HLMA408: Traitement des données

Échantillonnage aléatoire

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Sommaire

Sommaire

Retour sur l'étude `babies23.data`

À San Francisco[1], 1236 naissances ont été répertoriées au cours d'une année à la *Kaiser Foundation Health Plan*[2]

- Quelle proportion de mère fume du tabac?
- Parmi les fumeuses, quelle est la consommation quotidienne?
- Cette consommation influe-t-elle sur le développement de l'enfant?

Faire une enquête exhaustive est compliqué et prend du temps[3]
$\Longrightarrow$ sondage sur un échantillon

---

[1] D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.

[2] dont seulement 1226 ont donné leur consommation de tabac

[3] l'étude a pris un an!

Retour sur l'étude `babies23.data`

À San Francisco[1], 1236 naissances ont été répertoriées au cours d'une année à la *Kaiser Foundation Health Plan*[2]

- Quelle proportion de mère fume du tabac?
- Parmi les fumeuses, quelle est la consommation quotidienne?
- Cette consommation influe-t-elle sur le développement de l'enfant?

Faire une enquête exhaustive est compliqué et prend du temps[3]
$\Longrightarrow$ sondage sur un échantillon

---

(1) D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.

(2) dont seulement 1226 ont donné leur consommation de tabac

(3) l'étude a pris un an!

Questions abordées

- Comment **estimer** une moyenne sur un échantillon ?
- Comment **mesurer l'erreur** introduite par l'**échantillonnage** ?
- Comment choisir le nombre de naissances à recenser ?

Sommaire

Un peu de vocabulaire

- **Population (totale)**: N, peut être grand (voire infini)
- **Échantillon** : sous partie de la population
- **Paramètre** : grandeur définie sur la population que l'on cherche à estimer
- **Taille de l'échantillon**: nombre d'individus échantillonné, n
- **Statistique**: quantité aléatoire calculée sur l'échantillon (fluctue en fonction de l'échantillon)

Un peu de vocabulaire

- **Population (totale)**: N, peut être grand (voire infini)
- **Échantillon** : sous partie de la population
- **Paramètre** : grandeur définie sur la population que l'on cherche à estimer
- **Taille de l'échantillon**: nombre d'individus échantillonné, n
- **Statistique**: quantité aléatoire calculée sur l'échantillon (fluctue en fonction de l'échantillon)

Définition

Une **statistique** qui permet d'estimer un paramètre s'appelle un **estimateur** (du paramètre sous-jacent)

Rem: on utilise souvent la notation "chapeau", *e.g.*, $\hat{x}_n$ pour désigner un estimateur

Échantillonnage aléatoire simple

- Choisir n individus parmi les N de la population totale, de façon aléatoire et uniforme
- On s'interdit de choisir deux fois le même individu dans l'échantillon (sorte de tirage sans remise)
- en général $n \ll N$

Rem: différent de l'échantillonnage par strates (sondages, etc.)

Dénombrement

Prenons l'exemple de la base de données `babies23.data`:

- population totale: $N = 1226$ (valeurs manquantes éliminées)
- échantillon: $n = 91$ (choix arbitraire ici)

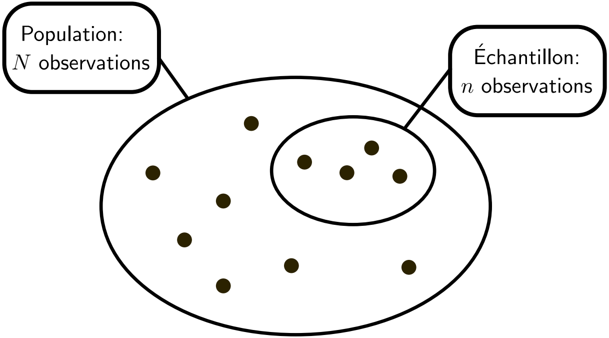
Combien y a-t-il d'échantillons possibles de taille 91 sur une population de taille 1226 ?

Dénombrement

Rappel: $N = 1226$, $n = 91$

On veut choisir n individus distincts parmi les N:

Il y a $\binom{1226}{91} \simeq 2.7 \times 10^{88}$ choix possibles.



