GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=14271

Cours 7

- Échantillonnage
- Méthodes d'estimation ponctuelle
  - méthode de maximum de vraisemblance (EMV)
- Propriétés de l'EMV
- Information (statistique / de Fisher)
- Intervalles de confiance (asymptotiques)

Probabilité par rapport à statistique

- Pour une *valeur connue* de p, on peut calculer la probabilité d'une issue possible
- Ce qui est la *probabilité*
- Dans de nombreuses situations pratiques, cependant, nous ne savons pas p, mais plutôt, nous disposons de données qui sera utilisée pour *l'estimation* de p
- Ce qui est la *statistique*

Échantillonage

- Le but d'une étude statistique est d'obtenir des connaissances sur l'ensemble de la *population*, c.-à-d. **l'estimation d'un paramètre**
- Puisque un dénombrement complet de la population est très souvent pratiquement impossible, il faut d'autres moyens plus pratique
- ⇒ Un **échantillonnage** consiste à choisir parmi les éléments de la population un certain nombre d'unités pour lesquelles nous obtiendrons des observations (données)
- Nos données sont considérées comme la suite d'un processus aléatoire : si la collecte de données ont été répétées, le résultat serait probablement différent, qui peuvent influer sur les conclusions tirées sur la base de données
- C.-à-d., nos conclusions sont sujettes à la *variation aléatoire*

Utilité d'échantillonnage

- Un jardinier possède deux millions de graines pratiquement identiques, qui donnent soit des fleurs blanches, soit des fleurs rouges
- Il désire connaître en avance *le pourcentage* des fleurs blanches (afin d'être en mesure de les vendre sans tromper ses clients)
- S'il veut être *absolument certain* du type de fleurs produit, il sera obligé de semer *toutes les graines*
- **Donc, il n'aura plus des graines à vendre ! !**
- $\Rightarrow$ Il faut un **échantillon**
- (Même si le processus n'est pas déstructif, il est le plus souvent *impossible ou irréalisable* (le temps, les coûts) de mesurer chaque individu de la population)

Représentativité

- Sur la base de ses observations, le jardinier fera une *estimation* du nombre de fleurs blanches/rouges parmi les deux millions de graines
- $\Rightarrow$ On *généralise* à l'ensemble de *la population* les connaissances acquises sur la base de *quelques observations*
- On ne peut pas être *absolument certain* de notre prédiction, puisque l'on ne considère qu'une fraction seulement de la population totale : $\Rightarrow$ Imprécision due à l'échantillonnage
- Généralement il y aura *un écart* entre les observations faites sur l'échantillon et celles effectuées sur *la totalité* de la population
- Mais : si l'échantillon est choisi *de façon scientifique*, il est possible de faire une évaluation *probabiliste*
- $\Rightarrow$ Possible *d'évaluer l'erreur*, et déterminer la *précision de l'estimation*

Méthodes d'échantillonnage

Échantillonnage arbitraire

- Impossible de quantifier les probabilités associées, donc difficile d'estimer les paramètres et l'écart-type d'estimation (erreur standard d'estimation (ES))
- p. ex. les dix premiers à entrer dans la salle
- ⇒ **PAS recommandé ! !**

Échantillonnage aléatoire

- Correspond à des méthodes de tirage où chaque unité de la population a *une probabilité positive et connue* d'être sélectionnée
- Ces méthodes permettent d'estimer les paramètres de la population, et aussi d'obtenir une mesure de l'ES
- Pour nous, les méthodes le plus important correspond à soit AVEC remise (indépendant), soit SANS remise (échantillonnage aléatoire simple (EAS))

Estimation

- La procédure d'utilisation des informations obtenues à partir de l'échantillon qui permet de déduire des résultats concernant l'ensemble de la population est appelée **estimation**
- La valeur *inconnue* d'une population (à estimer à partir d'un échantillon) est appelée un **paramètre**
- p. ex. : la moyenne (μ) ; la proportion (le pourcentage) (p)
- Le paramètre de la population est estimé à partir d'une **statistique** calculée sur la base d'un échantillon
  $\Rightarrow$ une statistique est *une fonction des données obtenues*
- Un **estimateur** est une statistique utilisée afin d'estimer (deviner la valeur d') un paramètre θ ; c.-à-d. il est une règle qui nous permet de calculer une approximation de θ basée sur les valeurs de l'échantillon $X_1, \ldots, X_n$
- Une **estimation** est une valeur observée (calculée) de l'estimateur sur un échantillon

