Chapitre 8. Estimation de paramètres.
Ces notes sont en cours d'évolution

1. Inférence statistique et modèles paramétriques
2. Estimation de paramètres
3. Vraisemblance et exhaustivité
4. Calcul de l'estimation
5. Loi multinomiale
6. Comportement asymptotique des EVM
7. Efficacité
8. Intervalles de confiance
9. Borne de Cramer-Rao. Familles exponentielles

1 Inférence statistique et modèles paramétriques

L'inférence statistique (et l'apprentissage machine) est une technique qui utilise les données observées pour faire une inférence sur la loi (le mécanisme) qui a généré ces données. On postule donc l'existence d'un tel mécanisme. Par exemple, étant donné $X_1, \cdots, X_n \sim P$, trouver la loi de probabilité P. Souvent, on se concentre sur un paramètre de cette loi : $P = P_\theta$. Une fois cette loi inférée, on peut s'en servir pour faire de l'estimation et/ou de la prédiction.

Cette démarche comprends trois volets :

1. l'estimation : au vu de l'observation $X = x$ (x est généralement un vecteur), attribuer une valeur à F_θ (fonction de répartition), c-à-d se donner une valeur θ_0 pour $\theta : \widehat{F} = F_{\theta_0}$.
2. les tests : juger si une hypothèse est raisonnable.
3. les intervalles de confiance. Ils permettent de se faire une idée sur la précision de l'estimation.

Le but de ce chapitre est d'explorer ces concepts. Mais commençons par quelques exemples.

1.1 Exemples introductifs

1.1.1 Modèle de Poisson

Rappelons qu'un processus de Poisson (qui modélise le nombre d'occurences d'événements "rares" par unité de mesure) est basé sur les trois postulats suivants :

1. le taux "d'arrivées" est constant (λ),
2. les événements qui arrivent dans des intervalles disjoints sont indépendants,
3. l'occurence d'événements simultanés est négligeable et la probabilité d'une arrivée dans un petit intervalle est proportionnelle à la longueur h de l'intervalle : $P = \lambda h$, et la probabilité de plus d'une arrivée est de l'ordre de $o(h)$.

La table ci-dessous (Particules α, Berkson, 1966) montre le nombre k d'émissions observées dans 1207 intervalles de 10 secondes chacun (manuel, page 256).

k	*Observé* N_k	*Espéré* $N \times Poi\,(\lambda = 8.367)$
0	1	0.28052
1	4	2.3471
2	13	9.8192
3	28	27.386
4	56	57.284
5	105	95.859
6	126	133.68
7	146	159.78
8	164	167.11
9	161	155.36
10	123	129.99
11	101	98.873
12	74	68.939
13	53	44.37
14	23	26.518
15	15	14.792
16	9	7.735
17	3	3.807
18	1	1.7696
19	1	0.77929
Total	$N = 1207$	1206. 5

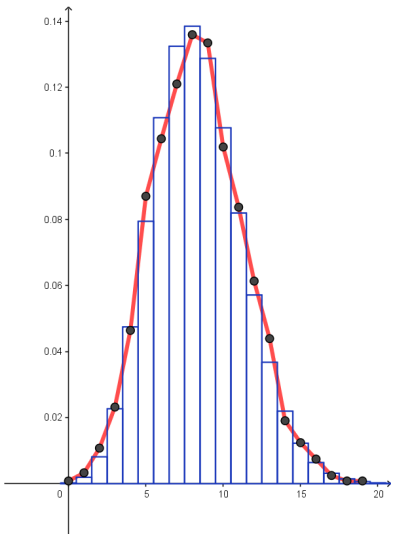
Par exemple (3è colonne),

$$1207 \times P\,(X = 3) = 1207 \times \frac{(8.367)^3}{3!} e^{-8.367} = 27.386$$

On a

$$\widehat{\lambda} = \frac{1}{N} \sum_k k N_k = \frac{T}{N} = \frac{10099}{1207} = 8.367$$

où T = nombre total d'observations. Un test du khi-carré (noter qu'il faut regrouper les cases $0, 1, 2$ ainsi que les cases 16 à 19) donne une valeur-$p = 0.85$. Le modèle ne semble pas contredit (l'histogramme est celui de la loi $P\,(\lambda = 8.367)$).



Ajustement avec $Poi\,(\lambda = 8.367)$

Ce qui est important ici, ce sont les remarques suivantes :

- Si on devait répéter l'expérience, on aurait une autre estimation de λ. On peut donc considérer l'estimation de λ comme résultant d'une variable aléatoire ayant une distribution d'échantillonnage.
- La valeur 8.367 est une réalisation de cette variable aléatoire.
- Quelle est la distribution d'échantillonnage ? Ceci est très important car on veut avoir une idée sur sa variabilité.
- Y-a-t-il un meilleur moyen d'estimer λ ? (nous avons pris la moyenne des observations).
- Que dire de la qualité/précision de l'ajustement ?

L'expérimentateur a groupé ses observations dans 16 cases C_1 (cellules $0-2$: 18 observations), C_2 (28 observations), ..., C_{16} (cellules $17-19$: 5 observations). Si le modèle est correct, la probabilité de tomber sur l'une de ces cases est $Poi\,(\lambda = 8.367)$:

$$\begin{aligned} P\,(X \in C_1) &= p_1 = \pi_0 + \pi_1 + \pi_2 \\ P\,(X \in C_2) &= p_2 = \pi_3 \\ &\vdots \\ P\,(X \in C_{15}) &= p_{15} = \pi_{16} \\ P\,(X \in C_{16}) &= p_{16} = \pi_{17} + \pi_{18} + \pi_{19} + \cdots \end{aligned}$$

où $\pi_k = P\,(X = k)$.

Si $X_1, \cdots, X_{1207}$ sont *iid* Poisson, $Y =$ nombre d'observations tombant dans C_i avec $Y \backsim Bin\,(n = 1207, p_i)$. Ce qui donne

$$E_i = E\,(Y) = 1207 p_i$$

(3ème colonne de la table, $E_i =$ espérés). La distribution conjointe de toutes les cases est une multinomiale avec $n = 1207$ et $p_1, \cdots, p_{16}$.

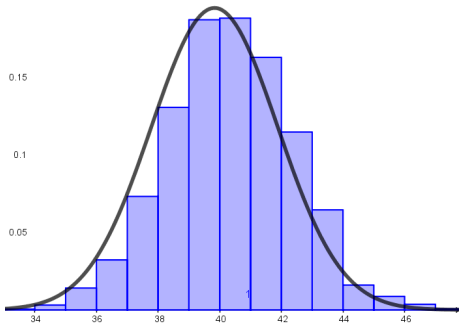
1.1.2 Modèle normal

Les données du tableau qui suivent ont été ètudiées par le statisticien belge A. Quetelet $(1796-1874)$. Elles concernent les mesures (en pouces) de tour de poitrine de 5738 soldats écossais.

Mesure	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Fréquence	3	18	81	185	420	749	1073	1079	934	658	370	92	50	21	4	1

La moyenne observée est $\overline{x} = 39.8318$ et l'écart-type $s = 2.0496$. On calcule les probabilités espérées avec le modèle $X \sim N\left(\mu = 39.83, \sigma^2 = 2.050^2\right)$ comme suit :

$$\begin{aligned} P\,(X \leq 33.5) &= \Phi\left(\frac{33.5 - 39.8318}{2.0496}\right) = 0.001 \\ P\,(33.5 < X \leq 34.5) &= \Phi\left(\frac{34.5 - 39.8318}{2.0496}\right) - \Phi\left(\frac{33.5 - 39.8318}{2.0496}\right) = 0.0037 \\ P\,(34.5 < X \leq 35.5) &= \Phi\left(\frac{35.5 - 39.8318}{2.0496}\right) - \Phi\left(\frac{34.5 - 39.8318}{2.0496}\right) = 0.01 \\ P\,(35.5 < X \leq 36.5) &= \Phi\left(\frac{36.5 - 39.8318}{2.0496}\right) - \Phi\left(\frac{35.5 - 39.8318}{2.0496}\right) = 0.03 \\ &\textit{etc.} \end{aligned}$$



Pour vérifier la qualité de l'ajustement, il faut faire un test (Khi-carré, Kolmogorov-Smirnov). Nous verrons ça dans le chapitre 9.

1.1.3 Modèle binomial

Une élection va se dérouler. Le pourcentage d'electeurs qui votent pour le candidat C est p (p inconnu jusqu'au jour du scrutin). On interroge n (par exemple $n = 1000$) personnes (dont les réponses sont supposées indépendantes). Le nombre X de personnes votant C dans cet échantillon $\backsim Bin(n, p)$. À partir de cette loi, l'observateur essaye de deviner p. Faire de l'inférence statistique, c'est faire des affirmations sur p au vu de X. Par exemple :

$$p > 50\%;$$
$$51\% < p < 53\%;$$
$$p = 52\%,$$

etc.

1.1.4 Exercice

Exercice 1 *Dans l'exemple des particules α (Berkson), les auteurs ont choisi $\overline{X}$ comme estimateur de λ ($\overline{x} = 8.367$ est son estimation). Nous savons que $E(X) = E\left(S^2\right) = \lambda$ dans le cas d'une loi de Poisson en plus du fait que les deux estimateurs convergent en probabilité vers λ. On aurait pu tout aussi bien utiliser S^2 pour estimer λ. Si vous deviez choisir, lequel prendriez-vous ? Argumentez.*

1.2 Modèle paramétrique. Estimation

Les exemples ci-dessus ont donné une idée sur ce qu'on entend par un modèle paramétrique et une estimation. Pécisons ces notions.

Définition 1 *Un modèle statistique $\mathcal{F}$ est une famille de lois (FR, densités, ...). Un modèle paramétrique est une famille $\mathcal{F}$ qui peut être paramétrée par un nombre fini de paramètres.*

Exemple 1 *Le modèle normal requiert 2 paramètres.*

$$\mathcal{F} = \left\{ f\left(x \left| \mu, \sigma^2 \right.\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right], \mu \in \mathbb{R}, \sigma > 0 \right\}$$

Plus généralement, on a

$$\boxed{\mathcal{F} = \{ f(x|\theta), \theta \in \Theta \}}$$

On appelle Θ l'espace des paramètres et θ le paramètre inconnu (qui peut être un vecteur).

Remarque 1 *Un modèle est dit non paramétrique si on ne peut pas le paramétrer avec un nombre fini de paramètres.*

Exemple 2 *Soit $X_1, \cdots, X_n$ iid de loi de Bernoulli $(\theta = p)$. On veut estimer p.*

Exemple 3 *Soit $X_1, \cdots, X_n$ iid de loi de Poi $(\theta = \lambda)$. On veut estimer λ.*

Exemple 4 *Soit $X_1, \cdots, X_n$ iid de loi de $N\left(\theta = \left(\mu, \sigma^2\right)\right)$. On veut estimer μ et σ^2.*

Une fois le(s) paramètre(s) estimé(s), il faut vérifier la qualité de l'ajustement (ceci sera vu au chapitre 9 quand on parlera des tests statistiques).

1.2.1 Exercices

Exercice 2 *On sait qu'une variable aléatoire a pour loi l'une des deux lois A ou B ci-dessous.*

θ	$P_\theta(X=1)$	$P_\theta(X=2)$	$P_\theta(X=3)$
A	$1/2$	$1/2$	0
B	0	$1/2$	$1/2$

1. *Quel est l'espace des paramètres Θ ?*
2. *Si on a observé $X = 1$, quelle est la vraie valeur de θ ?*
3. *Si on a observé $X = 2$, quelle est la vraie valeur de θ ?*

Exercice 3 *On sait qu'une population est décrite par une loi uniforme $U[-\theta; \theta]$. Est-il possible d'estimer θ en utilisant la moyenne ? La variance ?*

Exercice 4 *Un échantillon aléatoire $X_1, \cdots, X_n$ iid est tiré d'une loi de Bernoulli $(\theta = p)$. Cependant, vous n'observez que $T = \sum_{i=1}^n X_i$. Décrire le modèle statistique correspondant.*

Exercice 5 *Dire pourquoi on peut paramétrer une famille de Bernoulli (p) par $\theta = \ln \frac{p}{1-p}$. Une telle procédure s'appelle une reparamétrisation.*

Exercice 6 *Soit donné le modèle statistique $f_\theta(x) = N(\theta, 1)$ où $\theta \in \{0, 10\}$.*

1. *Pensez-vous qu'on puisse faire une inférence crédible sur θ avec une seule observation ?*
2. *Peut-on dire la même chose avec une seule observation si on doit choisir entre $f_\theta(x) = N(0,1)$ et $f_\theta(x) = N(1,1)$? Avec 100 observations ?*

Exercice 7 *On sait que $X \sim N\left(Y, \sigma^2\right)$ où $Y \sim N\left(0, d^2\right)$ avec σ^2 et d^2 inconnus.*

1. *Décrire un modèle statistique pour une observation (X, Y).*
2. *Si Y n'est pas observé, décrire un modèle statistique pour X.*

2 Estimation de paramètres

Les ingrédients nécessaires pour effectuer une estimation paramétrique (d'un paramètre inconnu θ) se résument ainsi :

1. un modèle (une loi) de probabilité $f(x|\theta)$;
2. un espace des paramètres : $\theta \in \Theta$;
3. un échantillon aléatoire (en général iid) de taille n généré par le modèle ;
4. des candidats estimateurs ne dépendant que de l'information fournie par l'échantillon (les données) ;
5. les propriétés des estimateurs pour évaluer les qualités de l'estimateur choisi (précision, efficacité, exhaustivité, etc.).

2.1 Quelques définitions

On répète n fois (n connu) une expérience de Bernoulli(p) (p inconnu). Le nombre X de succès est modélisé par une loi binomiale : $X \sim Bin\,(n,p)$.

$$P\,(X=x)=\binom{n}{x}p^x\,(1-p)^{n-x}\,,\qquad x=0,1,\cdots,n$$

On cherche à estimer le pourcentage p (théorique) d'une expérience au vu d'un échantillon de taille n dans lequel on a observé $X=x$ succès. Il est naturel d'estimer p par $\widehat{p}=X/n$ (fréquence relative du nombre de succès observés). Donc, avant l'observation de X, on prévoit d'estimer p par la fonction suivante de X :

$$\widehat{p}=\widehat{p}\,(X)=X/n$$

C'est donc un estimateur (X/n) et le résultat est une estimation (x/n). Espérance et variance :

$$E\,(\widehat{p})=E\left(\frac{X}{n}\right)=\frac{np}{n}=p:\ \textit{sans biais}$$
$$VAR\,(\widehat{p})=VAR\left(\frac{X}{n}\right)=\frac{np\,(1-p)}{n^2}=\frac{p\,(1-p)}{n}$$

Exemple 5 *Loi normale. Elle dépend de deux paramètres $\mu\in\mathbb{R}$ et $\sigma>0$:*

$$f\,(x\,|\mu,\sigma\,)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],\qquad x\in\mathbb{R}$$

L'utilisation de cette loi est justifiée par le Théorème Central Limite.

Exemple 6 *Loi Gamma. Elle dépend de deux paramètres α et $\lambda>0$:*

$$f\,(x\,|\alpha,\lambda\,)=\frac{\lambda\,(\lambda x)^{\alpha-1}}{\Gamma\,(\alpha)}e^{-\lambda x},\qquad x\geq 0$$

Approche utilisée. Les **données observées** sont des réalisations de variables aléatoires $X_1,X_2,\cdots,X_n$ dont la distribution conjointe dépend du paramètre inconnu θ (ce peut être un vecteur). En général, les X_i sont indépendants et de même loi $f\,(x\,|\theta\,)$. On dit que $(X_1,X_2,\cdots,X_n)$ est un échantillon aléatoire de taille n. Sa loi conjointe est donc

$$f\,(x_1\,|\theta\,)\,f\,(x_2\,|\theta\,)\cdots f\,(x_n\,|\theta\,)$$

On utilisera un **estimateur** de θ :

$$T\,(X_1,\cdots,X_n)$$

qui a donc une **distribution d'échantillonnage**. On évaluera la **variabilité** de l'estimateur : écart-type (donc erreur-type ou sdandard) et le biais.

Méthodes
- Moments (indépendante du modèle choisi).
- Maximum de vraisemblance (plus efficace).
- Bayesienne (décision en situation d'incertitude incorporant les observations).

Avant de developper ces méthodes, il est nécessaire de parler (rappeler) d'estimateurs et d'estimation.

Définition 2 *Statistique. Estimateur. Soit $X_1,\cdots,X_n$ iid de loi commune $f\,(x\,|\theta\,)$.*

On appelle statistique toute fonction

$$T=T\,(X_1,\cdots,X_n)$$

qui ne dépend d'aucun paramètre inconnu. La loi de T s'appelle distribution d'échantillonnage (noter que la loi de T dépend généralement de paramètres inconnus).

On appelle estimateur de θ toute statistique

$$\widehat{\theta} = \widehat{\theta}\left(X_1, \cdots, X_n\right)$$

utilisée pour estimer θ.

Plus généralement, si $g(\theta)$ est une fonction de θ, on appelle estimateur de $g(\theta)$ toute statistique utilisée pour estimer $g(\theta)$.

Exemple 7 *Si $X_i \backsim N\left(\mu, \sigma^2\right)$, et $T_1 = \overline{X}$, $T_2 = S^2$, la distribution d'échantillonnage de T_1 est $N\left(\mu, \sigma^2/n\right)$ et celle de T_2 est $\frac{\sigma^2}{n-1}\chi^2_{n-1}$.*

Exemple 8 *Soit $X_1, \cdots, X_n$ iid $\sim Exp(\lambda)$. Rappelons que la médiane m est la solution de l'équation*

$$F(m) = \frac{1}{2} = 1 - e^{-\lambda m} \Rightarrow m = \frac{1}{\lambda}\ln 2$$

On voit que m est une fonction du paramètre λ. Soit T = nombre de variables $X_i \geq m$. Alors T n'est pas une statistique puisqu'on ne peut pas la calculer à partir des observations $x_1, \cdots, x_n$. Il faut faire la distinction entre une variable aléatoire (m en est une) et une statistique (qui est aussi une variable aléatoire).

Définition 3 ***Biais***. *Un estimateur $\widehat{\theta}$ est dit sans biais pour θ si $E\left(\widehat{\theta}\right) = \theta$. Sinon, le biais est*

$$\boxed{Biais\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta}$$

Exemple 9 *$E\left(\overline{X}\right) = \mu \Rightarrow Biais\left(\overline{X}\right) = 0$ (sans biais). De même, si $X_1, \ldots, X_n$ sont iid, alors S^2 est sans biais pour σ^2.*

Exemple 10 *Le nombre de pannes hebdomadaires d'un certain type d'ordinateur est une variable aléatoire $X \sim Poi(\lambda)$. Un échantillon aléatoire d'observations $X_1, \cdots, X_n$ du nombre de pannes par semaine est choisi.*

1. *Trouver un estimateur sans biais de λ.*
2. *Le coût hebdomadaire des réparations est $C = X + X^2$. Trouver $E(C)$.*
3. *Trouver un estimateur sans biais de C.*

Réponses

1. *Puisque $E(X) = \lambda$, un estimateur sans biais de λ est $\overline{X}$.*
2. *On a*

$$E(C) = E\left(X + X^2\right) = E(X) + E\left(X^2\right) = E(X) + Var(X) + E(X)^2 = 2\lambda + \lambda^2.$$

   Rappel : $E(X) = Var(X) = \lambda$.

