Eléments de statistique
Leçons 3 et 4 - Estimation

Louis Wehenkel

Département d'Electricité, Electronique et Informatique - Université de Liège
B24/II.93 - L.Wehenkel@ulg.ac.be

MATH0487-1 : 3BacIng, 3BacInf - Octobre 2014

Find slides: http://montefiore.ulg.ac.be/∼lwh/Stats/

Objectifs de l'estimation statistique

- On souhaite développer des méthodes générales permettant d'exploiter un échantillon statistique afin d'inférer des informations relatives à la population.
- On considère dans cette leçon le problème d'estimer un paramètre numérique qui caractérise la population.
- On souhaite calculer à partir d'un échantillon statistique une valeur pour ce paramètre aussi proche que possible de sa vraie valeur, et on souhaite aussi déterminer un intervalle autour de cette valeur, qui caractérise la précision de la valeur estimée.

Formalisation du problème d'estimation de paramètres

- On suppose que la population est étudiée via une variable aléatoire $\mathcal{X}$ à valeurs réelles, dont la fonction de répartition est connue à la valeur près d'un paramètre numérique $\theta \in \Theta \subset \mathbb{R}$.
- Nous notons par $F_{\mathcal{X}}(x;\theta)$ cette fonction de répartition, en mettant en évidence la dépendance vis-à-vis du paramètre θ. Θ désigne l'ensemble des valeurs a priori possibles pour le paramètre θ.
- Nous disposons d'un échantillon $\mathbf{D}_n$ i.i.d. de la v.a. $\mathcal{X}$. Désignons par θ^* la vraie valeur du paramètre lui correspondant. Nous souhaitons calculer à partir de $\mathbf{D}_n$
  1. une valeur ponctuelle $\hat{\theta}$, en moyenne proche de θ^*;
  2. un intervalle $[\theta_{\text{inf}};\theta_{\text{sup}}]$, contenant θ^* avec une probabilité assez élevée (disons $1-\alpha$).
- Remarque: dans la version plus générale on considère une v.a. vectorielle $\boldsymbol{\mathcal{X}} \in \mathbb{R}^p$, et un paramètre vectoriel $\boldsymbol{\theta} \in \mathbb{R}^m$ (et en général $p \neq m$).

Discussion

- Le choix de la famille $\{F_{\mathcal{X}}(\cdot;\theta) : \theta \in \Theta\}$ est déterminant, tant au niveau de la forme analytique de ces fonctions de répartition, qu'au niveau de l'ensemble Θ de valeurs admises pour le paramètre.
- Le choix de cette famille $F_{\mathcal{X}}(\cdot;\theta)$ est idéalement fait préalablement à la collecte de données, ou si ce n'est pas possible, de façon indépendante du contenu de $\mathbf{D}_n$.
- Exemple:
  - Mesures répétées de la température $\mathcal{X}$ de l'eau dans une baignoire, avec erreurs de mesure $\sim \mathcal{N}(0;1)$.
  - On supposera que la température réelle $\theta^* \in [10;50]$, que n vaut 10, et que $\alpha = 0.05$.
  - Suggestion: prendre $\hat{\theta} = m_x$, $\theta_{\text{inf}} = m_x - \lambda$ et $\theta_{\text{sup}} = m_x + \lambda$, et déterminer la valeur la plus faible pour λ qui convient en fonction du risque α.

Remarques

- On considérera dans la suite soit le cas où $\mathcal{X}$ est continue, soit le cas où elle est discrète.
  - Si $\mathcal{X}$ est continue, on utilisera en lieu et place de la famille $\{F_{\mathcal{X}}(\cdot;\theta) : \theta \in \Theta\}$ une famille de densités $\{f_{\mathcal{X}}(\cdot;\theta) : \theta \in \Theta\}$.
  - Si $\mathcal{X}$ est discrète, on utilisera en lieu et place de la famille $\{F_{\mathcal{X}}(\cdot;\theta) : \theta \in \Theta\}$ une famille de lois $\{P_{\mathcal{X}}(\cdot;\theta) : \theta \in \Theta\}$.
- Exemple pour $\mathcal{X}$ discret (variable de Bernoulli):
  - La proportion $\theta^* \in [0 \ldots 1]$ de femmes dans une certaine population.
  - On dispose d'un échantillon i.i.d. de taille 100.
  - Suggestion: prendre $\hat{\theta} = m_x$, $\theta_{\text{inf}} = m_x - \lambda$ et $\theta_{\text{sup}} = m_x + \lambda$, et déterminer une valeur pour λ qui convient.

Estimateurs ponctuels

- Le but de cette partie de la leçon est d'étudier les estimateurs ponctuels d'un paramètre θ à une seule dimension ($m = 1$) caractérisant la loi $F_{\mathcal{X}}(\cdot;\theta)$ d'une v.a. $\mathcal{X}$ réelle ($p = 1$).
- Désignons par $T(\cdot)$ la fonction (appelée estimateur) qui calcule $\hat{\theta}$ à partir d'un échantillon $\mathbf{D}_n$ (que nous supposons i.i.d. selon $F_{\mathcal{X}}(\cdot;\theta)$).
- L'estimateur $T(\cdot)$ définit donc, pour une taille donnée n de l'échantillon, une variable aléatoire $\mathcal{T}_n$.
- Pour une valeur donnée de θ, la densité $f_{\mathcal{T}_n}(\cdot;\theta)$ (ou bien sa loi $P_{\mathcal{T}_n}(\cdot;\theta)$, si $\mathcal{T}_n$ est discrète) peut en principe être déduite de la loi $F_{\mathcal{X}}(\cdot;\theta)$ (pour toute valeur supposée du paramètre $\theta \in \Theta$).
- Dans ce qui suit nous nous intéressons à l'espérance et à la variance de ces familles de v.a. $\mathcal{T}_n, \forall n \in \mathbb{N}_0$.

Erreur quadratique moyenne, biais et variance

- Pour un échantillon donné $\mathbf{D}_n$ i.i.d. selon $F_{\mathcal{X}}(\cdot;\theta^*)$, l'erreur d'estimation de l'estimateur $T(\cdot)$ est la différence $T(\mathbf{D}_n) - \theta^*$.
- Pour une taille donnée n de l'échantillon, considérons la v.a. $\mathcal{T}_n$ (dont la loi dépend bien entendu de θ^*). L'erreur quadratique moyenne d'estimation via T est définie par

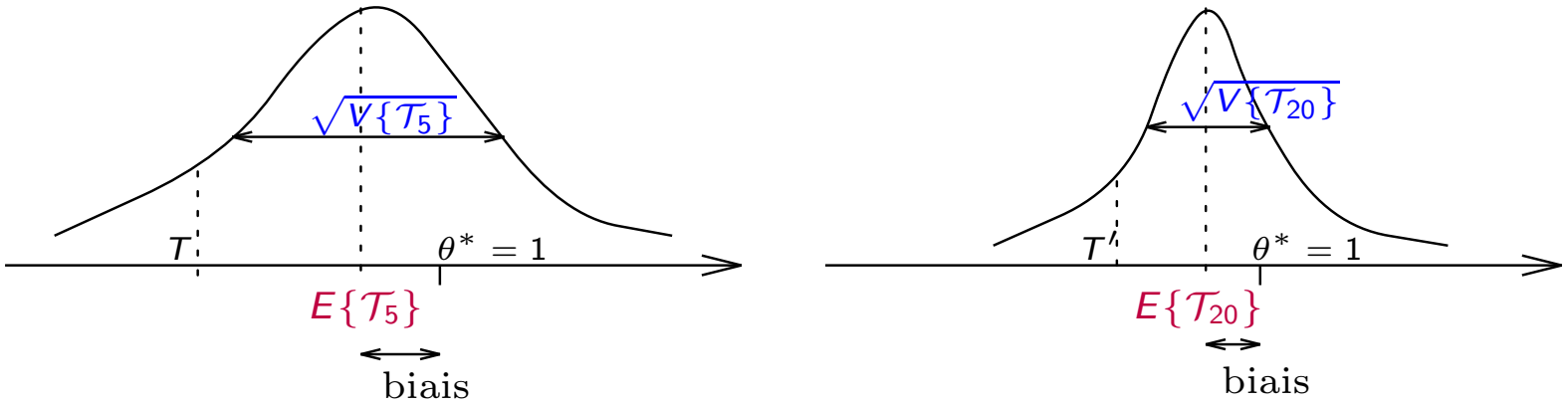
$$E\{(\mathcal{T}_n - \theta^*)^2\}.$$

- On peut décomposer $E\{(\mathcal{T}_n - \theta^*)^2\}$ en deux termes, comme suit

$$E\{(\mathcal{T}_n - \theta^*)^2\} = V\{\mathcal{T}_n\} + (E\{\mathcal{T}_n - \theta^*\})^2.$$

- $V\{\mathcal{T}_n\}$ est la **variance** de l'estimateur; ce terme mesure comment $\mathcal{T}_n$ fluctue autour de sa moyenne, d'un échantillon à l'autre.
- $E\{\mathcal{T}_n - \theta^*\}$ est le **biais** de l'estimateur; si cette grandeur est nulle, l'estimateur est dit 'non-biaisé', et son erreur quadratique moyenne se réduit alors à sa variance.

