Université d'Evry Val-d'Essonne Paris-Saclay **Année universitaire 2019-2020**
L2 SDV - MSV31

Estimation ponctuelle
Estimation par intervalle

VERSION PROVISOIRE du 27/09/19

Les savoir-faire pour les évaluations

1. Savoir ce qu'est un échantillon aléatoire, un échantillon, un estimateur, une estimation.
2. Savoir ce que sont des variables aléatoires i.i.d
3. Connaître les formules des estimateurs classiques (moyenne empirique, variance empirique, variance empirique corrigée)
4. Connaître espérance et variance de la moyenne empirique
5. Connaître l'espérance des variances empiriques (non corrigée et corrigée)
6. Connaître la définition du biais (et ce qu'est un estimateur sans biais ou asymptotiquement sans biais)
7. Connaîre l'erreur quadratique moyenne. Sa définition et le théorème et sa démonstration (= variance + biais au carré)
8. Savoir ce qu'est un estimateur consistant
9. Savoir que la méthode du maximum de vraissemblance permet de construire des estimateurs à partir de la loi de probabilité

Résumé et compléments

Calcul de biais et erreur quadratique moyenne

Définition

Pour $\hat{\theta}_n$ un estimateur de θ on définit :
le biais :

$$b(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$

Un estimateur est **sans biais** (ou non biaisé) lorsque $b(\hat{\theta}_n) = 0$.
Un estimateur est dit **asymptotiquement sans biais** si

$$\lim_{n \to +\infty} b(\hat{\theta}_n) = 0$$

On appelle **erreur quadratique moyenne** la quantité :

$$EQM(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = V(\hat{\theta}_n) + [b(\hat{\theta}_n)]^2$$

Un estimateur est dit **consistant** si :

$$\lim_{n \to \infty} EQM(\hat{\theta}_n) = 0$$

Preuve de le relation :$\mathbb{E}[(\hat{\theta}_n - \theta)^2] = V(\hat{\theta}_n) + [b(\hat{\theta}_n)]^2$

$$\mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] = \mathbb{E}\left[\left((\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]) + (\mathbb{E}[\hat{\theta}_n] - \theta)\right)^2\right]$$

$$\begin{aligned}\mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] &= \mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2 + 2(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta) + (\mathbb{E}[\hat{\theta}_n] - \theta)^2\right] \\ &= \mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2\right] + 2\mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta)\right] + \mathbb{E}\left[(\mathbb{E}[\hat{\theta}_n] - \theta)^2\right]\end{aligned}$$

Il reste à simplifier ces trois termes. Le premier terme est la définition de la variance de $\hat{\theta}_n$. Le dernier terme est l'espérance d'une constante puisque θ et $\mathbb{E}[\hat{\theta}_n]$ sont des nombres. Ainsi :

$$\mathbb{E}\left[(\mathbb{E}[\hat{\theta}_n] - \theta)^2\right] = (\mathbb{E}[\hat{\theta}_n] - \theta)^2 = b(\hat{\theta}_n)^2$$

Il reste à monter que le deuxième terme est nul.
Comme $(\mathbb{E}[\hat{\theta}_n] - \theta)$ est une constante on peut écrire :

$$\mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta)\right] = (\mathbb{E}[\hat{\theta}_n] - \theta)\mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])\right]$$

Or

$$\mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])\right] = \mathbb{E}[\hat{\theta}_n] - \mathbb{E}[\hat{\theta}_n] = 0$$

Soit $(X_1, ..., X_n)$ un échantillon aléatoire de **loi quelconque** tel que

$$\mathbb{E}[X_i] = \mu \quad \text{et} \quad V(X_i) = \sigma^2$$

Exemple avec la moyenne empirique :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

On calcule $\mathbb{E}[\overline{X}]$:

$$\mathbb{E}[\overline{X}] = \mathbb{E}[\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n] = \frac{1}{n}\mathbb{E}[X_1] + \cdots + \frac{1}{n}\mathbb{E}[X_n] = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu$$

ainsi, on obtient pour le biais

$$b(\overline{X}) = \mathbb{E}[\overline{X}] - \mu = 0$$

$\overline{X}$, la moyenne empirique, est un estimateur sans biais de l'espérance μ.

Pour calculer l'erreur quadratique moyenne, on calcule $V(\overline{X})$:

$$V(\overline{X}) = V(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n) = \left(\frac{1}{n}\right)^2 V(X_1) + \cdots + \left(\frac{1}{n}\right)^2 V(X_n) = \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Ainsi, on obtient pour l'erreur quadratique moyenne :

$$EQM(\overline{X}) = V(\overline{X}) + b(\overline{X})^2 = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$$

L'estimateur de la moyenne empirique est donc consistant.