Qualité d'un estimateur

- Pour répondre à la question : 'comment choisir entre des estimateurs candidats', on doit examiner ce qui fait un 'bon' estimateur
- On considère donc des qualités (statistiques) des estimateurs
- Certaines qualités importantes :
    - biais
    - variance
    - erreur quadratique moyenne (EQM)

Biais

- Le **biais** d'un estimateur T d'un paramètre θ est définit par :

$$b(T) = E[T] - \theta,$$

(c.-à-d. la différence entre *l'espérance* de la distribution d'échantillonnage de l'estimateur T et *la vraie valeur* du paramètre θ)

- Un estimateur est **sans biais** (ou **non biaisé**) si le biais égale à 0

$\boxed{\textbf{Exemple 7.1}}$ Quel est le biais de $\overline{X}$ en tant qu'estimateur de la moyenne de la population μ ...

Variance

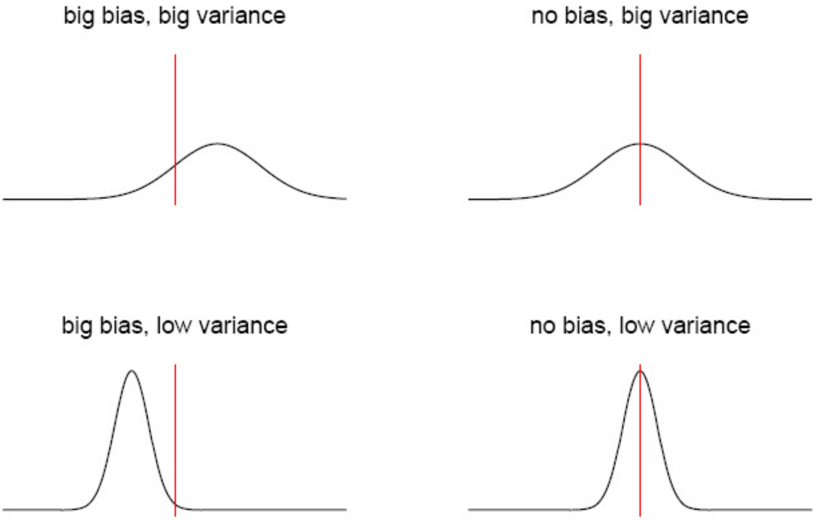
- Une autre qualité on peut considérer est le *variance* de l'estimateur :

$$Var(T) = E[(T - E[T])^2]$$

- Parmi deux estimateurs sans biais de θ, l'un sera *plus efficace* que l'autre si *sa variance est plus petite*

Exemple 7.2 Considérons maintenant la variance des estimateurs candidats de la moyenne de la population μ **...**

Biais et variance d'un estimateur T



Erreur Quadratique Moyenne (EQM)

- Une autre qualité que nous pouvons considérer est le **erreur quadratique moyenne (EQM)** d'un estimateur

$$EQM(T) = E[(T - \theta)^2]$$

- Ceci est différent de la variance lorsque l'estimateur T est biaisé
- Parfois, nous pourrions utiliser un estimateur qui a un peu de biais s'il a une variance beaucoup plus petite que la meilleure estimateur sans biais (*compromis biais-variance*)
- Il est simple à démontrer que l'EQM peut être exprimée comme une combinaison de biais et la variance :

$$EQM(T) = Var(T) + [b(T)^2]$$

Estimation (ponctuelle)