Illustration graphique: biais et variance



- Figure de gauche: illustration du biais et de la variance d'un estimateur $\mathcal{T}_n$: ici $n = 5$, la valeur de $\theta^* = 1$. Le biais est non-nul (négatif); la variance est plus grande que le biais au carré. Le point T représente une valeur estimée à partir d'un échantillon particulier de taille 5; elle sous-estime fortement la vraie valeur $\theta^* = 1$.
- Figure de droite: même population et même estimateur, avec $n = 20$: le biais et la variance sont réduits. La valeur T' déduite d'un échantillon de taille 20 est aussi plus proche de $\theta^* = 1$.
- La théorie de l'estimation statistique vise essentiellement à construire des estimateurs de faible biais, de faible variance, et tels que les deux diminuent aussi rapidement que possible lorsque la taille n de l'échantillon augmente.

Deux exemples

- Supposons que $\mathcal{X} \sim \mathcal{N}(\theta; 1)$ avec $\theta \in \mathbb{R}$.
  - Nous savons que m_x est un estimateur non-biaisé de θ.
  - Mais, ici Médiane$_x$ est aussi un estimateur non-biaisé de θ.
  - Cependant, m_x est plus précis que Médiane$_x$, car de plus faible variance.
  - On démontre en effet que m_x est (dans ce cas précis) l'estimateur non-biaisé de variance minimale (cf. suite du cours).
- Supposons que $\mathcal{X} \sim \mathcal{N}(0; \text{de variance } \theta)$ avec $\theta > 0$.
  - On sait que s_x^2 (resp. s_{n-1}^2) est un estimateur biaisé (resp. non-biaisé) de θ.
  - On peut cependant ici profiter du fait que par hypothèse $\mu_x = 0$, pour utiliser un autre estimateur, à savoir $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, qui est lui aussi non-biaisé mais est de variance plus faible que s_{n-1}^2.

Résumé/Remarques

- Pour un même problème d'estimation il est possible de construire différents estimateurs. La précision relative de ces différents estimateurs dépend des hypothèses de départ.
- Nous avons seulement suggéré la démarche pour choisir dans un certain contexte le bon estimateur, sur quelques exemples.
- En général, l'estimateur le plus précis peut être biaisé (il est souvent possible de 'troquer de façon bénéfique' une augmentation du biais pour une réduction plus forte de la variance). Mais, il n'y a pas de méthode générale pour construire un estimateur de précision maximale dans un contexte donné, et avec n fini.
- Cependant, il existe des méthodes génériques qui sont optimales lorsque $n \to \infty$ (i.e. en régime asymptotique) sous des conditions fort générales. C'est l'objet de la suite de ce cours d'en expliquer deux, à savoir la méthode du maximum de vraisemblance et l'approche bayesienne.

Méthode du Maximum de Vraisemblance

Premier exemple: variable de Bernoulli (cas discret le plus simple)

- Soit un échantillon $\mathbf{D}_n = (x_1, \ldots, x_n)$ avec $x_i \in \{0, 1\}$, et désignons par $\theta \in \Theta = [0; 1]$, la probabilité d'observer la valeur 1 d'une variable de Bernoulli $\mathcal{X}$.
- Sous l'hypothèse où l'échantillon est i.i.d. de $\mathcal{X}$, la probabilité d'observer cet échantillon particulier $\mathbf{D}_n$ vaut

$$P(\mathbf{D}_n|\theta) = \prod_{i=1}^{n}(x_i\theta + (1 - x_i)(1 - \theta)).$$

- La méthode du maximum de vraisemblance, consiste alors à estimer le paramètre θ par

$$\hat{\theta} = \arg\max_{\theta\in\Theta}\{P(\mathbf{D}_n|\theta)\},$$

c'est-à-dire à choisir la valeur de θ qui maximise la probabilité de nos observations de $\mathbf{D}_n$.

Méthode du Maximum de Vraisemblance

Variable de Bernoulli (voyons ce que cela donne)

- Désignons par n_0 le nombre de fois que nous observons la valeur 0 et par n_1 le nombre de fois que nous observons la valeur 1 parmi les $n = n_0 + n_1$ observations x_i de $\mathbf{D}_n$.
- On a par conséquent

$$P(\mathbf{D}_n|\theta) = \theta^{n_1}(1-\theta)^{n_0}.$$

- La valeur $\hat{\theta} \in [0;1]$ qui maximise $P(\mathbf{D}_n|\theta)$ peut-être obtenue par des techniques classiques (en calculant la dérivée par rapport à θ et en analysant les solutions de $\frac{\partial P(\mathbf{D}_n|\theta)}{\partial \theta} = 0$.).
- On trouve que

$$\hat{\theta} = \arg\max_{\theta\in[0;1]} \{\theta^{n_1}(1-\theta)^{n_0}\} = \frac{n_1}{n_0+n_1} = f_1 = m_x.$$

Méthode du Maximum de Vraisemblance

Second exemple: variable $\mathcal{N}(\mu;\sigma^2)$

- Soit un échantillon $\mathbf{D}_n = (x_1, \ldots, x_n)$ avec $x_i \in \mathbb{R}$, et désignons par $(\mu, \sigma^2) = (\theta_1, \theta_2) = \theta \in \Theta = \mathbb{R} \times \mathbb{R}^+$ le couple de paramètres de la loi $\mathcal{N}(\mu;\sigma^2)$ d'une variable gaussienne $\mathcal{X}$.
- Sous l'hypothèse où l'échantillon est i.i.d. de $\mathcal{X}$, la densité de probabilité d'observer cet échantillon particulier $\mathbf{D}_n$ vaut

$$f(\mathbf{D}_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left\{-\frac{(x_i-\theta_1)^2}{2\theta_2}\right\}.$$

- La méthode du maximum de vraisemblance, consiste alors à estimer le couple de paramètres $\theta = (\theta_1, \theta_2)$ par

$$\hat{\theta} = \arg\max_{\theta\in\Theta}\{f(\mathbf{D}_n|\theta)\},$$

c'est-à-dire à choisir la valeur de θ qui maximise la densité de probabilité de nos observations de $\mathbf{D}_n$.

Méthode du Maximum de Vraisemblance

Variable $\mathcal{N}(\mu;\sigma^2)$ (voyons ce que cela donne)

- Remarquons tout d'abord que maximiser $f(\mathbf{D}_n|\theta)$ revient aussi à maximiser $\log f(\mathbf{D}_n|\theta)$. On a

$$\log f(\mathbf{D}_n|\theta) = \sum_{i=1}^{n}\left(-\frac{1}{2}(\log(2\pi)+\log\theta_2) - \frac{(x_i-\theta_1)^2}{2\theta_2}\right)$$

- Il faut trouver $(\hat{\theta}_1,\hat{\theta}_2)$ pour annuler à la fois $\frac{\partial \log f(\mathbf{D}_n|\theta))}{\partial\theta_1}$ et $\frac{\partial \log f(\mathbf{D}_n|\theta)}{\partial\theta_2}$:

$\frac{\partial \log f(\mathbf{D}_n|\theta))}{\partial\theta_1} = +\frac{1}{2\theta_2}\sum_{i=1}^{n} 2(x_i-\theta_1)$

$\frac{\partial \log f(\mathbf{D}_n|\theta))}{\partial\theta_2} = \sum_{i=1}^{n}\left(-\frac{1}{2\theta_2}+\frac{(x_i-\theta_1)^2}{2\theta_2^2}\right)$

donc
$\frac{\partial \log f(\mathbf{D}_n|\theta))}{\partial\theta_1} = 0 \Rightarrow \hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = m_x$

donc
$\frac{\partial \log f(\mathbf{D}_n|\theta))}{\partial\theta_2} = 0 \Rightarrow \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-m_x)^2 = s_x^2$

Méthode du Maximum de Vraisemblance

Commentaires

- Dans la méthode du maximum de vraisemblance on choisit la valeur du (ou des) paramètre(s) qui maximise(nt) la probabilité d'observer l'échantillon $\mathbf{D}_n$ qu'on a sous la main (ou bien sa densité de probabilité, si la variable $\mathcal{X}$ est continue).
- Dans les deux exemples que nous avons traités, nous retrouvons les estimateurs usuels.
- Cependant, dans chacun de ces exemples, nous aurions pu imposer un ensemble de valeurs possibles Θ plus restrictif que celui (non contraignant) que nous avons utilisé pour faire nos calculs, ce qui aurait éventuellement conduit à des estimateurs différents.
- Nous avons traité le cas d'une v.a. $\mathcal{X}$ unidimensionnelle, mais on voit bien que l'idée pourrait s'appliquer aussi bien à des observations conjointes de plusieurs v.a.