Coefficient binomial

Nombre d'échantillons possibles:

$$\frac{1226 \times 1225 \times \cdots \times 1135}{91 \times 90 \times \cdots \times 1} = \frac{1226!}{1135! \times 91!}$$

Ce dernier nombre se note

$$\binom{1226}{91} \quad \text{ou} \quad C_{1226}^{91} \approx 2.7 \times 10^{88}$$

Rem: il y a $\binom{N}{n}$ échantillons possibles de taille n dans une population de N individus. En tirant un échantillon selon la loi uniforme, chaque échantillon a la même probabilité:

$$\frac{1}{\binom{N}{n}}$$

Code Python

```
df_babies = pd.read_csv("babies23.data",
                        skiprows=38, sep='\s+')
n_samples = 91
df_extract = df_babies.sample(n=n_samples)
```

Rem: voir le notebook `Echantillonage.ipynb`

Échantillonnage

- L'échantillonnage aléatoire simple met une structure aléatoire sur l'échantillon
- Différents échantillons ont des propriétés statistiques différentes, liées à la méthode d'échantillonnage
- Question : quelles sont ces propriétés ici?

Sommaire

Moyenne empirique

Notons x_i la variable qui vaut 1 (ou 0) si la i^{e} mère fume (ou non)

On cherche à estimer le **paramètre** taux de tabagisme chez la mère

$$\mu := \bar{x}_N = \frac{1}{1226}(x_1 + x_2 + \cdots + x_{1226}) \quad \text{inconnu}$$

à partir de l'échantillon prélevé

Technique classique: prendre la moyenne sur l'échantillon, qui vaut

$$\bar{x}_n := \frac{1}{91}(x_{i_1} + x_{i_2} + \cdots + x_{i_{91}})$$

où i_k est le numéro du k^{e} individu de l'échantillon

Rem: les i_k sont des variables aléatoires ici

Question: $\bar{x}_n$ (que l'on calcule) est-il éloigné de μ (inconnu) ?

Espérance

Rappel: les observations $x_{i_1}, x_{i_2}, \ldots, x_{i_{91}}$ sont aléatoires

Une moyenne par rapport à l'aléatoire s'appelle une **espérance**; celle du premier individu de notre échantillon vaut:

$$\begin{aligned}\mathbb{E}(x_{i_1}) &= \sum_{i=1}^{N} x_i \cdot \mathbb{P}(x_{i_1} = x_i) \\ &= \sum_{i=1}^{N} x_i \cdot \frac{1}{N} = \mu\end{aligned}$$

De même pour tous les individus de l'échantillon car ils sont supposés *i.i.d.* (**indépendants** et **identiquement** distribués)

$$\mathbb{E}(x_{i_1}) = \mathbb{E}(x_{i_2}) = \cdots = \mathbb{E}(x_{i_{91}}) = \mu$$

Espérance de la moyenne $\bar{x}_n$

L'espérance de la moyenne $\bar{x}_n$ sur l'échantillon est

$$\begin{aligned}\mathbb{E}(\bar{x}_n) &= \frac{1}{91}\Big(\mathbb{E}(x_{i_1}) + \mathbb{E}(x_{i_2}) + \cdots + \mathbb{E}(x_{i_{91}})\Big)\\ &= \frac{1}{91}\Big(91 \times \mu\Big) = \mu.\end{aligned}$$

En "espérance" (en moyenne vis-à-vis de l'aléa de l'échantillonnage), notre estimateur est égal au paramètre μ

Biais

Définition

Le **biais** d'un estimateur $\hat{x}_n$ de μ est noté $\mathbb{B}(\hat{x}_n)$ et vaut

$$\mathbb{B}(\hat{x}_n) := \mathbb{E}(\hat{x}_n) - \mu$$

Interprétation: le biais mesure l'erreur "systématique d'un estimateur"

Rem: pour la moyenne empirique $\mathbb{B}(\bar{x}_n) = 0$, on dit qu'elle est **sans biais** ou non biaisée

