

Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

09 - Modèle Probabiliste, Paramètre et Inférence

été 2022

Statistique et Paramètre

Statistique

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

Par exemple $t(x_1, \dots, x_n) = x_5$, $\frac{1}{n} \sum_{i=1}^n x_i$ 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 $\frac{p}{1-p}$

Modèle paramétrique

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

Modèle paramétrique

On dispose d'un échantillon $\{x_1, \dots, x_n\}$, correspondant à des réalisations de variables aléatoires X_1, \dots, 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 $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$

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

Identifiabilité

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

Modèle paramétrique identifiable ★★★

Exemple: 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

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

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

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

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

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
- ▶ $\mathbf{Y} = (Y_1, \dots, Y_n)$ est un échantillon aléatoire de n copies indépendantes de loi f_θ
- ▶ $\mathbf{y} = (y_1, \dots, y_n)$ les valeurs observées de $\mathbf{Y} = (Y_1, \dots, Y_n)$
- ▶ n la taille de l'échantillon

Estimateur - estimation

Un estimateur d'un paramètre θ est une variable aléatoire (fonction de l'échantillon \mathbf{Y}) et est noté $\hat{\theta}(\mathbf{Y})$.

La valeur estimée de $\hat{\theta}(\mathbf{Y})$ s'appelle aussi estimation et est notée $\hat{\theta}(\mathbf{y})$.

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

Modèle paramétrique

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

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

$$\hat{\theta}_1(\mathbf{Y}) = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \text{ et } \hat{\theta}_1(\mathbf{y}) = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

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

Biais

Biais d'un estimateur

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

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

Estimateur sans biais

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

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

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

Estimateur asymptotiquement sans biais

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

$$\lim_{n \rightarrow \infty} \text{bias}[\hat{\theta}(\mathbf{Y})] = 0$$

Biais

Example Y_1, \dots, Y_n de moyenne μ ,

- ▶ $\hat{\mu}(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n Y_i$ est un estimateur sans biais de μ
- ▶ $\tilde{\mu}(\mathbf{Y}) = \frac{1}{n+3} \sum_{i=1}^n Y_i$ est un estimateur asymptotiquement sans biais de μ

Example Y_1, \dots, Y_n de variance σ^2 ,

- ▶ $\hat{\sigma}^2(\mathbf{Y}) = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ estimateur sans biais de σ^2
- ▶ $\tilde{\sigma}^2(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$ est un estimateur asymptotiquement sans biais de σ^2

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

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

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

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

donc $\hat{F}(y)$ 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 $\hat{\theta}(\mathbf{Y})$ et on note $EQM(\hat{\theta}(\mathbf{Y}))$ la quantité

$$EQM(\hat{\theta}(\mathbf{Y})) = \mathbb{E}\left[(\hat{\theta}(\mathbf{Y}) - \theta)^2\right]$$

Erreur Quadratique Moyenne

$$EQM(\hat{\theta}(\mathbf{Y})) = \text{bias}(\hat{\theta}(\mathbf{Y}))^2 + \text{Var}(\hat{\theta}(\mathbf{Y}))$$

Consistance

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

Moyenne Quadratique (MSE)

Pour un estimateur sans biais

$$EQM(\hat{\theta}(\mathbf{Y})) = \text{Var}(\hat{\theta}(\mathbf{Y}))$$

Un estimateur asymptotiquement sans biais est consistant si

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}(\mathbf{Y})) = 0$$

Efficacité

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

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

Moyenne Quadratique (MSE)

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

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

Comme

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

$$\text{soit } EQM(\hat{\theta}_1(\mathbf{Y})) = \frac{\sigma^2}{n} \text{ et } EQM(\hat{\theta}_2(\mathbf{Y})) = \frac{\sigma^2}{2n}$$

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

Moyenne Quadratique (MSE) ★★★

Example: Y_1, \dots, Y_n de moyenne μ ,

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

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

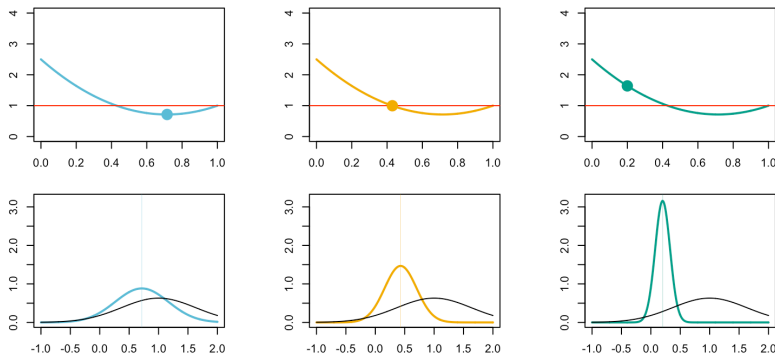
soit $EQM(\hat{\theta}_1(\mathbf{Y})) = \frac{\sigma^2}{n}$ et $EQM(\hat{\theta}_2(\mathbf{Y})) = (\alpha - 1)^2 \mu^2 + \frac{\alpha^2 \sigma^2}{2n}$, et