Méthode du Maximum de Vraisemblance

Formulation générique (cas continu, $\mathcal{X} \in \mathbb{R}$, $\Theta \subset \mathbb{R}$)

- On définit la fonction de vraisemblance, sous l'hypothèse i.i.d., par

$$\mathcal{L}(\mathcal{X}_1, \ldots, \mathcal{X}_n; \theta) \stackrel{\triangle}{=} \prod_{i=1}^{n} f_{\mathcal{X}}(\mathcal{X}_i; \theta).$$

  Il s'agit donc d'une fonction aléatoire du paramètre θ.

- De façon alternative, on peut définir la log-vraisemblance par

$$\ell(\mathcal{X}_1, \ldots, \mathcal{X}_n; \theta) \stackrel{\triangle}{=} \ln \mathcal{L}(\mathcal{X}_1, \ldots, \mathcal{X}_n; \theta) = \sum_{i=1}^{n} \ln f_{\mathcal{X}}(\mathcal{X}_i; \theta).$$

- Etant donné un échantillon $\mathbf{D}_n = (x_1, \ldots, x_n)$, on détermine

$$\hat{\theta}_{MV} = \arg \max_{\theta \in \Theta} \mathcal{L}(x_1, \ldots, x_n; \theta) = \arg \max_{\theta \in \Theta} \ell(x_1, \ldots, x_n; \theta).$$

Méthode du Maximum de Vraisemblance

Equations de vraisemblance

- En pratique, pour calculer $\hat{\theta}_{MV}$ lorsque $\Theta \subset \mathbb{R}$, on résout l'équation de vraisemblance

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n; \theta) = 0.$$

- Si θ est multidimensionnel, disons $\theta = (\theta_1, \ldots, \theta_m)^T$, on résout simultanément les m équations de vraisemblance

$$\begin{array}{rcl} \frac{\partial}{\partial \theta_1} \ln \mathcal{L}(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m) & = & 0 \\ \vdots & \vdots & \vdots \\ \frac{\partial}{\partial \theta_m} \ln \mathcal{L}(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m) & = & 0 \end{array}$$

pour trouver $\hat{\theta}_{MV} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^T_{MV} \in \mathbb{R}^m$.

Méthode du Maximum de Vraisemblance

Propriétés principales

- Asympotiquement (i.e. lorsque $n \to \infty$), la variable aléatoire $\hat{\theta}_{MV}$ suit une loi normale, dont la moyenne tend vers la vraie valeur θ^*, et dont la variance est minimale et tend vers 0.
- On dit que l'estimateur du maximum de vraisemblance est convergent et asympotiquement efficace.
- Remarques:
  1. Dans le cas où $\mathcal{X}$ est discrète, on remplace $f_{\mathcal{X}}(\cdot;\theta)$ par $P_{\mathcal{X}}(\cdot;\theta)$
  2. Tout cela se généralise de façon immédiate au cas où θ est un vecteur et au cas où $\mathcal{X}$ est un vecteur.
  3. La solution de $\arg\max_{\theta\in\Theta} \mathcal{L}(x_1,\ldots,x_n;\theta)$ n'est pas toujours unique.

Démarche bayesienne

Considérons le problème suivant

- Une tirelire contient deux sortes de pièces de monnaie qui se ressemblent très fortement, mais qui son issues de deux lots de fabrication différents. Le premier lot a donné lieu à des pièces qui tombent sur pile α_1 fois en moyenne, et le second a donné lieu à des pièces qui tombent sur pile α_2 fois.
- On tire une pièce au hasard, puis on la lance n fois, et on observe qu'elle tombe n_1 fois sur pile. On désigne par $\mathbf{D}_n = (x_1, \ldots, x_n)$ la suite de lancers, avec $x_i \in \{0, 1\}$, la valeur 1 signifiant "pile".
- En supposant que l'on connaît les valeurs α_1 et α_2 caractérisant les deux lots, et qu'on connaît aussi les proportions β_1 et β_2 de pièces des deux lots contenues dans la tirelire, on demande
  1. de déterminer si la pièce est issue de la population α_1 ou α_2,
  2. de calculer la probabilité que la même pièce tombe sur pile lors d'un $n+1$-ème lancer.

Démarche bayesienne

Principe de résolution

- Appelons θ la variable aléatoire discrète correspondant au lot, et dont les 2 valeurs possibles appartiennent à $\Theta = \{\alpha_1, \alpha_2\}$.
- On a donc $\forall \theta \in \Theta$ et $\forall \mathbf{D}_n \in \{0,1\}^n$ que

$$P(\theta, \mathbf{D}_n) = P(\theta)P(\mathbf{D}_n|\theta) = P(\theta)\prod_{i=1}^{n}(\theta x_i + (1-\theta)(1-x_i))\,.$$

- Pour choisir, nous pouvons, par exemple, déterminer la valeur la plus probable de $\theta \in \{\alpha_1, \alpha_2\}$ étant données les observations $\mathbf{D}_n = (x_1, \ldots, x_n)$, i.e. celle qui maximise

$$P(\theta|\mathbf{D}_n) = \frac{P(\theta, \mathbf{D}_n)}{\sum_{\theta \in \{\alpha_1,\alpha_2\}} P(\theta, \mathbf{D}_n)} = \frac{P(\theta)P(\mathbf{D}_n|\theta)}{\sum_{\theta \in \{\alpha_1,\alpha_2\}} P(\theta)P(\mathbf{D}_n|\theta)}.$$

Démarche bayesienne

Principe de résolution (suite)

- Pour prédire la valeur de la variable $\mathcal{X}_{n+1}$, nous devons déterminer la loi de probabilité

$$P(\mathcal{X}_{n+1}|\mathbf{D}_n) = \sum_{\theta\in\Theta} P(\mathcal{X}_{n+1},\theta|\mathbf{D}_n) = \sum_{\theta\in\Theta} P(\mathcal{X}_{n+1}|\theta,\mathbf{D}_n)P(\theta|\mathbf{D}_n).$$

- Notons que si nous connaissons θ, $\mathbf{D}_n$ n'est pas utile pour calculer $P(\mathcal{X}_{n+1}|\theta,\mathbf{D}_n)$; en d'autres mots, $P(\mathcal{X}_{n+1}|\theta,\mathbf{D}_n) = P(\mathcal{X}_{n+1}|\theta)$, c'est-à-dire que $\mathcal{X}_{n+1} \perp \mathbf{D}_n|\theta$.
- In fine, nous pouvons donc prédire si au lancer suivant la pièce tombe sur pile ou face en calculant

$$x^*_{n+1} = \arg \max_{x_{n+1}\in\{0,1\}} \left\{ \sum_{\theta\in\Theta} P(x_{n+1}|\theta)P(\theta|\mathbf{D}_n) \right\}.$$

Démarche bayesienne

Décryptons ce qui vient d'être dit...

- Pour résoudre nos deux sous-problèmes (identifier le lot, ou bien prédire le résultat le plus probable du prochain lancer), il nous faut essentiellement être en mesure de calculer $P(\theta|\mathbf{D}_n)$.
- Pour calculer $P(\theta|\mathbf{D}_n)$, nous exploitons d'une part la vraisemblance de l'échantillon, i.e. $P(\mathbf{D}_n|\theta)$, et d'autre part notre information à priori sur la probabilité de tomber sur l'un ou l'autre lot, i.e. $P(\theta)$.
- Le reste des développements, ne fait appel qu'à la 'gymnastique' habituelle du calcul de probabilités.
- NB: Si les deux lots sont équiprobables, 'choisir' θ revient alors à déterminer $\hat{\theta}$ selon la méthode du maximum de vraisemblance.
- Prédire, ici, ne consiste pas à choisir puis prédire, mais plutôt à prédire en prenant en compte les incertitudes qu'on aurait au cas où on serait forcé de choisir.

