Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

09 - Modèle Probabiliste, Paramètre et Inférence

été 2022

Statistique et Paramètre

Statistique

Étant donné un échantillon $\{x_1, \cdots, x_n\}$ une statistique est une fonction des observations, $t(\boldsymbol{x}) = t(x_1, \cdots, x_n)$..

Par exemple $t(x_1, \cdots, x_n) = x_5$, $\displaystyle\frac{1}{n}\sum_{i=1}^{n} x_i$ ou $\max\{x_i\} - \min\{x_i\}$.

Paramètre

Un paramètre est un nombre qui décrit la distribution de X. C'est un nombre fixe, et souvent inconnu.

Par exemple p pour une loi $\mathcal{B}(p)$ ou $\displaystyle\frac{p}{1-p}$

Modèle paramétrique

Formalisation du problème : nous supposons disposer de $Y_1, \cdots, Y_n$ copies indépendantes d'une variable aléatoire Y dont la densité est paramétré par un paramétre réel ($\theta \in \Theta \subset \mathbb{R}$) ou vectoriel ($\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$).

Modèle paramétrique

On dispose d'un échantillon $\{x_1, \cdots, x_n\}$, correspondant à des réalisation de variables aléatoires $X_1, \cdots, X_n$ indépendantes et de même loi $F_\theta \in \mathcal{F}$ où $\mathcal{F}$ est la famille de lois données, et où θ est inconnu.

Exemples :

- Loi de Bernoulli $Y \sim \mathcal{B}(p)$, $\theta = p \in (0, 1)$,
- Loi de Poisson $Y \sim \mathcal{P}(\lambda)$, $\theta = \lambda \in \mathbb{R}_+$,
- Loi normale $\mathcal{N}(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$

Modèle paramétrique identifiable ★★★

On notera que la paramétrisation de la famille $\mathcal{F}$ n'est pas unique

Exemples :

- Loi de Bernoulli $Y \sim \mathcal{B}(p)$, $\theta = p \in (0,1)$ ou $\theta = \dfrac{p}{1-p} \in \mathbb{R}_+$
- Loi de Poisson $Y \sim \mathcal{P}(\lambda)$, $\theta = \lambda \in \mathbb{R}_+$, ou $\theta = \log \lambda \in \mathbb{R}$

Identifiabilité

$$\theta_1 \neq \theta_2 \Longrightarrow F_{\theta_1} \neq F_{\theta_2} \text{ ou } F_{\theta_1} = F_{\theta_2} \Longrightarrow \theta_1 = \theta_2.$$

Modèle paramétrique identifiable ★★★

Example: Le modèle Gaussien, sur $\mathbb{R}$

$$\mathcal{F} = \left\{ f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right);\ \theta = (\mu, \sigma^2) \right\}.$$

où $\mu \in \mathbb{R}$ et $\sigma > 0$. Alors

$$f_{\theta_1} = f_{\theta_2}$$

$$\Longleftrightarrow \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 \right) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2 \right)$$

$$\Longleftrightarrow \frac{1}{\sigma_1^2}(x-\mu_1)^2 + \log\sigma_1 = \frac{1}{\sigma_2^2}(x-\mu_2)^2 + \log\sigma_2$$

$$\Longleftrightarrow x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} + \log\sigma_1 - \log\sigma_2 \right) = 0$$

Modèle paramétrique identifiable ★★★

Example: Le modèle mélange d'exponentielles, sur $\mathbb{R}_+$

$$\mathcal{F} = \left\{ f_\theta(x) = \alpha\lambda_1 e^{-\lambda_1 x} + (1-\alpha)\lambda_2 e^{-\lambda_2 x};\ \theta = (\alpha, \lambda_1, \lambda_2)\right\}.$$

où $\alpha \in (0,1)$ et $\lambda_1, \lambda_2 > 0$.

Soient $\theta_1 = (\alpha, \lambda_1, \lambda_2)$ et $\theta_2 = (1-\alpha, \lambda_2, \lambda_1)$,

$$\theta_1 \neq \theta_2 \text{ mais } f_{\theta_1} = f_{\theta_2}$$

Ce modèle n'est alors pas identifiable...

Modèle paramétrique

Étant donné un modèle paramétrique,

- θ est le paramètre (en général inconnu) de la loi F_θ
- Θ est l'espace des paramètres
- $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ est un échantillon aléatoire de n copies indépendantes de loi f_θ
- $\boldsymbol{y} = (y_1, \cdots, y_n)$ les valeurs observées de $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$
- n la taille de l'échantillon

Estimateur - estimation

Un estimateur d'un paramètre θ est une variable aléatoire (fonction de l'échantillon $\boldsymbol{Y}$) et est noté $\widehat{\theta}(\boldsymbol{Y})$.
La valeur estimée de $\widehat{\theta}(\boldsymbol{Y})$ s'appelle aussi estimation et est notée $\widehat{\theta}(\boldsymbol{y})$.

