses composantes suit une distribution normale univariée, c-a-d  $\forall u \in \mathbb{R}^p$ :

$$u^T X \sim \mathcal{N}(u^T \mu, u^T \Omega u)$$

## Combinaisons linéaires de normales

Les combinaisons linéaires de variables normales sont normales :

$$a_{r \times 1} + B_{r \times p} X \sim \mathcal{N}(a + B\mu, B\Omega B^T)$$

## indépendants

Si on a  $X_1, \dots, X_n$  indépendants  $\sim \mathcal{N}(\mu, \sigma^2)$  then  $X_{n \times 1} = (X_1, \dots, X_n)^T \sim \mathcal{N}_{n(\mu 1_n, \sigma^2 I_n)}$

Le  $1_n$  c'est pour transformer la moyenne en un vecteur de taille  $n$ .

Le  $I_n$  c'est pour avoir une matrice diagonale de taille  $n$  (parce que comme les variables sont indépendantes, la matrice de covariance est diagonale).

## Conditional

Let

$$X \sim \mathcal{N}(\mu_{p \times 1}, \Omega_{p \times p})$$

(en bref,  $X$  est une variable aléatoire normale de dimension  $p$ ).

Maintenant, si on connaît une ou plusieurs des composantes (normales, donc) de  $X$ , on peut calculer la distribution conditionnelle des autres composantes. Et ce sera une distribution normale aussi.

Mettons qu'on connaisse l'ensemble  $\mathcal{B}$  des composantes et qu'on cherche la distribution conditionnelle des autres, on obtient

$$(X_{\mathcal{A}} | X_{\mathcal{B}} = x_{\mathcal{B}}) \sim \mathcal{N}(\mu_A + \Omega_{A,B} \Omega_{B,B}^{-1} (x_B - \mu_B), \Omega_{A,A} - \Omega_{A,B} \Omega_{B,B}^{-1} \Omega_{B,A})$$

où  $\Omega_{A,B}$  est la matrice des covariances où on garde les lignes  $A$  et les colonnes  $B$ .

Parfois  $\Omega_{A,A}$  s'écrit  $\Omega_A$ .

## Transformations

p. ex. on veut savoir si le résultat du dé est pair ou non :

$$\{1, 2, 3, 4, 5, 6\} = g^{-1}(\mathcal{B}) \quad \{0, 1\} = \mathcal{B}$$

On prend un sous-ensemble de  $\mathcal{B}$ .

## Markov inequality

If  $X$  takes only real positive values. Let  $a \in \mathbb{R}^*$ . Then :

$$P(X > a) \leq \frac{E(X)}{a}$$

## Convolution

$$f_{Z(z)} = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) dy$$

En fait là on cherche tous les moyens d'arriver à un certain  $z$  donc on ne fait varier que  $y$ , et après on complète avec le  $x$  qui reste.

maintenant, si on cherche :

$$\begin{aligned} F_{Z(z)} &= P(Z \leq z) \\ &= \int_0^z \int_0^{z-y} f_X(z-y) f_Y(y) dx dy \end{aligned}$$

## Inequalities

Let  $X$  a random variable,  $a > 0$ ,  $h$  a non-negative function and  $g$  a convex function.

Basic inequality :

$$P(h(X) > a) \leq \frac{E(h(X))}{a}$$

Markov's inequality :

$$P(|X| > a) \leq \frac{E(|X|)}{a}$$

Chebyshov's inequality :

$$P(|X| > a) \leq \frac{E(X^2)}{a^2}$$

or

$$P(|X - E(X)| > a) \leq \frac{\text{var}(X)}{a^2}$$

Jensen's inequality :

$$E(g(X)) \geq g(E(X))$$

## Convergence

From the strongest to the weakest.

### in mean square

$$\lim_{n \rightarrow \infty} E((X_n - X)^2) = 0$$

where  $E(X_n^2), E(X^2) < \infty$

### in probability

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0$$

for all  $\varepsilon > 0$

(square it and use Markov's inequality to prove that mean square convergence implies convergence in probability)

### in distribution

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

for all  $x$  where  $F$  is continuous.

### conv. probability $\not\Rightarrow$ conv. mean square

On prend  $X_n$  telle que la proba d'être zéro est forte et la proba d'être très grand est faible et  $X = 0$ .

- $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$
- $\mathbb{P}(X_n = n) = \frac{1}{n}$

On a  $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - 0| > \varepsilon) = P\left(\frac{1}{n} > \varepsilon\right)$  qui tend vers 0 à l'infini donc  $X_n$  converge en probabilité vers 0.

$(E((X_n - X)^2) = E(X_n^2) = \frac{1}{n} \cdot n^2 = n)$  donc  $X_n$  ne converge pas en moyenne quadratique).

