3. Idées fondamentales de la statistique

Modèles statistiques

On étudie une **population** (ensemble d'individus ou d'éléments) à partir d'un **échantillon** (sous-ensemble).

- **modèle statistique** : X =la quantité étudiée (variable aléatoire) ; la loi F de X est supposée connue sauf un nombre fini de paramètres θ
- **échantillon** (doit être représentatif de la population) : "données" $x_1, \ldots, x_n$, souvent supposées comme étant une réalisation de $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ (**indépendantes et identiquement distribuées**) ou $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$
- **statistique** : une fonction $T_n = h(X_1, \ldots, X_n)$ des variables aléatoires $X_1, \ldots, X_n$
- **estimateur** : une statistique utilisée pour estimer certains paramètres de F
- **Notations** :

$T_n = h(X_1, \ldots, X_n)$	est la statistique (variable aléatoire)
$t_n = h(x_1, \ldots, x_n)$	est la **réalisation** de T_n au moyen des x_i
$\widehat{\theta}$ ou $\widehat{\theta}_n$	est un **estimateur** d'un paramètre θ

Remarques

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ et $x_1, \ldots, x_n$ une réalisation correspondante. Alors

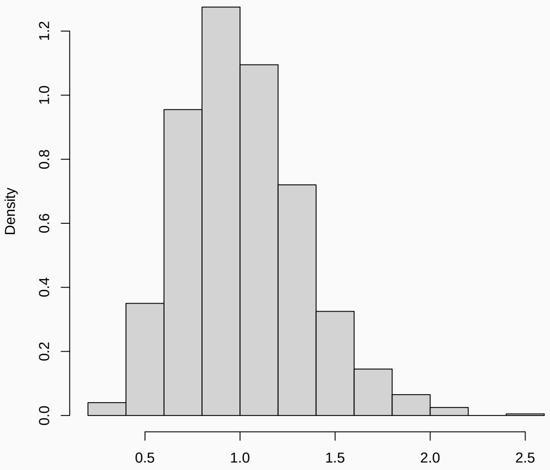
- $\widehat{\mu}_n = \overline{X}_n$ est une estimateur de μ, dont la valeur observée est $\overline{x}_n$
- $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$, est un estimateur de σ^2, dont la valeur observée est $n^{-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2$

Remarques :

- une statistique T_n étant fonction des variables aléatoires $X_1, \ldots, X_n$, c'est elle-même une variable aléatoire !
- La loi de T_n dépend de la loi des X_i, et est appelée **distribution d'échantillonnage de T_n**
- Si on ne peut pas déduire la loi de T_n de celle des X_i, on doit se contenter parfois de connaître $\mathbb{E}(T_n)$ et $\mathrm{var}(T_n)$, ou, si T_n est liée à $\overline{X}_n$, l'approximer à l'aide de $\mathbb{E}(X)$ et $\mathrm{var}(X)$ et le théorème central limite

Loi d'échantillonnage

Histogramme de 1000 réalisations de $\overline{X}_n = \frac{1}{10}(X_1 + \ldots + X_{10})$ où les $X_i \overset{iid}{\sim} \exp(1)$



Problèmes attaqués par la statistique

On suppose un **modèle** (c'est à dire une famille de distributions $f(x;\theta)$) et on souhaite

- **estimer** les paramètres θ de ce modèle
- poser des questions au sujet de la valeur de ces paramètres, par exemple **tester** si $\theta = 0$
- **prédire** les valeurs des observations futures

Il existe plusieurs méthodes pour estimer les paramètres d'un modèle. On va décrire les suivantes :

- **méthode des moments** (simple)
- **méthode du maximum de vraisemblance** (souvent utilisée car optimale dans beaucoup de situations)

3.1 Estimation de paramètres

Méthode des moments

- Supposons que l'échantillon tiré soit représentatif de la population
- Pour obtenir des estimateurs pour les paramètres inconnus de la population, on égalise les "moments" de l'échantillon ("empirique") à ceux de la population ("théorique")
- $k$ème moment
  - Population ("théorique") : $m_k = \mathbb{E}_\theta(X^k) = m_k(\theta)$. Comme la loi de X dépend de θ, $m_k = m_k(\theta)$
  - Echantillon ("empirique") : $\widehat{m}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$
  - L'estimateur des moments s'obtient on égalisant m_k et $\widehat{m}_k$, ce qui donne des équation(s) pour $\theta \in \mathbb{R}^p$
  - Par exemple $m_1(\theta) = \mathbb{E}_\theta(X) = \sum_{i=1}^n X_i/n$
- On a donc besoin d'autant de moments (supposés finies !) que de paramètres inconnus

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$. Estimer θ.

