Grand classique

Outils de mesure de la qualité d'un estimateur

Dans ce document, nous introduisons des outils, tous devenus hors programme, qui permettent de mesurer la qualité d'un estimateur.
On considère une variable aléatoire X définie sur un certain espace probabilisé ainsi que $(X_1, \dots, X_n)$ un n-échantillon de X (avec $n \in \mathbb{N}^*$).

Biais et risque quadratique d'un estimateur

L'idée générale est les réalisations de l'estimateur T_n soient proches de θ. Pour mesurer l'écart entre T_n et θ, on peut utiliser les quantités suivantes : $T_n - \theta$, $|T_n - \theta|$, $(T_n - \theta)^2$,...
Mais bien évidemment, ce qui nous intéresse le plus est l'écart moyen... D'où les définitions suivantes :

Définitions 1 — **Biais, risque quadratique**

Soit T_n un estimateur de θ.

D1 Si T_n admet une espérance, le **biais de l'estimateur T_n**, noté $b_\theta(T_n)$, est le réel défini par :

$$b_\theta(T_n) = \mathbb{E}_\theta(T_n - \theta)$$

D2 Si T_n admet une variance, le **risque quadratique de l'estimateur T_n**, noté $r_\theta(T_n)$, est le réel défini par :

$$r_\theta(T_n) = \mathbb{E}_\theta\left((T_n - \theta)^2\right)$$

Vocabulaire
- Si $b_\theta(T_n) = 0$, on dit que l'estimateur T_n est **sans biais**.
- Si $\lim\limits_{n\to+\infty} b_\theta(T_n) = 0$, on dit que l'estimateur T_n est **asymptotiquement sans biais**.

Propriétés 1

Avec les notations précédentes :

P1 Si T_n admet une espérance, alors : $b_\theta(T_n) = \mathbb{E}_\theta(T_n) - \theta$.

P2 Si T_n admet une variance, alors : $r_\theta(T_n) = \mathbb{V}_\theta(T_n) + \left(b_\theta(T_n)\right)^2$.

Remarque
Si T_n est sans biais, alors $r_\theta(T_n) = \mathbb{V}_\theta(T_n)$.

⋆ Démonstration :

P1. Supposons que T_n possède une espérance. Ainsi :

$$\begin{aligned} b_\theta(T_n) &= \mathbb{E}_\theta(T_n - \theta) \\ &= \mathbb{E}_\theta(T_n) - \theta \end{aligned}$$
↙ linéarité de l'espérance

P2. Supposons que T_n possède une variance. Ainsi :

$$\begin{aligned} r_\theta(T_n) &= \mathbb{E}_\theta\left((T_n - \theta)^2\right) \\ &= \mathbb{E}_\theta\left(T_n^2 - 2\theta T_n + \theta^2\right) \\ &= \mathbb{E}_\theta(T_n^2) - 2\theta\mathbb{E}_\theta(T_n) + \theta^2 \\ &= \mathbb{V}_\theta(T_n) + \left(\mathbb{E}_\theta(T_n)\right)^2 - 2\theta\mathbb{E}_\theta(T_n) + \theta^2 \\ &= \mathbb{V}_\theta(T_n) + \left(\mathbb{E}_\theta(T_n) - \theta\right)^2 \\ &= \mathbb{V}_\theta(T_n) + \left(b_\theta(T_n)\right)^2 \end{aligned}$$
↙ linéarité de l'espérance, toutes existent
↙ formule de Koenig-Huygens
↙ point précédent

⋆

Dans l'idéal, on cherche à obtenir un estimateur dont le biais et le risque quadratique sont les plus proches de 0 possible, l'idéal étant un estimateur sans biais et de variance minimale... Mais celui-ci n'existe pas toujours, ou n'est pas simple à trouver.
De façon générale, pour comparer deux estimateurs :

- s'ils ont même biais, on préférera celui de risque quadratique minimal,
- sinon, on pourra parfois préférer un estimateur biaisé à un estimateur sans biais si son risque quadratique est plus faible que la variance de l'estimateur sans biais.

Exemple 1

On suppose $n \geqslant 2$. Notons $V_n = \dfrac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2$ la variance empirique de X.