- La procédure d'utilisation des informations obtenues à partir de l'échantillon qui permet de déduire des résultats concernant l'ensemble de la population est appelée **estimation**
- La valeur *inconnue* d'une population (à estimer à partir d'un échantillon) est appelée un **paramètre**
- p. ex. : la moyenne (μ) ; la proportion (le pourcentage) (p)
- Le paramètre de la population est estimé à partir d'une **statistique** calculée sur la base d'un échantillon
  $\Rightarrow$ une statistique est *une fonction des données obtenues*
- Un **estimateur** est une statistique utilisée afin d'estimer (deviner la valeur d') un paramètre θ ; c.-à-d. il est une règle qui nous permet de calculer une approximation de θ basée sur les valeurs de l'échantillon $X_1, \ldots, X_n$
- Une **estimation** est une valeur observée (calculée) de l'estimateur sur un échantillon

Méthodes d'estimation ponctuelle

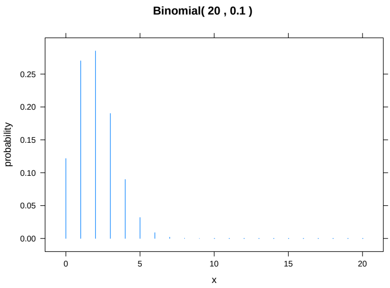
- *Méthode de maximum de vraisemblance* (qui donne souvent des estimateurs 'intuitifs')
- *Méthode des moments* – on va l'illustrer (<u>VIDÉO SEULEUMENT</u>), mais
  $\Rightarrow$ **cela ne fait <u>PAS</u> partie de l'examen**
- *Méthode des moindres carrés* (plus tard, avec 'régression')
- Méthode du minimum des déviations absolues
- Estimation bayesian

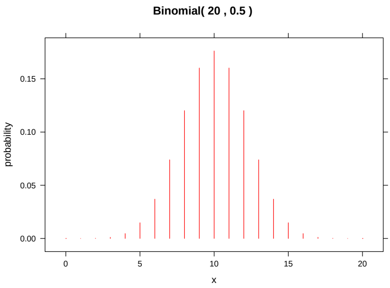
Vraisemblance

- Pour une valeur p connue, on peut exprimer la probabilité de n'importe quelles données possibles
- En revanche, on peut *considérer les observations comme connues* et considérer la probabilité en fonction du paramètre inconnu p
- La fonction de probabilité vue de cette façon est appelée la **vraisemblance**

Vraisemblance illustrée

20 lancements d'une pièce ; on observe $??$ piles





De ces deux distributions, de laquelle est-il le plus vraisemblable que l'échantillon soit issu ?

Définition de la vraisemblance

- **Définition** : Soit $x \sim f(x;\theta)$. La **vraisemblance** et **log vraisemblance** sont :

$$L(\theta) = f(x;\theta), \quad \ell(\theta) = \log L(\theta),$$

considérés comme des fonctions du paramètre θ.

- Soient $x = (x_1, \ldots, x_n)$ une réalisation des VAs $X_1, \ldots, X_n$. Alors

$$L(\theta) = f(x;\theta) = \prod_{j=1}^{n} f(x_j;\theta), \quad \ell(\theta) = \sum_{j=1}^{n} \log f(x_j;\theta),$$

où $f(x_j;\theta)$ est la loi de x_j.

- **<u>À NOTER</u> :** log $=$ log base e $=$ <u>log naturel</u>

Estimation par maximum de vraisemblance

- Une méthode d'estimation intuitive est **l'estimation par maximum de vraisemblance**
- Par exemple, l'estimateur le plus 'évident' p est $\hat{p} = X/n$ se révèle être **l'estimateur du maximum de vraisemblance (EMV / MLE)**
- En général, l'EMV est la valeur qui rend la probabilité aussi grande que possible – c'est la valeur qui *rend les données observées le plus probable*
- La manière habituelle de trouver l'EMV : le calcul – trouver la dérivée de la fonction de (log) vraisemblance, annuler et résoudre :

$$\frac{d \log L(\hat{\theta})}{d\theta} = 0, \quad \frac{d^2 \log L(\hat{\theta})}{d\theta^2} < 0$$

- (Cette méthode ne fonctionne pas dans tous les cas)
- Nous supposons que la première équation a une solution *unique* (ce n'est pas toujours vrai dans la réalité)