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

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

Moyenne Quadratique (MSE) ★★★

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



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

- ▶ $\text{bias}(\hat{\theta}_\alpha(\mathbf{Y}))$ augmente quand α diminue
- ▶ $\text{Var}(\hat{\theta}_\alpha(\mathbf{Y}))$ diminue quand α diminue

Moyenne Quadratique (MSE)

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

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

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

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

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

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

Moyenne Quadratique (MSE)

Aussi, le rapport d'efficacité vaut

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

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

En fait \hat{p}_2 domine \hat{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 la partie 12, on verra qu'on peut être amené à utiliser

$$\hat{p}_1 = \frac{S_n}{n} \text{ et } \hat{p}_3 = \frac{S_n + 2}{n + 4}$$

Maximum de Vraisemblance

Vraisemblance (Likelihood)

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

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

L'estimation du maximum de vraisemblance est

$$\hat{\theta}(\mathbf{y}) = \operatorname{argmin}_{\theta \in \Theta} \{\mathcal{L}(\theta|\mathbf{y})\} \text{ et } \hat{\theta}(\mathbf{Y}) = \operatorname{argmin}_{\theta \in \Theta} \{\mathcal{L}(\theta|\mathbf{Y})\}$$

Maximum de Vraisemblance

Log-Vraisemblance

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

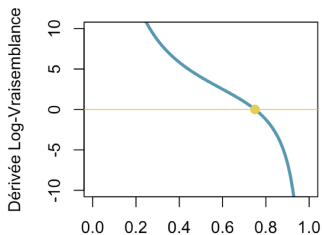
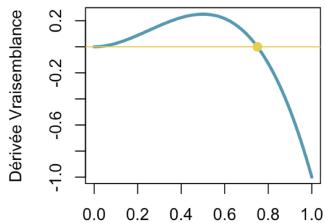
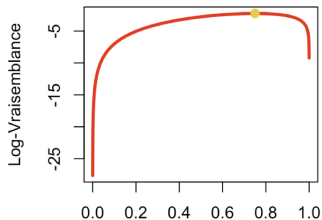
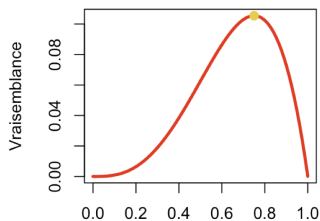
$$\log \mathcal{L}(\theta | \mathbf{y}) = \sum_{i=1}^n \log f_\theta(y_i)$$

L'estimation du maximum de vraisemblance est

$$\hat{\theta}(\mathbf{y}) = \operatorname{argmin}_{\theta \in \Theta} \{ \log \mathcal{L}(\theta | \mathbf{y}) \} \text{ et } \hat{\theta}(\mathbf{Y}) = \operatorname{argmin}_{\theta \in \Theta} \{ \log \mathcal{L}(\theta | \mathbf{Y}) \}$$

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

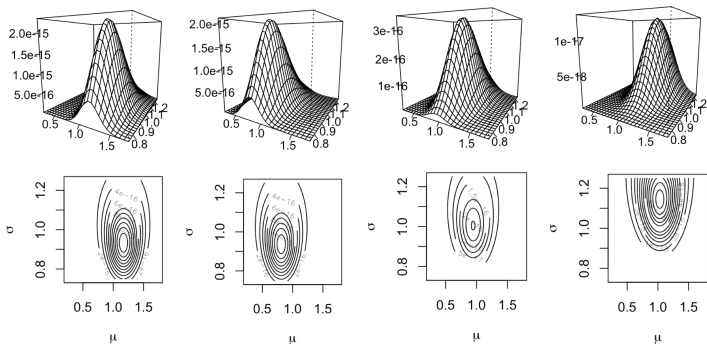
$$\mathbf{y} = \{0, 1, 1, 1\}, Y_i \sim \mathcal{B}(\theta).$$