3. *On veut un estimateur sans biais de $2\lambda + \lambda^2$. Essayons avec $2\overline{X} + \overline{X}^2$. On a*

$$\begin{aligned} E\left(2\overline{X} + \overline{X}^2\right) &= 2E\left(\overline{X}\right) + E\left(\overline{X}^2\right) = 2\lambda + Var\left(\overline{X}\right) + E\left(\overline{X}\right)^2 \\ &= 2\lambda + \frac{\lambda}{n} + \lambda^2 \end{aligned}$$

   On voit que cet estimateur est biaisé (il surestime en moyenne $2\lambda + \lambda^2$) : Biais $= \frac{\lambda}{n}$ (il est asymptotiquement sans biais : plus bas). Corrigeons le biais et prenons

$$T = 2\overline{X} + \overline{X}^2 - \overline{X}/n$$

   Alors :

$$E(T) = E\left(2\overline{X} + \overline{X}^2 - \frac{\overline{X}}{n}\right) = 2\lambda + \frac{\lambda}{n} + \lambda^2 - \frac{\lambda}{n} = 2\lambda + \lambda^2$$

Exemple 11 *Soit $X_1, \cdots, X_n$ iid $\backsim U\,]0,\theta[$, c-à-d*

$$f\left(x\,|\theta\right) = \begin{cases} 1/\theta & si \quad x \in\,]0,\theta[\\ \\ 0 & sinon \end{cases}$$

Si $\widehat{\theta} = 2\overline{X}$, on a

$$E\left(\widehat{\theta}\right) = 2E\left(\overline{X}\right) = 2\frac{\theta}{2} = \theta$$

Si $\widehat{\theta} = X_{(n)}$, on a

$$P\left(X_{(n)} \leq x\right) = \left(\frac{x}{\theta}\right)^n, \qquad si\ x \in\,]0,\theta[$$

ce qui donne

$$E\left(X_{(n)}\right) = \int_0^\theta \left(1 - \left(\frac{x}{\theta}\right)^n\right) dx = \theta - \frac{1}{n+1}\frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1}\theta$$

On a donc un biais :

$$Biais\left(\widehat{\theta}\right) = \frac{1}{n+1}\theta$$

Remarque 2 *Disposer d'un estimateur sans biais n'est pas toujours une bonne chose comme le montre l'exemple suivant.*

Exemple 12 *Soit $X \sim Poi\,(\lambda)$ et supposons qu'on désire estimer $\theta = P\,(X=0)^2 = e^{-2\lambda}$. On dispose d'une seule observation X. Un estimateur sans biais $T = T\,(X)$ doit vérifier*

$$\begin{aligned} E\,(T) &= e^{-\lambda}\sum_{x=0}^{\infty} T\,(x)\,\frac{\lambda^x}{x!} = e^{-2\lambda} \\ \Rightarrow &\sum_{x=0}^{\infty} T\,(x)\,\frac{\lambda^x}{x!} = e^{\lambda}e^{-2\lambda} = e^{-\lambda} \\ &= \sum_{x=0}^{\infty} (-1)^x\,\frac{\lambda^x}{x!} \end{aligned}$$

*Par identification (**unicité du développement**), la seule fonction $T\,(x)$ qui vérifie cette égalité est*

$$T\,(x) = (-1)^x$$

On déduit que le seul estimateur sans biais pour $e^{-2\lambda}$ est 1 si l'observation X est paire et -1 si elle est impaire. Pas fameux.

Exemple 13 *(Poisson, suite). Puisque*

$$P\,(X=0) = e^{-\lambda}$$

la statistique

$$T = e^{-\overline{X}}$$

est-elle sans biais pour

$$g\,(\lambda) = e^{-\lambda}\ ?$$

Pour savoir, il faut calculer $E\,(T)$ et vérifier l'égalité avec $e^{-\lambda}$. Ce genre de calcul n'est généralement pas facile, mais on peut profiter de propriétés spécifiques de sommes de variables iid de loi de Poisson. D'abord,

$$S = \sum_{i=1}^{n} X_i \sim Poi\,(n\lambda)$$

De plus, la FGM de S est donnée par

$$M_S(t) = E\left(e^{tS}\right) = e^{n\lambda\left(e^t-1\right)}$$

On utilise ceci pour déduire

$$E\left(e^{-\overline{X}}\right) = E\left(e^{-S/n}\right) = M_S(-1/n) = \exp\left[n\lambda\left(e^{-1/n}-1\right)\right] \neq e^{-\lambda}$$

Définition 4 *L'estimateur $\widehat{\theta}$ est dit asymptotiquement sans biais si*

$$\lim_{n\to\infty} E\left(\widehat{\theta}\right) = \theta$$

Définition 5 *Un estimateur $\widehat{\theta}$ est dit convergent (ou consistant) si*

$$\widehat{\theta} \xrightarrow{P} \theta$$

c-à-d si $\forall \epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta} - \theta\right| > \epsilon\right) = 0$$

On a les résultats suivants.

Théorème 1 *Slutsky. Si $\widehat{\theta}$ est convergent et g est continue en θ, alors $g\left(\widehat{\theta}\right)$ est convergent.*

Remarque 3 *On voit que la transformation d'un estimateur convergent par une fonction continue préserve cette propriété (convergence). Ceci n'est cependant pas vrai pour un estimateur sans biais : la transformée par une fonction continue d'un estimateur sans biais n'est pas nécessairement sans biais. En d'autres termes, si $E\left(\widehat{\theta}\right) = \theta$ et si $g(x)$ est continue, alors $E\left(g\left(\widehat{\theta}\right)\right)$ n'est pas nécessairement égal à $g(\theta)$.*

Théorème 2 *Si $\widehat{\theta}$ est dit asymptotiquement sans biais et $\lim_{n\to\infty} VAR\left(\widehat{\theta}\right) = 0$, alors $\widehat{\theta}$ est convergent pour θ.*
Preuve. *Chebyshev.* ■

Définition 6 ***E**rreur **Q**uadratique **M**oyenne EQM (sert notamment à mesurer la vitesse de convergence de $\widehat{\theta}$ vers θ).*

$$\boxed{EQM\left(\widehat{\theta}\right) = E\left[\left(\widehat{\theta} - \theta\right)^2\right] = VAR\left(\widehat{\theta}\right) + Biais\left(\widehat{\theta}\right)^2}$$

(on dit aussi risque quadratique pour EQM).

Exemple 14 *Considérons un échantillon aléatoire X_1, X_2, X_3 issu d'une loi de moyenne inconnue μ et de variance inconnue σ^2. On propose les estimateurs suivants de μ :*

$$\widehat{\mu}_1 = \frac{X_1 + X_2 + X_3}{3} = \overline{X}_3$$
$$\widehat{\mu}_2 = X_1 - X_2 + X_3$$
$$\widehat{\mu}_3 = \frac{2X_1 + 3X_2}{5}$$

Lesquels de ces estimateurs sont sans biais ? Trouver la variance et l'EQM de chacun. Lequel vous semble le meilleur ?

Réponse *On a*

$$E\left(\widehat{\mu}_1\right) = \frac{1}{3}E\left(X_1 + X_2 + X_3\right) = \mu$$
$$E\left(\widehat{\mu}_2\right) = E\left(X_1 - X_2 + X_3\right) = \mu - \mu + \mu = \mu$$
$$E\left(\widehat{\mu}_2\right) = \frac{2}{5}\mu + \frac{3}{5}\mu = \mu$$

Les trois estimateurs sont sans biais. Donc leurs variances sont égales à leurs EQM. On a

$$Var\left(\widehat{\mu}_1\right) = Var\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{3}{9}\sigma^2 = \frac{1}{3}\sigma^2$$
$$Var\left(\widehat{\mu}_2\right) = Var\left(X_1 - X_2 + X_3\right) = 3\sigma^2$$
$$Var\left(\widehat{\mu}_3\right) = Var\left(\frac{2X_1 + 3X_2}{5}\right) = \frac{4}{25}\sigma^2 + \frac{9}{25}\sigma^2 = \frac{13}{25}\sigma^2$$

Exemple 15 *Soit* $X_1, X_2, \cdots, X_n$ *iid* $\backsim$ *Bernoulli* $(\theta = p)$. *Prenons*

$$\widehat{\theta} = \widehat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

On a

$$E\left(\widehat{p}\right) = p$$
$$VAR\left(\widehat{p}\right) = \frac{p\left(1-p\right)}{n} \underset{n\to\infty}{\to} 0$$

Exemple 16 *Soit* $X_1, X_2, \cdots, X_n$ *iid* $\backsim$ *Bernoulli* $(\theta = p)$. *On a vu plus haut que la féquence relative* $\widehat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *est l'estimateur naturel de* θ *et qu'il est sans biais. Sa variance est*

$$Var\left(\widehat{\theta}\right) = \theta\left(1-\theta\right)/n$$

Considérons maintenant l'estimateur

$$T = a\overline{X} + \frac{1}{2}\left(1-a\right)$$

où $0 \leq a \leq 1$ *(c'est donc une moyenne pondérée de* $1/2$ *et de la fréquence relative observée* $\overline{X}$*). Par exemple, si on observe* 6 *succès dans* 10 *essais, alors*

$$T = t = \frac{a}{6} + \frac{1}{2}\left(1-a\right) = \frac{1}{2} + \frac{1}{10}a$$

est l'estimation d'un succès, alors que le premier estimateur donne $\overline{x} = 6/10$. *On a*

$$E\left(T\right) = aE\left(\overline{X}\right) + \frac{1}{2}\left(1-a\right) = a\theta + \frac{1}{2}\left(1-a\right)$$

et le biais est

$$Biais\left(T\right) = a\theta + \frac{1}{2}\left(1-a\right) - \theta = \left(\frac{1}{2} - \theta\right)\left(1-a\right)$$

Ainsi, il y a un biais si $a < 1$. *Cherchons les EQM*

$$EQM\left(\overline{X}\right) = Var\left(\overline{X}\right) = \theta\left(1-\theta\right)/n$$

et

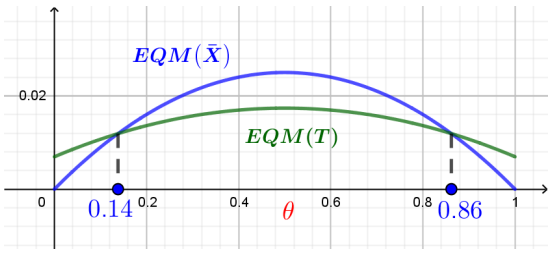
$$\begin{aligned} EQM(T) &= Var(T) + Biais^2(T) \\ &= Var\left(a\overline{X} + \frac{1}{2}(1-a)\right) + \left(\frac{1}{2} - \theta\right)^2 (1-a)^2 \\ &= a^2 Var\left(\overline{X}\right) + \left(\frac{1}{2} - \theta\right)^2 (1-a)^2 \\ &= a^2 \theta(1-\theta)/n + \left(\frac{1}{2} - \theta\right)^2 (1-a)^2 \end{aligned}$$

Prenons $n = 10$ *et* $a = n/(n+2)$. *Dans ce cas,*

$$EQM\left(\overline{X}\right) = \theta(1-\theta)/10$$

et (calcul)

$$EQM(T) = \frac{1}{144}\left(-6\theta^2 + 6\theta + 1\right)$$



On voit (graphe) que $EQM(T) < EQM\left(\overline{X}\right)$ *pour la plus grande partie intérieure de l'intervalle* $[0;1]$. *Si on a des raisons de croire que* θ *n'est pas proche de* 0 *ou de* 1, *l'estimateur* T *est peut être préférable à* $\overline{X}$ *(cette approche est liée à la méthode bayésienne).*

Vocabulaire La quantité

$$\sqrt{VAR\left(\widehat{\theta}\right)} = se\left(\widehat{\theta}\right)$$

est l'erreur-type (ou erreur standard). Elle sert à construire un intervalle de confiance pour θ. C'est donc une mesure de la précision de l'estimateur. Elle dépend de son biais et de sa variabilité (sa variance). Par exemple, $se\left(\overline{X}\right) = Var\left(\overline{X}\right) = \sigma/\sqrt{n}$ (pas de biais).

Définition 7 *L'estimateur* $\widehat{\theta}$ *est dit asymptotiquement normal si*

$$\frac{\widehat{\theta} - \theta}{se\left(\widehat{\theta}\right)} \overset{Loi}{\to} N(0,1)$$

Définition 8 *Intervalle de confiance par excès. Un intervalle* $I = I(X_1, \cdots, X_n)$, *indépendant de* θ, *est un intervalle de niveau* $1-\alpha$ *par excès si*

$$P_\theta(\theta \in I) \geq 1 - \alpha \qquad \forall \theta \in \Theta$$

Exemple 17 *Soit* $(X_1, \cdots, X_n)$ *un échantillon. Supposons* $VAR(X_1) = \sigma^2$ *connue. On cherche à estimer par un intervalle de confiance* $\theta = E_\theta(X_1)$. *Si on utilise* $\widehat{\theta} = \overline{X}$, *on a*

$$\begin{aligned} P_\theta\left(\left|\overline{X} - \theta\right| < \delta\right) &\geq 1 - \frac{\sigma^2}{n\delta^2} = 1 - \alpha \qquad \textit{(Chebyshev)} \\ \Rightarrow I(\theta) &= \left[\overline{X} - \frac{\sigma}{\sqrt{n\alpha}}; \overline{X} + \frac{\sigma}{\sqrt{n\alpha}}\right] \end{aligned}$$

est un $IC_{1-\alpha}(\theta)$. *Note. On a résolu* $1 - \frac{\sigma^2}{n\delta^2} = 1 - \alpha \Rightarrow \delta = \frac{\sigma}{\sqrt{n\alpha}}$.

Si on utilise le théorème central limite, on obtient un IC asymptotique.

2.1.1 Exercices

Exercice 8 *Puisque le but d'un estimateur*

$$\widehat{\theta} = T(X_1, \cdots, X_n)$$

est d'estimer le paramètre inconnu θ, pensez-vous vous qu'on puisse trouver un estimateur qui soit exact, c-à-d tel que $\widehat{\theta} = \theta$? Pourquoi (ou pourquoi pas) ? On pourra supposer, si cela aide, que les X_i suivent une loi continue donnée (par exemple la loi exponentielle).

Exercice 9 *Soit $X_1, \cdots, X_n$ iid $\sim Exp(\lambda = 1/\theta)$. On a $E\left(\overline{X}\right) = \lambda = 1/\theta$. Ainsi $\overline{X}$ est un estimateur sans biais de $1/\theta$. Il semble donc naturel d'utiliser $1/\overline{X}$ comme estimateur de θ. Vérifier que $E\left(1/\overline{X}\right) = n\theta/(n-1)$. Quel est le biais ? Comment corriger le biais ?*

Exercice 10 *Soit $X_1, \cdots, X_n iid \sim U\,]0, \theta[$. On a vu que*

$$E\left(X_{(n)}\right) = \frac{n}{n+1}\theta$$

Donc $X_{(n)}$ est asymptotiquement sans biais. Est-il convergent ?

Exercice 11 *On a vu que la transformation d'un estimateur convergent par une fonction continue préserve cette propriété (convergence). Illustrer par un exemple. Indication. On peut prendre une seule observation $X \sim U[0; \theta]$ et $\widehat{\theta} = 2X$. Choisir une fonction continue $g(x)$.*

Exercice 12 *Considérons un échantillon aléatoire X_1, X_2, X_3 issu d'une loi de moyenne inconnue μ et de variance inconnue σ^2. On propose l'estimateur suivant de μ :*

$$T = a + bX_1 + cX_2 + dX_3$$

où a, b, c, d sont des constantes. Pour quelles valeurs de ces constantes l'estimateur est-il sans biais ? Quelle est la variance (et l'EQM) de T ? Pour quelles valeurs des constantes a-t-on $Biais(T) = 0$?

Exercice 13 *Soit $X_1, \cdots, X_n$ iid $\sim N\left(\mu, \theta = \sigma^2\right)$. Nous savons que $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur sans biais de σ^2.*

1. *Quelle est l'erreur type $\sqrt{S^2}$?*
2. *(d'après examen) Soit $T = cS^2$ $(c > 0)$. Pour quelle valeur de c l'estimateur T a-t-il la plus petite EQM ? Indication. $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ avec $E\left((n-1)S^2/\sigma^2\right) = n-1$ et $Var\left((n-1)S^2/\sigma^2\right) = 2n-2$.*

3 Vraisemblance et exhaustivité

3.1 La fonction vraisemblance

On appelle **vraisemblance** de θ au vu de l'observation x (qui est généralement un vecteur) la loi de probabilité (densité ou fonction de masse) servant à définir le modèle, mais considérée comme fonction de θ. Cette approche est minimale puisqu'on ne considère aucune autre source d'information (cf Bayes). On adopte alors une nouvelle notation :

$$\boxed{L(\theta\,|x) = f(x\,|\theta) = f_\theta(x)}$$

C'est donc la probabilité d'obtenir l'observation x quand la valeur du paramètre est égale à θ (et non le contraire !). La fonction de vraisemblance induit une relation d'ordre dans l'espace Θ des paramètres. On préferera θ_1 à θ_2 si $L(\theta_1\,|x) > L(\theta_2\,|x)$ (au vu de la même observation x).

Sur le vocabulaire. On parle de "densité d'une observation x" et de "vraisemblance d'une valeur du paramètre θ" (pour indiquer le changement du point de vue).

– Le probabiliste a un point de vue
  – théorique (pas besoin de support dans le monde réel)
  – déductif : hypothèses ⇒ conclusions
  – les lois sont vues de manière générique : une densité est un représentant d'une famille indexée par le paramètre.
– Le statisticien a un point de vue
  – lié à une observation concrète (même si c'est une spéculation, donc non effectivement observée)
  – inductif : à partir de l'observation, il remonte à la loi qui l'a engendrée (conséquences → causes)
  – concerné par une famille composée de plusieurs lois (travail de détective).

Pour résumer, on dispose d'une fonction de 2 variables x et θ. Cette fonction est considérée par le
– probabiliste (qui fixe θ) comme une loi de probabilité $f_\theta(x)$ (ou $f(x|\theta)$)
– statisticien (qui fixe ("observe") x) comme une vraisemblance $L(\theta|x)$

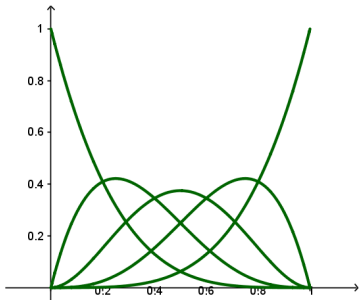
Ainsi, on parle toujours de la probabilité d'une observation et de la vraisemblance d'un paramètre (ou d'un modèle), pas l'inverse (sauf dans le cas de l'approche bayésienne).

Exemple 18 *On répète $n = 4$ fois (n fixé) une expérience de Bernoulli ($\theta = p$). On sait que le nombre X de succès $\sim Bin(n = 4, p)$:*

$$P(X = k|p) = \binom{4}{k} p^k (1-p)^{4-k}, \qquad k = 0, 1, 2, 3, 4$$

La vraisemblance reprend cette formule comme fonction de p :

$$L(p|k) = \binom{4}{k} p^k (1-p)^{4-k} = \text{ polynôme de degré 4 en } p$$



$L(p|k)\,; k = 0, \cdots 4$

On peut tracer les 5 courbes correspondant aux 5 observations $k = 0, \cdots, 4$.

Qu'observe t-on ?

1. Usuellement, on se fixe un modèle (quitte à le remettre en question).
2. Ensuite, au vu de l'observation $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$, on fait une inférence sur θ.
3. Le modèle étant fixé, il est plus pertinent de considérer que ce que l'on observe, ce n'est pas x, mais la fonction (courbe) $L(\theta|x)$ (ou $l(\theta|x)$) :

$$\boxed{\begin{array}{cc} Modèle & \\ + & = observation\ L(\theta|x) \\ observation\ x & \end{array}}$$

Ainsi, dans l'exemple précédent, dire "on a observé $X = 3$" ne refère en rien au modèle binomial. Il convient plutôt de dire qu'on a observé (sous réserve d'ajustement correct, c'est-à-dire selon le modèle) la courbe

$$L(p|3) = P(X = 3) = \binom{4}{3} p^3 (1-p)^{4-3} = 4p^3(1-p)$$

Un petit calcul montre que cette courbe atteint son maximum en $p = 3/4$. C'est la valeur la plus vraisemblable de p ayant fourni l'observation $X = 3$.

3.1.1 Le log-vraisemblance

On appelle log-vraisemblance du paramètre θ au vu de l'observation x le log de la vraisemblance :

$$l(\theta|x) = \ln L(\theta|x)$$

Pourquoi la log-vraisemblance ? Quatre raisons :

1. On voudra comparer les vraisemblances de différents modèles (but : choisir celui qui a la plus grande vraisemblance). Cette démarche sera justifiée plus loin dans le cours. On peut donc introduire $l(\theta|x)$ car la fonction ln est strictement croissante.
2. Si $x = (x_1, \cdots, x_n)$ est un "échantillon" (i.e. une suite de n réalisations indépendantes d'une variable aléatoire de loi $g_\theta(x)$), on a

$$L(\theta|x) = f(x|\theta) = P(X = x) \overset{ind.}{=} \prod_{i=1}^{n} g_\theta(x_i)$$

Chacun des

$$g_\theta(x_i) = probabilité = nombre \leq 1 \Rightarrow L(\theta|x)$$

nombre très petit. Par exemple si

$$g_\theta(x_i) = 0.1$$

et

$$n = 100 \Rightarrow L(\theta|x) = 10^{-100} \Rightarrow difficulté\ calculatoire$$

(underflow). Le passage au log fait disparaître cet inconvénient ($\ln 10^{-100} \approxeq -230$).