EMV, cont

- L'EMV $\hat{\theta}$ remplit la condition

$$L(\hat{\theta}) \geq L(\theta) \qquad \text{pour toute } \theta,$$

  ce qui équivaut à $\log L(\hat{\theta}) \geq \log L(\theta)$, car les valeurs maximales de $L(\theta)$ et $\log L(\theta)$ sont obtenues *à la même valeur θ*
- L'EMV peut :
  - exister et être unique,
  - ne pas être unique, ou
  - ne pas exister
- Dans la pratique, il est normalement nécessaire d'utiliser des algorithmes numériques pour obtenir $\hat{\theta}$ et $d^2 \log L(\hat{\theta})/d\theta^2$

Avantages/désavantages de la méthode

- Pour un échantillon 'suffisamment grand', l'EMV est :
  - non-biaisé
  - consistent
  - efficace (EQM minimal ; donc au moins puissant que l'estimateur EMM)
  - *normalement distribué*
  - donc, pratique pour l'inférence statistique
- En revanche, l'EMV :
  - pourrait être très biaisé si la taille de l'échantillon est petite
  - pourrait être très compliqué d'évaluer (il faut le faire numériquement)

PAUSE

Exemple

$\boxed{\textbf{Exemple 7.3}}$ Soit $X \sim Bin(n, p)$. Trouver l'EMV de p **...**

Exemple

Exemple 7.4 Soient $X_1, \ldots, X_n \sim$ iid $Pois(\lambda)$, $\lambda > 0$. Calculer :

1. $L(\lambda)$
2. $\log L(\lambda)$
3. $\hat{\lambda}_{EMV}$ ($+$ vérifier que l'extremum est un *maximum*)

Solution

1. $L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} \propto e^{-n\lambda}\lambda^{\sum y_i} \quad (= e^{-n\lambda}\lambda^{n\bar{y}})$

2. $\ell(\lambda) = n\bar{y}\log\lambda - n\lambda$

3. $\hat{\lambda}_{EMV} : \frac{d\ell(\lambda)}{d\lambda} = \frac{n\bar{y}}{\lambda} - n = 0 \implies \frac{\bar{y}}{\lambda} = 1 => \hat{\lambda}_{EMV} = \bar{y}$

- *Vérifier max :*

$$\frac{d^2\ell(\lambda)}{d\lambda^2} = \frac{d\ell(\lambda)}{d\lambda}\left[\frac{n\bar{y}}{\lambda} - n\right] = -\frac{n\bar{y}}{\lambda^2} < 0,$$

car $n\bar{y} > 0, \lambda > 0$, donc $\frac{n\bar{y}}{\lambda^2} > 0 \implies -\frac{n\bar{y}}{\lambda^2} < 0,$

alors l'extremum $(\hat{\lambda}_{EMV})$ est un *maximum*

Exemple

PAS important pour nous ! !

Example 7.5 Soient $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$. Trouver les EMV de μ et σ^2.

Solution : La densité normale est donnée par

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\},$$

donc le log vraisemblence pour un échantillon aléatoire (iid) $y_1, \ldots, y_n$ est

$$\ell(\mu, \sigma) = \log L(\mu, \sigma) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Solution, cont

En dérivant, on a $$\frac{\partial \ell(\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum(y_i-\mu) = 0 \qquad (*)$$

et $$\frac{\partial \ell(\mu,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(y_i-\mu)^2 = 0 \qquad (**)$$

En résolvant (*), on a (pour toute valeur σ^2) :

$$\sum(y_i-\mu) = 0 \quad => \quad \sum y_i = n\mu \quad => \quad \hat{\mu} = \sum y_i/n = \overline{y}$$

En résolvant (**) (en utilisant $\hat{\mu}$ au lieu de μ), on a :

$$-n\sigma^2+\sum(y_i-\hat{\mu})^2 = 0 \quad => \quad \sum(y_i-\hat{\mu})^2 = n\sigma^2 \quad => \quad \hat{\sigma}^2 = \frac{1}{n}\sum(y_i-\hat{\mu})^2$$