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 \mathbf{y}



Maximum de Vraisemblance

Équation de Vraisemblance ou Condition du Premier Ordre

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

$$\hat{\theta}(\mathbf{y}) = \underset{\theta \in \Theta}{\operatorname{argmin}} \{ \log \mathcal{L}(\theta | \mathbf{y}) \}$$

et il vérifie (moyennant quelques hypothèses supplémentaires)

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

(résultat admis)

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

Exemple 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

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

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

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

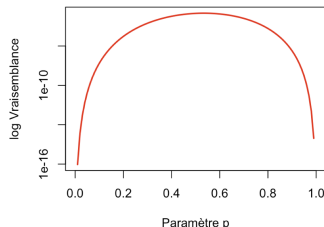
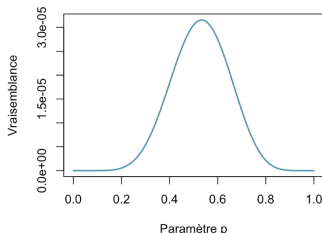
$$\left. \frac{\partial}{\partial p} \log \mathcal{L}(p; \mathbf{x}) \right|_{p=\hat{p}} = 0 \text{ si et seulement si } \frac{s_n}{\hat{p}} = \frac{n - s_n}{1 - \hat{p}}, \text{ soit } \hat{p} = \frac{s_n}{n} = \bar{x}$$

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

- ce que nous dit la pratique

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

```
1 > n=15
2 > x = c(1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0)
3 > vraisemblance = fonction(p) prod(dbinom(x,size = 1,
      prob = p))
4 > vect_p = seq(0,1,by=0.01)
5 > 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 $p \mapsto \mathcal{L}(p; \mathbf{x})$

```
1 > optim(par = .5, fn = function(z) -vraisemblance(z))
2 $par
3 [1] 0.5333252
4
5 $value
6 [1] -3.155276e-05
```

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

```
1 > mean(x)
2 [1] 0.5333333
```

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

- ce que nous disent les mathématiques

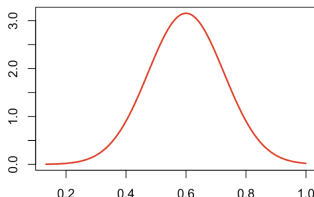
Comme $\hat{p}(\mathbf{x}) = \bar{x}$, on peut utiliser la loi des grands nombres,

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

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

Si $p = 60\%$ la distribution (approchée) de $\hat{p}(\mathbf{X})$ serait

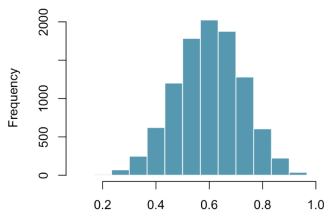
```
1 > u=seq(2/15,1,by=.001)
2 > plot(u,dnorm(u,.6,sqrt(.4*.6/15))
```



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

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

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



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

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

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

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

$$\log \theta = \sum_{i=1}^n \log f(x_i; \mu = \theta_1, \sigma^2 = \theta^2)$$

```
1 > logLik = function(t) -sum(log(dnorm(x, mean = t[1], sd
    = t[2])))
2 > (opt <- optim(par = c(150, 5), logLik))
3 $par
4 [1] 164.713474    5.632331
5
6 $value
7 [1] 352.5451
```

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

En fait, on peut montrer que

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

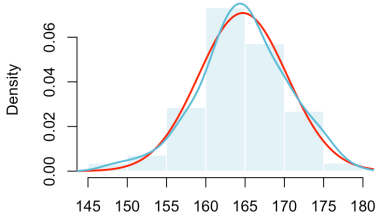
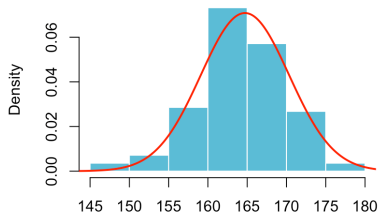
```
1 > mean(x)
2 [1] 164.7143
3 > sd(x)
4 [1] 5.659129
5 > sqrt((n-1)/n)*sd(x)
6 [1] 5.633808
```

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

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

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

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

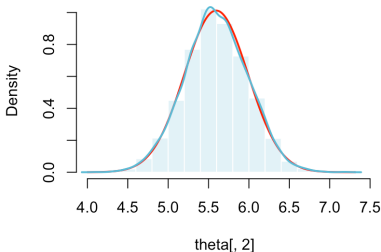
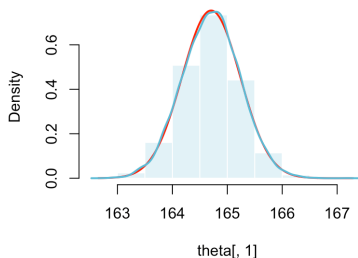


On peut aussi regarder la distribution de $\hat{\theta}_1$ et de $\hat{\theta}_2$, en faisant des simulations

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

► par bootstrap (rééchantillonnage)