(dans de nombreux ouvrages, $\widehat{\theta}$ désigne aussi bien $\widehat{\theta}(\boldsymbol{y})$ que $\widehat{\theta}(\boldsymbol{Y})$)

Modèle paramétrique

L'estimateur est une variable aléatoire $\widehat{\theta}(\boldsymbol{Y})$ et l'estimation est une constante $\widehat{\theta}(\boldsymbol{y})$

Example : observations suivant une loi $\mathcal{N}(\theta, 1)$

$$\widehat{\theta}_1(\boldsymbol{Y}) = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ et } \widehat{\theta}_1(\boldsymbol{y}) = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\widehat{\theta}_2(\boldsymbol{Y}) = \frac{\min\{Y_i\} + \max\{Y_i\}}{2} \text{ et } \widehat{\theta}_2(\boldsymbol{y}) = \frac{\min\{y_i\} + \max\{y_i\}}{2}$$

Biais

Biais d'un estimateur

On appelle biais d'un estimateur $\widehat{\theta}$ de θ la quantité

$$\text{bias}[\widehat{\theta}(\boldsymbol{Y})] = \mathbb{E}[\widehat{\theta}(\boldsymbol{Y})] - \theta$$

Estimateur sans biais

$\widehat{\theta}(\boldsymbol{Y})$ est un estimateur sans biais de θ si $\text{bias}[\widehat{\theta}(\boldsymbol{Y})] = 0$

Biais ★★★

Comme $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$, le biais est souvent une fonction de n.

Si $\text{bias}[\widehat{\theta}(\boldsymbol{Y})] \neq 0$, il arrive souvent que le biais soit petit quand n devient grand

Estimateur asymptotiqment sans biais

$\widehat{\theta}(\boldsymbol{Y})$ est un estimateur asymptotiquement sans biais de θ si

$$\lim_{n\to\infty} \text{bias}[\widehat{\theta}(\boldsymbol{Y})] = 0$$

Biais

Example $Y_1, \cdots, Y_n$ de moyenne μ,

- $\widehat{\mu}(\boldsymbol{Y}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i$ est un estimateur sans biais de μ
- $\widetilde{\mu}(\boldsymbol{Y}) = \dfrac{1}{n+3}\displaystyle\sum_{i=1}^{n} Y_i$ est un estimateur asymptotiquement sans biais de μ

Example $Y_1, \cdots, Y_n$ de variance σ^2,

- $\widehat{\sigma}^2(\boldsymbol{Y}) = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ estimateur sans biais de σ^2
- $\widetilde{\sigma}^2(\boldsymbol{Y}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ est un estimateur asymptotiquement sans biais de σ^2

Biais ★★★

Example $Y_1, \cdots, Y_n$, de loi F. Soit $x \in \mathbb{R}$,

$$\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbf{1}(Y_i \leq y)}_{X_i}$$

où les variables X_i sont des variables de Bernoulli $\mathcal{B}(p)$ où $p = F(y)$.

$$\mathbb{E}[\widehat{F}(y)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}[Y_i \leq y] = F(y)$$

donc $\widehat{F}(y)$ est un estimateur sans biais de $F(y)$, pour tout y.

Moyenne Quadratique (MSE)

Erreur Quadratique Moyenne

On appelle erreur quadratique moyenne d'un estimateur $\widehat{\theta}(\boldsymbol{Y})$ et on note $EQM(\widehat{\theta}(\boldsymbol{Y}))$ la quantité

$$EQM(\widehat{\theta}(\boldsymbol{Y})) = \mathbb{E}\Big[\big(\widehat{\theta}(\boldsymbol{Y}) - \theta\big)^2\Big]$$

Erreur Quadratique Moyenne

$$EQM(\widehat{\theta}(\boldsymbol{Y})) = \text{bias}\big(\widehat{\theta}(\boldsymbol{Y})\big)^2 + \text{Var}\big(\widehat{\theta}(\boldsymbol{Y})\big)$$

Consistance

Un estimateur $\widehat{\theta}(\boldsymbol{Y})$ est consistant si $\lim_{n\rightarrow\infty} EQM(\widehat{\theta}(\boldsymbol{Y})) = 0$

Moyenne Quadratique (MSE)

Pour un estimateur sans biais

$$EQM(\widehat{\theta}(\boldsymbol{Y})) = \text{Var}\big(\widehat{\theta}(\boldsymbol{Y})\big)$$

Un estimateur asymptotiquement sans biais est consistant si

$$\lim_{n\to\infty} \text{Var}\big(\widehat{\theta}(\boldsymbol{Y})\big) = 0$$

Efficacité

Soient $\widehat{\theta}_1(\boldsymbol{Y})$ et $\widehat{\theta}_2(\boldsymbol{Y})$ deux estimateurs de θ. $\widehat{\theta}_1(\boldsymbol{Y})$ est plus efficace que $\widehat{\theta}_2(\boldsymbol{Y})$ si $EQM(\widehat{\theta}_1(\boldsymbol{Y})) < EQM(\widehat{\theta}_2(\boldsymbol{Y}))$.

$$eff(\widehat{\theta}_1(\boldsymbol{Y}), \widehat{\theta}_2(\boldsymbol{Y})) = \frac{EQM(\widehat{\theta}_2(\boldsymbol{Y}))}{EQM(\widehat{\theta}_1(\boldsymbol{Y}))} = \text{rapport d'efficacité}$$

Moyenne Quadratique (MSE)

Example: $Y_1, \cdots, Y_n$ de moyenne μ (et de variance σ^2)

$$\widehat{\mu}_1(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ et } \widehat{\mu}_2(\boldsymbol{Y}) = \frac{2}{n}\sum_{i=1}^{n/2} Y_i$$

Comme

$$\begin{cases} \mathbb{E}[\widehat{\mu}_1(\boldsymbol{Y})] = \mu \text{ donc bias}\big(\widehat{\mu}_1(\boldsymbol{Y})\big) = 0 \text{ et Var}\big(\widehat{\mu}_1(\boldsymbol{Y})\big) = \dfrac{\sigma^2}{n} \\ \mathbb{E}[\widehat{\mu}_2(\boldsymbol{Y})] = \mu \text{ donc bias}\big(\widehat{\mu}_2(\boldsymbol{Y})\big) = 0 \text{ et Var}\big(\widehat{\mu}_2(\boldsymbol{Y})\big) = \dfrac{\sigma^2}{2n} \end{cases}$$

soit $EQM(\widehat{\theta}_1(\boldsymbol{Y})) = \dfrac{\sigma^2}{n}$ et $EQM(\widehat{\theta}_2(\boldsymbol{Y})) = \dfrac{\sigma^2}{2n}$

alors $\mathit{eff}(\widehat{\mu}_1(\boldsymbol{Y}), \widehat{\mu}_2(\boldsymbol{Y})) = 2$, autrement dit, le premier estimateur est deux fois plus efficace que le second.