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimer μ et σ^2.

Méthode des moments

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$. Estimer θ.

Méthode des moments

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimer μ et σ^2.

Méthode du maximum de vraisemblance

Définition: Soient $x_1, \ldots, x_n$ des données supposées être une réalisation d'un échantillon aléatoire $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$. La **vraisemblance** pour θ est la fonction

$$L(\theta) = f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

f est la fonction de densité ou de masse (on suppose qu'elle existe)

Définition: L'**estimateur du maximum de vraisemblance** (maximum likelihood) $\widehat{\theta}_{\mathrm{ML}}$ d'un paramètre θ est celui qui maximise la fonction de vraisemblance parmi tous les θ possibles :

$$L(\widehat{\theta}_{\mathrm{ML}}) \geq L(\theta) \quad \text{pour tout } \theta$$

Il est plus facile de maximiser $\ell(\theta) := \log L(\theta)$, souvent en résolvant $d\ell(\theta)/d\theta = 0$, et vérifiant qu'il s'agit bien d'un maximum (par exemple si la deuxiṁe dérivée est négative $d^2\ell(\theta)/d\theta^2 < 0$)

Exemple $x_1, \ldots, x_n$ réalisations d'une loi $\exp(\lambda)$ avec $\lambda > 0$. Estimer λ

Exemple : maximum de vraisemblance

Exemple Supposons que $x_1, \ldots, x_n$ soient des réalisations i.i.d. d'une loi exponentielle,

$$f(x;\lambda) = \lambda e^{-\lambda x},\ x \geq 0, \quad \lambda > 0.$$

Trouver $\widehat{\lambda}_{\mathrm{ML}}$.

Erreur quadratique moyenne

Définition: L'**erreur quadratique moyenne** de l'estimateur $\widehat{\theta}$ de θ est

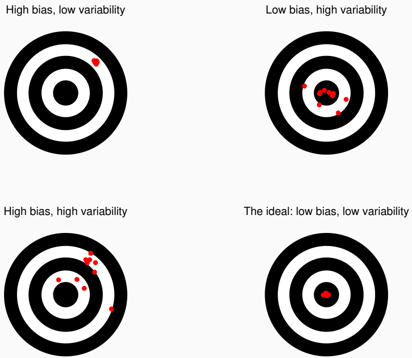
$$\mathrm{EQM}_\theta(\widehat{\theta}) = \mathbb{E}_\theta\{(\widehat{\theta}-\theta)^2\} = \cdots = \mathrm{Var}_\theta(\widehat{\theta}) + [b_\theta(\widehat{\theta})]^2,$$

où $b_\theta(\widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta}) - \theta$ est le **biais** de $\widehat{\theta}$

- La distribution de $\widehat{\theta}$ dépend de celle des X_i et donc de θ
- Si $\widehat{\theta}$ et $\widehat{\theta}'$ sont deux estimateurs du même paramètre θ et $\mathrm{EQM}_\theta(\widehat{\theta}) \leq \mathrm{EQM}_\theta(\widehat{\theta}')$, on préfère $\widehat{\theta}$
- si $b_\theta(\widehat{\theta}) < 0$, alors $\widehat{\theta}$ sous-estime θ
- si $b_\theta(\widehat{\theta}) > 0$, alors $\widehat{\theta}$ sur-estime θ
- si $b_\theta(\widehat{\theta}) \equiv 0$, alors $\widehat{\theta}$ est **non biaisé**, et $\mathrm{EQM}_\theta(\widehat{\theta}) = \mathrm{var}(\widehat{\theta})$

Exemple Soient $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. La médiane M_n et la moyenne $\overline{X}_n$ sont (à peu près) non-biaisés pour μ mais $\mathrm{var}(M_n) > \mathrm{var}(\overline{X}_n)$. Lequel des estimateurs $\overline{X}_n$ et M_n de μ est préférable ? Et si des valeurs aberrantes peuvent apparaître ?