### conv. in distribution $\not\Rightarrow$ conv. in probability

On prend  $X_0$  choisi uniformément entre 0 et 1 et  $X_{2n} = X_0$  (donc  $X_{2n}$  est constant). On prend  $X_{2n+1} = 1 - X_0$ .

$X_{2n}$  suit donc une distribution uniforme  $\sim U[0, 1]$  et  $X_{2n+1} \sim U[0, 1]$  aussi (montrons-le avec la CDF, si on veut la probabilité que  $X_{2n+1} \leq 0.2$  on veut la probabilité que  $1 - X_0 \leq 0.2$  on veut  $X_0 \geq 0.8$ , or  $X_0$  est uniforme donc  $\mathbb{P}(X_0 \leq 0.2) = \mathbb{P}(X_0 \geq 0.8)$ ). Donc  $X_n$  converge en distribution vers  $X_0$ .

Par contre,  $X_n$  ne converge pas en probabilité :

$$\begin{aligned} \mathbb{P}(|X_n - X_0| > \varepsilon) &= \mathbb{P}(|X_{2n} - X_0| > \varepsilon \wedge |X_{2n+1} - X_0| > \varepsilon) \\ &= \mathbb{P}(|X_0 - X_0| > \varepsilon \wedge |1 - X_0 - X_0| > \varepsilon) \\ &= 0 + q \end{aligned}$$

$q \neq 0$  car il y a des exemples où  $1 - 2X_0 \neq 0$  (par exemple si  $X_0 = 0.2$ ).

Donc  $X_n$  ne converge pas en probabilité.

## Law of large numbers

Let  $X_1, X_2, \dots, X_n$  be i.i.d. random variables with  $E(X_i) = \mu$  (finite), then the sample average  $X_n = \frac{X_1 + \dots + X_n}{n}$  converges in probability to  $\mu$ .

## Maximum and minimum distributions

$$\begin{aligned} & P(\min(X_1, \dots, X_n) < x) \\ &= 1 - P(\min(X_1, \dots, X_n) \geq x) \\ &= 1 - P(X_1 \geq x, \dots, X_n \geq x) \\ &= 1 - P(X_1 \geq x) \dots P(X_n \geq x) \\ &= 1 - (1 - F(x))^n \end{aligned}$$

## Statistics

$Y_1, \dots, Y_n$  are i.i.d. random variables and  $y_1, \dots, y_n$  are the observed values.

We will assume that the distribution of  $Y$  is  $f(y, \theta)$  where  $\theta$  is a parameter.

## Method of moments

We have this dataset from that we want to recover normal distribution parameters.

We know that theoretically  $E(Y) = \mu$  and that

$$\frac{1}{n} \sum_{i=1}^n y_i = \mu$$

. We also know that theoretically  $\text{var}(Y) = \sigma^2$  and that

$$\frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2 = \sigma^2$$

Therefore we can estimate  $\mu$  and  $\sigma^2$  from the dataset.

## Maximum likelihood estimation (MLE)

The **likelihood** is defined as  $L(\theta) = f(x_1 \cap x_2 \cap \dots \cap x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$  (since all the  $x_i$  are independent)

We want to find the value of  $\theta$  that maximizes the likelihood.

To do so, we usually take the log of the likelihood - because it's easier to work with and because the log is a monotonic function - and then we differentiate it with respect to  $\theta$  (e.g. with respect to  $\mu$  and  $\sigma^2$ ) and set it to zero, then solve for  $\theta$ .

En fait en faisant ça on maximise la probabilité que les données observées soient générées par le modèle.

## M-estimation

généralisation de la méthode du MLE

On utilise pas forcément L, mais n'importe quelle fonction  $\rho$ . Dans le cas du MLE  $\rho(x, \theta) = \log(f(x, \theta))$ .

$\rho$  est souvent concave (comme ça on a un seul maximum) et différentiable.

On peut p. ex prendre  $\rho(x, \theta) = -(x - \mu)^2$  pour la “least squares estimation”.

## Bias

$$\text{bias}(\theta) = E(\hat{\theta}) - \theta$$

p. ex. pour une distrib normale :

$$\text{bias}(\hat{\mu}) = E(\hat{\mu}) - \mu = E\left(\frac{1}{n} \sum_{i=1}^n y_i\right) - \mu = \mu - \mu = 0$$

mais

$$\text{bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \dots = -\frac{\sigma^2}{n}$$

For example, an estimator that is equally likely to underestimate  $g(\theta)$  by 1,000,000 units or to overestimate  $g(\theta)$  by 1,000,000 units would be an unbiased estimator, but it would never yield an estimate close to  $g(\theta)$ . Therefore, the mere fact that an estimator is unbiased does not necessarily imply that the estimator is good or even reasonable.