Variance de l'estimateur

Définition

La **variance** de l'estimateur $\hat{x}_n$ est définie comme

$$\mathbb{V}\mathrm{ar}(\hat{x}_n) = \mathbb{E}\left[\left(\hat{x}_n - \mathbb{E}(\hat{x}_n)\right)^2\right] = \mathbb{E}(\hat{x}_n^2) - (\mathbb{E}(\hat{x}_n))^2$$

Interprétation: la variance mesure la variation d'un estimateur autour de son espérance

Propriétés

Soit $\alpha \in \mathbb{R}$, et X,Y deux variables aléatoires indépendantes

$$\mathbb{V}\mathrm{ar}(X + \alpha) = \mathbb{V}\mathrm{ar}(X)$$

$$\mathbb{V}\mathrm{ar}(\alpha X) = \alpha^2\, \mathbb{V}\mathrm{ar}(X)$$

$$\mathbb{V}\mathrm{ar}(X + Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y)$$

Variance de la moyenne empirique

Variance (en supposant les x_{i_k} indépendants) :

$$\mathbb{V}\mathrm{ar}(\bar{x}_n) = \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{k=1}^{n} x_{i_k}\right)$$

Variance de la moyenne empirique

Variance (en supposant les x_{i_k} indépendants) :

$$\mathbb{V}\mathrm{ar}(\bar{x}_n) = \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{k=1}^{n} x_{i_k}\right) = \frac{1}{n^2}\,\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} x_{i_k}\right)$$

Variance de la moyenne empirique

Variance (en supposant les x_{i_k} indépendants) :

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\bar{x}_n) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{k=1}^{n} x_{i_k}\right) = \frac{1}{n^2}\,\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} x_{i_k}\right)\\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}\mathrm{ar}(x_{i_k}) = \frac{\mathbb{V}\mathrm{ar}(x_{i_1})}{n}\end{aligned}$$

Variance de la moyenne empirique

Variance (en supposant les x_{i_k} indépendants) :

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\bar{x}_n) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{k=1}^{n} x_{i_k}\right) = \frac{1}{n^2}\,\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} x_{i_k}\right) \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}\mathrm{ar}(x_{i_k}) = \frac{\mathbb{V}\mathrm{ar}(x_{i_1})}{n}\end{aligned}$$

<u>Conclusion</u>: la variance de la moyenne empirique est réduite d'un facteur n par rapport à celle d'une seule observation

Variance de la moyenne empirique

Variance (en supposant les x_{i_k} indépendants) :

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\bar{x}_n) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{k=1}^{n} x_{i_k}\right) = \frac{1}{n^2}\,\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} x_{i_k}\right) \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}\mathrm{ar}(x_{i_k}) = \frac{\mathbb{V}\mathrm{ar}(x_{i_1})}{n}\end{aligned}$$

<u>Conclusion</u>: la variance de la moyenne empirique est réduite d'un facteur n par rapport à celle d'une seule observation

<u>Rem</u> : $\sigma^2 := \mathbb{V}\mathrm{ar}\,(x_{i_1}) = \cdots = \mathbb{V}\mathrm{ar}\,(x_{i_n})$

Variance de la moyenne empirique

Variance (en supposant les x_{i_k} indépendants) :

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\bar{x}_n) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{k=1}^{n} x_{i_k}\right) = \frac{1}{n^2}\,\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} x_{i_k}\right) \\
&= \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{V}\mathrm{ar}(x_{i_k}) = \frac{\mathbb{V}\mathrm{ar}(x_{i_1})}{n}
\end{aligned}$$

<u>Conclusion</u>: la variance de la moyenne empirique est réduite d'un facteur n par rapport à celle d'une seule observation

<u>Rem</u> : $\sigma^2 := \mathbb{V}\mathrm{ar}\left(x_{i_1}\right) = \cdots = \mathbb{V}\mathrm{ar}\left(x_{i_n}\right) = \sum_{i=1}^{N}(x_i - \mu)^2\mathbb{P}(x_{i_1} = x_i)$

Variance de la moyenne empirique

Variance (en supposant les x_{i_k} indépendants) :