Biais et variance



- $\theta =$ "bulle centrale", supposée être la vraie valeur
- fléchettes rouges $=$ réalisations de $\widehat{\theta}$ qui estime θ

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\widehat{\mu} = \overline{X}_n$, et $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. On admet que $\text{var}_{\mu,\sigma^2}(\sum_{i=1}^n (X_i - \overline{X}_n)^2) = 2\sigma^4(n-1)$. Trouver le biais et la variance de $\widehat{\mu}$. Trouver les valeurs de a qui minimisent le biais, la variance, et la EQM, pour $\widehat{\sigma}^a := a\widehat{\sigma}_n^2$.

Biais et variance : exemple

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\widehat{\mu} = \overline{X}_n$, et $\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$. On admet que $\mathrm{var}_{\mu,\sigma^2}(\sum_{i=1}^n (X_i - \overline{X}_n)^2) = 2\sigma^4(n-1)$. Trouver le biais et la variance de $\widehat{\mu}$. Trouver les valeurs de a qui minimisent le biais, la variance, et la EQM, pour $\widehat{\sigma}^a := a\widehat{\sigma}_n^2$.

Retour sur le maximum de vraisemblance

- Soient $x_1, \ldots, x_n$ des données supposées être une réalisation d'un échantillon aléatoire $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ (densité / masse). La **vraisemblance** pour θ est la fonction

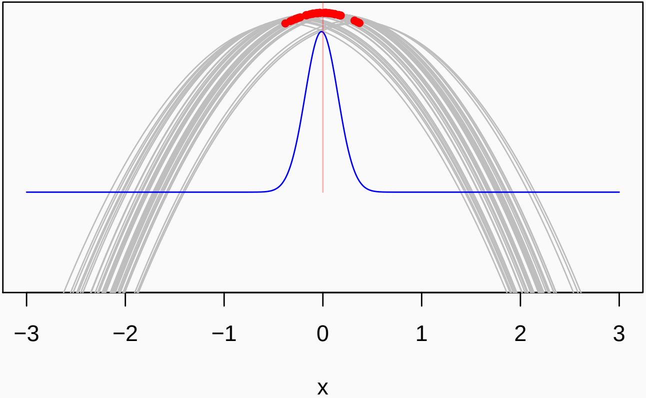
$$L(\theta) = f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

- $\widehat{\theta}_{\mathrm{ML}}$ satisfait

$$L(\widehat{\theta}_{\mathrm{ML}}) \geq L(\theta) \quad \text{pour tout } \theta$$

- **Interprétation** : dans le cas discret, on maximise $\mathrm{Pr}_\theta(X_1 = x_1, \ldots, X_n = x_n)$
- Il est plus facile de maximiser $\ell(\theta) := \log L(\theta)$, souvent en résolvant $d\ell(\theta)/d\theta = 0$, et vérifiant qu'il s'agit bien d'un maximum

La vraisemblance est une fonction aléatoire



Fonctions de vraisemblances correspondantes à 50 échantillons de taille $n = 40$. La vraie valeur est $\theta = 0$ et les 50 maxima sont en rouge.

Distribution asymptotique/approximative de $\widehat{\theta}_{\mathrm{ML}}$

- Dans des modèles "réguliers" (normal, exponentiel, Poisson, ...) on a $\ell'(\widehat{\theta}_{\mathrm{ML}}) = 0$
- Comme $\ell(\theta)$ est une somme de variables aléatoires iid, le théorème central limite s'applique
- Grâce à la méthode delta $\widehat{\theta}_{\mathrm{ML}} \overset{\mathrm{app}}{\sim} \mathcal{N}(\theta, 1/I_n(\theta))$ avec **l'information de Fisher**

$$I_n(\theta) = -\mathbb{E}_\theta(\ell''(\theta)) = \mathbb{E}_\theta(J_n(\theta)), \qquad J_n(\theta) = -\ell''(\theta)$$

  $1/I_n(\theta)$ est la **variance asymptotique** de $\widehat{\theta}_{\mathrm{ML}}$, $I_n(\theta)$ est la courbure
- **L'information observée** est $J_n(\widehat{\theta}_{\mathrm{ML}})$
- **Attention !** Certains modèles, tel que $U[0, \theta]$, ne sont pas réguliers

Exemple $x_1, \dots, x_n$ réalisations d'une loi $\exp(\lambda)$ avec densité $\lambda e^{-\lambda x} I(x \geq 0)$, $\lambda > 0$

Exemple

Exemple $x_1, \ldots, x_n$ réalisations d'une loi $\exp(\lambda)$ avec densité $\lambda e^{-\lambda x} I(x \geq 0)$, $\lambda > 0$