3. Maximiser la vraisemblance mènera souvent à calculer sa dérivée. Il est plus facile de dériver une somme qu'un produit.
4. On sera mené à considérer les vraisemblances comme des variables aléatoires et il n'existe pas de théorème simple sur les produits de variables aléatoires. Les principaux théorèmes de la théorie des probabilités portent sur les sommes : lois des grands nombres, TCL, etc.

3.1.2 Exercice

Exercice 14 *On prend un échantillon de sang de n individus pour tester la présence d'un anticorps.*

1. *Quel est le modèle statistique approprié pour cette expérience ?*
2. *Si dans un échantillon de $n = 10$, trois des tests sont positifs, quel est le graphe de la fonction vraisemblance ?*

3.2 Exhaustivité

Soit $X_1, \cdots, X_n$ iid de loi $f(x|\theta)$. La question que l'on se pose dans cette section est la suivante : existe-t-il une statistique

$$T(X_1, \cdots, X_n) = (T_1(X_1, \cdots, X_n), \ldots, T_k(X_1, \cdots, X_n))$$

contenant toute l'information sur θ (selon le modèle) ?

Définition 9 *La statistique $T(X_1, \cdots, X_n)$ est dite* ***exhaustive*** *(sufficient en anglais, Fisher 1922) pour θ (qui peut être un vecteur) si la loi conditionnelle de $(X_1, \cdots, X_n)$ sachant $T = t$ ne dépend pas de θ, quel que soit t.*

En d'autres termes, il n'y a pas besoin d'en savoir plus sur la loi des X_i, T suffit (selon le modèle). Il est évident que l'échantillon lui-même $X = (X_1, \cdots, X_n)$ est une statistique exhaustive pour θ puisque $P(X = x \,|X = x) = 1 \ \forall\theta$. De plus, une statistique exhaustive n'est pas unique. On cherche alors une statistique exhaustive T qui réduit la dimension, en d'autres termes qui réduit la dimension n de l'échantillon à une dimension $k < n$. Enfin, il est important de comprendre qu'une statistique exhaustive dépend du modèle. Des modèles différents peuvent produire des statistiques exhaustives différentes.

Exemple 19 *Loi binomiale. Soit $X_1, \cdots, X_n$ iid $\sim$ Bernoulli $(\theta = p)$. Montrons que $T = \sum_{i=1}^n X_i$ est exhaustive pour p. On a*

$$\begin{aligned} P(X_1 = x_1, \cdots, X_n = x_n \,|T = t) &= \frac{P(X_1 = x_1, \cdots, X_n = x_n, T = t)}{P(T = t)} \\ &= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t}\theta^t (1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} \end{aligned}$$

car on a une séquence particulière de t succès (numérateur). Ainsi, l'ordre dans lequel on obtient les succès n'a aucune importance (n'apporte pas d'information) pour θ.

Exemple 20 *Loi de Poisson. Soit $X_1, \cdots, X_n$ iid $\sim Poi(\theta = \lambda)$. Alors $T = \sum_{i=1}^n X_i$ est exhaustive pour λ. On a $T \sim Poi(n\lambda)$ et*

$$\begin{aligned} P(X_1 = x_1, \cdots, X_n = x_n \,|T = t) &= \frac{P(X_1 = x_1, \cdots, X_n = x_n, T = t)}{P(T = t)} \\ &= \frac{e^{-n\lambda}\lambda^t}{x_1! \cdots x_n!} \times \frac{t!}{e^{-n\lambda}(n\lambda)^t} \\ &= \frac{t!}{x_1! \cdots x_n! n^t} \end{aligned}$$

est indépendante de λ. Noter que les x_i $(= 0, 1, \cdots, t)$ sont liés par la relation $x_1 + \cdots + x_n = t$. On reconnaît une loi multinomiale (quels sont ses paramètres ?).

Exemple 21 *Loi de Poisson. Prenons $n = 2 : X_1, X_2$ iid $\sim Poi(\theta = \lambda)$. Soit $T = X_1 + 2X_2$ et supposons qu'on a observé $X_1 = 1$ et $X_2 = 1$, ce qui donne $T = 3$. Alors*

$$\begin{aligned} P(X_1 = 1, X_2 = 1 \,|T = 3) &= \frac{P(X_1 = 1, X_2 = 1)}{P(X_1 = 1, X_2 = 1) + P(X_1 = 3, X_2 = 0)} \\ &= \frac{\left(\frac{\lambda e^{-\lambda}}{1!}\right)^2}{\left(\frac{\lambda e^{-\lambda}}{1!}\right)^2 + \left(\frac{\lambda^3 e^{-\lambda}}{3!}\right)\left(\frac{\lambda^0 e^{-\lambda}}{0!}\right)} \\ &= \frac{\lambda^2 e^{-2\lambda}}{\frac{1}{6}\lambda^2 e^{-2\lambda}(\lambda + 6)} \\ &= \frac{6}{\lambda + 6} \end{aligned}$$

C'est une fonction de λ. Donc T n'est pas exhaustive pour λ. Ceci est intuitivement clair car on a accordé plus de poids à X_2 qu'à X_1.

Une définition équivalente de l'exhaustivité est la suivante.

Définition 10 *La statistique $T(X_1, \cdots, X_n)$ est exhaustive pour θ si et seulement si pour toute autre statistique $U(X_1, \cdots, X_n)$ la loi conditionnelle de U sachant $T = t$ ne dépend pas de $\theta, \forall t$. En particulier, $T(X_1, \cdots, X_n)$ est exhaustive pour θ si et seulement si la loi conditionnelle de X_i sachant $T = t$ ne dépend pas de θ, pour chaque $i = 1, \cdots, n$.*

Exemple 22 *Soit* $X_1, \cdots, X_n$ *iid* $\sim U[0;\theta]$ *et prenons* $T = X_{(n)} = \max(X_1, \cdots, X_n)$. *On a*

$$f(x_i|t) = \frac{1}{t}, \quad 0 \leq x_i \leq t,\ t = x_{(n)}$$

On en déduit que la statistique T *est exhaustive pour* θ. *Noter que* $X_i|T \sim U[0;T]$. *Vérifier l'exhaustivité avec la définition initiale.*

Les définitions ci-dessus ne sont pas très pratiques pour trouver des statistiques exhaustives. Le résultat suivant fournit une façon utile de le faire.

Théorème 3 *(de factorisation de Fisher-Neyman). La statistique* $T(X_1, \cdots, X_n)$ *est* ***exhaustive*** *pour* θ *si et seulement si la loi conjointe peut se décomposer comme suit :*

$$\boxed{f(x_1, \cdots, x_n|\theta) = g(T(x_1, \cdots, x_n)|\theta)\, h(x_1, \cdots, x_n)}$$

où $h(x_1, \cdots, x_n)$ *ne dépend pas de* θ.
Preuve. *Voir manuel (dans le cas unidimensionnel discret).* ■

Exemple 23 *Bernoulli* $(\theta = p)$. *On a*

$$P(X_i = x) = \theta^x (1-\theta)^{1-x}, \qquad x = 0, 1$$

et

$$\begin{aligned} f(x_1, \cdots, x_n|\theta) &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}, \qquad x_i = 0, 1 \\ &= \theta^{\sum_{i=1}^n x_i} (1-\theta)^{\sum_{i=1}^n (1-x_i)} \\ &= \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i} (1-\theta)^n \\ &= g(x_1, \cdots, x_n|\theta) \times 1 \end{aligned}$$

avec

$$g(t|\theta) = \left(\frac{\theta}{1-\theta}\right)^t (1-\theta)^n$$

Exemple 24 *Loi normale. Soit* $X_1, \cdots, X_n$ *iid* $\backsim N(\mu, \sigma^2)$. *Ici,* $\theta = \mu$ *ou* $\theta = (\mu, \sigma^2)$ *selon que* σ^2 *est connu ou non. On a*

$$\begin{aligned} f(x_1, \cdots, x_n|\theta) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &\overset{calculs}{=} \exp\left[-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right](2\pi\sigma^2)^{-n/2} \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right] \end{aligned}$$

Si σ^2 *est connu,* $\theta = \mu$, $T(X_1, \cdots, X_n) = \overline{X}$ *et*

$$g(t|\mu) = \exp\left[-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right]$$

et $\overline{X}$ *est exhaustive pour* μ. *Si* σ^2 *est inconnu,* $\theta = (\mu, \sigma^2)$ *et*

$$f(x_1, \cdots, x_n|\mu, \sigma^2) = g(\overline{x}, s^2, \theta) \times (2\pi\sigma^2)^{-n/2}$$

Donc $T = (\overline{X}, S^2)$ *est exhaustive pour* $\theta = (\mu, \sigma^2)$.

Exemple 25 *Reprenons la loi de Poisson. Soit* $X_1, \cdots, X_n$ *iid* $\backsim Poi(\lambda)$, *i.e.*

$$f(x_1, \cdots, x_n | \lambda) = \left(\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}\right) \frac{1}{x_1! \cdots x_n!}$$

Ici, $T(x_1, \cdots, x_n) = \sum_{i=1}^n x_i$, $g(t|\lambda) = \lambda^t e^{-n\lambda}$ *et* $h(x_1, \cdots, x_n) = \frac{1}{x_1! \cdots x_n!}$.

Le théorème suivant nous dit que l'EQM (en particulier la variance pour un estimateur sans biais) d'un estimateur est minimale si on utilise une statistique exhaustive.

Théorème 4 *(Rao-Blackwell). Soit* $\widehat{\theta}$ *un estimateur de* θ *tel que* $E\left(\widehat{\theta}^2\right) < \infty$ *(existe) pour tout* θ. *Soit* T *une statistique exhaustive pour* θ *et soit* $\widetilde{\theta} = E\left(\widehat{\theta}|T\right)$. *Alors,* $\forall \theta \in \Theta$,

$$\boxed{E\left(\left(\widetilde{\theta} - \theta\right)^2\right) \leq E\left(\left(\widehat{\theta} - \theta\right)^2\right)}$$

L'inégalité étant stricte pour $\widetilde{\theta} \neq \widehat{\theta}$.
Preuve. *On a*

$$E\left(\widetilde{\theta}\right) = E\left(E\left(\widehat{\theta}|T\right)\right) = E\left(\widehat{\theta}\right)$$

En d'autres termes, $\widetilde{\theta}$ *et* $\widehat{\theta}$ *ont le même biais (s'il y en a) pour* θ. *Il suffit donc de regarder les variances respectives pour comparer les EQM :*

$$\begin{aligned} Var\left(\widehat{\theta}\right) &= Var\left(E\left(\widehat{\theta}|T\right)\right) + E\left(Var\left(\widehat{\theta}|T\right)\right) \\ &= Var\left(\widetilde{\theta}\right) + E\left(Var\left(\widehat{\theta}|T\right)\right) > Var\left(\widetilde{\theta}\right) \end{aligned}$$

La dernière inégalité est vraie si $\widetilde{\theta} \neq \widehat{\theta}$. *On a égalité si*

$$\widehat{\theta} = g(T) \Rightarrow \widetilde{\theta} = E\left(\widehat{\theta}|T\right) = \widetilde{\theta} = E(g(T)|T) = g(T) = \widehat{\theta}$$

■

Ce théorème donne des arguments pour l'emploi des statistiques exhaustives.

Proposition 1 *Si* T *est exhaustive pour* θ, *alors* $\widehat{\theta}_{EMV}$ *est une fonction de* T :

$$\boxed{\widehat{\theta}_{EMV} = \varphi(T)}$$

Preuve. *En effet,*

$$L(\theta) = g(T|\theta)\, h(x_1, \cdots, x_n)$$

et le maximum ne dépend que de $g(T|\theta)$. ■

Ce résultat nous dit qu'on peut extraire une statistique exhaustive à partir de la fonction vraisemblance.

Définition 11 *Une statistique* T *est dite* ***exhaustive minimale*** *pour* θ *si elle est exhaustive et si, pour toute autre statistique exhaustive* U *pour* θ, *il existe une fonction* g *telle que* $T = g(U)$.

Exemple 26 *Loi normale* $N\left(\theta = \mu, \sigma^2\right)$, σ^2 *connu. Il est facile de vérifier que toute fonction vraisemblance est un multiple de*

$$\exp\left(-\frac{n}{2\sigma^2}(\overline{x} - \theta)^2\right)$$

Elle prend son maximum en $\theta = \overline{x}$. *Ceci nous dit qu'on obtient* $\overline{x}$ *à partir de toute fonction vraisemblance (selon ce modèle) et par conséquent* $\overline{X}$ *est exhaustive minimale.*

On a les deux résultats suivants.

Proposition 2 *Si T et U sont exhaustives minimales pour θ, alors il existe une bijection g telle que $T = g(U)$.*

Proposition 3 *Si T est exhaustive minimale pour θ, alors toute statistique exhaustive pour θ de la forme $U = g(T)$ est également minimale.*

3.3 Exercices

Exercice 15 *Loi de Poisson. Prenons $n = 2 : X_1, X_2$ iid $\sim Poi(\theta = \lambda)$. Soit $T = X_1 + 2X_2$ et supposons qu'on a observé $X_1 = 1$ et $X_2 = 1$, ce qui donne $T = 3$. Est-il vrai que $T \sim Poi(\lambda = 3)$? Indication. Quelle est la FGM de T ?*

Exercice 16 *Loi exponentielle. Prenons $n = 2 : X_1, X_2$ iid $\sim Exp(\theta = 1/\lambda)$. Montrer que $T = X_1 + X_2$ est exhaustive pour θ. Indication. Quelle est la loi de T ?*

4 Calcul de l'estimation d'un paramètre

4.1 La méthode des moments

Soit X une variable aléatoire. Le moment d'ordre k de X est défini (sous réserve de convergence absolue) par

$$\boxed{\mu_k = E\left(X^k\right), \qquad k = 1, 2, \cdots}$$

Le moment empirique d'ordre k de X est défini par

$$\boxed{\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad k = 1, 2, \cdots}$$

où $X_1, \cdots, X_n$ sont iid de même loi que X.

On considère $\widehat{\mu}_k$ comme un estimateur de μ_k. On vérifie facilement que

$$\boxed{E(\widehat{\mu}_k) = \mu_k} \quad \text{(sans biais)}$$

En règle générale, on utilise les moments d'ordres les plus bas possibles pour estimer

$$\begin{array}{l} \theta_1 = g_1(\mu_1, \cdots, \mu_k) \\ \qquad \vdots \qquad\qquad \Rightarrow \widehat{\theta}_i = g_i(\widehat{\mu}_1, \cdots, \widehat{\mu}_k), \qquad i = 1, 2, \cdots, k \\ \theta_k = g_k(\mu_1, \cdots, \mu_k) \end{array}$$

Remarque 4 *Si les fonctions g_i sont continues, alors les estimateurs $\widehat{\theta}_i$ sont convergents pour θ_i :*

$$\widehat{\theta}_i = g_i(\widehat{\mu}_1, \cdots, \widehat{\mu}_k) \xrightarrow{P} g_i(\mu_1, \cdots, \mu_k) \quad \textit{quand } n \to \infty$$

Étapes du calcul

1. Calculer les moments d'ordre petit (en général autant qu'il y a de paramètres à estimer).
2. Les moments sont des fonctions des paramètres. On résout le système (en général non linéaire).
3. On remplace les moments théoriques par les moments empiriques.

Illustrons la méthode des moments par des exemples.

4.1.1 Quelques exemples

Exemple 27 *Estimation de la variance de population. Puisque* $Var(X) = \mu_2 - \mu_1^2$, *on utilise donc*

$$\widehat{Var}(X) = \widehat{\mu}_2 - \widehat{\mu}_1^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}^2$$

Exemple 28 *Soit* $(X_1, \cdots, X_n)$ *un échantillon aléatoire de loi*

$$f_X(x|\theta) = (\theta+1)\, x^\theta, \qquad 0 < x < 1$$

On a

$$\mu = \mu_1 = E(X) = \int_0^1 x(\theta+1)\, x^\theta dx = \frac{\theta+1}{\theta+2}$$

Un petit calcul donne

$$\theta = \frac{2\mu - 1}{1-\mu}$$

On remplace μ *par* $\widehat{\mu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ *pour obtenir*

$$\widehat{\theta} = \frac{2\overline{X} - 1}{1 - \overline{X}}$$

Remarquons que la loi de $\widehat{\theta}$ *(échantillonnage) est trop compliquée.*

Exemple 29 *Considérons le cas de* $X \backsim Poi(\lambda)$. *On sait que* $E(X) = \lambda$ *et, par conséquent,*

$$\widehat{\lambda} = \widehat{\mu}_1 = \overline{X}$$

Soit $S = \sum_{i=1}^n X_i$. *Alors, on a*

$$\widehat{\lambda} = \frac{S}{n}$$

C'est une variable aléatoire dont la distribution est la distribution d'échantillonnage. Maintenant, $X_i \backsim Poi(\lambda_0)$ *(pour une certaine valeur inconnue de* λ*). Donc* $S \backsim Poi(n\lambda_0)$ *(selon le modèle), ce qui donne*

$$P\left(\widehat{\lambda} = x\right) = P(S = nx) = \frac{(n\lambda_0)^{nx}\, e^{-n\lambda_0}}{(nx)!}$$

où x *est tel que* $nx = k \in \mathbb{N}$. *On a*

$$E\left(\widehat{\lambda}\right) = \frac{E(S)}{n} = \lambda_0$$
$$VAR\left(\widehat{\lambda}\right) = \frac{VAR(S)}{n^2} = \frac{\lambda_0}{n}$$

Si $n \uparrow$, *on voit que* S *est asymptotiquement normale et on en conclut que la loi asymptotique de* $\widehat{\lambda}$ *est* $N\left(\mu = \lambda_0, \sigma^2 = \frac{\lambda_0}{n}\right)$. *En conclusion :*

1. $\widehat{\lambda}$ *est sans biais*
2. *l'erreur type est* $\sqrt{\lambda_0/n} = \sigma_{\widehat{\lambda}}$ *(inconnue car dépend de* λ_0*). On estime* λ_0 *en mettant* $\widehat{\lambda}$ *à sa place (ceci sera justifié plus tard) :*

$$s_{\widehat{\lambda}} = \sqrt{\widehat{\lambda}/n} = \text{ erreur standard estimée}$$

Ceci nous permet d'avoir une idée de l'erreur commise. Puisque $\widehat{\lambda} \backsim$*normale, il est peu probable que notre estimation de* λ *dépasse* 2 *erreurs type (au niveau de* 95%*). Il reste à determiner si la loi de Poisson est le bon modèle (problème d'ajustement, chapitre 9).*

A.N. Amiante (Rice, page 261),ou particules α (plus haut).

Exemple 30 *Loi normale $N\left(\mu,\sigma^2\right)$. On a*

$$\begin{array}{ll} \mu_1=\mu & \\ \mu_2=\mu^2+\sigma^2 & \Rightarrow \quad \sigma^2=\mu_2-\mu^2 \end{array}$$

et, par conséquent,

$$\widehat{\mu}=\overline{X}$$
$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

La distribution d'échantillonnage de $\overline{X}$ est $N\left(\mu,\sigma^2/n\right)$ et on a

$$n\widehat{\sigma}^2/\sigma^2\backsim\chi_{n-1}^2$$

De plus, $\overline{X}$ et $\widehat{\sigma}^2$ sont indépendantes.

A.N. Tour de poitrine soldats écossais (plus haut).