```
1 > theta = matrix(NA, 1000, 2)
2 > for(i in 1:nrow(theta)){
3 +   xs = sample(x, size = n, replace = TRUE)
4 +   logLik = function(t) -sum(log(dnorm(xs, mean = t[1],
5 +   sd = t[2])))
6 +   theta[i,] <- optim(par = c(150, 5), logLik)$par
7 + }
7 > hist(theta[, 1])
```



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

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

```
1 > mean(theta[,1])
2 [1] 164.7059
3 > sd(theta[,1])
4 [1] 0.5289618
5 > mean(theta[,1]) + c(-1.96,1.96)*sd(theta[,1])
6 [1] 163.6691 165.7427
7 > quantile(theta[,1],c(.025,.975))
8      2.5%      97.5%
9 163.6716 165.7399
```

Aussi, $\mathbb{P}(\mu \in [163.7; 165.7]) \sim 95\%$

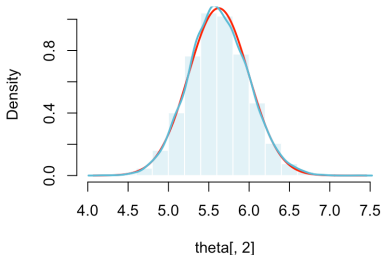
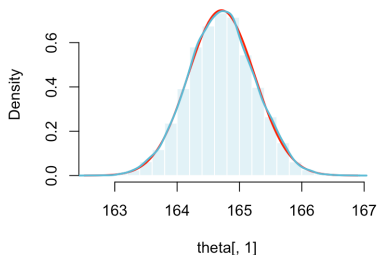
```
1 > mean(theta[,2]) + c(-1.96,1.96)*sd(theta[,2])
2 [1] 4.822267 6.367143
3 > quantile(theta[,2],c(.025,.975))
4      2.5%      97.5%
5 4.818738 6.360761
```

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

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

► par simulation d'échantillons gaussiens

```
1 > theta = matrix(NA, 1000, 2)
2 > for(i in 1:nrow(theta)){
3 +   xs = rnorm(n, mean(x), sd(x))
4 +   logLik = function(t) -sum(log(dnorm(xs, mean = t[1],
5     sd = t[2])))
6 +   theta[i,] <- optim(par = c(150, 5), logLik)$par
7 + }
7 > hist(theta[, 1])
```



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

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

```
1 > mean(theta[,1])
2 [1] 164.7104
3 > sd(theta[,1])
4 [1] 0.5345402
5 > mean(theta[,1]) + c(-1.96,1.96)*sd(theta[,1])
6 [1] 163.6627 165.7581
7 > quantile(theta[,1],c(.025,.975))
8      2.5%      97.5%
9 163.6500 165.7482
```

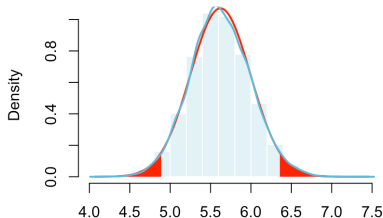
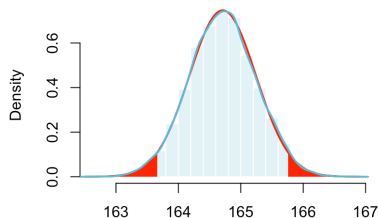
Aussi, $\mathbb{P}(\mu \in [163.6; 165.7]) \sim 95\%$

```
1 > mean(theta[,2]) + c(-1.96,1.96)*sd(theta[,2])
2 [1] 4.892208 6.352456
3 > quantile(theta[,2],c(.025,.975))
4      2.5%      97.5%
5 4.904579 6.373500
```

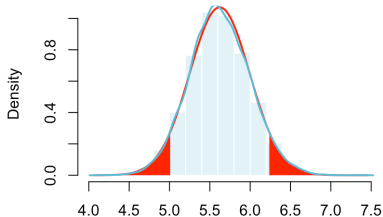
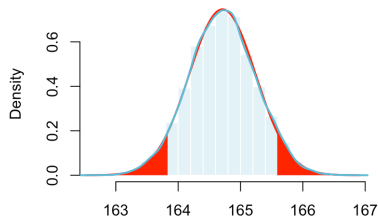
Aussi, $\mathbb{P}(\sigma \in [4.90; 6.36]) \sim 95\%$

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

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