Moyenne Quadratique (MSE) ★★★

Example: $Y_1, \cdots, Y_n$ de moyenne μ,

$$\widehat{\mu}_1(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n Y_i = \overline{Y} \text{ et } \widehat{\mu}_\alpha(\boldsymbol{Y}) = \frac{\alpha}{n}\sum_{i=1}^n Y_i = \alpha\overline{Y},\ \alpha \in [0,1]$$

$$\begin{cases}\mathbb{E}[\widehat{\mu}_1(\boldsymbol{Y})] = \mu \text{ donc bias}\big(\widehat{\mu}_1(\boldsymbol{Y})\big) = 0 \text{ et Var}\big(\widehat{\mu}_1(\boldsymbol{Y})\big) = \dfrac{\sigma^2}{n}\\ \mathbb{E}[\widehat{\mu}_\alpha(\boldsymbol{Y})] = \alpha\mu\ \text{ bias}\big(\widehat{\mu}_\alpha(\boldsymbol{Y})\big) = (\alpha-1)\mu \text{ et Var}\big(\widehat{\mu}_\alpha(\boldsymbol{Y})\big) = \dfrac{\alpha^2\sigma^2}{n}\end{cases}$$

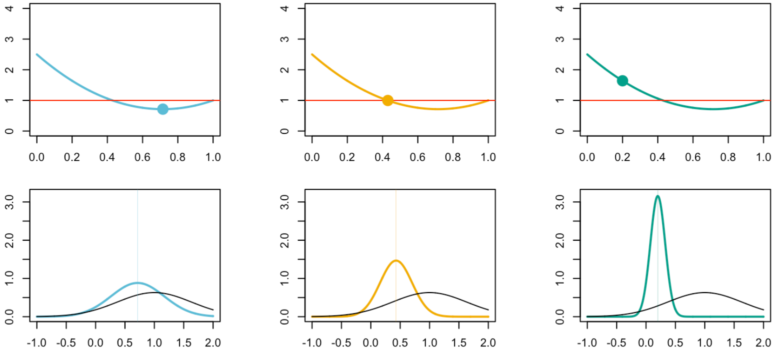
soit $EQM(\widehat{\theta}_1(\boldsymbol{Y})) = \dfrac{\sigma^2}{n}$ et $EQM(\widehat{\theta}_2(\boldsymbol{Y})) = (\alpha-1)^2\mu^2 + \dfrac{\alpha^2\sigma^2}{2n}$, et

$$eff(\widehat{\mu}_1(\boldsymbol{Y}), \widehat{\mu}_\alpha(\boldsymbol{Y})) = \frac{(\alpha-1)^2\mu^2 n + \alpha^2\sigma^2}{\sigma^2} = \alpha^2 + (\alpha-1)^2 n \cdot cv^2$$

en notant $cv = \mu/\sigma^2$.

Moyenne Quadratique (MSE) ★★★

Il existe des α tels que $eff(\widehat{\mu}_1(\boldsymbol{Y}), \widehat{\mu}_\alpha(\boldsymbol{Y})) < 1$.



$EQM(\widehat{\theta}_\alpha(\boldsymbol{Y})) = \text{bias}\big(\widehat{\theta}_\alpha(\boldsymbol{Y})\big)^2 + \text{Var}\big(\widehat{\theta}_\alpha(\boldsymbol{Y})\big)$, et on observe que

- $\text{bias}\big(\widehat{\theta}_\alpha(\boldsymbol{Y})\big)$ augmente quand α diminue
- $\text{Var}\big(\widehat{\theta}_\alpha(\boldsymbol{Y})\big)$ diminue quand α diminue

Moyenne Quadratique (MSE)

Example: $Y_1, \cdots, Y_n$ des variables $\mathcal{B}(p)$. Soit $S_n = Y_1 + \cdots + Y_n$. $S_n \sim \mathcal{B}(n, p)$ donc $\mathbb{E}[S_n] = np$ et $\text{Var}[S_n] = np(1-p)$.

$$\widehat{p}_1 = \frac{S_n}{n} \text{ et } \widehat{p}_2 = \frac{S_n + 1}{n+2}$$

$$\mathbb{E}[\widehat{p}_1] = p \text{ et } \text{Var}(\widehat{p}_1) = \frac{p(1-p)}{n}$$

Comme c'est un estimateur sans biais de p, $EQM(\widehat{p}_1) = \dfrac{p(1-p)}{n}$

$$\mathbb{E}[\widehat{p}_2] = \frac{np+1}{n+2} \text{ et } \text{Var}(\widehat{p}_2) = \frac{\text{Var}\,(S_n)}{(n+2)^2} = \frac{np(1-p)}{(n+2)^2}$$

$$EQM(\widehat{p}_2) = \left[\frac{np+1}{n+2} - p\right]^2 + \frac{np(1-p)}{(n+2)^2} = \frac{(1-2p)^2 + np(1-p)}{(n+2)^2}$$

Moyenne Quadratique (MSE)

Aussi, le rapport d'efficacité vaut

$$eff(\widehat{p}_1, \widehat{p}_2) = \frac{EQM(\widehat{p}_2)}{EQM(\widehat{p}_1)} = \frac{n}{(n+2)^2}\left[n + \frac{(1-2p)^2}{p(1-p)}\right]$$

Si $p \sim 1/2$, ce rapport vaut $n^2/(n+2)^2 < 1$.