Exemple 31 *Loi exponentielle ($\lambda=1/\theta$).*

$$f\left(x\right)=\frac{1}{\theta}e^{-x/\theta},\qquad x\geq 0,\theta>0$$

On a

$$\mu_1=\mu=\theta_1\Rightarrow\widehat{\theta}_1=\widehat{\mu}=\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$$

On peut également utiliser $\mu_2=E\left(X^2\right)=2\theta^2$:

$$\widehat{\mu}_2=\sum_{i=1}^{n}X_i^2=2\theta^2\Rightarrow\widehat{\theta}_2=\sqrt{\frac{\sum_{i=1}^{n}X_i^2}{2n}}$$

Un autre candidat est la médiane théorique (est-il acceptable ?) :

$$\frac{1}{2}=F\left(m\right)=1-e^{-m/\theta}\Rightarrow m=\theta\ln 2$$

La médiane échantillonnale est la valeur du milieu. Si la taille de l'échantillon est impaire ($n=2k+1$),

$$\widehat{M_d}=X_{(k+1)}$$

et

$$\widehat{\theta}_3=\frac{X_{(k+1)}}{\ln 2}$$

Exemple 32 *Loi $U\left]-\theta,\theta\right[$. Ici, $E\left(X\right)=0$ et donc μ_1 ne sert pas (ne donne pas d'information sur θ). On prend $\mu_2=E\left(X^2\right)=\theta^2/3$.*

Exemple 33 *Loi Gamma(α,λ). On a*

$$\begin{array}{ccccc} \mu_1=\alpha/\lambda & & \lambda=\mu_1/\left(\mu_2-\mu_1^2\right) & & \widehat{\lambda}=\overline{X}/\widehat{\sigma}^2 \\ \mu_2=\alpha\left(\alpha+1\right)/\lambda^2 & \overset{calculs}{\Rightarrow} & \alpha=\mu_1^2/\left(\mu_2-\mu_1^2\right) & \Rightarrow & \widehat{\alpha}=\overline{X}^2/\widehat{\sigma}^2 \end{array}$$

AN. Quantité de précipitations lors d'orages, d'ouragans, etc. (Rice p. 263). Nous y reviendrons plus loin.

Examinons les distributions d'échantillonnage de $\widehat{\alpha}$ et $\widehat{\lambda}$. Une étude théorique est impossible car les fonctions sont compliquées. On procède alors par simulation (BOOTSTRAP) :

1. *on génère beaucoup d'échantillons (disons* 1000*) de taille* n *chacun de loi Gamma avec les valeurs des paramètres (supposés connus :* α_0, λ_0*).*
2. *on calcule ensuite les estimations de* α *et* λ*.*
3. *Ceci donne un histogramme des valeurs de* λ *qui devrait nous fournir une idée sur la distribution d'échantillonnage de* $\widehat{\lambda}$*. Noter que cette opération requiert des valeurs connues : on substitue les estimations.*
4. *On calcule les erreurs type* $s_{\widehat{\alpha}}$ *et* $s_{\widehat{\lambda}}$*.*

Nous reprendrons ceci plus loin.

Pour conclure cette section, on peut dire que la méthode des moments donne un estimateur convergent (utiliser la loi faible des grands nombres : $\widehat{\mu}_k \xrightarrow{P} \mu_k$ et Slutsky : $g\left(\widehat{\mu}_k\right) \xrightarrow{P} g\left(\mu_k\right)$ si g est continue en μ_k).

4.1.2 Exercices

Exercice 17 *Montrer que l'estimateur*

$$\widehat{Var}\left(X\right) = \widehat{\mu}_2 - \widehat{\mu}_1^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}^2$$

est biaisé pour $Var\left(X\right)$*. Comment le modifier pour supprimer le biais ?*

Exercice 18 *Soit* $X_1, \cdots, X_n$ *iid* $\sim Exp\left(\theta = \lambda\right)$*. Nous savons que* $E\left(X_i\right) = 1/\lambda$ *et que* $S = \sum_{i=1}^{n} X_i \sim Gamma\left(n, \lambda\right)$ *(vérifiez !).*

1. *Montrer que* $E\left(1/\overline{X}\right) = \frac{n}{n-1}\lambda$*. L'estimateur* $T = 1/\overline{X}$ *est donc biaisé, mais convergent, pour* λ*.*
2. *Montrer que la fonction de densité de* $T = 1/\overline{X}$ *est donnée par*

$$f_T\left(t\right) = \frac{n^n \lambda^n}{(n-1)! t^{n+1}} \exp\left(-\frac{n\lambda}{t}\right), \quad t > 0$$

4.2 La méthode du maximum de vraisemblance

Exemple (Motivation). On dispose d'une pièce de monnaie biaisée. On sait qu'en moyenne, la proportion des Piles est l'un des trois nombres : $p = 0.2, p = 0.3$ ou $p = 0.8$. On veut déterminer lequel. On fait l'expérience suivante : on lance la pièce $n = 2$ fois et on observe le nombre de piles. Le modèle mathématique est les suivant. On a un échantillon aléatoire (X_1, X_2) d'une loi de Bernoulli(p), où

$$p \in \Theta = \{0.2, 0.3, 0.8\}$$

(Θ = espace des paramètres). La méthode des moments est inadéquate car elle donne

$$\overline{X} = \frac{X_1 + X_2}{2} \Rightarrow \widehat{p} = 0 \text{ ou } \widehat{p} = 0.5 \text{ ou } \widehat{p} = 1$$

qui n'appartiennent pas à Θ. Considérons la fonction de masse conjointe de (X_1, X_2) :

(x_1, x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	
$p_{(x_1,x_2)}$					Total
$p = 0.2$	$\boxed{0.64}$	0.16	0.16	0.04	1
$p = 0.3$	0.49	$\boxed{0.21}$	$\boxed{0.21}$	0.09	1
$p = 0.8$	0.04	0.16	0.16	$\boxed{0.64}$	1

Décision

Nombre de piles observé	$\widehat{p}$ *le plus vraisemblable*
0	0.2
1	0.3
2	0.8

Cette méthode est utile partout en statistique : estimation, ajustement, ... Elle a d'intéressantes propriétés, notamment pour les échantillons de grandes tailles.

Soit $X_1, \cdots, X_n$ iid avec loi (densité ou fonction de masse)

$$f(x_1, \cdots, x_n \,|\theta)$$

Pour $X_1 = x_1, \cdots, X_n = x_n$ observés (fixes), la fonction vraisemblance de θ est

$$L(\theta \,|x_1, \cdots, x_n) = L(\theta) = f(x_1, \cdots, x_n \,|\theta) \overset{ind.}{=} f(x_1 \,|\theta) \cdots f(x_n \,|\theta)$$

L'estimation du maximum de vraisemblance (ou de vraisemblance maximale) est la valeur de θ, si elle existe, qui maximise $L(\theta)$:

$$\boxed{\widehat{\theta} = \arg\max L(\theta)}$$

Son objectif est de rendre les données les plus vraisemblables possible.

Il est souvent préférable de prendre le logarithme (log-vraisemblance) :

$$l(\theta \,|x_1, \cdots, x_n) = l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i \,|\theta)$$

4.2.1 Quelques exemples

Exemple 34 *$X \backsim NB(r = 3, \theta = p)$. Rappelons que*

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, \cdots$$

Si on observe $x = 5$, on a

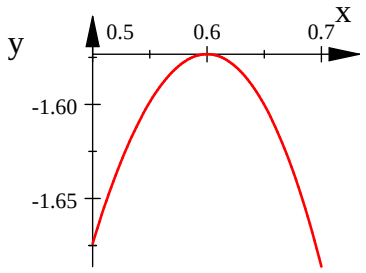
$$P(X = 5) = \binom{4}{2} p^3 (1-p)^2$$

et donc

$$L(p) = \binom{4}{2} p^3 (1-p)^2 = \textit{polynôme de degré } 5$$

et

$$l(p) = \ln \binom{4}{2} + 3 \ln p + 2 \ln (1-p)$$



$$l(p) = \ln \binom{4}{2} + 3 \ln p + 2 \ln (1-p)$$

Un petit calcul montre que le maximum est atteint pour $p = \frac{3}{5}$ (ce qui est visible sur le graphe).

Exemple 35 *Soit* $X \backsim Poi(\theta = \lambda)$:

$$p_X(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \cdots$$

On a

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = e^{-n\lambda} \frac{\lambda^{x_1 + \cdots + x_n}}{x_1! \cdots x_n!}$$
$$\Rightarrow l(\lambda) = -n\lambda + n\overline{x} \ln \lambda - \ln(x_1! \cdots x_n!)$$

et

$$\frac{\partial}{\partial \lambda} l(\lambda) = -n + \frac{n\overline{x}}{\lambda} = 0 \Rightarrow \lambda = \overline{x}$$

(point critique stationnaire). On a

$$\frac{\partial^2}{\partial \lambda^2} l(\lambda) = -\frac{n\overline{x}}{\lambda^2} < 0$$

ce qui confirme que nous avons un maximum en $\lambda = \overline{x}$*. En conclusion l'EVM (estimateur de vraisemblance maximale) est*

$$\widehat{\lambda} = \overline{X}$$

Noter que c'est le même estimateur que celui donné par la méthode des moments.

Exemple 36 *Soit* $X \backsim Bernoulli(\theta = p)$:

$$p_X(x) = P(X = x) = p^x (1 - p)^{1 - x}, \quad x = 0, 1.$$

On a

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1 - p)^{1 - x_i} = p^{\sum_{i=1}^{n} x_i} (1 - p)^{n - \sum_{i=1}^{n} x_i} = p^{n\overline{x}} (1 - p)^{n(1 - \overline{x})}, \qquad x_i = 0, 1.$$
$$\Rightarrow l(p) = n\overline{x} \ln p + n(1 - \overline{x}) \ln(1 - p)$$

et

$$\frac{\partial}{\partial p} l(p) = \frac{n\overline{x}}{p} - \frac{n(1 - \overline{x})}{1 - p} = 0 \Rightarrow p = \overline{x}$$

Donc (vérifier qu'on a bien un maximum) :

$$\widehat{p} = \overline{X}$$

Exemple 37 *Soit* $X \backsim N(\mu, \sigma^2)$:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right]$$
$$\Rightarrow l(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

On a

$$\frac{\partial l}{\partial \mu}(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \qquad (*)$$

$$\frac{\partial l}{\partial \sigma^2}(\mu, \sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \qquad (**)$$

Ce qui donne (calcul) :

$$(*) \Rightarrow \widehat{\mu} = \overline{X}$$
$$(**) \Rightarrow \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \frac{n - 1}{n} S^2$$

Ce sont les mêmes estimateurs que ceux obtenus avec la méthode des moments.

Exemple 38 *Soit* $X \backsim Gamma(\alpha, \lambda)$. *Rappelons que la densité est donnée par*

$$f(x|\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}, \qquad x \geq 0, \alpha, \lambda > 0$$

Ici, les calculs (omis) donnent :

$$l(\alpha,\lambda) = n\alpha \ln\lambda + (\alpha-1)\sum_{i=1}^{n}\ln x_i - \lambda\sum_{i=1}^{n}x_i - n\ln\Gamma(\alpha)$$

On a

$$\frac{\partial l}{\partial \alpha}(\alpha,\lambda) = n\ln\lambda + \sum_{i=1}^{n}\ln x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0 \qquad (*)$$

$$\frac{\partial l}{\partial \lambda}(\alpha,\lambda) = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n}x_i = 0 \qquad (**)$$

On a alors

$$(**) \Rightarrow \lambda = \frac{\alpha}{\overline{x}}$$

(donc $\widehat{\lambda} = \widehat{\alpha}/\overline{X}$*). Substituons ceci dans* $(*)$:

$$n\ln\alpha - n\ln\overline{x} + \sum_{i=1}^{n}\ln x_i - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0$$

Cette équation ne peut pas être résolue exactement (contient une digamma : $\Gamma'(\alpha)/\Gamma(\alpha)$*). Il faut donc une solution numérique (Newton-Raphson, la plus populaire). Pour initialiser la méthode, on peut utiliser l'estimation de la méthode des moments.*

Remarque*. La méthode EVM donne de meilleurs résultats (moins dispersés) que celle des moments. Pour avoir une idée de la distribution d'échantillonnage, il faut faire une simulation.*

Exemple 39 *Durée de survie. Censure de type* 2. *La durée de vie* X *d'un composant suit une loi* $X \backsim Exp(\lambda = 1/\theta)$. *On teste* n *composants choisis au hasard et on observe les* r *premières pannes* $(1 \leq r \leq n)$ *:*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(r)} \leq \cdots \leq X_{(n)}$$

La vraisemblance de $\left(X_{(1)}, X_{(2)}, \cdots, X_{(r)}\right)$ *est (preuve omise)*

$$\begin{aligned} L\left(\theta \left| x_{(1)}, x_{(2)}, \cdots, x_{(r)}\right.\right) &= \frac{n!}{(n-r)!}\exp\left(-\sum_{i=1}^{r}\frac{x_{(i)}}{\theta}\right)\exp\left(-\frac{(n-r)x_{(r)}}{\theta}\right)\left(\frac{1}{\theta}\right)^{r} \\ &= \frac{n!}{(n-r)!\theta^{r}}\exp\left(-\frac{\sum_{i=1}^{r}x_{(i)} + (n-r)x_{(r)}}{\theta}\right) \end{aligned}$$

Noter que

$$T = \sum_{i=1}^{r}X_{(i)} + (n-r)X_{(r)}$$

est la durée totale de survie jusqu'au r*ième composant. On cherche* $\widehat{\theta}_{EMV}$*. On a*

$$l(\theta) = constante - r\ln\theta - \frac{T}{\theta}$$

$$\frac{\partial l}{\partial \theta}(\theta) = -\frac{r}{\theta} + \frac{T}{\theta^2} = 0$$

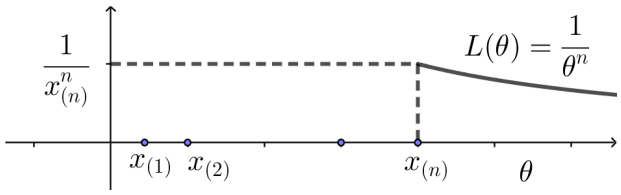
$$\Rightarrow \boxed{\widehat{\theta} = \frac{T}{r}}$$

En particulier, si $r = n$ *(pas de censure), alors* $\widehat{\theta} = \overline{X}$.

Exemple 40 *Modèle $X \backsim U\,]0;\theta[$, $\theta > 0$. On a $f(x\,|\theta) = 1/\theta$ pour $x \in\,]0;\theta[$. Que vaut θ ? On prend un échantillon iid $X_1, \cdots, X_n$ de loi $f(x\,|\theta)$, puis on ordonne les X_i : $X_{(1)} \leq \cdots \leq X_{(n)}$. Puisque les $x_{(i)} \in\,]0;\theta[$, il est clair que l'EVM de θ est donné par la plus grande observation*

$$\widehat{\theta} = X_{(n)}$$

Cependant, le modèle a des déficiences intuitives (voir graphique) :



- *ne satisfait pas les conditions de régularité*
- *θ apparait dans le domaine de la loi de X*
- *$l'(\theta)$ ne fonctionne pas (θ est lié à $X_{(n)}$)*

Cependant, nous avons déjà vu que la distribution d'échantillonnage de $\widehat{\theta} = X_{(n)}$ est

$$P\left(\widehat{\theta} \leq x\right) = \left(\frac{x}{\theta}\right)^n \qquad \text{pour } x \in\,]0;\theta[$$
$$\Rightarrow f_{\widehat{\theta}}(x) = n\left(\frac{x}{\theta}\right)^{n-1}\frac{1}{\theta} \qquad \text{pour } x \in\,]0;\theta[$$

et

$$P\left(n\left(\widehat{\theta}-\theta\right) \leq x\right) = P\left(\widehat{\theta} \leq \frac{x}{n}+\theta\right) = \left(\frac{\frac{x}{n}+\theta}{\theta}\right)^n = \left(1+\frac{x/\theta}{n}\right)^n \underset{n\to\infty}{\to} e^{x/\theta}$$

On a donc une convergence en loi vers une loi exponentielle concentrée sur $]-\infty;0[$.

Exemple 41 *Régression linéaire simple (RLS). On dispose d'un modèle de la forme*

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \cdots, n$$

où les Y_i sont des variables aléatoires indépendantes, e_i sont iid $\sim N\left(0, \sigma^2\right)$ et les x_i sont les covariables (non aléatoires) fixées par le chercheur. Les données viennent sous forme de couples (x_i, Y_i). Il est clair que

$$Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

et que les Y_i sont indépendantes (mais non identiquement distribués) et donc leur loi conjointe est le produit des lois marginales. On veut estimer le vecteur des paramètres (inconnus) du modèle

$$\theta = \left(\beta_0, \beta_1, \sigma^2\right)$$

On a alors

$$L\left(\theta\,|y_1, \cdots, y_n\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right]$$

et

$$l(\theta) = n \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

4.2.2 Principe du maximum de vraisemblance

Si $x = (x_1, \cdots, x_n)$ et $y = (y_1, \cdots, y_n)$ sont deux observations telles que $L(\theta\,|x)$ est proportionnel à $L(\theta\,|y)$ (on écrit $L(\theta\,|x) \propto L(\theta\,|y)$), c-à-d s'il existe une constante $c(x, y)$ indépendante de θ telle que

$$L(\theta\,|x) = c(x, y)\, L(\theta\,|y) \quad \forall \theta \in \Theta$$

alors les conclusions (inférence) tirées à partir de x ou de y sont les mêmes (les deux échantillons contiennent la même information sur θ).

Une autre formulation de ce principe, dite principe d'exhaustivité, nous dit que si la statistique $T(X_1, \ldots, X_n)$ est exhaustive pour θ, $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ sont deux observations iid de $f(x\,|\theta)$ (même modèle) telles que $T(x) = T(y)$, alors les deux échantillons contiennent la même information sur θ. Dans ce cas, T est minimale pour θ.

Exemple 42 *Soit $X_1, \cdots, X_n$ iid $\backsim N\left(\theta = \mu, \sigma^2\right)$ (σ^2 connu). Alors*

$$L(\mu\,|x) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$
$$L(\mu\,|y) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

On a alors

$$\begin{aligned}
\frac{L(\mu\,|x)}{L(\mu\,|y)} &= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu)^2 - \sum_{i=1}^{n}(y_i-\mu)^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\mu)^2 - (y_i-\mu)^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\mu^2 - 2\mu x_i + x_i^2 - \left(\mu^2 - 2\mu y_i + y_i^2\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\mu^2 - 2\mu x_i + x_i^2 - \mu^2 + 2\mu y_i - y_i^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2\mu x_i + x_i^2 + 2\mu y_i - y_i^2\right]\right\} \\
&= c(x, y) \Leftrightarrow \sum_{i=1}^{n}\left[-x_i + y_i\right] = 0 \\
&\Leftrightarrow \overline{x} = \overline{y}
\end{aligned}$$

Dans ce cas ($\overline{x} = \overline{y}$),

$$c(x, y) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[x_i^2 - y_i^2\right]\right\}$$

Noter que si $\overline{x} = \overline{y}$, $c(x, y)$ peut aussi s'écrire (exercice) :

$$c(x, y) = \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^2 - \sum_{i=1}^{n}\left(\frac{y_i - \overline{y}}{\sigma}\right)^2\right]\right\}$$

Ainsi, le principe nous dit qu'on doit tirer les mêmes conclusions sur $\theta = \mu$ à partir de deux échantillons vérifiant $\overline{x} = \overline{y}$.

En clair, le principe nous dit que tout ce qu'on peut dire sur la valeur inconnue du paramètre θ doit être uniquement basé sur $L(\theta|x)$ et rien d'autre. Certains statisticiens considèrent que c'est une restriction trop sévère. Considérons l'exemple suivant.