```
1 > var(theta[,1])
2 [1] 0.2857332
3 > var(x)/n
4 [1] 0.2859441
5 > var(theta[,2])
6 [1] 0.1387653
7 > var(x)/(2*n)
8 [1] 0.1429721
```

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

```
1 > cor(theta)
2           [,1]      [,2]
3 [1,] 1.000000000 0.003202517
4 [2,] 0.003202517 1.000000000
```

Vraisemblance

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

```
1 > library(MASS)
2 > fitdistr(x,"normal")
3      mean      sd
4 164.7142857 5.6338083
5 ( 0.5323448) ( 0.3764247)
```

on retrouve

```
1 > mean(x)
2 [1] 164.7143
3 > sd(x)
4 [1] 5.659129
5 > sd(x)*sqrt((length(x)-1)/length(x))
6 [1] 5.633808
```

On retrouve ici, numériquement

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

Les valeurs entre parenthèses sont les écart-types des estimateurs

```
1 > fitdistr(x,"normal")
2      mean      sd
3 164.7142857  5.6338083
4 ( 0.5323448) ( 0.3764247)
```

On peut noter que

```
1 > 5.6338083 /sqrt(length(x))
2 [1] 0.5323448
3 > 5.6338083 /sqrt(2*length(x))
4 [1] 0.3764247
```

car

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

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

Méthode des Moments

Méthode des Moments

Soit $\mathbf{Y} = (y_1, \dots, y_n)$ 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 $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_d)$ la solution du système d'équations

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

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

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

Exemple 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 $\widehat{m}_1 = \bar{x}$ donc $\widehat{p}(x) = \bar{x}$

- ce que nous disent les mathématiques

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

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

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

$$\begin{cases} \hat{n}\hat{p} = \bar{x} = \frac{1}{n} \sum_{i=1}^n y_i \\ \hat{n}\hat{p}(1 - \hat{p}) = s^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \end{cases}$$

soit

$$\hat{p} = \frac{\bar{y} - s^2}{\bar{y}} \text{ et } \hat{n} = \frac{\bar{y}^2}{\bar{y} - s^2}$$

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

Maximum de Vraisemblance vs Méthode des Moments

Exemple 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 ?

```
1 > 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

$$\bar{y} = \frac{\hat{\theta}}{\hat{\theta} + 1} \text{ soit } \hat{\theta} = \frac{\bar{y}}{1 - \bar{y}}$$

Maximum de Vraisemblance vs Méthode des Moments

- maximum de vraisemblance

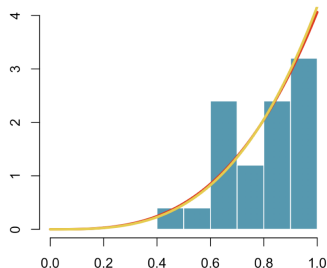
$$\log \mathcal{L}(\theta; \mathbf{y}) = n \log(\theta) + (\theta - 1) \sum_{i=1}^n \log(y_i)$$

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

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

```
1 > (a=mean(y)/(1-mean(y)))  
2 [1] 4.062526  
3 > (b=-25/sum(log(y)))  
4 [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)

```
1 > f = function(x, theta) theta*x^(theta-1)
2 > fitdistr(y, f, start = list(theta = 1))
3     theta
4     4.1672932
5     (0.8334586)
```

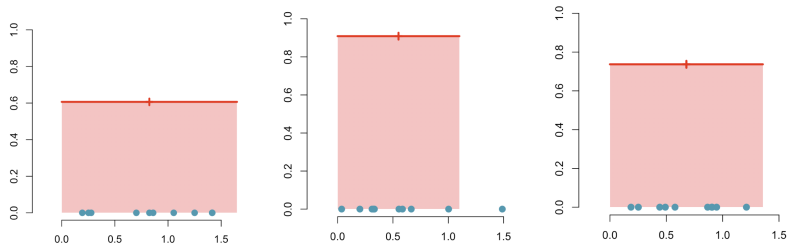
qui coïncide avec $b = -25 / \sum(\log(y))$.

Méthode des Moments

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

$$\bar{y} = \hat{\theta}/2 \text{ i.e. } \hat{\theta} = 2\bar{y}$$

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



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

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 $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n Y_k$ centrée converge vers une loi normale :

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

Méthode des Moments ★★★

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

Delta-Method

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

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

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).