- Justifions que V_n possède une espérance et déterminons-la.

$$\begin{aligned} V_n &= \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}(X_k^2 - 2\overline{X_n}X_k + \overline{X_n}^2) \\ &= \frac{1}{n}\sum_{k=1}^{n}X_k^2 - 2\overline{X_n}\frac{1}{n}\sum_{k=1}^{n}X_k + \overline{X_n}^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}X_k^2 - \overline{X_n}^2 \end{aligned}$$

↳ $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n}X_k$

Or :

- ✓ pour tout $k \in [\![1;n]\!]$, X_k^2 admet une espérance car X admet une variance ;
- ✓ $\overline{X_n}^2$ admet une espérance (car $\overline{X_n}$ admet une variance comme combinaison linéaire de telles variables aléatoires).

Par conséquent, V_n est une combinaison linaire de variables aléatoires admettant une espérance ; V_n admet donc une espérance et :

$$\begin{aligned} \mathbb{E}(V_n) &= \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n}X_k^2 - \overline{X_n}^2\right) \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k^2) - \mathbb{E}(\overline{X_n}^2) \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X) - \mathbb{E}(\overline{X_n}^2) \\ &= \mathbb{E}(X^2) - \mathbb{E}(\overline{X_n}^2) \\ &= \mathbb{V}(X) + \mathbb{E}(X)^2 - \mathbb{V}(\overline{X_n}) - \mathbb{E}(\overline{X_n}) \\ &= \frac{n-1}{n}\mathbb{V}(X) \end{aligned}$$

↳ linéarité de l'espérance, toutes existent

↳ les X_k ont même loi que X

↳ formule de Koenig-Huygens

↳ $\mathbb{E}(\overline{X_n}) = \mathbb{E}(X)$ et $\mathbb{V}(\overline{X_n}) = \frac{1}{n}\mathbb{V}(X)$

Remarque
Je ne détaille pas ces deux calculs, très classiques.

Conclusion : la variance empirique est un estimateur de $\mathbb{V}(X)$ d'espérance $\frac{n-1}{n}\mathbb{V}(X)$.

- Donnons, à partir de V_n, un estimateur V_n' qui soit un estimateur sans biais de $\mathbb{V}(X)$.
Posons, pour $n \geqslant 2$:

$$V_n' = \frac{n}{n-1}V_n \qquad = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2$$

- ✓ V_n' est un estimateur de $\mathbb{V}(X)$ (car $X_1, \dots, X_n$ sont indépendantes et de même loi ; et que l'expression de V_n' ne fait pas apparaître $\mathbb{V}(X)$),
- ✓ et d'après le calcul précédent :

$$\mathbb{E}(V_n') = \mathbb{V}(X)$$

Conclusion : $\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2$ est un estimateur sans biais de $\mathbb{V}(X)$.

☞ Pour info...
De nombreux logiciels informatiques (et des calculatrices) utilisent l'estimateur V_n' (**variance empirique corrigée**) pour la variance plutôt que V_n.

Estimateur convergent

Définition 2 — **Estimateur convergent**

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de θ.
On dit que $(T_n)_{n\in\mathbb{N}^*}$ est convergente (ou plus simplement que l'estimateur T_n est convergent) lorsque :

$$\forall \theta \in \Theta,\ \forall \varepsilon > 0,\ \lim_{n\to+\infty}\mathbb{P}_\theta\big(|T_n - \theta| \geqslant \varepsilon\big) = 0$$

Confusion d'objets !
On verra parfois la confusion entre estimateur et suite d'estimateurs.

Autrement dit :
T_n est convergent ssi $T_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$

Propriété 2 — **Condition suffisante d'estimateur convergent**

Soit T_n un estimateur de θ admettant une variance.
Si $\lim\limits_{n\to+\infty} r_\theta(T_n) = 0$, alors T_n est convergent.

⋆ Démonstration : Supposons que $\lim\limits_{n\to+\infty} r_\theta(T_n) = 0$. Montrons : $\forall \varepsilon > 0,\ \lim\limits_{n\to+\infty} \mathbb{P}_\theta\big([|T_n - \theta| \geqslant \varepsilon]\big) = 0$.
Soit $\varepsilon > 0$. Soit $n \in \mathbb{N}^*$.
La variable aléatoire $(T_n - \theta)^2$:

- ✓ est à valeurs positives,
- ✓ admet une espérance (car T_n admet une variance).