Exemple 43 *On lance une pièce de monnaie n fois et on observe r piles. Le modèle pour cette expérience est bien entendu le modèle binomial $X \sim Bin(n, \theta = p)$. La fonction vraisemblance est*

$$L(\theta|X = r) = \binom{n}{r}\theta^r(1-\theta)^{n-r} \qquad (*)$$

Nous savons (petit calcul) que cette fonction atteint son maximum en la fréquence relative observée : $\theta = r/n$. Maintenant, supposons que la même pièce est lancée jusqu'à ce qu'on obtienne r piles. Le modèle correspondant à cette expérience est bien entendu le modèle négatif binomial $Y \sim NB(r, \theta = p)$:

$$P(Y = k) = \binom{k-1}{r-1}\theta^r(1-\theta)^{k-r}$$

Si on a obtenu les r piles au bout de n lancers ($k = n$), on a alors

$$L(\theta|Y = n) = P(Y = n) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r} \qquad (**)$$

Comparons $(*)$ et $(**)$:

$$\frac{L(\theta|X = r)}{L(\theta|Y = n)} = \frac{\binom{n}{r}}{\binom{n-1}{r-1}} = cte$$

*Le principe du maximum de vraisemblance nous dit qu'**on doit tirer les mêmes conclusion sur** θ, **indépendamment du fait que les données ont été obtenues de deux façons complètement différentes**. Si on tient compte des aspects spécifiques des modèles utilisés, on peut faire des inférences différentes pour les deux situations : par exemple tester l'hypothèse $\theta = \theta_0$ peut être fait de différentes façons selon le modèle considéré.*

On pourrait faire une inférence basée sur la fonction de vraisemblance comme suit. Considérons l'ensemble (région de vraisemblance)

$$C(x) = \{\theta \in \Theta : L(\theta|x) \geq c\}$$

où $c > 0$ est une constante donnée. On a donc éliminé d'emblée toutes les valeurs de θ qui sont inférieures à c. Ceci signifie que nous rejetons les valeurs non conformes aux données observées x. La taille de l'ensemble $C(x)$ peut alors servir de mesure de notre ignorance sur la vraie valeur de θ. Reste le problème du choix de la constante c. Nous verrons (intervalles de confiance, plus loin dans ce chapitre) comment exploiter les propriétés du modèle pour y arriver. Mais donnons un exemple.

Exemple 44 *Reprenons l'exemple de la loi normale. Soit $X_1, \cdots, X_n$ iid $\backsim N(\theta = \mu, \sigma^2)$ (σ^2 connu). On a vu que la fonction vraisemblance est*

$$L(\theta|x) = \left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right]$$

Maintenant,

$$\begin{aligned}(x_i-\theta)^2 &= ((x_i-\overline{x})+(\overline{x}-\theta))^2 = (x_i-\overline{x})^2 + 2(x_i-\overline{x})(\overline{x}-\theta) + (\overline{x}-\theta)^2 \\ \Rightarrow \sum_{i=1}^{n}(x_i-\theta)^2 &= \sum_{i=1}^{n}(x_i-\overline{x})^2 + 2(\overline{x}-\theta)\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{=0} + \sum_{i=1}^{n}(\overline{x}-\theta)^2 \\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\theta)^2 = (n-1)s^2 + n(\overline{x}-\theta)^2\end{aligned}$$

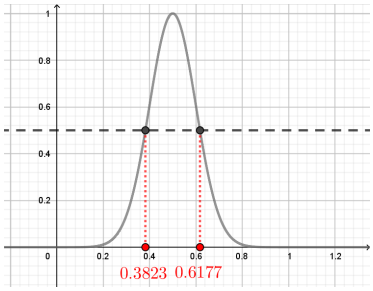
On a alors

$$L(\theta|x) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right] \exp\left(-\frac{n-1}{2\sigma^2}s^2\right)$$
$$= Cte \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right]$$

On peut donc utiliser

$$\boxed{L(\theta|x) = \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right]}, \ \sigma^2 \text{ connu}$$

On voit (sans calcul) que le maximum est atteint en $\theta = \overline{x}$. A titre numérique, prenons $n = 100$, $\sigma^2 = 1$ et supposons qu'on a observé $\overline{x} = 0.5$.



Si on prend $c = 0.5$, l'intervalle de vraisemblance pour cette valeur de c est donné par (voir figure)

$$C(x) = \{L(\theta|x = (x_1, \ldots, x_{100})) \geq 0.5\} = [0.3823; 0.6177]$$

Remarque 5 *Retour sur l'exhaustivité. Nous avons vu (théorème de factorisation) qu'une statistique $T = T(X_1, \cdots, X_n)$ est **exhaustive** pour θ si et seulement si la loi conjointe peut se décomposer comme suit :*

$$f(x_1, \cdots, x_n|\theta) = g(T|\theta)\, h(x_1, \cdots, x_n)$$

Le principe du maximum de vraisemblance nous dit de ne pas tenir compte de $h(x_1, \cdots, x_n)$. On voit qu'en général, on a réduit la dimension du problème.

Exemple 45 *Soit $\Theta = \{\theta_1, \theta_2\}$ et supposons qu'on doive décider entre les deux modèles :*

	X	1	2	3	4
$\theta = \theta_1$	$f(x\|\theta_1)$	1/2	1/6	1/6	1/6
$\theta = \theta_2$	$f(x\|\theta_2)$	1/4	1/4	1/4	1/4
	$\Lambda = \frac{f(x\|\theta_1)}{f(x\|\theta_2)}$	2	2/3	2/3	2/3

On voit que le rapport de vraisemblance Λ est le même si on observe $X = 2, 3$ ou 4. Par conséquent, la statistique T définie par

$$T(1) = 0$$
$$T(2) = T(3) = T(4) = 1$$

est exhaustive pour θ. On ne considère que 2 valeurs, au lieu de 4 initialement.

L'intérêt est évidemment de trouver une statistique exhaustive qui réduise au maximum la dimension du problème, c-à-d une statistique exhaustive minimale.

4.2.3 Principe d'invariance

Ce principe important s'énonce comme suit.

Principe d'invariance. Si $\widehat{\theta}$ est l'EVM de θ et $g(\theta)$ est une fonction de θ, alors $g\left(\widehat{\theta}\right)$ est l'EVM de $g(\theta)$.

Il est utile si l'EVM de θ est plus facile à calculer que celui de $g(\theta)$ (fonction compliquée).

Exemple 46 *Soit $X \backsim Exp(\lambda)$. Ceci est un cas particulier de $Gamma(\alpha = 1, \lambda)$. Un exemple précédent a donné $\widehat{\lambda} = 1/\overline{X}$. Il arrive souvent qu'on reparamètre cette loi comme suit : $Exp(\lambda = 1/\theta)$, c-à-d*

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta}, \qquad x \geq 0$$

Faisons un calcul direct de $\widehat{\theta}$:

$$L(\theta) = \frac{1}{\theta^n} \exp\left(-\sum_{i=1}^{n} x_i/\theta\right) = \frac{1}{\theta^n} \exp(-n\overline{x}/\theta)$$
$$l(\theta) = -n \ln\theta - \frac{n\overline{x}}{\theta}$$

et

$$\frac{\partial l}{\partial \theta}(\theta) = -\frac{n}{\theta} + n\frac{\overline{x}}{\theta^2} = 0$$

dont la solution est

$$\theta = \overline{x}$$

On a donc

$$\widehat{\theta} = \overline{X} = 1/\widehat{\lambda}$$

Exemple 47 *Si $X_1, \cdots, X_n$ iid $\sim Bernoulli(\theta = p)$, nous savons que la fréquence relative $\widehat{\theta} = \overline{X}$ est l'EVM de p. On déduit du principe d'invariance que $\sqrt{\overline{X}(1-\overline{X})}$ est l'EVM de $\sigma = \sqrt{p(1-p)}$.*

4.2.4 Exercices

Exercice 19 *Trouver les EMV de l'exemple de la régression linéaire simple.*

Exercice 20 *Soit $X_1, \cdots, X_n$ iid $\sim Poi(\theta)$.*

1. *Calculer $\dfrac{L(\theta|x)}{L(\theta|y)}$.*
2. *En déduire que $T = \sum_{i=1}^{n} X_i$ est une statistique exhaustive minimale pour θ.*

5 Loi multinomiale

On dispose de m ($m \geq 2$) boites. On effectue une suite de n épreuves indépendantes et identiques où, à chaque épreuve, le résultat tombe dans l'une des boîtes :

Résultats	**	*		***	$\cdots$	**	****
Boîte	1	2	3	4	$\cdots$	$m-1$	m

Soit X_i = nombre de succès dans la boîte i ($i = 1, \cdots, m$) : $X_i \backsim Bin(n, p_i)$ où $p_i = P$(tomber dans la boîte i). La loi conjointe des X_i est donc une loi multinomiale :

$$p_{(X_1,\cdots,X_m)}(x_1, \cdots, x_m | p_1, \cdots, p_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m},$$

avec

$$x_i = 0, \cdots, n\,;$$
$$\sum_{i=1}^{m} x_i = n\,;$$
$$0 \leq p_i \leq 1\,;$$
$$\sum_{i=1}^{m} p_i = 1$$

Noter que les X_i sont dépendants et négativement corrélés (cas $m = 2 : x_1 + x_2 = n$ est l'équation d'une droite avec pente négative : angle de l'inclinaison est obtus, donc cosinus négatif. Une preuve dans le cas général est donnée plus bas.)

On veut estimer les p_i avec la méthode des EVM :

$$L(p_1, \cdots, p_m) = \frac{n!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i^{x_i}$$
$$\Rightarrow l(p_1, \cdots, p_m) = \ln n! - \sum_{i=1}^{m} \ln x_i! + \sum_{i=1}^{m} x_i \ln p_i$$

Nous avons ici un problème de minimisation avec contrainte ($\sum_{i=1}^{m} p_i = 1$). Une façon de le traiter est celle des multiplicateurs de Lagrange. On forme la fonction

$$g(p_1, \cdots, p_m, \lambda) = l(p_1, \cdots, p_m) + \lambda \left(\sum_{i=1}^{m} p_i - 1 \right)$$

On a

$$\frac{\partial g}{\partial p_i}(p_1, \cdots, p_m, \lambda) = \frac{x_i}{p_i} + \lambda = 0, \qquad i = 1, \cdots, m$$
$$\frac{\partial g}{\partial \lambda}(p_1, \cdots, p_m, \lambda) = \sum_{i=1}^{m} p_i - 1 = 0$$

On tire (petit calcul) les estimateurs :

$$\widehat{p}_i = X_i/n, \qquad i = 1, \cdots, m$$

Donc l'EMV de p_i = fréquence relative dans la boite i.

5.1 Distribution d'échantillonnage des $\widehat{p}_i$

Puisque $X_i \sim Bin(n, p_i)$, on a (TCL)

$$\frac{X_i - np_i}{\sqrt{np_i(1-p_i)}} \overset{Loi}{\to} Z \sim N(0,1)$$

et (méthode delta-TCL)

$$\frac{\widehat{p}_i - p_i}{\sqrt{p_i(1-p_i)/n}} \overset{Loi}{\to} Z \sim N(0,1)$$
$$\Rightarrow \sqrt{n}(\widehat{p}_i - p_i) \overset{Loi}{\to} Z \sim N(0, p_i(1-p_i))$$

De plus, on a

$$E(\widehat{p}_i) = p_i$$
$$VAR(\widehat{p}_i) = p_i(1-p_i)/n \underset{n\to\infty}{\to} 0$$

Ce sont donc des estimateurs sans biais et convergents.

5.2 Rappel. Matrice de variance-covariance*

Soit $X = (X_1, \cdots, X_p)$, ou en notation matricielle,

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

un vecteur aléatoire général. Soit

$$E(X) = (E(X_1), \cdots, E(X_p)) = (\mu_1, \cdots, \mu_p) = \mu$$

Pour toute matrice aléatoire $M = (M_{ij})_{p \times q}$, on définit

$$E(M) = (E(M_{ij}))_{p \times q}$$

La matrice de variance-covariance de X est définie par

$$\begin{aligned} Cov(X) &= (Cov(X_i, X_j))_{p \times p} \\ &\overset{def}{=} \left(E \left[\underbrace{(X-\mu)}_{p \times 1} \underbrace{(X-\mu)^t}_{1 \times p} \right] \right) \text{ (symétrique)} \end{aligned}$$

Rappelons que

$$Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)], \qquad i, j = 1, \cdots, p$$

est l'élément dans la position (i, j) dans la matrice de variance-covariance. On note parfois

$$Cov(X_i, X_j) = \sigma_{ij}$$

En particulier, $Cov(X_i, X_i) = VAR(X_i) = \sigma_i^2$. Maintenant,

$$\begin{aligned} Cov(X) &= E\left[(X-\mu)(X-\mu)^t\right] \\ &= E\left[XX^t - X\mu^t - \mu X^t + \mu\mu^t\right] \\ &= E\left(XX^t\right) - E(X)\mu^t - \mu E\left(X^t\right) + \mu\mu^t \\ &= E\left(XX^t\right) - \mu\mu^t \end{aligned}$$

De plus, si les vecteurs $p-$dimensionnels X et Y sont indépendants,

$$Cov(X + Y) = Cov(X) + Cov(Y)$$

5.3 Cas de la loi multinomiale

Soit le vecteur aléatoire Y (de taille m) tel que $P(Y = e_j) = p_j$ avec

$$e_j = (0, \cdots, 1, 0, \cdots, 0), \qquad j = 1, \cdots, m$$

(base canonique de $\mathbb{R}^m$). Pour une seule épreuve, Y indique la boîte observée. Soit $Y_1, \cdots, Y_n$ iid de même loi que Y.

Exemple 48 *On lance un dé $n = 5$ fois ($m = 6$). On a par exemple les résultats suivants :*

	Réalisation						
	Boîtes						
Épreuves	1	2	3	4	5	6	
Y_1	1	0	0	0	0	0	e_1
Y_2	0	0	0	0	1	0	e_5
Y_3	0	1	0	0	0	0	e_2
Y_4	0	1	0	0	0	0	e_2
Y_5	0	1	0	0	0	0	e_2
$\sum_{j=1}^5 Y_j$	1	3	0	0	1	0	$e_1 + 3e_2 + e_5$

Ainsi, en général,

$$(X_1, \cdots, X_m) = \sum_{j=1}^{n} Y_j \backsim mult\,(n, p_1, \cdots, p_m)$$

avec

$$E\,(Y_i) = \sum_{j=1}^{m} e_j p_j = \begin{bmatrix} p_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ p_2 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ p_m \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix} = p, \qquad i = 1, \cdots, n$$

et

$$\begin{aligned} E\left(YY^t\right) &= \sum_{j=1}^{m} \left(e_j e_j^t\right) p_j \\ &= \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \\ 0 & & & 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \\ 0 & & & p_m \end{bmatrix} \\ &= \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & & \\ \vdots & & \ddots & \\ 0 & & & p_m \end{bmatrix} = diag\,(p_1, \cdots, p_m) \end{aligned}$$

Donc

$$\begin{aligned} Cov\left(YY^t\right) &= E\left(YY^t\right) - E\,(Y)\,E\left(Y^t\right) \\ &= \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & & \\ \vdots & & \ddots & \\ 0 & & & p_m \end{bmatrix} - \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix} \begin{bmatrix} p_1 & p_2 & \cdots & p_m \end{bmatrix} \\ &= \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & & \\ \vdots & & \ddots & \\ 0 & & & p_m \end{bmatrix} - \begin{bmatrix} p_1^2 & p_1 p_2 & \cdots & p_1 p_m \\ p_1 p_2 & p_2^2 & & \\ \vdots & & \ddots & \\ p_1 p_m & & & p_m^2 \end{bmatrix} \\ &= \begin{bmatrix} p_1\,(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_1 p_2 & p_2\,(1-p_2) & & \\ \vdots & & \ddots & \\ -p_1 p_m & & & p_m\,(1-p_m) \end{bmatrix} \end{aligned}$$

Alors,

$$\begin{aligned} Cov\,(X_1, \cdots, X_m) &= Cov\left(\sum_{j=1}^{n} Y_j\right) \overset{ind.}{=} \sum_{j=1}^{n} Cov\,(Y_j) = nCov\,(Y) \\ &= n \begin{bmatrix} p_1\,(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_1 p_2 & p_2\,(1-p_2) & & \\ \vdots & & \ddots & \\ -p_1 p_m & & & p_m\,(1-p_m) \end{bmatrix} \end{aligned}$$

On voit que

$$\begin{aligned} VAR\,(X_j) &= np_j\,(1-p_j) \\ Cov\,(X_i, X_j) &= -np_i p_j < 0 \end{aligned}$$

5.4 Calcul de $Cov(X_i, X_j)$: approche probabiliste

Pour ceux qui ont du mal avec l'approche matricielle ci-dessus, on peut facilement retrouver ces résultats par une approche directe. D'abord, il est clair que chacune des variables X_j est une binomiale

$$X_j \backsim Bin(n, p_j)$$

et donc

$$VAR(X_j) = np_j(1 - p_j)$$

Il nous reste à trouver $Cov(X_i, X_j)$ pour $i \neq j$. Considérons les boîtes i et j (disons B_i et B_j) et convenons qu'on a un succès si on tombe sur la boite i ou sur la boite j. La variable qui décrit le nombre de succès (tomber sur B_i ou sur B_j) est clairement $X_i + X_j$. De plus, c'est une variable de loi binomiale :

$$(X_i + X_j) \backsim Bin(n, p_i + p_j)$$

Sa variance est alors

$$VAR(X_i + X_j) = n(p_i + p_j)(1 - (p_i + p_j)) \qquad (*)$$

Par ailleurs, on a

$$\begin{aligned} VAR(X_i + X_j) &= VAR(X_i) + VAR(X_j) + 2Cov(X_i, X_j) \\ &= np_i(1 - p_i) + np_j(1 - p_j) + 2Cov(X_i, X_j) \qquad (**) \end{aligned}$$

Égalant (*) et (**), on a

$$n(p_i + p_j)(1 - (p_i + p_j)) = np_i(1 - p_i) + np_j(1 - p_j) + 2Cov(X_i, X_j)$$

Un petit calcul donne

$$Cov(X_i, X_j) = -np_ip_j$$

comme attendu.

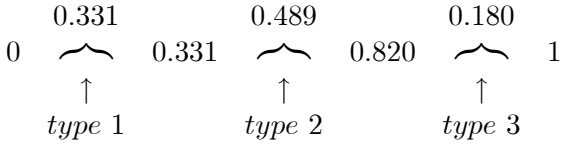
Exercice 21 *Quel est le coefficient de corrélation ρ_{ij} ? Que vaut-il dans le cas $m = 2$ (deux boîtes) ? Ce résultat était-il prévisible ?*

5.5 Simulation d'une multinomiale

Comment générer une multinomiale ? Prenons un cas numérique. Soit

$$(X_1, X_2, X_3) \backsim mult(n = 1029, p_1 = 0.331, p_2 = 0.489, p_3 = 0.180)$$

Idée. On fait une seule épreuve. Tous les systèmes informatiques ont un générateur de nombres pseudo-aléatoires de loi $U \backsim [0; 1]$.



Algorithme

1. $x_1 = 0$, $x_2 = 0$, $x_3 = 0$ (initialisation des compteurs)
2. $j = 1$ (initialisation de la première épreuve)
3. Générer $U \backsim [0; 1]$
4. Si $U \in [0; 0.331]$, $x_1 = x_1 + 1$
   Si $U \in]0.331; 0.820]$, $x_2 = x_2 + 1$
   Si $U \in]0.820; 1]$, $x_3 = x_3 + 1$

5. Si $j = 1029$: arrêt
6. $j = j + 1$ et aller en 3.

Au bout du compte, le vecteur (x_1, x_2, x_3) est l'effectif des 1029 épreuves. Par exemple, avec (l'utilitaire d'analyse d') excel, j'ai obtenu les résultats suivants : $x_1 = 349$, $x_2 = 479$, $x_3 = 201$, ce qui donne les estimations

$$\begin{aligned}\widehat{p}_1 &= 349/1029 = 0.339\\ \widehat{p}_2 &= 479/1029 = 0.466\\ \widehat{p}_3 &= 201/1029 = 0.196\end{aligned}$$

Exemple 49 *(Rice page 273. Hardy-Weinberg). On considère les génotypes AA, Aa, aa avec probabilités associées* $(1-\theta)^2$, $2\theta(1-\theta)$, θ^2 *(noter la dépendance du paramètre* θ*). Ici,* θ = *probabilité de transmission du gène a pour un parent.*

Estimation de θ. *On prend un échantillon aléatoire de taille* 1029. *Reprenons les données de Rice :*

Génotype	*AA*	*Aa*	*aa*
Fréquence	342	500	187

On a une réalisation de

$$(X_1, X_2, X_3) \backsim mult\left(n = 1029, p_1(\theta) = (1-\theta)^2, p_2(\theta) = 2\theta(1-\theta), p_3(\theta) = \theta^2\right)$$

Puisque $X_i \backsim Bin(n, p_i)$, *une façon naïve d'estimer* θ *est d'utiliser la convergence suivante :*

$$\begin{aligned}\frac{X_3}{n} = \widehat{p}_3 \Rightarrow \widehat{p}_3 &\xrightarrow{P} p_3(\theta) = \theta^2\\ \Rightarrow \widehat{\theta}_{Naive} &= \sqrt{\widehat{p}_3} \xrightarrow{P} \theta\end{aligned}$$

car la fonction $x \to \sqrt{x}$ *est continue. On prendrait alors*

$$\widehat{\theta}_{Naive} = \sqrt{\frac{187}{1029}} = 0.4263$$

comme estimation.