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\bar{x}_n) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{k=1}^{n} x_{i_k}\right) = \frac{1}{n^2}\,\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} x_{i_k}\right) \\ &= \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{V}\mathrm{ar}(x_{i_k}) = \frac{\mathbb{V}\mathrm{ar}(x_{i_1})}{n}\end{aligned}$$

Conclusion: la variance de la moyenne empirique est réduite d'un facteur n par rapport à celle d'une seule observation

Rem : $\sigma^2 := \mathbb{V}\mathrm{ar}\left(x_{i_1}\right) = \cdots = \mathbb{V}\mathrm{ar}\left(x_{i_n}\right) = \sum_{i=1}^{N}(x_i - \mu)^2\mathbb{P}(x_{i_1} = x_i)$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Écart quadratique moyen

Définition

L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

Écart quadratique moyen

Définition

L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

Écart quadratique moyen

Définition

L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu\big)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)\big)^2$$

Écart quadratique moyen

Définition

L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu\big)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)\big)^2$$

$$= \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n)\big)^2 + \underbrace{\mathbb{E}\big(\mathbb{B}(\hat{x}_n)\big)^2} + 2\,\underbrace{\mathbb{E}\big((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n)\big)}$$

Écart quadratique moyen

Définition

L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

$$\begin{aligned}\mathbb{E}(\hat{x}_n - \mu)^2 =& \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu\big)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)\big)^2 \\ =& \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n)\big)^2 + \underbrace{\mathbb{E}\big(\mathbb{B}(\hat{x}_n)\big)^2} + 2\,\underbrace{\mathbb{E}\big((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n)\big)}\end{aligned}$$

Écart quadratique moyen

Définition

L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

$$\begin{aligned}
\mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu\big)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)\big)^2 \\
&= \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n)\big)^2 + \underbrace{\mathbb{E}\big(\mathbb{B}(\hat{x}_n)\big)^2}_{\mathbb{B}(\hat{x}_n)^2} + 2\,\underbrace{\mathbb{E}\big(\big(\hat{x}_n - \mathbb{E}(\hat{x}_n)\big)\mathbb{B}(\hat{x}_n)\big)}_{\mathbb{B}(\hat{x}_n)\mathbb{E}\left(\hat{x}_n - \mathbb{E}(\hat{x}_n)\right)}
\end{aligned}$$

Écart quadratique moyen

Définition

L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

$$\begin{aligned}\mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu\big)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)\big)^2 \\ &= \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n)\big)^2 + \underbrace{\mathbb{E}\big(\mathbb{B}(\hat{x}_n)\big)^2}_{\mathbb{B}(\hat{x}_n)^2} + 2\underbrace{\mathbb{E}\big((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n)\big)}_{\mathbb{B}(\hat{x}_n)\mathbb{E}\left(\hat{x}_n - \mathbb{E}(\hat{x}_n)\right) = 0}\end{aligned}$$

Écart quadratique moyen

Définition

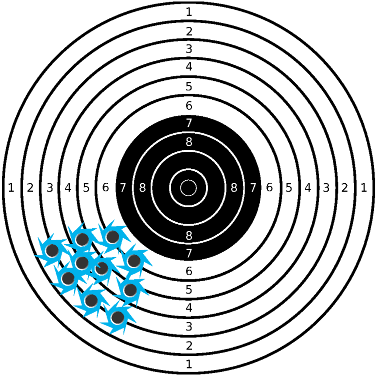
L'**écart quadratique moyen** d'un estimateur $\hat{x}_n$ d'un paramètre μ est donné par: $\mathbb{E}(\hat{x}_n - \mu)^2$