Démarche bayesienne

Comparaison avec la méthode du maximum de vraisemblance

- Supposons que $n = 3$, $\mathbf{D}_3 = (0, 1, 1)$, que $\alpha_1 = 0.4$, $\alpha_2 = 0.6$, et que $\beta_1 = 0.7$ et que $\beta_2 = 0.3$.
- On a $P(\mathbf{D}_3|\theta = \alpha_1) = (0.6)^1(0.4)^2 = 0.096$ et $P(\mathbf{D}_3|\theta = \alpha_2) = (0.4)^1(0.6)^2 = 0.144$.
- Par la méthode du maximum de vraisemblance on choisirait θ qui maximise $P(\mathbf{D}_3|\theta)$ donc $\theta = \alpha_2$, et on prédirait que $P(x_4 = 1|\theta = \alpha_2) = \alpha_2 = 0.6$.
- L'approche bayesienne choisirait la valeur de θ qui maximise $P(\theta|\mathbf{D}_3)$: pour $\theta = \alpha_1 = 0.4 : P(\theta)P(\mathbf{D}_3|\theta) = \beta_1 P(\mathbf{D}_3|\theta = \alpha_1) = (0.7)(0.096) = 0.0672$ et $\theta = \alpha_2 = 0.6 : P(\theta)P(\mathbf{D}_3|\theta) = \beta_2 P(\mathbf{D}_3|\theta = \alpha_2) = (0.3)(0.144) = 0.0432$.
- Donc $P(\theta = \alpha_1|\mathbf{D}_3) = \frac{0.0672}{0.0672+0.0432} = 0.609$ et $P(\theta = \alpha_2|\mathbf{D}_3) = 0.391$. On choisirait donc plutôt $\theta = \alpha_1$.
- On prédirait que $P(x_4 = 1|\mathbf{D}_3) = \alpha_1 P(\theta = \alpha_1|\mathbf{D}_3) + \alpha_2 P(\theta = \alpha_2|\mathbf{D}_3)$ ce qui donne $P(x_4 = 1|\mathbf{D}_3) = (0.4)(0.609) + (0.6)(0.391) = 0.4782$.
- **Homework:** refaire le même calcul en supposant que $n = 6$, et que $\mathbf{D}_6 = (0, 1, 1, 1, 0, 1)$, toutes autres choses restant égales par ailleurs.

Démarche bayesienne: en général

- La démarche bayesienne consiste à
  1. postuler la loi a priori $f_\theta(\cdot)$ dont le support est l'ensemble Θ, mais qui n'est pas nécessairement uniforme sur Θ, afin de décrire l'incertitude qu'on a sur la valeur du paramètre θ;
  2. calculer par la formule de Bayes la densité de probabilité conditionnelle $f_{\theta|\mathbf{D}_n}(\cdot)$, en supposant que l'échantillon $\mathbf{D}_n$ est i.i.d. une fois fixé θ;
  3. exploiter cette densité conditionnelle pour déterminer une valeur du paramètre θ, soit selon la formule

$$\hat{\theta}_{MAP} = \arg \max_{\theta' \in \Theta} f_{\theta|\mathbf{D}_n}(\theta'),$$

soit selon la formule

$$\hat{\theta}_{EXP} = \arg \min_{\theta' \in \Theta} E\{(\theta' - \theta)^2 | \mathbf{D}_n\} = E\{\theta|\mathbf{D}_n\} = \int_{\theta' \in \Theta} \theta' f_{\theta|\mathbf{D}_n}(\theta') \mathrm{d}\theta';$$

  4. utiliser $f_{\theta|\mathbf{D}_n}(\cdot)$ pour faire des prédictions, et pour construire des intervalles de confiance sur θ au niveau $1 - \alpha$ souhaité.

Application: estimation bayesienne d'une proportion

- Ici θ désigne le paramètre p_1 d'une variable de Bernoulli.
- En l'absence de plus d'informations, choisissons comme loi $f_\theta(\theta)$ la loi uniforme sur $\Theta = [0;1]$. $f_\theta(\theta) = 1$.
- On dispose d'un échantillon $\mathbf{D}_n$ i.i.d., et désignons par n_1 le nombre de fois qu'on observe la valeur 1 et $n_0 = n - n_1$ le nombre de fois qu'on observe la valeur 0.
- On calcule

$$f_{\theta|\mathbf{D}_n}(\theta) = \frac{f_\theta(\theta) P_{\mathbf{D}_n|\theta}(\mathbf{D}_n)}{\int_{\theta\in\Theta} f_\theta(\theta) P_{\mathbf{D}_n|\theta}(\mathbf{D}_n) \mathrm{d}\theta} = \frac{\theta^{n_1}(1-\theta)^{n_0}}{\int_{\theta'\in\Theta} \theta^{n_1}(1-\theta)^{n_0}\mathrm{d}\theta}.$$

- NB: la loi $f_{\theta|\mathbf{D}_n}(\theta)$ est une loi "Bêta de type I" $\mathcal{B}e(n_1+1, n_0+1)$, et on montre que $\int_{\theta'\in\Theta} \theta^{n_1}(1-\theta)^{n_0}\mathrm{d}\theta = \frac{\Gamma(n_1+1)\Gamma(n_0+1)}{\Gamma(n+2)}$ (voir slides suivants).

Estimation bayesienne d'une proportion

Encart sur les lois $\mathcal{B}e(n,p)$ et la fonction $\Gamma(t)$

- Pour rappel: la fonction $\Gamma(t)$ est **définie** par

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \mathrm{d}x.$$

NB: on a $\Gamma(n) = (n-1)!, \forall n \in \mathbb{N}_0$.

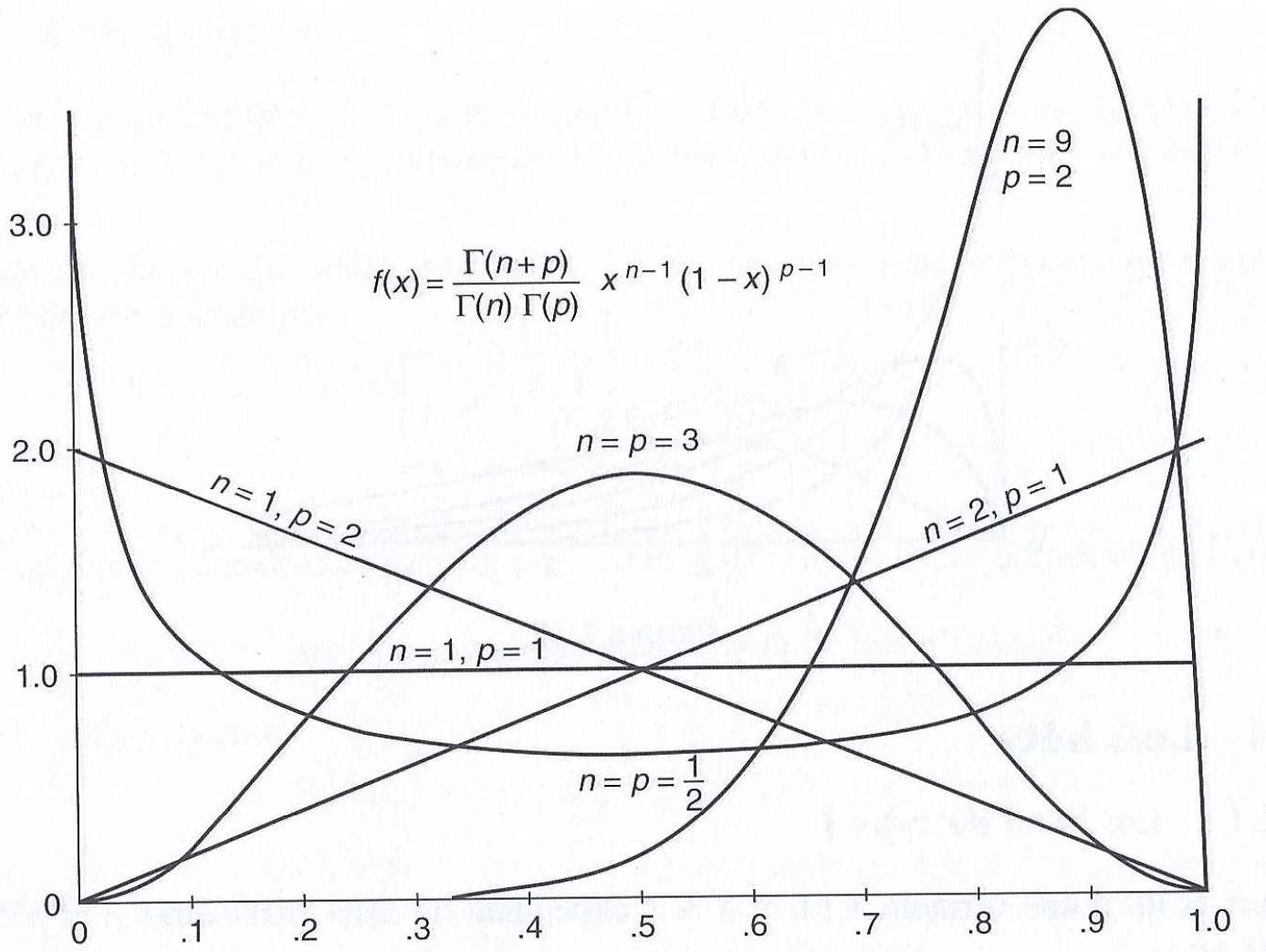
- **Par définition**, on dit que $\mathcal{X} \in]0;1[$ suit une loi $\mathcal{B}e(n,p)$, avec $n,p>0$, si sa densité s'écrit comme suit:

$$f_{\mathcal{X}}(x) = \frac{\Gamma(n+p)}{\Gamma(n)\Gamma(p)} x^{n-1}(1-x)^{p-1}.$$

On a $E\{\mathcal{X}\} = \frac{n}{n+p}$, et $\forall n,p>1$ le mode $\arg\max_{x\in]0;1[} f_{\mathcal{X}}(x) = \frac{n-1}{n+p-2}$.