En fait $\widehat{p}_2$ domine $\widehat{p}_1$ si

$$p \in \left(\frac{1}{2} - \sqrt{\frac{n+1}{2n+1}}, \frac{1}{2} + \sqrt{\frac{n+1}{2n+1}}\right)$$

Dans le partie 12, on verra qu'on peut être amené à utiliser

$$\widehat{p}_1 = \frac{S_n}{n} \text{ et } \widehat{p}_3 = \frac{S_n + 1}{n+2}$$

Maximum de Vraisemblance

Vraisemblance (Likelihood)

Soit $\boldsymbol{Y} = (y_1, \cdots, y_n)$ un échantillon i.i.d. de variables de loi f_θ. La fonction de vraisemblance est

$$\mathcal{L}(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_\theta(y_i)$$

L'estimation du maximum de vraisemblance est

$$\widehat{\theta}(\boldsymbol{y}) = \underset{\theta\in\Theta}{\text{argmin}}\big\{\mathcal{L}(\theta|\boldsymbol{y})\big\} \text{ et } \widehat{\theta}(\boldsymbol{Y}) = \underset{\theta\in\Theta}{\text{argmin}}\big\{\mathcal{L}(\theta|\boldsymbol{Y})\big\}$$

Maximum de Vraisemblance

Log-Vraisemblance

Soit $\boldsymbol{Y} = (y_1, \cdots, y_n)$ un échantillon i.i.d. de variables de loi f_θ. La fonction de log-vraisemblance est

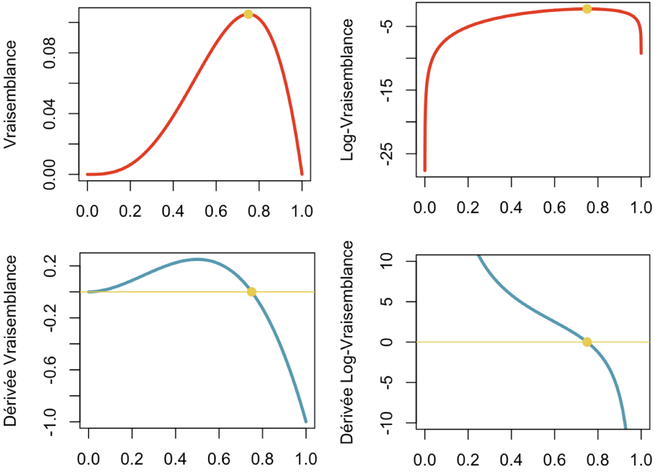
$$\log \mathcal{L}(\theta|\boldsymbol{y}) = \sum_{i=1}^{n} \log f_\theta(y_i)$$

L'estimation du maximum de vraisemblance est

$$\widehat{\theta}(\boldsymbol{y}) = \underset{\theta\in\Theta}{\text{argmin}}\Big\{\log \mathcal{L}(\theta|\boldsymbol{y})\Big\} \text{ et } \widehat{\theta}(\boldsymbol{Y}) = \underset{\theta\in\Theta}{\text{argmin}}\Big\{\log \mathcal{L}(\theta|\boldsymbol{Y})\Big\}$$

Vraisemblance, cas $\mathcal{B}(p)$

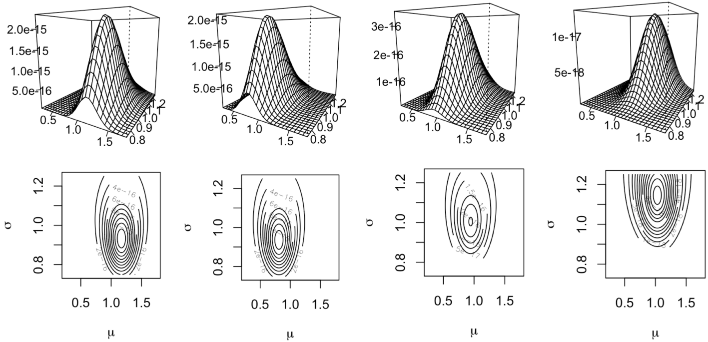
$$\boldsymbol{y} = \{0, 1, 1, 1\}, \; Y_i \sim \mathcal{B}(\theta).$$



Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

Pour les paramètres univariés, on peut visualiser la vraisemblance, mais c'est plus compliqué en dimension plus grande...

Vraisemblance $\mathcal{L}(\mu, \sigma^2)$ pour 4 échantillons $\boldsymbol{y}$



Maximum de Vraisemblance

Équation de Vraisemblance ou Condition du Premier Ordre

Soit $\boldsymbol{Y} = (y_1, \cdots, y_n)$ un échantillon i.i.d. de variables de loi f_θ. L'estimation du maximum de vraisemblance est

$$\widehat{\theta}(\boldsymbol{y}) = \underset{\theta\in\Theta}{\text{argmin}}\left\{\log\mathcal{L}(\theta|\boldsymbol{y})\right\}$$

et il vérifie (moyennant quelques hypothèses supplémentaires)

$$\left.\frac{\partial\log\mathcal{L}(\theta|\boldsymbol{y})}{\partial\theta}\right|_{\theta=\widehat{\theta}(\boldsymbol{y})} = 0$$

(résultat admis)

Vraisemblance, cas $\mathcal{B}(p)$

Example 1 : on a fait un sondage sur 15 personnes pour savoir s'ils appréciaient le cours de MAT4681, quelle est l'estimation par maximum de vraisemblance de la proportion de gens satisfaits ?