Rem: l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

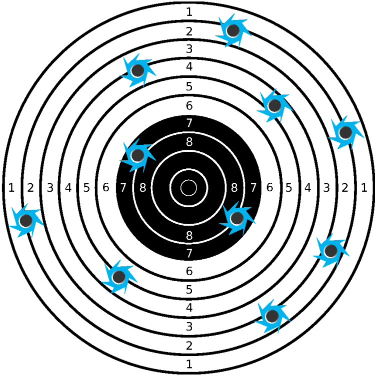
Propriété : $\boxed{\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2}$

$$\begin{aligned}
\mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu\big)^2 = \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)\big)^2 \\
&= \mathbb{E}\big(\hat{x}_n - \mathbb{E}(\hat{x}_n)\big)^2 + \underbrace{\mathbb{E}\big(\mathbb{B}(\hat{x}_n)\big)^2}_{\mathbb{B}(\hat{x}_n)^2} + 2\,\underbrace{\mathbb{E}\big(\big(\hat{x}_n - \mathbb{E}(\hat{x}_n)\big)\mathbb{B}(\hat{x}_n)\big)}_{\mathbb{B}(\hat{x}_n)\mathbb{E}\left(\hat{x}_n - \mathbb{E}(\hat{x}_n)\right) = 0} \\
&= \mathbb{V}\mathrm{ar}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2
\end{aligned}$$

Biais ou variance?

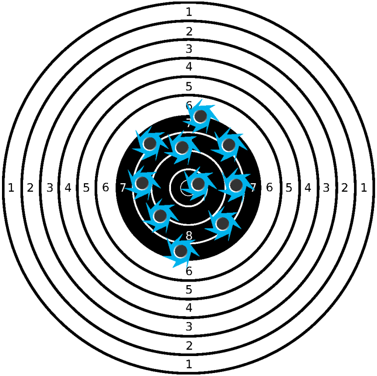


Erreurs systématiques
(fort biais)

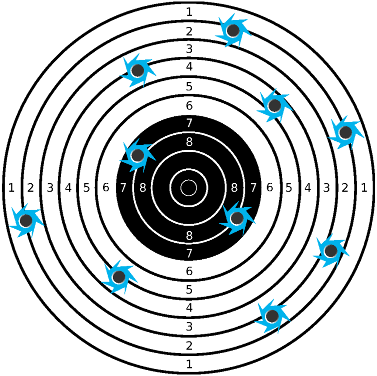


Erreurs stochastiques
(forte variance)

Biais ou variance?

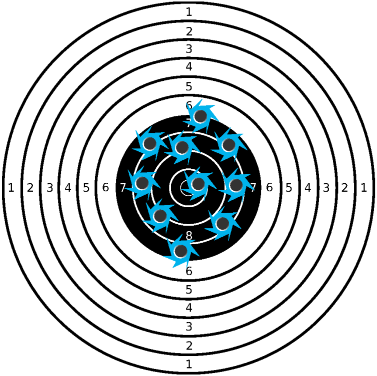


Biais nul
faible variance

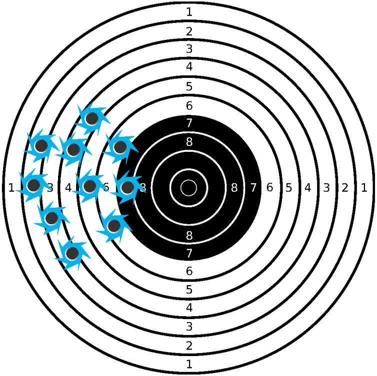


Biais nul
forte variance

Biais ou variance?



Biais nul
faible variance



Biais important
faible variance

Estimation avec ou sans biais de σ^2

Définition

La variance empirique est définie par

$$s_n^2(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

Biais: $\mathbb{E}(s_n^2(\mathbf{x})) = \frac{n-1}{n}\sigma^2$ (*cf.* calcul en TD)

Rem: l'estimateur[4] de σ^2, $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ est sans biais, mais dans le cas gaussien, son risque quadratique (qui vaut $\frac{2\sigma^4}{n-1}$) est plus grand que celui de $s_n(\mathbf{x})$ (qui vaut $\frac{2n-1}{n^2}\sigma^4$)

[4] appelé parfois l'estimateur sans biais de la variance

Sommaire

Théorème Central Limite (TCL)

Si la taille de l'échantillon est grande la moyenne empirique $\bar{x}_n$ est distribuée approximativement suivant une loi gaussienne :