- Remarque: lorsque $n=p=1$ on obtient la loi uniforme $f_{\mathcal{X}}(x)=1$ sur $[0;1]$.

Allures de quelques lois Bêta



Estimation bayesienne d'une proportion

Analyse du résultat

- Dans notre cas, partant d'une loi a priori uniforme:

$$f_\theta(\theta) \sim \mathcal{B}e(1,1)$$

nous obtenons, pour un échantillon composé de n_1 valeurs à 1 et n_0 valeurs à 0, une loi a posteriori

$$f_{\theta|\mathbf{D}_{n_1,n_0}}(\theta) \sim \mathcal{B}e(1+n_1, 1+n_0).$$

- On en déduit

$$\hat{\theta}_{MAP} = \frac{n_1}{n_1+n_0} = \hat{\theta}_{MV},$$

et

$$\hat{\theta}_{EXP} = \frac{n_1+1}{n_1+n_0+2} \quad \text{(plus proche de } 1/2 \text{ que } \hat{\theta}_{MAP}\text{)}.$$

Estimation bayesienne d'une proportion

Généralisation

- Supposons qu'au lieu de partir d'une loi a priori f_θ uniforme, nous postulions que
$$f_\theta(\theta) \sim \mathcal{B}e(m_1, m_0), m_1, m_0 > 0$$
Par exemple, si nous pensons que θ doit être proche de 0.5 (comme dans un jeu de pile ou face), nous pourrions utiliser $m_1 = m_0 = m$, en ajustant m.
- Nous obtenons alors une loi a posteriori
$$f_{\theta|\mathbf{D}_{n_1,n_0}}(\theta) \sim \mathcal{B}e(m_1 + n_1, m_0 + n_0).$$
- On en déduit
$$\hat{\theta}_{MAP} = \frac{n_1 + m_1 - 1}{n_1 + n_0 + m_1 + m_0 - 2},$$
et
$$\hat{\theta}_{EXP} = \frac{n_1 + m_1}{n_1 + n_0 + m_1 + m_0}.$$

Estimation bayesienne d'une proportion

Comparaison avec la méthode du MV

- Les deux estimateurs

$$\hat{\theta}_{MAP} = \frac{n_1 + m_1 - 1}{n_1 + n_0 + m_1 + m_0 - 2} \neq \hat{\theta}_{MV},$$

et

$$\hat{\theta}_{EXP} = \frac{n_1 + m_1}{n_1 + n_0 + m_1 + m_0} \neq \hat{\theta}_{MV},$$

exploitent l'information a priori disant que θ est proche $\frac{m_1}{m_1+m_0}$.

- Cela se traduit par des estimateurs, en moyenne plus proches de $\frac{m_1}{m_1+m_0}$ que $\hat{\theta}_{MV}$, et de plus faible variance que celui-ci.
- Les deux estimateurs bayesiens on une variance d'autant plus faible que m_1 et m_0 sont grands.
- Cependant, cela ce traduit aussi par un biais en général non-nul, d'autant plus élevé que la vraie valeur de θ est différente de $\frac{m_1}{m_1+m_0}$.

Démarche bayesienne (discussion)

Propriétés des estimateurs bayesiens

- Dans l'exemple précédent, on observe que quelque soit $m_1, m_2 > 0$, lorsque n_1 et n_2 deviennent grands, les estimateurs bayesiens $\hat{\theta}_{MAP}$ et $\hat{\theta}_{EXP}$ rejoignent l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$.
- En général, on montre que si f_θ est non-nulle partout sur Θ, cela reste encore vrai.
- Par conséquent, asympotiquement (lorsque $n \to \infty$), les trois estimateurs sont équivalents, et donc les deux estimateurs bayesiens sont également convergents et asympotiquement efficaces, et cela indépendamment du choix de f_θ de support égal à Θ.

Démarche bayesienne (discussion)

Caractère décomposable des estimateurs bayesiens

- L'approche baysienne possède une propriété intéressante:
- Supposons qu'on dispose d'un dataset supplémentaire $\mathbf{D}'_{n'_1,n'_0}$, en plus de $\mathbf{D}_{n_1,n_0}$.
- On peut alors calculer $f_{\theta|\mathbf{D}_{n_1,n_0},\mathbf{D}'_{n'_1,n'_0}}$ de plusieurs façons:
  1. On fusionne les deux datasets, et on applique au dataset fusionné la méthode bayesienne en partant de la loi a priori f_θ.
  2. On applique f_θ au premier des deux datasets, $\mathbf{D}_{n_1,n_0}$, puis on applique la loi $f_{\theta|\mathbf{D}_{n_1,n_0}}$ au second dataset $\mathbf{D}'_{n'_1,n'_0}$.
  3. On applique f_θ au second des deux datasets, $\mathbf{D}'_{n'_1,n'_0}$, puis on applique la loi $f_{\theta|\mathbf{D}'_{n'_1,n'_0}}$ au premier dataset $\mathbf{D}_{n_1,n_0}$.
- HOMEWORK: vérifier qu'on obtient bien le même résultat.

Stratification: explication détaillée (1)

Considérons que sur $(\Omega, \mathcal{E}, P)$ nous cherchons à étudier une v.a. $\mathcal{X}$ dont nous souhaitons estimer $E\{X\}$, en nous servant de la connaissance d'une variable auxiliaire $\mathcal{Z}$ binaire, dont nous connaissons la loi de façon exacte.

Notons par $P(\mathcal{Z}=1)$ (resp. $P(\mathcal{Z}=0)=1-P(\mathcal{Z}=1)$) le paramètre de la variable $\mathcal{Z}$, et par $E\{\mathcal{X}|\mathcal{Z}=1\}$ (resp. $E\{\mathcal{X}|\mathcal{Z}=0\}$) l'espérance conditionnelle de $\mathcal{X}$ sachant que $\mathcal{Z}=1$ (resp. sachant que $\mathcal{Z}=0$), ainsi que par $V\{\mathcal{X}|\mathcal{Z}=1\}$ (resp. $V\{\mathcal{X}|\mathcal{Z}=0\}$) la variance conditionnelle correspondante.

Nous savons que (application du théorème de l'espérance totale):

$$E\{\mathcal{X}\} = E\{E\{\mathcal{X}|\mathcal{Z}\}\} = P(\mathcal{Z}=1)E\{\mathcal{X}|\mathcal{Z}=1\} + P(\mathcal{Z}=0)E\{\mathcal{X}|\mathcal{Z}=0\},$$

et que (application du théorème de la variance totale):

$$\begin{array}{rcl} V\{\mathcal{X}\} & = & E\{V\{\mathcal{X}|\mathcal{Z}\}\} + V\{E\{\mathcal{X}|\mathcal{Z}\}\} \\ & = & P(\mathcal{Z}=1)V\{\mathcal{X}|\mathcal{Z}=1\} + P(\mathcal{Z}=0)V\{\mathcal{X}|\mathcal{Z}=0\} + \\ & & P(\mathcal{Z}=1)(E\{\mathcal{X}\} - E\{\mathcal{X}|\mathcal{Z}=1\})^2 + P(\mathcal{Z}=0)(E\{\mathcal{X}\} - E\{\mathcal{X}|\mathcal{Z}=0\})^2. \end{array}$$

Estimation stratifiée de $E\{\mathcal{X}\}$: nous supposons que nous pouvons tirer $n_1 \in \mathbb{N}_0$ observations de $\mathcal{X}$ distribuées i.i.d. selon $F_{\mathcal{X}|\mathcal{Z}=1}(x)$ et, de façon indépendante, aussi $n_0 \in \mathbb{N}_0$ observations de $\mathcal{X}$ distribuées i.i.d. selon $F_{\mathcal{X}|\mathcal{Z}=0}(x)$, ce qui nous permet de calculer les deux moyennes d'échantillon correspondantes notées $m_{x|z=1}$ et $m_{x|z=0}$, qui fournissent des estimateurs non-biaisés de $E\{\mathcal{X}|\mathcal{Z}=1\}$ et $E\{\mathcal{X}|\mathcal{Z}=0\}$, estimateurs qui sont respectivement de variance $V\{\mathcal{X}|\mathcal{Z}=1\}/n_1$ et $V\{\mathcal{X}|\mathcal{Z}=0\}/n_0$.