Exemple avec la variance empirique :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

On reproduit le même type de calcul que pour la moyenne empirique, c'est seulement un peu plus compliqué.

On commence par simplifier S^2 :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}n\overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\overline{X} + \overline{X}^2$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2$$

En utilisant la formule de la variance (théorème de König-Huygens) on a :

$$V(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2$$

et donc

$$\mathbb{E}[X_1^2] = V(X_1) + \mathbb{E}[X_1]^2 = \sigma^2 + \mu^2$$

On a de même :

$$\mathbb{E}[\overline{X}^2] = V(\overline{X}) + \mathbb{E}[\overline{X}]^2 = \frac{\sigma^2}{n} + \mu^2$$

On rassemble maintenant les calculs :

$$\mathbb{E}[S^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\overline{X}^2] = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

Ainsi

$$\mathbb{E}[S^2] = \left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\sigma^2$$

On voit alors que l'estimateur variance empirique est biaisé puisque :

$$b(S^2) = \mathbb{E}[S^2] - \sigma^2 = -\frac{\sigma^2}{n}$$

D'où l'idée de construire un estimateur de la variance corrigé, appelé variance empirique corrigée et notée S^{*2} et tel que :

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

On a alors un estimateur sans biais de la variance puisque :

$$b(S^{*2}) = \mathbb{E}[S^{*2}] - \sigma^2 = \mathbb{E}[\frac{n}{n-1}S^2] - \sigma^2 = \frac{n}{n-1}\mathbb{E}[S^2] - \sigma^2 = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 - \sigma^2 = 0$$

Intervalle de probabilité pour la moyenne avec une loi normale de variance connue

On observe des réalisations $x_1, ..., x_n$ des variables aléatoires $X_1, ..., X_n$ qui ont pour loi $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 connu mais μ inconnu.
Le but du jeu est de construire un intervalle autour de $\overline{x}$ (puisque la meilleure estimation de la moyenne μ est $\overline{x}$) dans lequel μ a une forte probabilité d'être présent.

L'intervalle de confiance sera alors du type $[\overline{x} - c, \overline{x} + c]$ avec c un nombre à trouver. Cet intervalle est symétrique autour de $\overline{x}$ puisqu'il n'y a aucune raison de privilégier un côté plus que l'autre et que $\overline{X} - \mu$ est une loi symétrique. Comme on évalue une probabilité, on utilise l'estimateur de la moyenne empirique $\overline{X}$ et on cherche c tel que

$$P(\overline{X} - c < \mu < \overline{X} + c) = 1 - \alpha$$

On peut remanier cette inégalité :

$$P(-c < \mu - \overline{X} < c) = 1 - \alpha$$

ou encore

$$P(|\mu - \overline{X}| < c) = 1 - \alpha$$

ou encore

$$P(|\overline{X} - \mu| < c) = 1 - \alpha$$

On sait cependant que

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

On va utiliser cette variable pour trouver c.

On recherche ainsi u tel que

$$P(|\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}| < u) = 1 - \alpha$$

Avec $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$ on écrit :

$$P(|Z| < u) = 1 - \alpha$$

qui s'écrit aussi

$$P(-u < Z < u) = 1 - \alpha$$

ou bien

$$P(Z < u) - P(Z < -u) = 1 - \alpha$$

ou bien par symétrie de la loi centrée

$$P(Z < u) - P(Z > u) = 1 - \alpha$$

ou bien

$$P(Z < u) - (1 - P(Z < u)) = 1 - \alpha$$

ce qui donne finalement

$$P(Z < u) = 1 - \frac{\alpha}{2}$$

Avec cette relation, on peut dire que u est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. On l'écrit $u = q_{1-\frac{\alpha}{2}}$. Il est connu dans les tables ou par des logiciels (par exemple R).

Ainsi

$$P(|\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}| < q_{1-\frac{\alpha}{2}}) = 1-\alpha$$

On va alors transformer l'inégalité pour faire apparaître un intervalle autour de μ.

$$P(-q_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < q_{1-\frac{\alpha}{2}}) = 1-\alpha$$

ou

$$P(-q_{1-\frac{\alpha}{2}} < \frac{\mu-\overline{X}}{\frac{\sigma}{\sqrt{n}}} < q_{1-\frac{\alpha}{2}}) = 1-\alpha$$

ou

$$P(-q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu-\overline{X} < q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

et on obtient

$$P(\overline{X}-q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

soit

$$P(\mu \in [\overline{X}-q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}+q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]) = 1-\alpha$$

L'intervalle de probabilité recherché est alors

$$[\overline{X}-q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}+q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

et l'intervalle de confiance

$$[\overline{x}-q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x}+q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$