- ce que nous dit la théorie

$$\mathcal{L}(p;\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{s_n}(1-p)^{n-s_n},\ s_n = \sum_{i=1}^{n} x_i$$

$$\log \mathcal{L}(p;\boldsymbol{x}) = s_n \log(p) + (n-s_n)\log(1-p)$$

$$\frac{\partial}{\partial p}\log \mathcal{L}(p;\boldsymbol{x}) = \frac{\partial}{\partial p} s_n x \log(p) + (n-s_n)\log(1-p) = \frac{s_n}{p} - \frac{n-s_n}{1-p}$$

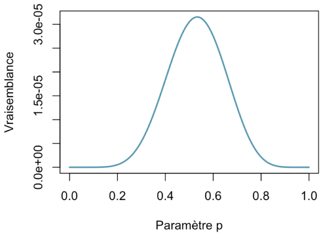
$$\left.\frac{\partial}{\partial p}\log \mathcal{L}(p;\boldsymbol{x})\right|_{p=\widehat{p}} = 0 \text{ si et seulement si } \frac{s_n}{\widehat{p}} = \frac{n-s_n}{1-\widehat{p}},\ \text{soit } \widehat{p} = \frac{s_n}{n} = \overline{x}$$

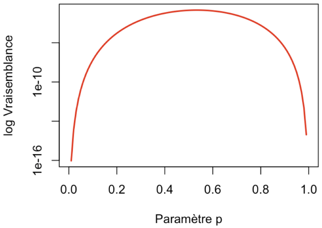
Vraisemblance, cas $\mathcal{B}(p)$

- ce que nous dit la pratique

Traçons la fonction de (log)vraisemblance $p \mapsto \mathcal{L}(p; \boldsymbol{x})$

```
> n=15
> x = c(1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0)
> vraisemblance = function(p) prod(dbinom(x,size = 1,
    prob = p))
> vect_p = seq(0,1,by=0.01)
> plot(vect_p,Vectorize(vraisemblance)(vect_p))
```





Vraisemblance, cas $\mathcal{B}(p)$

- ce que nous dit la pratique

On peut chercher le maximum de la fonction $p \mapsto \mathcal{L}(p; \boldsymbol{x})$

```
> optim(par = .5,fn = function(z) -vraisemblance(z))
$par
[1] 0.5333252

$value
[1] -3.155276e-05
```

La théorie nous avait dit que ce maximum a une forme particulière, $\widehat{p} = \overline{x}$

```
> mean(x)
[1] 0.5333333
```

Vraisemblance, cas $\mathcal{B}(p)$

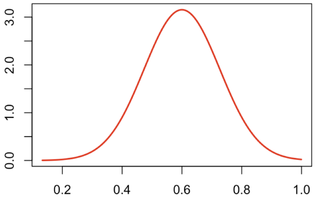
- ce que nous disent les mathématiques

Comme $\widehat{p}(\boldsymbol{x}) = \overline{x}$, on peut utiliser la loi des grands nombres,

$$Z_n = \sqrt{n}\frac{\widehat{p}(\boldsymbol{X}) - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

mais ici $n = 15$ (approximation Gaussienne peut être mauvaise)
Si $p = 60\%$ la distribution (approchée) de $\widehat{p}(\boldsymbol{X})$ serait

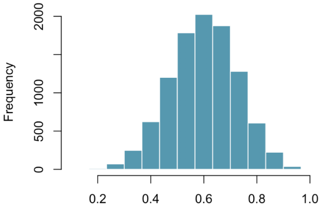
```
> u=seq(2/15,1,by=.001)
> plot(u,dnorm(u,.6,sqrt(.4*.6/15)))
```



Vraisemblance, cas $\mathcal{B}(p)$

- ce que nous disent les simulations, si on suppose $\theta = 60\%$

```
> theta=rep(NA,1e4)
> for(s in 1:1e4){
+     x=sample(0:1,size = n,prob = c(.4,.6),replace=
   TRUE)
+     neglogL = function(p) -sum(log(dbinom(x,size =
   1,prob = p)))
+     theta[s] = optim(par = .5,fn = neglogL)$par
+ }
> hist(theta)
```



Vraisemblance, cas $\mathcal{N}(\mu,\sigma^2)$

Soit $\{x_1,\cdots,x_n\}$ la taille (en cm) de 112 élèves de sexe féminin

```
> x = Davis$height[Davis$sex == "F"]
```

Supposons que les x_i sont des réalisations de variables indépendantes $X_i \sim \mathcal{N}(\mu,\sigma^2)$.

$$\log \boldsymbol{\theta} = \sum_{i=1}^{n} \log f(x_i;\mu = \boldsymbol{\theta}_1,\sigma^2 = \boldsymbol{\theta}^2)$$

```
> logLik = function(t) -sum(log(dnorm(x,mean = t[1],sd
    = t[2])))
> (opt <- optim(par = c(150,5),logLik))
$par
[1] 164.713474   5.632331

$value
[1] 352.5451
```

Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

En fait, on peut montrer que

$$\widehat{\boldsymbol{\theta}}_1 = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x} \text{ et } \widehat{\boldsymbol{\theta}}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2} = \sqrt{\frac{n-1}{n}} \cdot \widehat{\sigma}$$

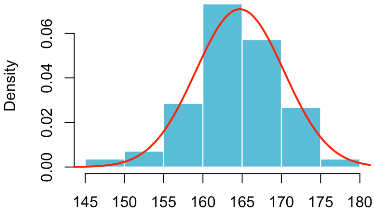
```
> mean(x)
[1] 164.7143
> sd(x)
[1] 5.659129
> sqrt((n-1)/n)*sd(x)
[1] 5.633808
```

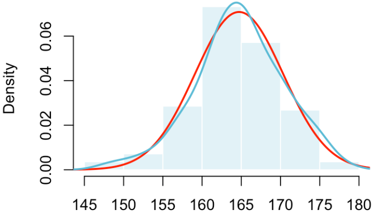
On peut aussi comparer la densité de la loi $\mathcal{N}(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2^2)$ avec

- l'histogramme de $\{x_1, \cdots, x_n\}$
- une estimation de la densité de $\{x_1, \cdots, x_n\}$

Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

```
> hist(x, probability=TRUE)
> plot(density(x))
> curve(dnorm(x,opt$par[1],opt$par[2]), from = min(x),
   to = max(x), col = "red")
```



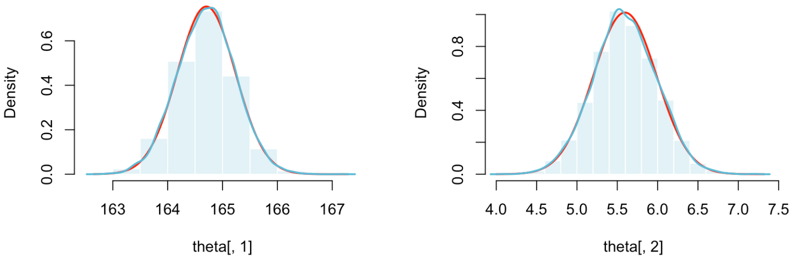


On peut aussi regarder la distribution de $\widehat{\boldsymbol{\theta}}_1$ et de $\widehat{\boldsymbol{\theta}}_2$, en faisant des simulations

Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

- par bootstrap (rééchantillonnage)

```
> theta = matrix(NA, 1000,2)
> for(i in 1:nrow(theta)){
+ xs = sample(x, size = n, replace = TRUE)
+ logLik = function(t) -sum(log(dnorm(xs, mean = t[1],
    sd = t[2])))
+ theta[i,] <- optim(par = c(150,5),logLik)$par
+ }
> hist(theta[,1])
```



Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

Les estimateurs $\widehat{\boldsymbol{\theta}}_1$ et $\widehat{\boldsymbol{\theta}}_2$ semblent avoir une distribution normale.

```
> mean(theta[,1])
[1] 164.7059
> sd(theta[,1])
[1] 0.5289618
> mean(theta[,1]) + c(-1.96,1.96)*sd(theta[,1])
[1] 163.6691 165.7427
> quantile(theta[,1],c(.025,.975))
    2.5%     97.5%
163.6716 165.7399
```

Aussi, $\mathbb{P}\big(\mu \in [163.7; 165,.7]\big) \sim 95\%$

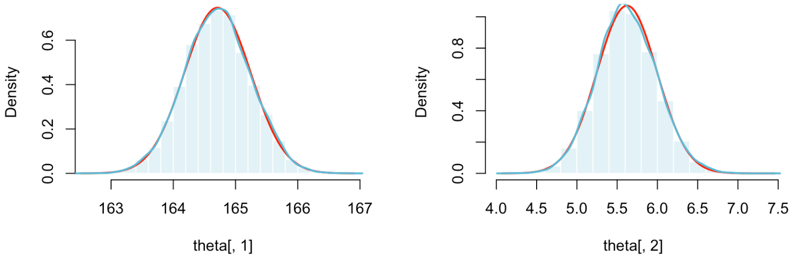
```
> mean(theta[,2]) + c(-1.96,1.96)*sd(theta[,2])
[1] 4.822267 6.367143
> quantile(theta[,2],c(.025,.975))
    2.5%     97.5%
4.818738 6.360761
```

Aussi, $\mathbb{P}\big(\sigma \in [4.82; 6.36]\big) \sim 95\%$

Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

- par simulation d'échantillons gaussienns

```
> theta = matrix(NA, 1000,2)
> for(i in 1:nrow(theta)){
+ xs = rnorm(n, mean(x),sd(x))
+ logLik = function(t) -sum(log(dnorm(xs, mean = t[1],
    sd = t[2])))
+ theta[i,] <- optim(par = c(150,5),logLik)$par
+ }
> hist(theta[,1])
```



Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

Les estimateurs $\widehat{\boldsymbol{\theta}}_1$ et $\widehat{\boldsymbol{\theta}}_2$ semblent avoir une distribution normale.

```
> mean(theta[,1])
[1] 164.7104
> sd(theta[,1])
[1] 0.5345402
> mean(theta[,1]) + c(-1.96,1.96)*sd(theta[,1])
[1] 163.6627 165.7581
> quantile(theta[,1],c(.025,.975))
    2.5%     97.5%
163.6500 165.7482
```

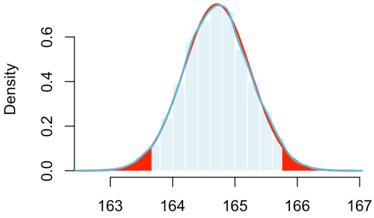
Aussi, $\mathbb{P}\big(\mu \in [163.6; 165, .7]\big) \sim 95\%$