Cette façon de faire ignore les informations fournies par les autres boîtes : p_1 *et* p_2 *dépendent également de* θ.

Cherchons l'EMV.

$$\begin{aligned}L(\theta) &= \frac{n!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i(\theta)^{x_i}\\ &= cte \times \left[(1-\theta)^2\right]^{x_1} \left[2\theta(1-\theta)\right]^{x_2} \left[\theta^2\right]^{x_3}\\ &= cte \times (1-\theta)^{2x_1+x_2} \times \theta^{x_2+2x_3}\end{aligned}$$

et

$$l(\theta) = cte + (2x_1 + x_2)\ln(1-\theta) + (x_2 + 2x_3)\ln\theta$$

Donc

$$\frac{\partial}{\partial\theta} l(\theta) = \frac{x_2 + 2x_3}{\theta} - \frac{2x_1 + x_2}{1-\theta} = 0$$

Après calculs,

$$\widehat{\theta} = \frac{X_2 + 2X_3}{2X_1 + 2X_2 + 2X_3} = \frac{X_2 + 2X_3}{2n}$$

L'estimation est alors (on utilise la même notation que celle de l'estimateur)

$$\widehat{\theta} = \frac{500 + 2 \times 187}{2 \times 1029} = \frac{437}{1029} = 0.424\,7$$

Remarque. Dans le calcul de l'EMV, nous avons cherché la dérivée $\frac{\partial}{\partial\theta} l(\theta)$ sans tenir compte de la contrainte $p_1 + p_2 + p_3 = 1$. Pourquoi est-ce justifié ?

Examinons la dispersion de $\widehat{\theta}$ (combien de décimales retenir ?). On a

$$\begin{aligned} E\left(\widehat{\theta}\right) &= E\left(\frac{X_2 + 2X_3}{2n}\right) = \frac{1}{2n}\left(E(X_2) + 2E(X_3)\right) \\ &= \frac{1}{2n}\left(np_2(\theta) + 2np_3(\theta)\right) \\ &= \frac{1}{2}\left(2\theta(1-\theta) + 2\theta^2\right) \\ &= \theta \qquad \textit{(sans biais)} \end{aligned}$$

et

$$\begin{aligned} VAR\left(\widehat{\theta}\right) &= VAR\left(\frac{X_2 + 2X_3}{2n}\right) = \frac{1}{4n^2}\left[VAR(X_2) + 4VAR(X_3) + 2 \times 2Cov(X_2, X_3)\right] \\ &= \frac{1}{4n^2}\left[np_2(\theta)(1 - p_2(\theta)) + 4np_3(\theta)(1 - p_3(\theta)) - 4np_2(\theta)p_3(\theta)\right] \\ &\overset{calculs}{=} \frac{\theta(1-\theta)}{2n} = g(\theta) \end{aligned}$$

Ainsi, $\widehat{\theta}$ est un estimateur sans biais de θ et sa variance dépend de θ (inconnu). Il nous faut donc estimer cette variance :

$$\widehat{VAR}\left(\widehat{\theta}\right) = \widehat{\sigma}^2(\theta) = \frac{\widehat{\theta}\left(1 - \widehat{\theta}\right)}{2n} = \frac{0.4247 \times (1 - 0.4247)}{2 \times 1029}$$

ce qui nous donne

$$\widehat{\sigma}(\theta) = se_{\widehat{\theta}} = 0.0109$$

Ce résultat peut donc servir pour établir un intervalle de confiance pour θ.

Remarque 6 *On peut utiliser une simulation (bootstrap) pour générer beaucoup d'échantillons de $\widehat{\theta}$, puis évaluer $VAR\left(\widehat{\theta}\right)$ (l'algorithme sera expliqué plus loin).*

Quelle est la variance de $\widehat{\theta}_{Naive}$? On a

$$\widehat{\theta}_{Naive} = \sqrt{\widehat{p}_3} = \sqrt{X_3/n}$$

C'est une fonction non linéaire de X_3. Une approche est la méthode delta (TCL). Rappelons ce résultat (cf. chapitre 5).

Proposition 4 *Si*

$$\sqrt{n}(X_n - \mu) \overset{Loi}{\to} N\left(0, \sigma^2\right)$$

et si $g'(\mu)$ existe ($g'(\mu) \neq 0$), alors

$$\sqrt{n}(g(X_n) - g(\mu)) \overset{Loi}{\to} N\left(0, \sigma^2 g'(\mu)^2\right)$$

Rappelons aussi la conséquence suivante.

Corollaire 1 *Si*

$$\sqrt{n}\left(X_n-\mu\right)\overset{Loi}{\to}N\left(0,\sigma^2\right)$$

et $E\left(X_n\right)=\mu$, $VAR\left(X_n\right)=\sigma^2/n$, *alors*

$$E\left(g\left(X_n\right)\right)\approxeq g\left(\mu\right)+\frac{1}{2}\left|g''\left(\mu\right)\right|\sigma^2/n$$

Ce qui donne

$$\left|Biais\left(g\left(X_n\right)\right)\right|\approxeq\frac{1}{2}\left|g''\left(\mu\right)\right|\sigma^2/n$$

Exemple 50 *Revenons à l'exemple des génotypes AA, Aa, aa. On a vu que*

$$\begin{aligned}n&=1029,\\p_1\left(\theta\right)&=\left(1-\theta\right)^2,\\p_2\left(\theta\right)&=2\theta\left(1-\theta\right),\\p_3\left(\theta\right)&=\theta^2\end{aligned}$$

et que

$$\widehat{\theta}_{Naive}=\sqrt{\widehat{p}_3}=\sqrt{X_3/n}\approxeq0.4363$$

Pour comparaison :

$$\widehat{\theta}_{EVM}=\frac{X_2+2X_3}{2n},VAR\left(\widehat{\theta}_{EVM}\right)=\frac{\theta\left(1-\theta\right)}{2n},$$

$$\widehat{VAR}\left(\widehat{\theta}_{EVM}\right)=\frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{2n}\approxeq\boxed{\left(0.0109\right)^2}$$

TCL :

$$\sqrt{n}\left(\widehat{p}_3-p_3\right)\overset{Loi}{\to}N\left(0,p_3\left(1-p_3\right)\right)$$

avec

$$\begin{aligned}E\left(\widehat{p}_3\right)&=E\left(X_3/n\right)=np_3/n=p_3\qquad\textit{(sans biais)}\\VAR\left(\widehat{p}_3\right)&=\frac{1}{n^2}VAR\left(X_3\right)=\frac{np_3\left(1-p_3\right)}{n^2}=\frac{p_3\left(1-p_3\right)}{n}\end{aligned}$$

Soit $g\left(x\right)=\sqrt{x},x>0$. *On a* $g'\left(x\right)=\frac{1}{2\sqrt{x}}$, *ce qui donne*

$$\begin{aligned}VAR\left(\sqrt{\widehat{p}_3}\right)&\approxeq\frac{p_3\left(1-p_3\right)}{n}\left(\frac{1}{2\sqrt{p_3}}\right)^2=\frac{1-p_3}{4n}=\frac{1-\theta^2}{4n}\\&\Rightarrow\widehat{VAR}\left(\widehat{\theta}_{Naive}\right)\approxeq\frac{1-\widehat{\theta}^2_{Naive}}{4n}=\boxed{\left(0.014\right)^2}\end{aligned}$$

Estimons le biais. On a

$$\begin{aligned}g''\left(x\right)=\frac{-1}{4x^{3/2}}&\Rightarrow\left|Biais\left(\widehat{\theta}_{Naive}\right)\right|\approxeq\frac{p_3\left(1-p_3\right)}{2n}\times\frac{1}{4p_3^{3/2}}=\frac{1-\theta^2}{8n\theta}\\&\Rightarrow\left|\widehat{Biais}\left(\widehat{\theta}_{Naive}\right)\right|\approxeq\frac{1-\widehat{\theta}^2_{Naive}}{8n\widehat{\theta}_{Naive}}=0.0002\qquad\textit{(négligeable)}\end{aligned}$$

5.6 Bootstrap paramétrique

- Méthode pour approximer la distribution d'échantillonnage d'un estimateur.
- Elle utilise la puissance de calcul des ordinateurs.
- Il y a peu de développements théoriques.
- Basée sur la simulation de Monte-Carlo.
- Applicable pour de grands échantillons (méthode asymptotique).

Exemple 51 *Rice. Données Illinois (voir pages 264-265). 227 orages.* $n = 227, X =$ *quantité de pluie en pouces. Modèle :* $X \backsim Gamma\,(\alpha, \lambda)$. *Méthode des moments :*

$$\widehat{\lambda} = \overline{X}/\widehat{\sigma}^2, \qquad \widehat{\alpha} = \widehat{\lambda} \times \overline{X}$$

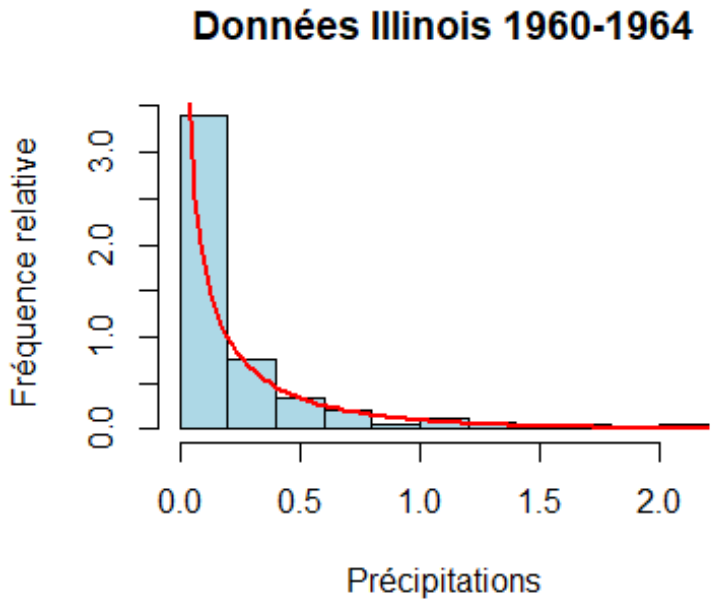
avec

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Calculs :

$$\overline{X} = 0.224, \qquad \widehat{\sigma}^2 = 0.1238$$
$$\Rightarrow \widehat{\lambda} = 1.674, \qquad \widehat{\alpha} = 0.375$$

Voir figure 8.3 page 264 : histogramme + densité Gamma $\left(\widehat{\alpha}, \widehat{\lambda}\right)$.



Code R utilisé pour cette figure (les données sont dans un fichier illinois, qu'on peut obtenir du site web du manuel) :

```
hist(illinois,xlab = 'Précipitations', ylab = 'Fréquence relative',
... main = "Données Illinois 1960-1964", probability = T, col='lightblue')
curve(dgamma(x, shape = 0.375, rate = 1.674),lwd = 2, col = 'red', add = T)
```

Exemple 52 *(suite de l'exemple précédent) Quelle est la distribution d'échantillonnage de* $\widehat{\alpha}$ *et de* $\widehat{\lambda}$ *?*

Si nous connaissons les vraies valeurs α_0 *et* λ_0*, on pourrait effectuer une simulation de Monte Carlo de taille* B *:*

1. $B = 1$
2. $X_1, \cdots, X_{227}$ *iid Gamma* (α_0, λ_0)
3. *Calcul de* $\widehat{\alpha}$ *et* $\widehat{\lambda}$
4. *Si* $B = 1000$*, fin*
5. $B = B + 1$

À la fin du programme, on a obtenu

$$\widehat{\alpha}_1, \cdots, \widehat{\alpha}_B$$
$$\widehat{\lambda}_1, \cdots, \widehat{\lambda}_B$$

Un histogramme des $\widehat{\alpha}_i$ $(i = 1, \cdots, B)$ *donne une bonne approximation de la distribution d'échantillonnage de* $\widehat{\alpha}$ *(idem pour* $\widehat{\lambda}$*).*

Mais nous ne connaissons pas α_0 *et* λ_0*.*

Algorithme <u>*BOOTSTRAP*</u>*.*

1. $B = 1$
2. $X_1^*, \cdots, X_{227}^*$ *iid Gamma* $(\alpha = 0.375, \lambda = 1.674)$ *(échantillon bootstrap)*
3. *Calcul de* α^* *et* λ^*
4. *Si* $B = 1000$*, fin*
5. $B = B + 1$

L'estimation bootstrap de l'erreur standard de $\widehat{\alpha}$ *est*

$$s_{\widehat{\alpha}} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (\alpha_i^* - \overline{\alpha}^*)}$$

Rice obtient

$$s_{\widehat{\alpha}} = 0.06, \qquad s_{\widehat{\lambda}} = 0.34$$

<u>Voici un code R</u> pour un jeu de données artificiel (le remplacer par celui qui vous intéresse).

```
# Fonction échantillonnage
boot.param <- function(X, B, N) {
  n <- length(X)
  out <- numeric(N)
  for(i in 1 :N) out[i] <- B(sample(X, n, replace = TRUE))
  return(out)
}
# Jeu de données
X <- c(131.7, 183.7, 73.3, 10.7, 150.4, 42.3, 22.2, 17.9, 264.0, 154.4, 4.3, 256.6, 61.9, 10.8, 48.8, 22.5, 8.8, 150.6, 103.0, 85.9)
# Appel de la fonction
out <- boot.param(X, var, 50000) # On échantillonne 50000 fois (remplacer par la valeur qui vous convient)
hist(out, freq = FALSE, xlab = expression(s[boot]^2), col = "white", border = "blue", main = "") # Histogramme
abline(v = var(X), lwd = 2,col = 'red') # Variance échantillonnale
print(quantile(out, c(0.05, 0.95))) # Intervalle de confiance à 95%
# Fin du programme
```

Exemple 53 *Génétique (suite). Rappelons :*

$$(X_1, X_2, X_3) \backsim mult\left(n = 1029, p_1 = (1-\theta)^2, p_2 = 2\theta(1-\theta), p_3 = \theta^2\right), \quad \theta \in]0; 1[\,.$$

$$EVM : \widehat{\theta} = \frac{X_2 + 2X_3}{2n};$$
$$données : x_1 = 342, x_2 = 500, x_3 = 187$$
$$\widehat{\theta}(x_1, x_2, x_3) = 0.4247;$$
$$VAR\left(\widehat{\theta}\right) = \frac{\theta(1-\theta)}{2n},$$
$$\widehat{VAR}\left(\widehat{\theta}\right) = (0.0109)^2$$

On peut estimer l'écart type de $\widehat{\theta}$ par le bootstrap.

1. $B = 1$
2. $(X_1^*, X_2^*, X_3^*) \backsim mult\left(n = 1029, \left(1-\widehat{\theta}\right)^2, 2\widehat{\theta}\left(1-\widehat{\theta}\right), \widehat{\theta}^2\right)$ *avec* $\widehat{\theta} = 0.4247$
3. $\theta^* = \frac{x_2+2x_3}{2n}$
4. *Si* $B = 1000$, *fin*
5. $B = B + 1$

On a ainsi $\theta_1^*, \cdots, \theta_{1000}^*$ *et*

$$s_{\widehat{\theta}} \approxeq \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\theta_i^* - \overline{\theta}^*\right)^2}$$

6 Comportement asymptotique des EVM

La quantité d'information dans la réalisation d'un événement A ($p = P(A)$) est donnée par $-\ln p$ (si $p \approxeq 1$, il y a peu d'information, mais si $p \approxeq 0$, il y a beaucoup d'information). Dans le cas d'une variable aléatoire X, de fonction de densité (ou de masse) $f(x|\theta)$, on cherche la quantité d'information contenue dans une observation.

Définition 12 *La fonction $\frac{\partial}{\partial\theta}\ln f(x|\theta)$ s'appelle la fonction score.*

Remarque 7 *Cette fonction regarde le taux de variation par rapport à θ (dans une observation).*

Exemple 54 *Soit $X \backsim Exp(\lambda)$. On a $f(x|\lambda) = \lambda e^{-\lambda x}$ ($x \geq 0$) et*

$$\ln f(x|\lambda) = \ln\lambda - \lambda x \Rightarrow \frac{\partial}{\partial\lambda}\ln f(x|\lambda) = \frac{1}{\lambda} - x$$

Exemple 55 *Soit $X \backsim Poi(\lambda)$. On a $f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$ ($x = 0, 1, \cdots$) et*

$$\ln f(x|\lambda) = -\lambda + x\ln\lambda - \ln x! \Rightarrow \frac{\partial}{\partial\lambda}\ln f(x|\lambda) = -1 + \frac{x}{\lambda}$$

Définition 13 *Soit $X_1, \cdots, X_n$ un échantillon iid de loi $f(x|\theta)$. La **fonction score de l'échantillon** est définie par*

$$\boxed{\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_i|\theta) = \frac{\partial}{\partial\theta}l(\theta)}$$

Rappelons que $\widehat{\theta}_{EMV}$ est solution de $\frac{\partial}{\partial\theta}l(\theta) = 0$ (quand elle existe).

Définition 14 *L'information de Fisher est définie par*

$$\boxed{I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right]}$$

sous réserve de convergence. C'est donc le second moment de la fonction score.