Nous construisons l'estimateur stratifié défini par $m_x^{s,n_1,n_0} = P(\mathcal{Z}=1)m_{x|z=1} + P(\mathcal{Z}=0)m_{x|z=0}$.

Stratification: explication détaillée (2)

Biais de l'estimateur stratifié: puisque $E\{m_{x|z=1}\} = E\{\mathcal{X}|\mathcal{Z}=1\}$, et que $E\{m_{x|z=0}\} = E\{\mathcal{X}|\mathcal{Z}=0\}$, nous avons que $E\{m_x^{s,n_1,n_0}\} = P(\mathcal{Z}=1)E\{\mathcal{X}|\mathcal{Z}=1\} + P(\mathcal{Z}=0)E\{\mathcal{X}|\mathcal{Z}=0\} = E\{\mathcal{X}\}$, vu le TH de l'espérance totale. En d'autres mots, l'estimateur stratifié est un estimateur non biaisé de $E\{\mathcal{X}\}$.

Variance de l'estimateur stratifié: comme il s'agit d'une combinaison linéaire de deux v.a. indépendantes (les deux sondages relatifs aux deux strates étant indépendants), la variance est obtenue par la formule suivante:

$$V\{m_x^{s,n_1,n_0}\} = (P(\mathcal{Z}=1))^2 V\{m_{x|z=1}\} + (P(\mathcal{Z}=0))^2 V\{m_{x|z=0}\}.$$

i.e.

$$V\{m_x^{s,n_1,n_0}\} = \frac{(P(\mathcal{Z}=1))^2}{n_1} V\{\mathcal{X}|\mathcal{Z}=1\} + \frac{(P(\mathcal{Z}=0))^2}{n_0} V\{\mathcal{X}|\mathcal{Z}=0\}.$$

Supposons alors que $n_1 = nP(\mathcal{Z}=1)$ et que $n_0 = nP(\mathcal{Z}=0)$, i.e. que nous allouons les n échantillons de façon proportionnelle à la loi de probabilité de la variable auxiliaire $\mathcal{Z}$. Nous obtenons, dans ces conditions, que

$$V\{m_x^s\} = \frac{P(\mathcal{Z}=1)}{n} V\{\mathcal{X}|\mathcal{Z}=1\} + \frac{P(\mathcal{Z}=0)}{n} V\{\mathcal{X}|\mathcal{Z}=0\} = \frac{1}{n} E\{V\{\mathcal{X}|\mathcal{Z}\}\}.$$

Or $E\{V\{\mathcal{X}|\mathcal{Z}\}\} \leq V\{\mathcal{X}\}$ (cf. TH de la variance totale; inégalité stricte si $E\{\mathcal{X}|\mathcal{Z}=1\} \neq E\{\mathcal{X}|\mathcal{Z}=0\}$.

Par conséquent, l'estimateur stratifié de façon proportionnelle est de variance plus faible qu'un estimateur non-stratifié utilisant $n_1 + n_0 = n$ échantillons tirés i.i.d. selon $F_{\mathcal{X}}$. (Qui serait de variance $V\{\mathcal{X}\}/n$).

Stratification: explication détaillée (3)

Vu ce qui précède, on peut se demander si l'allocation proportionnelle est la meilleure façon d'allouer les sondages.

En fait, elle n'est pas optimale en général, mais pour encore améliorer l'estimateur stratifié (par rapport à l'allocation proportionnelle), il faut disposer en plus de l'information a priori en ce qui concerne les variances conditionnelles $V\{\mathcal{X}|\mathcal{Z}=1\}$ et $V\{\mathcal{X}|\mathcal{Z}=0\}$.

Intuitivement, si la variance conditionnelle de $\mathcal{X}$ pour une des strates est très faible, il devrait en effet être plus productif d'allouer la majorité des échantillons à l'autre strate.

Mathématiquement, l'allocation optimale des nombres de sondages aux deux strates peut se formuler comme un problème d'optimisation: étant donné un nombre total d'observations n, déterminer sur base de la connaissance de $V\{\mathcal{X}|\mathcal{Z}=1\}$ et $V\{\mathcal{X}|\mathcal{Z}=0\}$ et de $P(\mathcal{Z}=1)$ et $P(\mathcal{Z}=0)$, la proportion idéale p_1 d'échantillons à allouer à la strate $\mathcal{Z}=1$ (le reste, $p_0=1-p_1$ étant alloué à la strate $\mathcal{Z}=0$). Il s'agit donc de minimiser la variance de l'estimateur stratifié en choisissant $p_1 \in [0;1]$, i.e. de façon à minimiser

$$V\{m_x^s\} = \frac{(P(\mathcal{Z}=1))^2}{p_1 n} V\{\mathcal{X}|\mathcal{Z}=1\} + \frac{(P(\mathcal{Z}=0))^2}{(1-p_1)n} V\{\mathcal{X}|\mathcal{Z}=0\}$$

les autres grandeurs ($P(\mathcal{Z}=1)$ et $P(\mathcal{Z}=0)$, n, et $V\{\mathcal{X}|\mathcal{Z}=1\}$ et $V\{\mathcal{X}|\mathcal{Z}=0\}$) étant fixées et connues a priori.

Stratification: explication détaillée (4)

Le résultat de ce calcul d'optimisation (que nous ne détaillons pas) nous donne que la proportion optimale à allouer aux deux strates valent respectivement:

$$p_1^* = \frac{P(\mathcal{Z}=1)\sqrt{V\{\mathcal{X}|\mathcal{Z}=1\}}}{E\{\sqrt{V\{\mathcal{X}|\mathcal{Z}\}}\}} \quad \text{et} \quad p_0^* = 1 - p_1^* = \frac{P(\mathcal{Z}=0)\sqrt{V\{\mathcal{X}|\mathcal{Z}=0\}}}{E\{\sqrt{V\{\mathcal{X}|\mathcal{Z}\}}\}}.$$

Ces valeurs se traduisent par une variance de l'estimateur stratifié de façon optimale qui vaut

$$V\{m_x^{s,*}\} = \frac{(E\{\sqrt{V\{\mathcal{X}|\mathcal{Z}\}}\})^2}{n}.$$

Notez bien que

$$(E\{\sqrt{V\{\mathcal{X}|\mathcal{Z}\}}\})^2 \leq E\{V\{\mathcal{X}|\mathcal{Z}\}\} \qquad \text{(cf. inégalité de Jensen; voir cours de probas).}$$

CONCLUSION: L'allocation proportionnelle est optimale seulement si la variance conditionnelle $V\{\mathcal{X}|\mathcal{Z}\}$ est une v.a. constante (i.e. que $V\{\mathcal{X}|\mathcal{Z}=1\} = V\{\mathcal{X}|\mathcal{Z}=0\}$).

REMARQUE: Pour une valeur de n fixée, les nombres d'échantillons 'optimaux' (ou bien ceux définis selon l'allocation proportionnelle) ne sont pas nécessairement des nombres entiers. Dans ce cas il faut les 'arrondir' vers le haut; on payera ainsi un faible prix en termes de nombre total de sondés, en bénéficiant d'une réelle augmentation de précision.

Stratification: explication détaillée (5)

GENERALISATIONS:
1. Le raisonnement que nous venons de faire peut être étendu de façon directe au cas où la variable auxiliaire $\mathcal{Z}$ possède plus que deux valeurs possibles, disons $\{z_1, \ldots, z_k\}$. Le résultat général est que 1. l'allocation proportionnelle est toujours meilleure que le sondage indépendant de $\mathcal{Z}$, et que 2. l'allocation optimale des observations doit se faire en fonction du rapport (supposé connu avant d'entamer le sondage)

$$\frac{P(\mathcal{Z} = z_i)\sqrt{V\{\mathcal{X}|\mathcal{Z} = z_i\}}}{E\{\sqrt{V\{\mathcal{X}|\mathcal{Z}\}}\}}.$$

2. HOMEWORK1: on peut se convaincre que les idées développées dans cette section sont directement transposables au cas où la population mère est finie et que les tirages sont effectués sans remise.