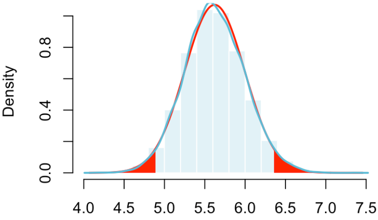
```
> mean(theta[,2]) + c(-1.96,1.96)*sd(theta[,2])
[1] 4.892208 6.352456
> quantile(theta[,2],c(.025,.975))
    2.5%     97.5%
4.904579 6.373500
```

Aussi, $\mathbb{P}\big(\sigma \in [4.90; 6.36]\big) \sim 95\%$

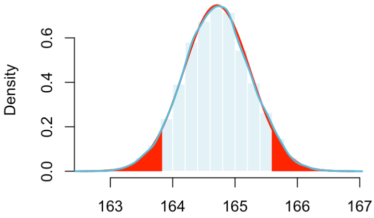
Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

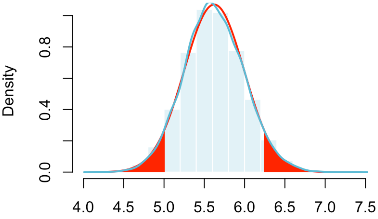
$\mathbb{P}(\mu \in [163.6; 165, .7]) \sim 95\%$ et $\mathbb{P}(\sigma \in [4.90; 6.36]) \sim 95\%$





$\mathbb{P}(\mu \in [164.0; 165, .4]) \sim 90\%$ et $\mathbb{P}(\sigma \in [5.14; 6.10]) \sim 90\%$





Vraisemblance, cas $\mathcal{N}(\mu, \sigma^2)$

En fait, on pourrait montrer que

$$\text{Var}[\widehat{\boldsymbol{\theta}}_1] = \frac{\sigma^2}{n} \text{ et } \text{Var}[\widehat{\boldsymbol{\theta}}_2] = \frac{\sigma^2}{2n}$$

```
> var(theta[,1])
[1] 0.2857332
> var(x)/n
[1] 0.2859441
> var(theta[,2])
[1] 0.1387653
> var(x)/(2*n)
[1] 0.1429721
```

$$\text{et } \text{Cov}[\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2] = 0$$

```
> cor(theta)
            [,1]        [,2]
[1,] 1.000000000 0.003202517
[2,] 0.003202517 1.000000000
```

Vraisemblance

Sous R, on peut utiliser la fonction `fitdistr` de `library(MASS)`,

```
> library(MASS)
> fitdistr(x,"normal")
      mean           sd
  164.7142857     5.6338083
 (   0.5323448) (   0.3764247)
```

on retrouve

```
> mean(x)
[1] 164.7143
> sd(x)
[1] 5.659129
> sd(x)*sqrt((length(x)-1)/length(x))
[1] 5.633808
```

On retrouve ici, numériquement

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \text{ et } \widehat{\theta}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{n-1}{n}} \cdot \widehat{\sigma}$$

Vraisemblance ★★★

Les valeurs entre parenthèses sont les écart-types des estimateurs

```
> fitdistr(x,"normal")
      mean           sd
  164.7142857     5.6338083
 (  0.5323448) (  0.3764247)
```

On peut noter que

```
> 5.6338083 /sqrt(length(x))
[1] 0.5323448
> 5.6338083 /sqrt(2*length(x))
[1] 0.3764247
```

car

$$\text{Var}(\widehat{\boldsymbol{\theta}}_1) = \frac{\sigma^2}{n} \text{ et } \text{Var}(\widehat{\boldsymbol{\theta}}_2) = \frac{\sigma^2}{2n}$$

pour un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, et $\boldsymbol{\theta} = (\mu, \sigma)$.

Méthode des Moments

Méthode des Moments

Soit $\boldsymbol{Y} = (y_1, \cdots, y_n)$ un échantillon i.i.d. de variables de loi f_θ. Soient $m_k(\theta) = \mathbb{E}[Y^k]$ où $Y \sim f_\theta$, et $\widehat{m}_k = \frac{1}{n}\sum_{i=1}^n y_i^k$ le moment empirique. Soit $\widehat{\theta} = (\widehat{\theta}_1, \cdots, \widehat{\theta}_d)$ la solution du système d'équations

$$\begin{cases} m_1(\widehat{\theta}) = \widehat{m}_1 \\ \quad\vdots \\ m_d(\widehat{\theta}) = \widehat{m}_d \end{cases}$$

Note: on peut parfois considérer les moments centrés (i.e. $\text{Var}[Y]$ au lieu de $\mathbb{E}[Y^2]$)

Méthode des moments, cas $\mathcal{B}(p)$

Example 2 : on a fait un sondage sur 15 personnes pour savoir s'ils appréciaient le cours de MAT4681, quelle est l'estimation par la méthode des moments de la proportion de gens satisfaits ?