Remarque 8 *Le carré est introduit pour éliminer le signe négatif introduit par le logarithme.*

Proposition 5 *Si*

$$1.\ \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f(x|\theta)\, dx = 0$$

et

$$2.\ \int_{\mathbb{R}} \frac{\partial^2}{\partial\theta^2} f(x|\theta)\, dx = 0$$

alors

$$\boxed{1.\ I(\theta) = VAR\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)}$$

$$\boxed{2.\ I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right)}$$

Preuve. *Noter que si on peut interchanger les opérations d'intégration et de dérivation, les hypothèses sont automatiquement vérifées (conditions de régularité) puisque l'intégrale de la densité vaut* 1*. On a*

$$\begin{aligned} 1.\ E\left[\frac{\partial}{\partial\theta}\ln f(X|\theta)\right] &= \int_{\mathbb{R}}\left[\frac{\partial}{\partial\theta}\ln f(x|\theta)\right] f(x|\theta)\, dx \\ &= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} f(x|\theta)\, dx \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f(x|\theta)\, dx = 0 \end{aligned}$$

ce qui montre le premier résultat. Pour le second, on a

$$\begin{aligned} 2.\ E\left[\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right] &= E\left[\frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta} f(X|\theta)}{f(X|\theta)}\right] \\ &= E\left[\frac{f(X|\theta)\frac{\partial^2}{\partial\theta^2} f(X|\theta) - \left(\frac{\partial}{\partial\theta} f(X|\theta)\right)^2}{f(X|\theta)^2}\right] \\ &= \int_{\mathbb{R}} \frac{f(x|\theta)\frac{\partial^2}{\partial\theta^2} f(x|\theta) - \left(\frac{\partial}{\partial\theta} f(x|\theta)\right)^2}{f(x|\theta)^2} f(x|\theta)\, dx \\ &= \underbrace{\int_{\mathbb{R}} \frac{\partial^2}{\partial\theta^2} f(x|\theta)\, dx}_{0} - \int_{\mathbb{R}} \left(\frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\right)^2 f(x|\theta)\, dx \\ &= -I(\theta) \end{aligned}$$

■

Exemple 56 *Soit* $X \backsim Exp(\lambda)$*. On a* $f(x|\lambda) = \lambda e^{-\lambda x}$ *(*$x \geq 0$*),* $E(X) = \frac{1}{\lambda}$*,* $VAR(X) = 1/\lambda^2$*,* $\ln f(x|\lambda) = \ln\lambda - \lambda x \Rightarrow$

$$\frac{\partial}{\partial\lambda}\ln f(x|\lambda) = \frac{1}{\lambda} - x$$

et

$$\frac{\partial^2}{\partial\lambda^2}\ln f(x|\lambda) = -\frac{1}{\lambda^2}$$

Par ailleurs,

$$E\left(\frac{\partial}{\partial\lambda}\ln f\left(X\left|\lambda\right.\right)\right)=E\left(\frac{1}{\lambda}-X\right)=\frac{1}{\lambda}-\frac{1}{\lambda}=0$$

Donc

$$I\left(\lambda\right)=VAR\left(\frac{1}{\lambda}-X\right)=VAR\left(X\right)=\frac{1}{\lambda^2}$$

et aussi

$$I\left(\lambda\right)=-E\left(-\frac{1}{\lambda^2}\right)=\frac{1}{\lambda^2}$$

Exemple 57 *Soit $X\backsim Poi\left(\lambda\right)$. On a $f\left(x\left|\lambda\right.\right)=\frac{\lambda^x}{x!}e^{-\lambda}$ $(x=0,1,\cdots)$, $\ln f\left(x\left|\lambda\right.\right)=-\lambda+x\ln\lambda-\ln x!$,*

$$\frac{\partial}{\partial\lambda}\ln f\left(x\left|\lambda\right.\right)=-1+\frac{x}{\lambda}$$
$$\frac{\partial^2}{\partial\lambda^2}\ln f\left(x\left|\lambda\right.\right)=-\frac{x}{\lambda^2}$$

Alors,

$$E\left(-1+\frac{X}{\lambda}\right)=-1+\frac{\lambda}{\lambda}=0$$
$$I\left(\lambda\right)=VAR\left(-1+\frac{X}{\lambda}\right)=\frac{1}{\lambda^2}VAR\left(X\right)=\frac{1}{\lambda}$$
$$I\left(\lambda\right)=-E\left(-\frac{X}{\lambda^2}\right)=\frac{\lambda}{\lambda^2}=\frac{1}{\lambda}$$

Exemple 58 *Soit $X\backsim N\left(\theta,\sigma^2\right)$. On a*

$$\ln f\left(x\left|\theta\right.\right)=-\ln\left(\sigma\sqrt{2\pi}\right)-\frac{1}{2\sigma^2}\left(x-\theta\right)^2$$
$$\Rightarrow\frac{\partial}{\partial\theta}\ln f\left(x\left|\theta\right.\right)=\frac{x-\theta}{\sigma^2}$$

et

$$\frac{\partial^2}{\partial\theta^2}\ln f\left(x\left|\theta\right.\right)=\frac{-1}{\sigma^2}$$

On a

$$I\left(\theta\right)\overset{def}{=}E\left[\left(\frac{\partial}{\partial\theta}\ln f\left(X\left|\theta\right.\right)\right)^2\right]=E\left(\frac{\left(X-\theta\right)^2}{\sigma^4}\right)$$
$$=\frac{1}{\sigma^4}E\left(\left(X-\theta\right)^2\right)=\frac{\sigma^2}{\sigma^4}=\frac{1}{\sigma^2}$$

et

$$E\left[\frac{\partial}{\partial\theta}\ln f\left(X\left|\theta\right.\right)\right]=E\left(\frac{X-\theta}{\sigma^2}\right)=0$$

Donc

$$I\left(\theta\right)=VAR\left(\frac{X-\theta}{\sigma^2}\right)=\frac{1}{\sigma^2}$$

et

$$I\left(\theta\right)=-E\left(-\frac{1}{\sigma^2}\right)=\frac{1}{\sigma^2}$$

Exemple 59 *Soit $X \backsim Bernoulli(\theta)$. On a*

$$f(x|\theta) = \theta^x (1-\theta)^{1-x}, \qquad x = 0, 1$$
$$\ln f(x|\theta) = x \ln \theta + (1-x) \ln (1-\theta)$$

Ce qui entraine

$$\left(\frac{\partial}{\partial \theta} \ln f(x|\theta) \right)^2 = \left(\frac{x}{\theta} - \frac{1-x}{1-\theta} \right)^2 = \begin{cases} \frac{1}{(1-\theta)^2} & si \quad x = 0 \\ \\ \frac{1}{\theta^2} & si \quad x = 1 \end{cases}$$

et alors

$$I(\theta) \overset{def}{=} E\left[\left(\frac{\partial}{\partial \theta} \ln f(X|\theta) \right)^2 \right] = \frac{(1-\theta)}{(1-\theta)^2} + \frac{\theta}{\theta^2}$$
$$= \frac{1}{1-\theta} + \frac{1}{\theta} = \frac{1}{\theta(1-\theta)}$$

On a aussi

$$\frac{\partial^2}{\partial \theta^2} \ln f(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

et

$$E\left[\frac{\partial}{\partial \theta} \ln f(X|\theta) \right] = E\left(\frac{X}{\theta} - \frac{1-X}{1-\theta} \right)$$
$$= \frac{\theta}{\theta} - \frac{1-\theta}{1-\theta} = 0$$

$$\Rightarrow I(\theta) = VAR\left(\frac{\partial}{\partial \theta} \ln f(X|\theta) \right)$$
$$= VAR\left(\frac{X}{\theta} - \frac{1-X}{1-\theta} \right)$$
$$= VAR\left(\frac{X-\theta}{\theta(1-\theta)} \right)$$
$$= \frac{1}{\theta^2 (1-\theta)^2} VAR(X)$$
$$= \frac{\theta(1-\theta)}{\theta^2 (1-\theta)^2}$$
$$= \frac{1}{\theta(1-\theta)}$$

Aussi :

$$I(\theta) = -E\left(\frac{\partial^2}{\partial \theta^2} \ln f(X|\theta) \right)$$
$$= -E\left(-\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2} \right)$$
$$= -\left(-\frac{\theta}{\theta^2} - \frac{1-\theta}{(1-\theta)^2} \right) = \frac{1}{\theta} + \frac{1}{1-\theta}$$
$$= \frac{1}{\theta(1-\theta)}$$

Exemple 60 *Soit* $X \backsim$ *Bernoulli* $\left(p=\theta^2\right)$. *On a*

$$f\left(x\left|\theta\right.\right)=\left(\theta^2\right)^x\left(1-\theta^2\right)^{1-x},\quad x=0,1$$

Le calcul (omis) donne

$$\left(\frac{\partial}{\partial\theta}\ln f\left(x\left|\theta\right.\right)\right)^2=\left(\frac{2x}{\theta}-2\theta\frac{1-x}{1-\theta^2}\right)^2=\begin{cases}\frac{4\theta^2}{(1-\theta^2)^2} & si\quad x=0\\ \frac{4}{\theta^2} & si\quad x=1\end{cases}$$

Alors

$$\begin{aligned}I\left(\theta\right)&\overset{def}{=}E\left[\left(\frac{\partial}{\partial\theta}\ln f\left(X\left|\theta\right.\right)\right)^2\right]\\&=E\left(\frac{2X}{\theta}-2\theta\frac{1-X}{1-\theta^2}\right)^2\\&\overset{calculs}{=}\frac{4}{1-\theta^2}\end{aligned}$$

Comparons avec l'exemple précédent où

$$I\left(\theta\right)=\frac{1}{\theta\left(1-\theta\right)}$$

On a

$$\arg\min_{\theta\in]0;1[}\frac{1}{\theta\left(1-\theta\right)}=\frac{1}{2}$$

Si $\theta\approxeq 0$ *(succès peu probable) ou si* $\theta\approxeq 1$ *(succès très probable), alors*

$$\frac{1}{\theta\left(1-\theta\right)}\approxeq\infty$$

(information moyenne à faible). Maintenant, pour

$$I\left(\theta\right)=\frac{4}{1-\theta^2},$$

si

$$\begin{aligned}\theta&\approxeq 0,\\I\left(\theta\right)&\approxeq 4\end{aligned}$$

et si

$$\begin{aligned}\theta&\approxeq 1,\\I\left(\theta\right)&\uparrow\infty\end{aligned}$$

La symétrie de l'exemple précédent est perdue.

Théorème 5 *Sous certaines conditions de régularité, l'EMV* $\widehat{\theta}$ *de* θ *vérifie*

$$\boxed{\sqrt{n}\left(\widehat{\theta}-\theta\right)\overset{Loi}{\rightarrow}N\left(0,I^{-1}\left(\theta\right)\right)}\qquad quand\ n\rightarrow\infty$$

et $\widehat{\theta}$ *est convergent* $\left(\lim_{n\rightarrow\infty}P\left(\left|\widehat{\theta}-\theta\right|>\varepsilon\right)=0\right)$.

Remarque 9 *Pour n suffisamment grand, on a l'approximation*

$$\widehat{\theta} \backsimeq N\left(\theta, \frac{1}{nI(\theta)}\right)$$
$$VAR\left(\widehat{\theta}\right) \approxeq \frac{1}{nI(\theta)}$$

On peut vérifier qu'alors

$$E\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = 0$$

Corollaire 2 *Si $I(\theta)$ est continue, alors (Slutsky)*

$$\boxed{\sqrt{nI\left(\widehat{\theta}\right)}\left(\widehat{\theta}-\theta\right) \overset{Loi}{\to} N(0,1)} \qquad \text{quand } n \to \infty$$

Exemple 61 *Soit $X \backsim Exp(\lambda)$. On a $f(x|\lambda) = \lambda e^{-\lambda x}$ $(x \geq 0)$, $E(X) = 1/\lambda$, $VAR(X) = 1/\lambda^2$, $\ln f(x|\lambda) = \ln\lambda - \lambda x$, $\frac{\partial}{\partial\lambda}\ln f(x|\lambda) = \frac{1}{\lambda} - x$*

$$\Rightarrow \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$$

$$I(\theta) = VAR\left(\frac{1}{\lambda} - X\right) = VAR(X) = \frac{1}{\lambda^2}$$

Alors

$$\sqrt{n}\left(\widehat{\lambda}-\lambda\right) \overset{Loi}{\to} N\left(0,\lambda^2\right) \qquad \text{quand } n \to \infty$$

Autre façon.

$$\sqrt{n}\left(\overline{X} - \frac{1}{\lambda}\right) \overset{Loi}{\to} N\left(0, \frac{1}{\lambda^2}\right) \qquad \text{quand } n \to \infty \text{ par TCL}$$

Si

$$g(x) = \frac{1}{x},$$

alors

$$g'(x) = \frac{-1}{x^2} \Rightarrow g'\left(\frac{1}{\lambda}\right) = -\lambda^2$$

et (TCL-delta) :

$$\sqrt{n}\left(\widehat{\lambda}-\lambda\right) \overset{Loi}{\to} N\left(0, \frac{1}{\lambda^2}\lambda^4\right) = N\left(0,\lambda^2\right) \qquad \text{quand } n \to \infty$$

Exemple 62 *Soit $X \backsim Poi(\lambda)$. On a vu (plus haut) que $I(\lambda) = 1/\lambda$. On conclut que*

$$\sqrt{n}\left(\widehat{\lambda}-\lambda\right) \overset{Loi}{\to} N(0,\lambda) \qquad \text{quand } n \to \infty$$

Exemple 63 *Soit $Y = (X_1, X_2, X_3) \backsim mult\left(n, (1-\theta)^2, 2\theta(1-\theta), \theta^2\right)$. On a vu que*

$$\widehat{\theta}_{EVM} = \frac{X_2 + 2X_3}{2n}$$

et que

$$(X_1, X_2, X_3) = \sum_{i=1}^{n} Y_i$$

où les Y_i $(i = 1, 2, 3)$ sont iid de loi

Y_i	e_1	e_2	e_3
P	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2

On a (en posant $y = (y_1, y_2, y_3)$)

$$\ln f(y|\theta) = \begin{cases} 2\ln(1-\theta) & si \quad y = e_1 \\ \ln 2 + \ln\theta + \ln(1-\theta) & si \quad y = e_2 \\ 2\ln\theta & si \quad y = e_3 \end{cases}$$

Donc

$$\frac{\partial}{\partial\theta}\ln f(y|\theta) = \begin{cases} -2/(1-\theta) & si \quad y = e_1 \\ 1/\theta - 1/(1-\theta) & si \quad y = e_2 \\ 2/\theta & si \quad y = e_3 \end{cases}$$

$$\begin{aligned} I(\theta) &\overset{def}{=} E\left[\left(\frac{\partial}{\partial\theta}\ln f(Y|\theta)\right)^2\right] \\ &= \left(\frac{2}{1-\theta}\right)^2(1-\theta)^2 + \left(\frac{1}{\theta} - \frac{1}{1-\theta}\right)^2(2\theta(1-\theta)) + \left(\frac{2}{\theta}\right)^2\theta^2 \\ &\overset{Calculs}{=} \frac{2}{\theta(1-\theta)} \end{aligned}$$

On a donc

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \overset{Loi}{\to} N\left(0, I^{-1}(\theta) = \frac{\theta(1-\theta)}{2}\right) \qquad quand\ n \to \infty$$

$$VAR\left(\widehat{\theta}\right) \approxeq \frac{2}{\theta(1-\theta)}\ (= valeur\ exacte)$$

$$E\left(\frac{\partial}{\partial\theta}\ln f(Y|\theta)\right) = -\frac{2}{1-\theta}(1-\theta)^2 + \left(\frac{1}{\theta} - \frac{1}{1-\theta}\right)(2\theta(1-\theta)) + \frac{2}{\theta}\theta^2$$
$$\overset{Calculs}{=} 0$$

$$\frac{\partial^2}{\partial\theta^2}\ln f(y|\theta) = \begin{cases} -2/(1-\theta)^2 & si \quad y = e_1 \\ -1/\theta^2 - 1/(1-\theta)^2 & si \quad y = e_2 \\ -2/\theta^2 & si \quad y = e_3 \end{cases}$$

$$\begin{aligned} I(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(Y|\theta)\right] \\ &= \frac{2}{(1-\theta)^2}(1-\theta)^2 + \left(\frac{1}{\theta^2} + \frac{1}{(1-\theta)^2}\right)(2\theta(1-\theta)) + \frac{2}{\theta^2}\theta^2 \\ &\overset{Calculs}{=} \frac{2}{\theta(1-\theta)} \end{aligned}$$

("Une fonction score est toujours centrée en 0 ").

7 Efficacité

Soit donné deux estimateurs $\widetilde{\theta}$ et $\widehat{\theta}$ d'un paramètre θ. On définit l'efficacité de $\widehat{\theta}$ relativement à $\widetilde{\theta}$ par

$$\boxed{eff\left(\widehat{\theta},\widetilde{\theta}\right)=\frac{VAR\left(\widetilde{\theta}\right)}{VAR\left(\widehat{\theta}\right)}}$$

Si

$$eff\left(\widehat{\theta},\widetilde{\theta}\right)<1$$

alors $\widetilde{\theta}$ a une plus petite variance que $\widehat{\theta}$. Cette mesure est utile pour des estimateurs dont le **biais est négligeable par rapport à la variance** (ou deux estimateurs dont le biais est comparable).

Note. Une définition similaire utilise l'EQM au lieu de la variance.

Dans le cas où on dispose seulement de résultats asymptotiques du type

$$\sqrt{n}\left(\widetilde{\theta}-\theta\right)\overset{Loi}{\rightarrow}N\left(0,c^2\right)\qquad\text{quand } n\rightarrow\infty$$
$$\sqrt{n}\left(\widehat{\theta}-\theta\right)\overset{Loi}{\rightarrow}N\left(0,d^2\right)\qquad\text{quand } n\rightarrow\infty$$

on définit l'**efficacité relative asymptotique** par

$$\boxed{eff\left(\widehat{\theta},\widetilde{\theta}\right)=\frac{c^2}{d^2}}$$

Remarque 10 *Quel est le rapport $\frac{\widehat{n}}{\widetilde{n}}$ pour que $\widehat{\theta}_{\widehat{n}}$ et $\widetilde{\theta}_{\widetilde{n}}$ aient la même variance ? On a*

$$VAR\left(\widehat{\theta}_{\widehat{n}}\right)=\frac{d^2}{\widehat{n}}\qquad et\qquad VAR\left(\widetilde{\theta}_{\widetilde{n}}\right)=\frac{c^2}{\widetilde{n}}$$
$$\Rightarrow\frac{d^2}{\widehat{n}}=\frac{c^2}{\widetilde{n}}\Leftrightarrow\boxed{\frac{\widehat{n}}{\widetilde{n}}=\frac{d^2}{c^2}}$$

Ainsi, l'efficacité relative asymptotique représente le ratio des tailles échantillonnales nécéssaires pour que les deux estimateurs aient la même variance.

Exemple 64 *Soit X = cosinus de l'angle suivi par un électron lors de la désintégration d'un muon. Le modèle proposé est*

$$f\left(x\left|\alpha\right.\right)=\frac{1+\alpha x}{2},\qquad pour\ x\in[-1;1]\ \ et\ \alpha\in[-1;1]$$

La log-vraisemblance est donnée par

$$l\left(\alpha\right)=\sum_{i=1}^{n}\ln\frac{1+\alpha x_i}{2}=\sum_{i=1}^{n}\ln\left(1+\alpha x_i\right)-n\ln 2$$

On a

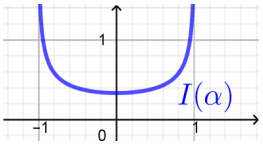
$$\frac{\partial l}{\partial\alpha}=\sum_{i=1}^{n}\frac{x_i}{1+\alpha x_i}=0$$

Il n'est pas possible de trouver une solution fermée pout déterminer $\widehat{\alpha}$. Il faut une solution numérique (itérative, avec une valeur initiale calculée avec la méthode des moments : cf cours d'analyse numérique). La fonction score est

$$\frac{\partial}{\partial\alpha}\ln f\left(x\left|\alpha\right.\right)=\frac{x}{1+\alpha x}$$

et

$$I(\alpha) \overset{def}{=} E\left[\left(\frac{X}{1+\alpha X}\right)^2\right] = \int_{-1}^{1}\left(\frac{x}{1+\alpha x}\right)^2 \frac{1+\alpha x}{2}dx$$
$$= \frac{1}{2}\int_{-1}^{1}\frac{x^2}{1+\alpha x}dx$$
$$\overset{calcul}{=} \begin{cases} \frac{1}{2}\dfrac{\ln\frac{1+\alpha}{1-\alpha}-2\alpha}{\alpha^3} & si \quad \alpha \neq 0 \\ \\ \frac{1}{3} & si \quad \alpha = 0 \end{cases}$$



Noter que

$$\lim_{\alpha\to 0}\frac{1}{2}\frac{\ln\frac{1+\alpha}{1-\alpha}-2\alpha}{\alpha^3} = \frac{1}{3}$$

(règle de l'Hospital). On a un minimum absolu en $\alpha = 0$ *(l'intégrale se calcule par un changement de variable). On a aussi*

$$E(X) = \int_{-1}^{1} x\frac{1+\alpha x}{2}dx = \frac{1}{2}\int_{-1}^{1} xdx + \frac{\alpha}{2}\int_{-1}^{1} x^2dx = \frac{\alpha}{3}$$

et

$$E\left(X^2\right) = \int_{-1}^{1} x^2\frac{1+\alpha x}{2}dx = \frac{1}{2}\int_{-1}^{1} x^2dx + \frac{\alpha}{2}\underbrace{\int_{-1}^{1} x^3dx}_{=0} = \frac{1}{3} = E_{\alpha=0}\left(X^2\right)$$

(ne dépend pas de α*). Donc*

$$\sqrt{n}\left(\widehat{\alpha}-\alpha\right) \overset{Loi}{\to} N\left(0, I^{-1}(\alpha)\right) \qquad quand\ n\to\infty \qquad (*)$$

Considérons l'estimateur par la méthode des moments (MM) :

$$\mu_1 = E(X) = \frac{\alpha}{3} \Rightarrow \widetilde{\mu}_1 = \overline{X} = \frac{\widetilde{\alpha}}{3} \Rightarrow \boxed{\widetilde{\alpha} = 3\overline{X}}$$

On a également

$$VAR(X) = \frac{1}{3} - \left(\frac{\alpha}{3}\right)^2 = \boxed{\frac{3-\alpha^2}{9}}$$

Par TCL, on a

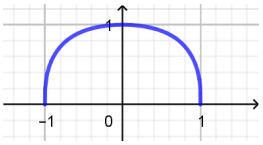
$$\sqrt{n}\left(\overline{X}-\frac{\alpha}{3}\right) \overset{Loi}{\to} N\left(0, \frac{3-\alpha^2}{9}\right) \qquad quand\ n\to\infty$$

Donc

$$\sqrt{n}\left(\widetilde{\alpha}-\alpha\right) \overset{Loi}{\to} N\left(0, 3-\alpha^2\right) \qquad quand\ n\to\infty$$

et

$$eff\left(\widetilde{\alpha},\widehat{\alpha}\right) = \frac{VAR\left(\widehat{\alpha}\right)}{VAR\left(\widetilde{\alpha}\right)} = \frac{I^{-1}(\alpha)}{3-\alpha^2} = \begin{cases} \dfrac{2\alpha^3}{(3-\alpha^2)\left(\ln\frac{1+\alpha}{1-\alpha}-2\alpha\right)} & si \quad \alpha \neq 0 \\ \\ 1 & si \quad \alpha = 0 \end{cases}$$



On a les quelques valeurs numériques suivantes :

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
$eff\left(\widetilde{\alpha},\widehat{\alpha}\right)$	1	0.997	0.989	0.975	0.9523	0.922	0.878	0.817	0.727	0.582	0.464	0.290

Quand α s'approche de 1, l'estimateur $\widetilde{\alpha}$ devient moins bon relativement à $\widehat{\alpha}$. L'estimateur $\widetilde{\alpha}$ avec 100 observations, ou bien l'estimateur $\widehat{\alpha}$ avec 46 observations ont à peu près la même efficacité lorsque $\alpha = 0.95$ ($0.464 \approxeq 46\%$). C'est uniquement dans le cas $\alpha = 0$ que les deux estimateurs sont de même efficacité.