Théorème

Soient $x_1, \ldots, x_n$ des variables aléatoires indépendantes, distribuées suivant la même loi, d'espérance μ et de variance σ^2; Alors, si n est grand ($n \geq 30$), la variable

$$Z = \frac{\bar{x}_n - \mu}{\sigma / \sqrt{n}}$$

suit approximativement une loi normale centrée, réduite $\mathcal{N}(0, 1)$

Rem: $\mathbb{V}\mathrm{ar}(\bar{x}_n) = \frac{\sigma^2}{n}$ donc $\mathbb{V}\mathrm{ar}(Z) = 1$ et $\mathbb{E}(Z) = 0$

Échantillonnage aléatoire simple (retour)

Pour un tel échantillon, les variables x_{i_k}, $k = 1, \ldots, n$ ont bien même loi, mais ne sont pas indépendantes (tirage sans remise)

Si $n \ll N$, on peut considérer que la dépendance est tellement faible qu'elle est négligeable (et le TCL s'applique)

Bilan. La moyenne empirique $\bar{x}_n$ est distribuée approximativement suivant une loi gaussienne si n (taille de l'échantillon) est grand

Échantillonnage aléatoire simple (retour)

Pour un tel échantillon, les variables x_{i_k}, $k = 1, \ldots, n$ ont bien même loi, mais ne sont pas indépendantes (tirage sans remise)

Si $n \ll N$, on peut considérer que la dépendance est tellement faible qu'elle est négligeable (et le TCL s'applique)

Bilan. La moyenne empirique $\bar{x}_n$ est distribuée approximativement suivant une loi gaussienne si n (taille de l'échantillon) est grand

Intervalles de confiance

Si l'on souhaite estimer μ, un premier intervalle de confiance est

$$\left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]$$

Cela nous donne un intervalle de confiance au niveau $68\,\%$:

$$\mathbb{P}\left(\mu \in \left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]\right)$$

Intervalles de confiance

Si l'on souhaite estimer μ, un premier intervalle de confiance est

$$\left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]$$

Cela nous donne un intervalle de confiance au niveau $68\,\%$:

$$\mathbb{P}\left(\mu \in \left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]\right) = \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right)$$

Intervalles de confiance

Si l'on souhaite estimer μ, un premier intervalle de confiance est

$$\left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]$$

Cela nous donne un intervalle de confiance au niveau $68\,\%$:

$$\begin{aligned}\mathbb{P}\left(\mu \in \left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]\right) &= \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right) \\ &= \mathbb{P}\left(\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-1, +1]\right)\end{aligned}$$

Intervalles de confiance

Si l'on souhaite estimer μ, un premier intervalle de confiance est

$$\left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]$$

Cela nous donne un intervalle de confiance au niveau $68\,\%$:

$$\begin{aligned}
\mathbb{P}\left(\mu \in \left[\bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\right]\right) &= \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right) \\
&= \mathbb{P}\left(\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-1, +1]\right) \\
&= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \\
&\approx 0.68
\end{aligned}$$

Intervalles de confiance : précision fixée

On souhaite estimer μ, avec un intervalle de confiance de la forme

$$\left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]$$

et qu'avec probabilité $1-\alpha$ le vrai paramètre soit dedans:

$$\mathbb{P}\left(\mu \in \left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]\right)$$

Intervalles de confiance : précision fixée

On souhaite estimer μ, avec un intervalle de confiance de la forme

$$\left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]$$

et qu'avec probabilité $1-\alpha$ le vrai paramètre soit dedans:

$$\mathbb{P}\left(\mu \in \left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]\right) = \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\delta\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right)$$

Intervalles de confiance : précision fixée

On souhaite estimer μ, avec un intervalle de confiance de la forme

$$\left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]$$

et qu'avec probabilité $1-\alpha$ le vrai paramètre soit dedans:

$$\begin{aligned}
\mathbb{P}\left(\mu \in \left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]\right) =& \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\delta\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right) \\
=& \mathbb{P}\left(\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta]\right)
\end{aligned}$$

Intervalles de confiance : précision fixée

On souhaite estimer μ, avec un intervalle de confiance de la forme

$$\left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]$$

et qu'avec probabilité $1-\alpha$ le vrai paramètre soit dedans:

$$\begin{aligned}
\mathbb{P}\left(\mu \in \left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]\right) &= \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\delta\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right) \\
&= \mathbb{P}\left(\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta]\right) \\
&= \Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1
\end{aligned}$$