À NOTER : cet estimateur **est diffŕent** de l'estimateur non biaisé $s^2 = \dfrac{1}{n-1}\sum(y_i-\hat{\mu})^2$

Solution, cont

Il faut vérifier que le log vraisemblance est un *maximum* (non min) pour la paire des valeurs $(\hat{\mu}, \hat{\sigma}^2)$: test de derivée seconde :

$$\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu^2} \cdot \frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial (\sigma^2)^2} - \left(\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu \partial (\sigma^2)} \right)^2 > 0$$

$$\text{ET} \quad \frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu^2} < 0 \quad ; \quad \frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial (\sigma^2)^2} < 0$$

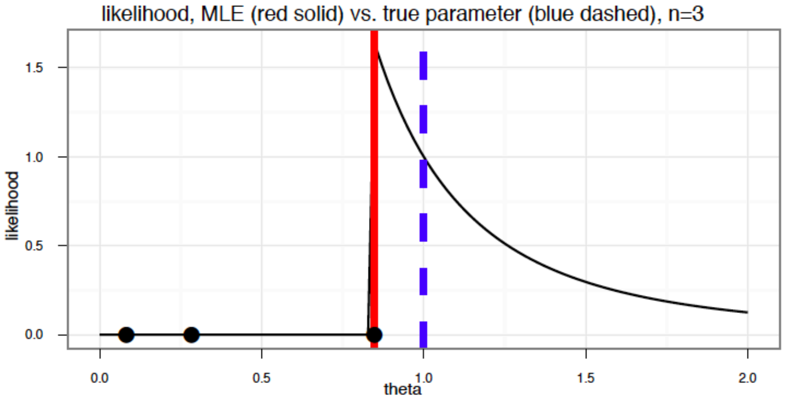
$$\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu \partial (\sigma^2)} = \frac{-1}{\sigma^4} \sum_{i=1}^{n} (y_i - \hat{\mu}) = 0 \quad ; \quad \frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu^2} = \frac{-n}{\hat{\sigma}^2} < 0$$

$$\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial (\sigma^2)^2} = \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = \frac{n^3}{2 \sum_{i=1}^{n} (y_i - \hat{\mu})^2} - \frac{n^3}{\sum_{i=1}^{n} (y_i - \hat{\mu})^2}$$

$$= \frac{-n^3}{2} < 0$$

Exemple uniforme – **le calcul ne marche pas ! !**

Example 7.6 Soient $y_1, \ldots, y_n$ un échantillon aléatoire tirée de la distribution uniforme $(0, \theta]$, dont la densité est $f(y) = 1/\theta$, $0 < y \leq \theta$ ($= 0$ sinon). Trouver l'EMV $\hat{\theta}$ de θ ...



Information (statistique)

- L'*information observée* $J(\theta)$ et l'*information espérée* (aussi appelée *Fisher information*) $I(\theta)$ sont :

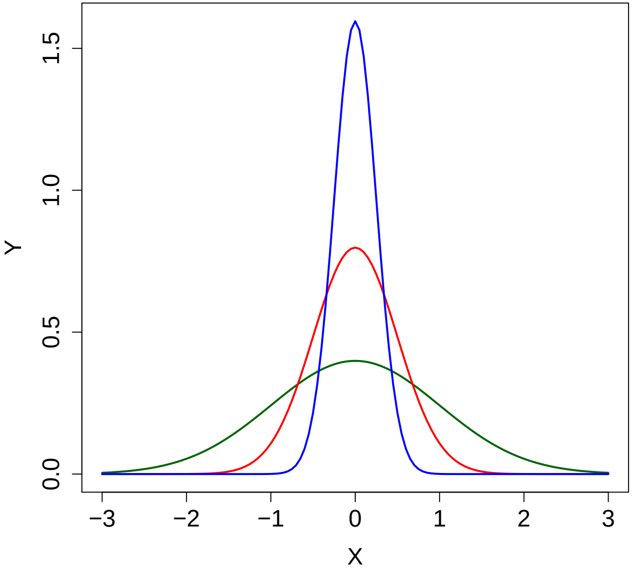
  - $J(\theta) = \dfrac{-d^2\ell(\theta)}{d\theta^2}$