En conclusion, on préfèrera utiliser $\widehat{\alpha}$ car on ne connaît pas la vraie valeur de α.

8 Intervalles de confiance

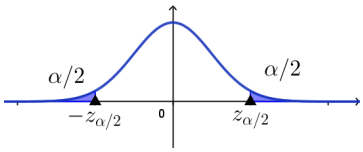
Soit $X_1, \cdots, X_n$ iid de loi P_θ (dépend du paramètre θ). But : construire un intervalle aléatoire dont la probabilité de contenir la vraie valeur de θ vaut $1-\alpha$ (valeur fixée).

8.1 Moyenne et variance d'une loi normale

8.1.1 Écart type connu

$X_1, \cdots, X_n$ iid $N\left(\mu, \sigma_0^2\right)$, avec σ_0 connu, μ inconnu. On sait que $\overline{X} \backsim N\left(\mu, \sigma_0^2/n\right)$ et que $Z = \sqrt{n}\frac{\overline{X}-\mu}{\sigma_0} \backsim N(0,1)$ (loi connue). On a

$$\begin{aligned} 1-\alpha &= \Phi\left(z_{\alpha/2}\right) - \Phi\left(-z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\overline{X}-\mu}{\sigma_0} \leq z_{\alpha/2}\right) \\ &= P\left(\overline{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \end{aligned}$$



L'intervalle de confiance pour μ de niveau $1-\alpha$ est donc

$$\boxed{IC_{1-\alpha}(\mu) = \underbrace{\overline{X}}_{\text{aléatoire}} \pm \underbrace{z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\text{fixe: } ME}}$$

(ME = marge d'erreur).

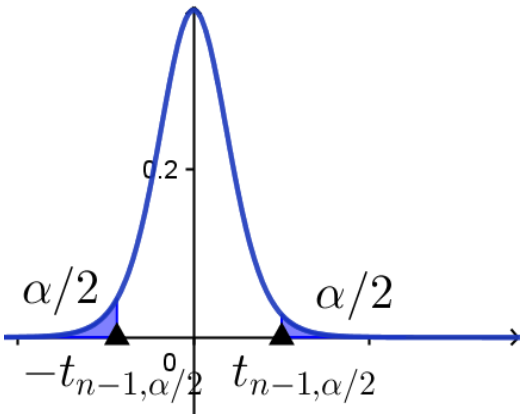
8.1.2 Écart type inconnu

Moyenne $X_1, \cdots, X_n$ iid $N\left(\mu, \sigma^2\right)$, avec σ inconnu, μ inconnu. On sait que $Z = \sqrt{n}\frac{\overline{X}-\mu}{S} \backsim t_{n-1}$ (loi connue) avec (variance empirique)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

On a

$$1-\alpha = P\left(-t_{n-1,\alpha/2} \leq \sqrt{n}\frac{\overline{X}-\mu}{S} \leq t_{n-1,\alpha/2}\right)$$



Ce qui donne l'$IC_{1-\alpha}$ pour μ :

$$\boxed{IC_{1-\alpha}(\mu) = \underbrace{\overline{X}}_{aléatoire} \pm \frac{t_{n-1,\alpha/2}}{\sqrt{n}} \underbrace{S}_{aléatoire}}$$

Variance Cherchons maintenant l'$IC_{1-\alpha}$ pour σ^2. On a

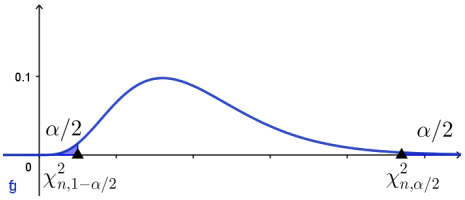
$$n\frac{\widehat{\sigma}^2}{\sigma^2} \backsim \chi^2_{n-1}$$

avec $\widehat{\sigma}^2$ = estimateur EMV :

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

On a

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{n-1,1-\alpha/2} \leq n\frac{\widehat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) \\ &= P\left(n\frac{\widehat{\sigma}^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq n\frac{\widehat{\sigma}^2}{\chi^2_{n-1,1-\alpha/2}}\right) \end{aligned}$$



et l'$IC_{1-\alpha}$ est

$$IC_{1-\alpha}\left(\sigma^2\right) = \boxed{\left]n\frac{\widehat{\sigma}^2}{\chi^2_{n-1,\alpha/2}}; n\frac{\widehat{\sigma}^2}{\chi^2_{n-1,1-\alpha/2}}\right[}$$

Cet intervalle n'est pas symétrique par rapport à σ^2 (contrairement aux deux précédents).

Remarque 11 *Si $\mu = \mu_0$ est connu, on utilise $\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sigma^2} \backsim \chi^2_n$ (noter le n) et l'$IC_{1-\alpha}$ est*

$$IC_{1-\alpha}\left(\sigma^2\right) = \left]\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{n,\alpha/2}}; \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{n,1-\alpha/2}}\right[$$

8.2 Proportion

Soit $X_1, \cdots, X_n$ iid $Bernoulli\,(\theta)$ $(0 < \theta < 1)$. On a $Y = \sum_{i=1}^{n} X_i \backsim Bin\,(n, \theta)$,

$$\begin{aligned} f\,(x\,|\theta\,) &= \theta^x\,(1-\theta)^{1-x}\,, \qquad x = 0, 1 \\ \ln f\,(x\,|\theta\,) &= x \ln\theta + (1-x)\ln(1-\theta) \\ \frac{\partial}{\partial\theta}\ln f\,(x\,|\theta\,) &= \frac{x}{\theta} - \frac{1-x}{1-\theta} \\ \frac{\partial^2}{\partial\theta^2}\ln f\,(x\,|\theta\,) &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \end{aligned}$$

Et donc

$$\begin{aligned} I\,(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2}\ln f\,(X\,|\theta\,)\right] = -E\left[-\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2}\right] \\ &= \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta\,(1-\theta)} \end{aligned}$$

Maintenant,

$$\begin{aligned} \frac{\partial}{\partial\theta} l\,(\theta\,|x_1, \cdots, x_n\,) &= \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1-x_i) \\ &= \frac{y}{\theta} - \frac{n-y}{1-\theta} = 0 \\ &\Rightarrow \boxed{\widehat{\theta} = \frac{Y}{n}} \end{aligned}$$

(proportion de succès). Donc

$$\boxed{IC_{1-\alpha}\,(\theta) \approxeq \widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{n}}}$$

C'est l'IC pour une proportion.

Application numérique. $n = 100, \widehat{\theta} = 0.45,\ 1-\alpha = 0.95 = 19/20$

$$IC_{0.95}\,(\theta) \approxeq 0.45 \pm 1.96\sqrt{\frac{0.45\,(1-0.45)}{100}} = [0.35; 0.55]$$

ICI, la marge d'erreur est

$$ME = 1.96\sqrt{\frac{0.45\,(1-0.45)}{100}} \approxeq 0.098$$

Quelle devrait être la valeur de n pour obtenir une $ME \leq 3\%$ de niveau $19/20$? On a : 95% des intervalles contiennent la vraie valeur de θ. Dans la pratique, on ne fait qu'une seule expérience.

$$1.96\sqrt{\frac{\theta\,(1-\theta)}{n}} = 0.03 \Rightarrow n = 1057$$

8.3 Autres exemples

Exemple 65 *$X_1, \cdots, X_n$ iid $\sim Exp(\lambda)$. On a $f(x|\lambda) = \lambda e^{-\lambda x}$ $(x > 0)$, $E(X) = 1/\lambda$, $\widehat{\lambda} = 1/\overline{X}$. On sait aussi que*

$$\begin{aligned} Exp(\lambda) &= Gamma(\alpha = 1, \lambda) \Rightarrow \sum_{i=1}^{n} X_i \backsim Gamma(\alpha = n, \lambda) \\ &= Gamma(\alpha = n, 1) \times \frac{1}{\lambda} \qquad (car\ \lambda = \ paramètre\ d'echelle) \\ &= Gamma\left(\alpha = \frac{2n}{2}, \frac{1}{2}\right) \times \frac{1}{2\lambda} \\ &= \frac{1}{2\lambda}\chi^2_{2n} \end{aligned}$$

Donc

$$2\lambda \sum_{i=1}^{n} X_i \backsim \chi^2_{2n}$$

et

$$\begin{aligned} 1 - \alpha = P\left(\chi^2_{2n,1-\alpha/2} \leq 2\lambda n\overline{X}. \leq \chi^2_{2n,\alpha/2}\right) \\ \Rightarrow IC_{1-\alpha}(\lambda) = \left]\frac{\chi^2_{2n,1-\alpha/2}}{2n\overline{X}}; \frac{\chi^2_{2n,\alpha/2}}{2n\overline{X}}\right[\end{aligned}$$

En général, on ne connait pas la distribution de $\widehat{\theta}$. Il faut passer par les méthodes asymptotiques (grands échantillons). On part du fait que

$$\sqrt{nI(\theta)}\left(\widehat{\theta} - \theta\right) \overset{Loi}{\rightarrow} N(0,1)$$

puis (Slutsky)

$$\sqrt{nI\left(\widehat{\theta}\right)}\left(\widehat{\theta} - \theta\right) \overset{Loi}{\rightarrow} N(0,1)$$

Finalement,

$$\boxed{IC_{1-\alpha}(\theta) : \widehat{\theta} \pm z_{\alpha/2}\frac{1}{\sqrt{nI\left(\widehat{\theta}\right)}}}$$

dont la probabilité tend vers $1 - \alpha$ quand $n \to \infty$.

Exemple 66 *Soit $X_1, \cdots, X_n$ iid $Exp(\lambda)$. On a $f(x|\lambda) = \lambda e^{-\lambda x}$ $(x > 0)$, $E(X) = 1/\lambda$, $\widehat{\lambda}_{EVM} = 1/\overline{X}$ et $I(\lambda) = 1/\lambda^2$. Alors*

$$IC_{1-\alpha}(\lambda) \approxeq \frac{1}{\overline{X}} \pm z_{\alpha/2}\frac{1}{\sqrt{n\overline{X}}}$$

Exemple 67 *Soit $X_1, \cdots, X_n$ iid $Poi(\lambda)$. On a $f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$ $(x = 0, 1, \cdots)$, $E(X) = \lambda$, $\widehat{\lambda}_{EVM} = \overline{X}$ et $I(\lambda) = 1/\lambda$. Alors*

$$IC_{1-\alpha}(\lambda) \approxeq \overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}}{n}}$$

Exemple 68 *$Y = (X_1, X_2, X_3) \backsim mult\left(n = 1029, (1-\theta)^2, 2\theta(1-\theta), \theta^2\right)$, $0 < \theta < 1$, $\widehat{\theta}_{EVM} = \frac{X_2 + 2X_3}{2n}$, $I(\theta) = \frac{2}{\theta(1-\theta)}$*

$$IC_{1-\alpha}(\theta) \approxeq \widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\theta}\left(1 - \widehat{\theta}\right)}{2n}}$$

Application numérique.

Génotype	AA	Aa	aa
Fréquence	342	500	187

$n = 1029$, $\widehat{\theta} = 0.4247$, $1 - \alpha = 0.95$, $z_{\alpha/2} = 1.96$. *Alors*

$$\begin{aligned} IC_{1-\alpha}(\theta) &\approxeq 0.4247 \pm 1.96\sqrt{\frac{0.4247(1-0.4247)}{2(1029)}} \\ &= [0.40; 0.45] \end{aligned}$$

8.4 Transformation stabilisant la variance

Nous traiterons ceci à travers un exemple. Soit $X \backsim Poi(\theta)$, $\theta > 0$, $\widehat{\theta} = \overline{X}$. On a

$$\begin{aligned} \sqrt{n}\left(\overline{X} - \theta\right) &\overset{Loi}{\to} N(0, \theta) \\ \frac{\sqrt{n}}{\sqrt{\overline{X}}}\left(\overline{X} - \theta\right) &\overset{Loi}{\to} N(0, 1) \end{aligned}$$

Ce qui donne

$$IC_{1-\alpha}(\theta) \approxeq \overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}}{n}}$$

Supposons qu'il existe une fonction g telle que $\theta g'(\theta)^2 \approxeq 1$ (stabilisation de la variance pour la loi de Poisson), alors (Cramer-delta)

$$\sqrt{n}\left(g\left(\overline{X}\right) - g(\theta)\right) \overset{Loi}{\to} N\left(0, \theta g'(\theta)^2\right) = N(0, 1)$$

On a

$$\theta g'(\theta)^2 \approxeq 1 \Rightarrow g'(\theta) \approxeq \theta^{-1/2} \Rightarrow g(\theta) \approxeq 2\sqrt{\theta}$$

Donc

$$\sqrt{n}2\left(\sqrt{\overline{X}} - \sqrt{\theta}\right) \overset{Loi}{\to} N(0, 1)$$

Et on a

$$\begin{aligned} IC_{1-\alpha}\left(\sqrt{\theta}\right) &\approxeq \sqrt{\overline{X}} \pm z_{\alpha/2}\frac{1}{2\sqrt{n}} \\ &= \left[\max\left(0, \sqrt{\overline{X}} - z_{\alpha/2}\frac{1}{2\sqrt{n}}\right), \sqrt{\overline{X}} + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right] \end{aligned}$$

et

$$IC_{1-\alpha}(\theta) = \left[\max\left(0, \sqrt{\overline{X}} - z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)^2, \left(\sqrt{\overline{X}} + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)^2\right]$$

9 Autres résultats théoriques

9.1 Borne de Cramer-Rao

Théorème 6 *Borne de Cramer-Rao. Soit donné un échantillon aléatoire iid $X_1, \cdots, X_n$ de loi $f(x|\theta)$. Soit $T = t(X_1, \cdots, X_n)$ un estimateur sans biais de $\tau(\theta)$ (une fonction donnée de θ). Alors, sous certaines conditions de régularité sur $f(x|\theta)$, on a*

$$\boxed{VAR(T) \geq \frac{(\tau'(\theta))^2}{nI(\theta)}} \qquad \forall n$$

($\tau'(\theta) = \frac{d\tau(\theta)}{d\theta}$). En particulier (pour $\tau(\theta) = \theta$)

$$\boxed{VAR(T) \geq \frac{1}{nI(\theta)}} \qquad \forall n$$

Preuve. *Omise.* ■

Exemple 69 *Soit $X \backsim Poi(\lambda = \theta)$, $\tau(\theta) = \theta$ et $T = \overline{X}$. On a $VAR\left(\overline{X}\right) = \frac{\theta}{n}$, $I(\theta) = \frac{1}{\theta}$ et*

$$\frac{1}{nI(\theta)} = \frac{\theta}{n} \Rightarrow VAR\left(\overline{X}\right) = \frac{1}{nI(\theta)}$$

La borne minimale est atteinte. On dit que l'estimateur $\overline{X}$ est $\boxed{\boldsymbol{UMVUE}}$ *(Uniformly Minimum Variance Unbiased Estimator).*

Définition 15 *Un estimateur sans biais qui atteint la borne de C-R est dit* ***efficace****.*

Remarque 12 *Attention ! Il se peut qu'il y ait un estimateur* ***biaisé*** *de variance plus petite.*

Exemple 70 *Soit $X \backsim N(\mu, 1)$. On a*

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2\right], \qquad x \in \mathbb{R}$$

et

$$\frac{\partial}{\partial\mu} \ln f(x|\mu) = x - \mu$$

$$I(\mu) = VAR\left[\frac{\partial}{\partial\mu} \ln f(x|\mu)\right] = VAR(X - \mu) = 1$$

EVM :

$$\frac{\partial}{\partial\mu} l(\mu) = \sum_{i=1}^{n} (x_i - \mu) = n\overline{x} - n\mu = 0$$

$$\Rightarrow \widehat{\mu} = \overline{X}$$

Puisque $E(\widehat{\mu}) = \mu$, cet estimateur est sans biais. On a

$$VAR(\widehat{\mu}) = VAR\left(\overline{X}\right) = \frac{1}{n} = \frac{1}{nI(\theta)}$$

La borne est atteinte. Donc $\overline{X}$ est UMVUE pour μ.

Exemple 71 *Soit $X \backsim U]0;\theta[$. On a vu que $\widehat{\theta}_{EMV} = X_{(n)} = \max(X_1, \cdots, X_n)$ et que*

$$F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^n \qquad et \qquad f_{X_{(n)}}(x) = n\left(\frac{x}{\theta}\right)^{n-1}\frac{1}{\theta} \qquad pour\ x \in]0;\theta[$$

On a

$$E\left(X_{(n)}\right) = \int_0^\theta \left(1 - \left(\frac{x}{\theta}\right)^n\right) dx = \frac{n}{n+1}\theta \qquad (biaisé)$$

$$\Rightarrow \widehat{\theta} = \frac{n+1}{n} X_{(n)} \qquad (sans\ biais)$$

On a

$$E\left(X_{(n)}^2\right) = \int_0^\theta x^2 \cdot n\left(\frac{x}{\theta}\right)^{n-1} \frac{1}{\theta} dx = \frac{n}{n+2}\theta^2$$
$$\Rightarrow E\left(\widehat{\theta^2}\right) = \left(\frac{n+1}{n}\right)^2 \frac{n}{n+2}\theta^2 = \theta^2 \frac{(n+1)^2}{n(n+2)}$$

Ce qui donne

$$VAR\left(\widehat{\theta}\right) = \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \frac{\theta^2}{n(n+2)}$$

Existe-t-il un autre estimateur sans biais de θ dont la variance est inférieure à celle de θ ? La réponse à cette question utilise la notion de ***famille exhaustive et complète*** *(STT3700).*

9.2 Familles exponentielles

Une famille flexible et très importante de lois de probabilité comprenant de nombreux modèles de probabilité courants est la famille exponentielle.

Définition 16 *Soit $X \sim f(x|\theta)$. On dit que X est membre de la famille exponentielle si $f(x|\theta)$ peut s'écrire sous la forme*

$$f(x|\theta) = c(\theta)\, h(x) \exp\left(\sum_{i=1}^{k} q_i(\theta)\, t_i(x)\right)$$

et si son support ne dépend pas de θ. Si, de plus, Θ contient un rectangle ouvert, la famille est dite ***régulière****.*

Exemple 72 *$X \sim Bin(n, \theta)$, $\theta \in \Theta = [0; 1]$. On a*

$$f(x|\theta) = \binom{n}{x}\theta^x (1-\theta)^{n-x} = \theta^n \binom{n}{x} \exp\left(\left(\ln\frac{\theta}{1-\theta}\right)x\right)$$

Donc $f(x|\theta)$ appartient à la famille exponentielle. De plus, elle est régulière de toute évidence.

L'intérêt des familles exponentielles est fourni dans le résultat suivant.

Théorème 7 *Soit $X \sim f(x|\theta)$ appartenant à la famille exponentielle et soit $X_1, \ldots, X_n iid \sim f(x|\theta)$. Alors*

$$T = \left(\sum_{i=1}^{k} t_1(X_i), \ldots, \sum_{i=1}^{k} t_k(X_i)\right)$$

est une statistique exhaustive pour θ.

9.2.1 Exercice

Exercice 22 *Montrer que les lois exponentielle et normale appartiennent à la famille exponentielle.*