- ce que nous dit la théorie

$$\mathbb{E}(X) = m_1(p) = p$$

or $\widehat{m}_1 = \overline{x}$ donc $\widehat{p}(x) = \overline{x}$

- ce que nous disent les mathématiques

Comme $\widehat{p}(\boldsymbol{x}) = \overline{x}$, on peut utiliser la loi des grands nombres,

$$Z_n = \sqrt{n}\frac{\widehat{p}(\boldsymbol{X}) - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

mais ici $n = 15$ (approximation Gaussienne peut être mauvaise)

Méthode des moments, cas $\mathcal{B}(n,p)$

Que se passe-t-il si $Y_i \sim \mathcal{B}(n,p)$, où n est aussi inconnu ?

$$\mathbb{E}[Y] = np \text{ et } \text{Var}[Y] = np(1-p)$$

On va alors résoudre

$$\begin{cases} \widehat{n}\widehat{p} = \overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i \\ \widehat{n}\widehat{p}(1-\widehat{p}) = s^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \big(y_i - \overline{y}\big)^2 \end{cases}$$

soit

$$\widehat{p} = \frac{\overline{y} - s^2}{\overline{y}} \text{ et } \widehat{n} = \frac{\overline{y}^2}{\overline{y} - s^2}$$

Note: il est possible d'avoir $\widehat{p} < 0$

Maximum de Vraisemblance vs Méthode des Moments

Example 3 : On observe des données modélisées par une loi de densité $\mapsto \theta y^{\theta-1}$ pour $y \in [0,1]$. Quels sont les estimateurs de θ?

```
> y = c(0.685, 0.754, 0.853, 0.973, 0.633, 0.97,
   0.984, 0.888, 0.876, 0.451, 0.637, 0.609, 0.898,
   0.761, 0.928, 0.819, 0.91, 0.998, 0.758, 0.931,
   0.981, 0.642, 0.885, 0.553, 0.686)
```

- méthode des moments

$$\mathbb{E}[Y] = \int_0^1 y \cdot \theta y^{\theta-1} dy = \theta \int_0^1 y^{\theta} dy = \theta \left[\frac{y^{\theta+1}}{\theta+1} \right]_0^1 = \frac{\theta}{\theta+1}$$

L'estimateur par la méthode des moments vérifie

$$\overline{y} = \frac{\widehat{\theta}}{\widehat{\theta}+1} \text{ soit } \widehat{\theta} = \frac{\overline{y}}{1-\overline{y}}$$

Maximum de Vraisemblance vs Méthode des Moments

- maximum de vraisemblance

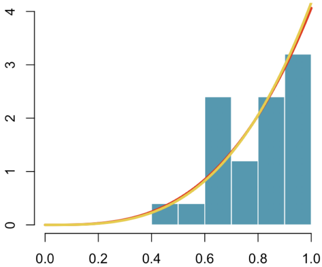
$$\log \mathcal{L}(\theta; \boldsymbol{y}) = n \log(\theta) + (\theta - 1) \sum_{i=1}^{n} \log(y_i)$$

$$\frac{\partial \log \mathcal{L}(\theta; \boldsymbol{y})}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \log(y_i)$$

$$\left.\frac{\partial \log \mathcal{L}(\theta; \boldsymbol{y})}{\partial \theta}\right|_{\theta = \widehat{\theta}} = 0 \text{ si } \frac{n}{\theta} = -\sum_{i=1}^{n} \log(y_i), \text{ i.e. } \widehat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log(y_i)}$$

```
> (a=mean(y)/(1-mean(y)))
[1] 4.062526
> (b=-25/sum(log(y)))
[1] 4.166513
```

Les deux densités sont très proches



Maximum de Vraisemblance vs Méthode des Moments

On peut aussi utiliser `fitdistr` pour l'estimateur du maximum de vraisemblance (en indiquant une valeur initiale pour l'algorithme)

```
> f = function(x, theta) theta*x^(theta-1)
> fitdistr(y, f, start = list(theta = 1))
     theta
  4.1672932
 (0.8334586)
```

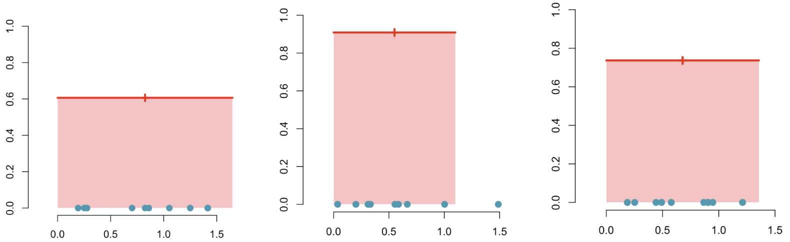
qui coïncide avec `b=-25/sum(log(y))`.

Méthode des Moments

Example: $\{y_1, \cdots, y_n\}$ de loi $\mathcal{U}([0,\theta])$, $\mathbb{E}[Y] = \theta/2$, alors

$$\overline{y} = \widehat{\theta}/2 \text{ i.e. } \widehat{\theta} = 2\overline{y}$$

Même si $y_i \leq \theta$ (par hypothèse), on peut avoir $\widehat{\theta} < y_j$



Note: estimateur du maximum de vraisemblance pour $\mathcal{U}([0,\theta])$?

Méthode des Moments

Théorème Central Limite

Soit (Y_n) une suite de variables aléatoires réelles indépendantes et de même loi admettant une espérance μ et une variance σ^2. La moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$ centrée converge vers une loi normale :

$$\sqrt{n}[\overline{X}_n - \mu] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

Méthode des Moments ★★★

L'estimateur de la méthode des moments sera approximativement Gaussien grace au théorème suivant

Delta-Method

Comme $\sqrt{n}[\overline{X}_n - \mu] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, alors

$$\sqrt{n}[g(\overline{X}_n) - g(\mu)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2[g'(\mu)]^2)$$

pour toute fonction g telle que $g'(\mu)$ existe et est non-nulle.

Note avoir la distribution d'un estimateur est important pour construire un intervalle de confiance (voir partie 10).