Ainsi, d'après l'inégalité de Markov :

$$\forall a > 0,\ \mathbb{P}_\theta\big([(T_n - \theta)^2 \geqslant a]\big) \leqslant \frac{\mathbb{E}_\theta\big(T_n - \theta)\big)}{a}$$

Avec $a = \varepsilon^2 > 0$:

$$\mathbb{P}_\theta\big([(T_n - \theta)^2 \geqslant \varepsilon^2]\big) \leqslant \frac{r_\theta(T_n)}{\varepsilon^2}$$

Mais, par stricte croissance de $\sqrt{\cdot}$ sur $\mathbb{R}^+$ et comme $|\varepsilon| = \varepsilon$ (car $\varepsilon > 0$) :

$$[(T_n - \theta)^2 \geqslant \varepsilon^2] = [|T_n - \theta| \geqslant \varepsilon]$$

On a ainsi établi :

$$\forall n \in \mathbb{N}^*,\ 0 \leqslant \mathbb{P}_\theta\big([|T_n - \theta| \geqslant \varepsilon]\big) \leqslant \frac{r_\theta(T_n)}{\varepsilon^2}$$

Or $\lim\limits_{n\to+\infty} r_\theta(T_n) = 0$. Donc, par théorème d'encadrement :

$$\lim_{n\to+\infty} \mathbb{P}_\theta\big([|T_n - \theta| \geqslant \varepsilon]\big) = 0$$

Conclusion : si $\lim\limits_{n\to+\infty} r_\theta(T_n) = 0$, alors T_n est convergent. ⋆

Un estimateur sans biais garantit l'absence d'erreur en moyenne, mais peut produire de très mauvaises estimations ponctuelles. Au contraire, un estimateur convergent, biaisé ou non, sera d'autant plus fiable que la taille de l'échantillon est grande. En général, on préfère ces derniers.

Propriété 3 — **Qualité de la moyenne empirique**

Soient X une variable aléatoire admettant une espérance et une variance ainsi que $(X_1, \dots, X_n)$ un n-échantillon de X.
La moyenne empirique de $X_1, \dots, X_n$ est un estimateur sans biais et convergent de $\mathbb{E}(X)$.

⋆ Démonstration :

- Remarquons déjà que :
  - ✓ $X_1, \dots, X_n$ sont indépendantes et suivent la même loi,
  - ✓ $\overline{X_n}$ est fonction de $X_1, \dots, X_n$ dont l'expression ne fait pas apparaître $\mathbb{E}(X)$.

  Conclusion : $\overline{X_n}$ est un estimateur de $\mathbb{E}(X)$.
- Soit $n \in \mathbb{N}^*$. La variable aléatoire $\overline{X_n}$ est une combinaison linéaire de variables aléatoires admettant une espérance, elle admet donc une espérance. Puis :

$$\begin{aligned}\mathbb{E}(\overline{X_n}) &= \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) && \text{↙ linéarité de l'espérance, toutes existent}\\ &= \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(X_k) && \text{↙ } \forall k \in [\![1;n]\!],\ \mathbb{E}(X_k) = \mathbb{E}(X)\\ &= \mathbb{E}(X)\end{aligned}$$

  Conclusion : $\overline{X_n}$ est un estimateur sans biais de $\mathbb{E}(X)$.
- Enfin, puisque les variables aléatoires $X_1, \dots, X_n$ sont indépendantes, de même espérance et de même variance (car ont la même loi), d'après la loi faible des grands nombres :

$$\forall \varepsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}\big([|\overline{X_n} - \mathbb{E}(X)| \geqslant \varepsilon]\big) = 0$$

  Conclusion : $\overline{X_n}$ est un estimateur convergent de $\mathbb{E}(X)$.

⋆

Propriété 4 — **Qualité de la variance empirique corrigée**

Soient X une variable aléatoire admettant un moment d'ordre 4 ainsi que $(X_1, ..., X_n)$ un n-échantillon de X.
La variance empirique corrigée de $X_1, ..., X_n$ (V'_n de l'exemple 1) est un estimateur sans biais et convergent de $\mathbb{V}(X)$.

Remarque

Cet estimateur permettrait d'obtenir une estimation du second paramètre d'une loi normale par exemple.

⋆ Démonstration :

- On vérifie déjà que V'_n est un estimateur de $\mathbb{V}(X)$.
- Par construction $\mathbb{E}(V'_n) = \mathbb{V}(X)$: $V_n{}'$ est sans biais.
- Et on a :

$$r_\theta(V'_n) = \mathbb{V}(V'_n) = ... = \frac{1}{n}\mathbb{E}\left((X - \mathbb{E}(X))^4\right) - \frac{n-3}{n(n-1)}\mathbb{V}(X)^2$$

Ainsi : $\lim\limits_{n\to+\infty} r_\theta(V'_n) = 0$ et on conclut sur la convergence en utilisant l'inégalité de Markov et le théorème d'encadrement...

⋆ **☞ Pour info...**

On pourra aller s'entraîner sur le sujet `ESSEC 2009 E2` : ici.