Intervalles de confiance : précision fixée

On souhaite estimer μ, avec un intervalle de confiance de la forme

$$\left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]$$

et qu'avec probabilité $1-\alpha$ le vrai paramètre soit dedans:

$$\begin{aligned}
\mathbb{P}\left(\mu \in \left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]\right) =& \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\delta\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right) \\
=& \mathbb{P}\left(\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta]\right) \\
=& \Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1 \\
=& 1 - \alpha
\end{aligned}$$

Intervalles de confiance : précision fixée

On souhaite estimer μ, avec un intervalle de confiance de la forme

$$\left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]$$

et qu'avec probabilité $1-\alpha$ le vrai paramètre soit dedans:

$$\begin{aligned}
\mathbb{P}\left(\mu \in \left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]\right) =& \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\delta\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right) \\
=& \mathbb{P}\left(\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta]\right) \\
=& \Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1 \\
=& 1 - \alpha \\
\iff \delta =& \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)
\end{aligned}$$

Intervalles de confiance : précision fixée

On souhaite estimer μ, avec un intervalle de confiance de la forme

$$\left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]$$

et qu'avec probabilité $1-\alpha$ le vrai paramètre soit dedans:

$$\begin{aligned}
\mathbb{P}\left(\mu \in \left[\bar{x}_n - \delta\frac{\sigma}{\sqrt{n}}, \bar{x}_n + \delta\frac{\sigma}{\sqrt{n}}\right]\right) =& \mathbb{P}\left(\mu - \bar{x}_n \in \left[-\delta\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}}\right]\right)\\
=& \mathbb{P}\left(\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta]\right)\\
=& \Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1\\
=& 1 - \alpha\\
\iff \delta =& \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)
\end{aligned}$$

IC est alors

$$\left[\bar{x}_n - \frac{\delta\sigma}{\sqrt{n}}; \bar{x}_n + \frac{\delta\sigma}{\sqrt{n}}\right]$$

Qu'est-ce qu'un niveau de confiance ?

Que veut dire la phrase : “le niveau de confiance de l'intervalle $[\ldots;\ldots]$ est de $95\,\%$” ?

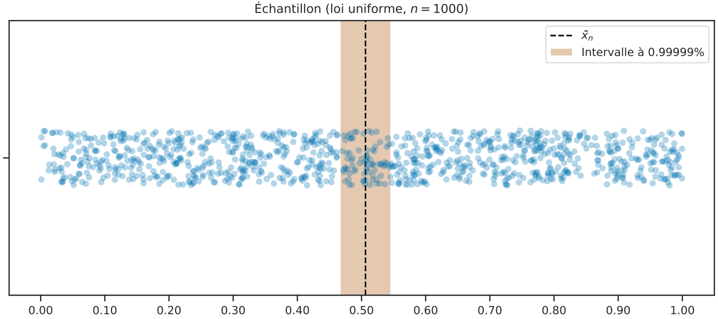
Si l'on prend beaucoup d'échantillons de taille n, on peut calculer autant de valeurs de $\bar{x}_n$ que l'on a d'échantillons, et donc autant d'intervalles de confiance que d'échantillons.
La phrase ci-dessus indique que la moyenne μ (inconnue) sur la population se trouve dans $95\,\%$ de ces intervalles de confiance.

Qu'est-ce qu'un niveau de confiance ?

Que veut dire la phrase : "le niveau de confiance de l'intervalle $[\ldots;\ldots]$ est de $95\,\%$" ?

Si l'on prend beaucoup d'échantillons de taille n, on peut calculer autant de valeurs de $\bar{x}_n$ que l'on a d'échantillons, et donc autant d'intervalles de confiance que d'échantillons.
La phrase ci-dessus indique que la moyenne μ (inconnue) sur la population se trouve dans $95\,\%$ de ces intervalles de confiance.

Exemple numérique



Bibliographie I

- Nolan, D. and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.