3. Lorsque $\mathcal{Z}$ est une variable continue, il est nécessaire de la discrétiser (idéalement de façon optimale) pour exploiter cette idée de stratification.

CONCLUSION: Le sondage stratifié est absolument recommandé, dès lors que les information disponibles le rendent possible!

HOMEWORK2: considérons le cas où en plus de connaître a priori les valeurs de $P(\mathcal{Z}) = z_i$ et les valeurs de $V\{\mathcal{X}|\mathcal{Z} = z_i\}, \forall i = 1, \ldots, k$, on sait aussi a priori que l'espérance conditionnelle de $\mathcal{X}$ est constante (i.e. que $E\{\mathcal{X}|\mathcal{Z} = z_i\} = E\{\mathcal{X}\}, \forall i = 1, \ldots, k$); quel serait alors la façon la plus efficace de sonder et de construire l'estimateur ?

Estimation par intervalle

- Il est souvent plus réaliste et plus intéressant de fournir un renseignement du type $a < \theta < b$, plutôt que de simplement écrire $\hat{\theta} = c$.
- Fournir un tel intervalle $[a; b]$ s'appelle donner une estimation par intervalle de θ, ou une estimation ensembliste.
- Dans le cas où le paramètre θ est vectoriel, on construira une "région", et on parle d'estimation par région.

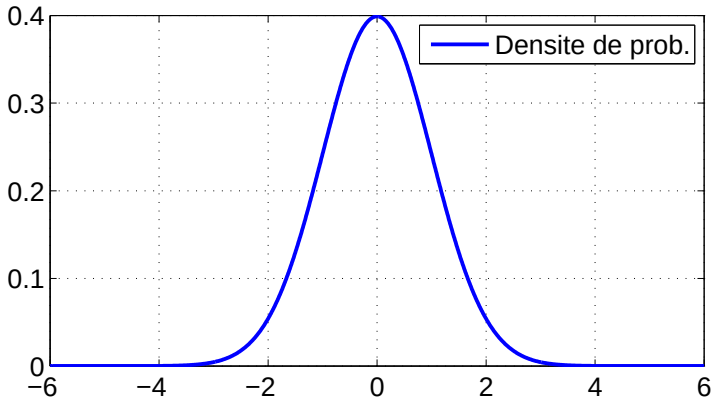
Discussion

- De façon générale, il n'est pas possible de déterminer un sous-ensemble non-trivial de Θ qui contienne avec certitude la vraie valeur du paramètre de population θ étudié.
- Par contre, pour une valeur supposée θ du paramètre, on peut en principe à partir de la loi $F_{\mathcal{X}}(x;\theta)$, de la connaissance de n, et de la définition de l'estimateur $T(\cdot)$ utilisé, déterminer un intervalle $[t_1; t_2]$ de probabilité $1-\alpha$ pour T, i.e. tel que

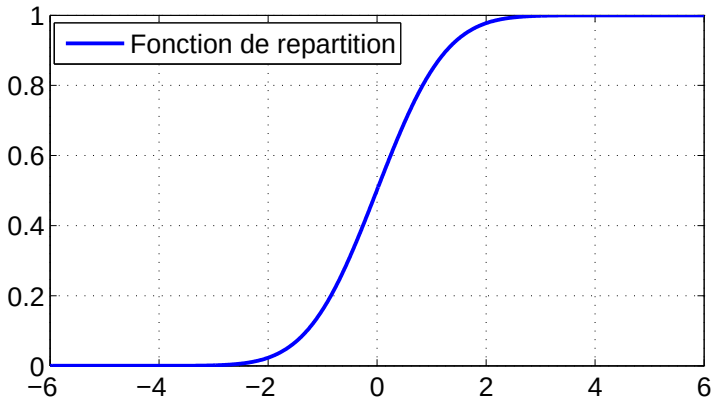
$$P(\mathcal{T}_n \in [t_1; t_2]) = 1 - \alpha.$$

  Notons par $t_1(\theta)$ et $t_2(\theta)$ les bornes ainsi définies (fonctions aussi de α et de n, ainsi que de l'estimateur $T(\cdot)$ utilisé).
- NB: en pratique on utilise généralement des intervalles symétriques centrés sur θ, i.e. $t_1 = \theta - \Delta\theta$ et $t_2 = \theta + \Delta\theta$.

Rappel: loi Gaussienne centrée réduite




```
x = -6:0.01:6;
y = 1 / (sqrt(2 * pi)) * exp(-0.5 * x .^2);
plot(x, y, 'LineWidth', 2);
hold on;
```



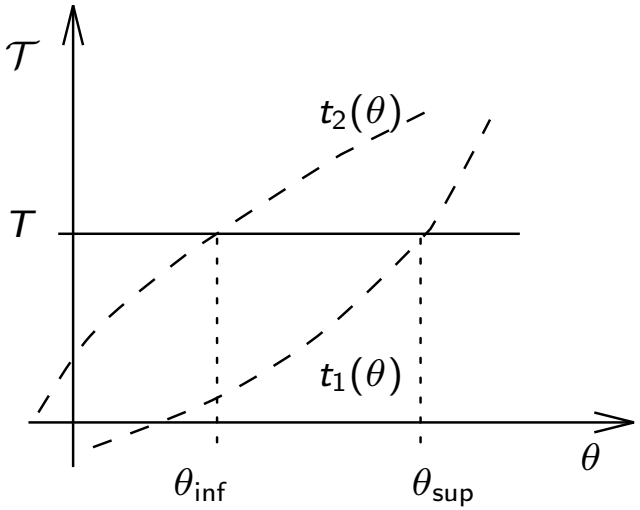

```
z = zeros(size(x));
for i=1:length(x)
    z(i) = sum(y(1:i));
end
z = z * (x(2) - x(1));
plot(x, z, 'LineWidth', 2);
```

$$f_{\mathcal{X}}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad F_{\mathcal{X}}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z, \quad \text{et } u_{\alpha/2} = F_{\mathcal{X}}^{-1}(1-\frac{\alpha}{2}), \text{ p.ex. } u_{0.025} = 1.96.$$

Exemple de calcul d'intervalle de probabilité

- Supposons que $\mathcal{X} \sim \mathcal{N}(\theta; \sigma^2)$ et que nous utilisons comme estimateur $T(\cdot)$ de θ la moyenne d'échantillon m_x, et que la taille d'échantillon vaut n.
- On a $m_x \sim \mathcal{N}(\theta; \sigma^2_{m_x})$, avec $\sigma_{m_x} = \frac{\sigma}{\sqrt{n}}$.
- On peut donc, par exemple, affirmer que $P(m_x \in [\theta - 1.96\frac{\sigma}{\sqrt{n}}; \theta + 1.96\frac{\sigma}{\sqrt{n}}) = 1 - 0.05 = 0.95$.
  (Puisque $u_{0.025} = 1.96$.)
- On aurait donc ici que our $\alpha = 0.05$, $t_1(\theta) = \theta - 1.96\frac{\sigma}{\sqrt{n}}$, et $t_2(\theta) = \theta + 1.96\frac{\sigma}{\sqrt{n}}$.
- Pour autant qu'on connaisse a priori la valeur de σ, on peut donc utiliser ces formules pour déterminer les fonctions $t_1(\theta)$ et $t_2(\theta)$.
- Si on ne connait pas la valeur de σ, il faut faire autrement...

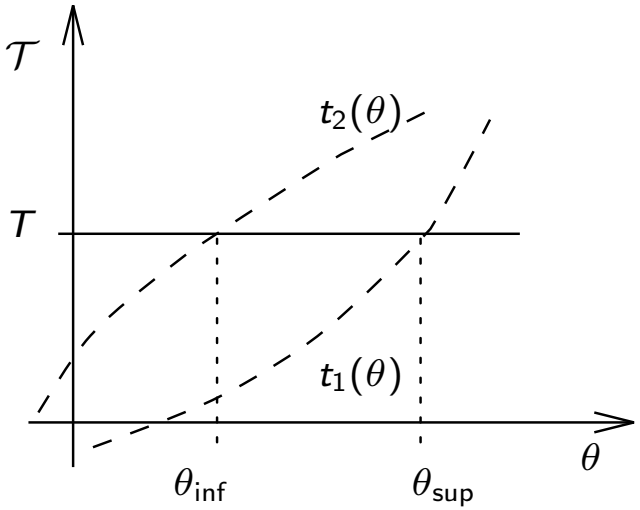
Principe général de construction d'intervalles de confiance



NB: $\theta_{\text{inf}} = t_2^{-1}(T)$ et $\theta_{\text{sup}} = t_1^{-1}(T)$.