  - $I(\theta) = E\{J(\theta)\} = E\left\{\dfrac{-d^2\ell(\theta)}{d\theta^2}\right\}$

- Elles sont des mesures de la *courbature* de $-\ell(\theta)$:

  plus les valeurs de $J(\theta)$ et $I(\theta)$ sont *grandes*, plus $\ell(\theta)$ et $L(\theta)$ sont *concentratés*

Exemple : distributions normales



Propriétés de l'EMV

- **Convergent** : $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1, \forall \epsilon > 0$
- **Invariance** : si $\hat{\theta}$ est l'EMV pour le paramètre θ, alors $h(\hat{\theta})$ est l'EMV pour le paramètre $h(\theta)$
- **Asymptotiquement sans biais** : $b(\theta) \to 0$ lorsque $n \to \infty$ (pour les échantillons 'petits' l'EMV pourrait être biaisé)
- **Efficacité asymptotique optimale** : aucun estimateur asymptotiquement sans biais peut avoir une variance plus petite que celle de l'EMV
- **Normalité asymptotique** : la distribution de $\hat{\theta}_n$ lorsque $n \to \infty$ est *la distribution normale*; cela nous donne une base pour la statistique inferentielle à partir de l'EMV (p. ex. IC)
- **IC approximatif** (niveau $1-\alpha$) pour θ : $\boxed{\hat{\theta} \pm z_{1-\alpha/2} \Big/ \sqrt{J(\hat{\theta})}}$

Conditions de régularité **(NON-EXAMINÉES)**

Les conditions techniques (pas très intéressantes ! !) dont la démonstration de normalité asymptotique dépend :

- La vraie valeur θ_0 de θ est un point *interieur* de l'espace du paramètre Θ, qui a *dimension finie* et qui est *compact* (sans 'trous' / contient les points limites)
- Pour deux valeurs de θ différentes, les densités sont *distinctes* (identifiabilité condition)
- Il existe une boule autour de θ_0 dans laquelle les *3 dérivées* de ℓ existent presque sûrement (c.-à-d. la probabilité $= 1$), et dont l'espérance de la 3$^{\text{ème}}$ dérivée est *bornée uniformement pour θ dans la boule*
- Il est valable de *changer l'ordre de dérivation et integration* (on peut dériver sous l'intégrale)

Exemple

Exemple 7.7 Soient $X_1, \ldots, X_n \sim$ iid *Bernoulli*(p). Calculer :

1. $L(p)$
2. $\ell(p)$
3. $\hat{p}_{EMV}$
4. $J(p)$
5. $I(p)$

Exemple 7.7, cont.

6. un IC approximatif de 95% pour p pour les données avec :

- $n = 10$ (nombre de piles $= 9$)

- $n = 20$ (nombre de piles $= 16$)

- $n = 100$ (nombre de piles $= 67$)

Exemple

Exemple 7.8 Soient $X_1, \ldots, X_n \sim$ iid $Pois(\lambda)$, $\lambda > 0$. Calculer :

1. $\hat{\lambda}_{EMV}$ en supposant $\sum X_i > 0$
2. $\hat{\lambda}_{EMV}$ en supposant $\sum X_i = 0$
3. l'EMV de $P(X = 0)$
4. $J(\lambda)$
5. $I(\lambda)$
6. un IC approximatif de 95% pour λ ...

Avertissement

- L'estimation par la méthode de maximum de vraisemblance est *séduisante* :
  - conceptuellement simple
  - interprétation intuitive
- Cependant, quelques difficultés ;
  conditions de régularité sur la fonction de vraisemblance qu'on ne peut pas ignorer :
  - difficiles à établir
  - difficiles à interpréter
  - difficiles à vérifier pour les cas réels
- Donc, même si très souvent utile, l'EMV n'est pas une panacée qui rendrait caduques les autres méthodes d'estimation