1. $\forall \theta \in \Theta$ (et pour la valeur utilisée de n et α, et pour notre estimateur $T(\cdot)$) on détermine les valeurs $t_1(\theta)$ et $t_2(\theta)$.
2. On en déduit l'intervalle de confiance $[\theta_{\text{inf}}; \theta_{\text{sup}}]$, de valeurs de θ telles que $T \in [t_1(\theta), t_2(\theta)]$, où $T = \mathcal{T}(\mathbf{D}_n)$.
3. Suggestion: faire ce graphique, lorsque $\mathcal{X} \sim \mathcal{N}(0; 1)$, et $n = 4$, avec les fonctions t_1 et t_2 indiquées au slide précédent, et en supposant que $m_x = 0.1$.

Sensibilité de l'intervalle de confiance par rapport n et α



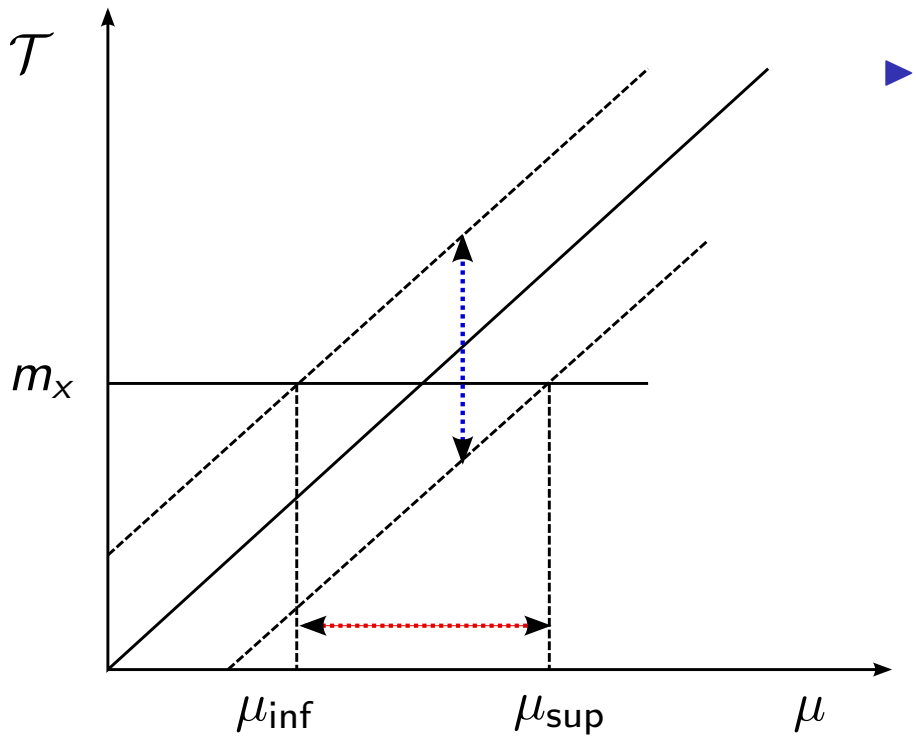
1. Lorsque l'on fait décroître α la courbe $t_1(\theta)$ descend, et la courbe $t_2(\theta)$ monte: l'intervalle de confiance (pour une même valeur observée de T) s'élargit donc. Typiquement, lorsque $\alpha = 0$ l'intervalle devient trivial, c'est-à-dire qu'il devient identique à l'ensemble initialement postulé Θ.
2. Si l'estimateur T est tel que son biais et sa variance diminuent lorsque n augmente (normalement c'est le cas), alors si on augmente la taille n de l'échantillon, les deux courbes $t_1(\theta)$ et $t_2(\theta)$ se rapprochent de la bissectrice $T = \theta$, et la largeur de l'intervalle de confiance diminue donc (pour une même observation T).

Interprétation

- Pour une valeur fixée quelconque de $\theta^* \in \Theta$, nous avons la garantie que pour une proportion $1 - \alpha$ des échantillons que nous pourrions observer, cette procédure produira un intervalle qui contient la valeur θ^*. (Pourquoi, au fait ?).
- Par conséquent, l'intervalle de confiance est une statistique bi-dimensionnelle extraite d'un échantillon dont la probabilité de contenir la vraie valeur θ^* du paramètre θ est de $1 - \alpha$, quelque soit la vraie valeur θ^* de ce paramètre.
- Autrement dit, le statisticien (ou bien l'ordinateur) qui affirme systématiquement que $\theta^* \in$ "cet intervalle de confiance" se trompera seulement avec une probabilité de α (pour autant que les autres hypothèses soient bien vérifiées).

Exemple: paramètre μ d'une loi $\mathcal{N}(\mu; \sigma^2)$

Cas 1: σ est connu.



- On utilise l'estimateur m_x.
- $m_x \sim \mathcal{N}(\mu; \frac{\sigma^2}{n})$
- L'intervalle de probabilité de m_x à $1-\alpha$ est

$$\mu - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq m_x \leq \mu + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

et l'intervalle de confiance de μ

$$m_x - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq m_x + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

P.ex., lorsque $1-\alpha = 0.95$, $u_{\alpha/2} = 1.96$

Exemple: paramètre μ d'une loi $\mathcal{N}(\mu; \sigma^2)$

Cas 2: σ est inconnu.

- On utilise le fait que $T = \frac{m_x - \mu}{s_{n-1}/\sqrt{n}}$ suit une loi de Student à $(n-1)$ degrés de liberté. (Voir http://en.wikipedia.org/wiki/Student's_t-distribution.)
- L'intervalle de probabilité de T à $1-\alpha$ est

$$-t_{\alpha/2} \leq T \leq t_{\alpha/2}$$

et l'intervalle de confiance de μ

$$m_x - t_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq m_x + t_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}.$$

- Remarque: lorsque $n > 30$, la loi de Student à $n-1$ degrés de liberté se confond avec la loi normale centrée réduite, et l'intervalle de confiance obtenu est alors identique à celui du "Cas 1" ou on remplace σ par son estimée s_{n-1}.

Exemple: paramètre μ d'une loi quelconque (σ fini)

- Lorsque n est suffisamment grand, le théorème central-limite implique que m_x suit encore une loi normale centrée sur μ et de variance σ^2/n.
- Les deux estimateurs précédents peuvent alors être utilisés, selon que l'on connait σ ou non.
- Attention: la valeur de n à partir de laquelle on a des résultats exacts dépend de la famille $F_{\mathcal{X}}(\cdot;\theta)$ de lois de la variable parente $\mathcal{X}$.
- Des raisonnements analogues permettent aussi de construire des intervalles de confiance d'autres types de paramètres, p.ex. la variance, les percentilles, etc.

Exemple: intervalle de confiance d'une proportion

- On utilise comme estimateur du paramètre p d'une variable de Bernoulli la fréquence relative f déduite de l'échantillon.
- On a que $nf \sim \mathcal{B}(n,p)$ (loi binomiale); et la loi de f se rapproche d'une loi normale $\mathcal{N}(p; \frac{p(1-p)}{n})$ lorsque $\min\{np, n(1-p)\} \geq 5$.
- On a donc l'intervalle de probabilité $1-\alpha$ pour f :

$$p - u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq f \leq p + u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

- NB: la largeur de l'intervalle de probabilité dépend de la valeur du paramètre à estimer, ce qui rend le calcul de l'intervalle de confiance plus compliqué (résolution d'une équation du second degré).
- Cependant, on peut utiliser la formule approchée suivante:

$$f - u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}.$$

Démarche bayesienne

Construction d'intervalles de probabilité pour θ

- On calcule la loi a posteriori $f_{\theta|\mathbf{D}}$ à partir de f_θ de $\mathbf{D}$ et de $f_{\mathcal{X}}(\cdot|\theta)$ comme expliqué précédemment.
- On détermine un intervalle de probabilité pour θ au niveau $(1-\alpha)$, à partir de la loi a posteriori $f_{\theta|\mathbf{D}}$.
- Lorsque n est suffisamment grand, et pour les choix de f_θ non-nuls sur Θ, on obtient des intervalles identiques à ceux qu'on peut obtenir par l'approche classique décrite ci-avant.
- Lorsque n est faible, les intervalles obtenus sont sensibles au choix de f_θ, ce qui peut être une bonne chose si le choix de f_θ repose sur des connaissances physiques du problème, ou bien une mauvaise chose, si ce choix est fait de façon arbitraire...


Brett Ryder