Introduction aux statistiques inférentielles

Mathieu Fauvernier

Equipe Biostatistique-Santé, UMR CNRS 5558, Université Lyon 1
Service de Biostatistique, Hospices Civils de Lyon

Section 1

Introduction et notions d'estimateur

Introduction

Les statistiques inférentielles permettent d'**inférer**, c'est à dire de **déduire**, des quantités d'intérêt (moyenne, variance, ...) d'une **population** d'étude à partir de données issues d'un **échantillon représentatif** de cette population.

On considère des données $(x_i)_{i \leq n} = (x_1, \ldots, x_n)$ comme **réalisation** de variables aléatoires $(X_i)_{i \leq n} = (X_1, \ldots, X_n)$.

Introduction

Les variables aléatoires $(X_i)_{i<=n}$ seront considérées comme **indépendantes et identiquement distribuées (i.i.d.)** selon une loi P_θ qui dépend d'un vecteur de paramètres $\boldsymbol{\theta}$.

Exemples

- Loi normale : $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)$
- Loi exponentielle : $X_i \sim \mathcal{E}(\lambda)$, $\boldsymbol{\theta} = (\lambda)$
- Loi de Poisson : $X_i \sim \mathcal{P}(\lambda)$, $\boldsymbol{\theta} = (\lambda)$
- Loi Bernoulli : $X_i \sim \mathcal{B}(p)$, $\boldsymbol{\theta} = (p)$

Introduction

L'ensemble des valeurs possibles pour $\boldsymbol{\theta}$ sont contenues dans un ensemble $\mathcal{I}$ (on note $\boldsymbol{\theta} \in \mathcal{I}$).

On s'intéresse à l'ensemble des lois définies par tous les $\boldsymbol{\theta}$ possibles, noté $(P_\theta)_{\boldsymbol{\theta} \in \mathcal{I}}$ et appelé **modèle paramétrique**.

Si $\mathcal{I}$ est fini ou dénombrable, on parlera de modèle discret, et si $\mathcal{I}$ est non-dénombrable on parlera de modèle continu.

Estimateur

Un **estimateur** de $\boldsymbol{\theta}$, noté $\hat{\boldsymbol{\theta}}$ ou $\hat{\boldsymbol{\theta}}_n$, est une fonction des variables $(X_i)_{i<=n}$ de loi P_θ

$$\hat{\boldsymbol{\theta}} = f(X_1, \ldots, X_n)$$

Comme fonction de variables aléatoires, **un estimateur est une variable aléatoire**

Nota bene : Un estimateur peut dépendre des $(X_i)_{i<=n}$, de la taille de l'échantillon n, **mais jamais de $\boldsymbol{\theta}$ lui-même**

Estimateur

1. $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur appelé **moyenne arithmétique empirique**
2. $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est un estimateur appelé **variance empirique non-corrigée**
3. $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est un estimateur appelé **variance empirique corrigée**
4. $\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$ est un estimateur appelé **écart-type empirique non-corrigé**
5. $s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$ est un estimateur appelé **écart-type empirique corrigé**

Il existe une infinité d'estimateurs possibles et dans la suite nous allons voir sur quels critères nous pouvons choisir les plus pertinents.

L'estimateur de la moyenne

L'estimateur de la moyenne se note

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

Comme on l'a vu, il s'agit d'une **variable aléatoire**.

On peut donc s'intéresser à son espérance et sa variance !

L'estimateur de la moyenne

Sachant que, les X_i sont i.i.d, avec $E(X_i) = \mu$ et $Var(X_i) = \sigma^2$

$$E\left(\bar{X}_n\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \quad ?$$

$$Var\left(\bar{X}_n\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \quad ?$$

L'estimateur de la moyenne

$$
\begin{aligned}
E\left(\bar{X}_n\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n\mu = \mu
\end{aligned}
$$

L'estimateur de la moyenne

$$
\begin{aligned}
E\left(\bar{X}_n\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu
\end{aligned}
$$

Ainsi, l'espérance de l'estimateur de la moyenne est la moyenne que l'on cherche à estimer ! On dit que l'estimateur de la moyenne est **sans biais**, car centré sur la bonne valeur.

L'estimateur de la moyenne

$$
\begin{aligned}
Var\left(\bar{X}_n\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \qquad \text{les } \boldsymbol{X_i} \text{ sont indépendants} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}
$$

L'estimateur de la moyenne

$$\begin{aligned} Var\left(\bar{X}_n\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \qquad \text{les } \boldsymbol{X_i} \text{ sont indépendants} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Ainsi, la variance de l'estimateur de la moyenne est de plus en plus petite lorsque n augmente (elle tend vers zéro).

L'estimateur de la moyenne

Nous connaissons l'espérance et la variance mais peut-on **connaître la loi de notre estimateur** ?

Section 2

Loi des grands nombres et théorème central limite

La loi des grands nombres

Soient $X_1, \ldots, X_n$ des v.a.r indépendantes et de même loi.

On note $\mu = E(X_1)$ et $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$

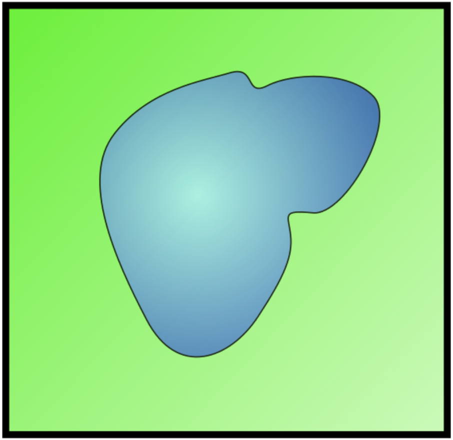
Alors, pour tout $\epsilon > 0$,

$$P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \underset{n \to \infty}{\rightarrow} 0$$

On dit que $\bar{X}_n$ converge en probabilité vers la vraie moyenne μ

Application de la loi des grands nombres : méthodes de Monte-Carlo

Quelle est la superficie de ce lac ?



Application de la loi des grands nombres : méthodes de Monte-Carlo

Si l'on connaît la superficie du terrain, on peut facilement avoir une approximation de la superficie du lac grâce à la loi des grands nombres.

Application de la loi des grands nombres : méthodes de Monte-Carlo

Si l'on connaît la superficie du terrain, on peut facilement avoir une approximation de la superficie du lac grâce à la loi des grands nombres.

Supposons que l'on tire n coups de canon, de manière aléatoire, sur le terrain. Si l'on note k le nombre de boulets tombés dans l'eau au final, on a :

$$\frac{\text{superficie}_{\text{lac}}}{\text{superficie}_{\text{terrain}}} \approx \frac{k}{n}$$

Application de la loi des grands nombres : méthodes de Monte-Carlo

Notons X_i la variable de Bernoulli $\mathcal{B}(p)$ telle que $X_i = 1$ si le i^e boulet atterrit dans l'eau et $X_i = 0$ sinon.

Application de la loi des grands nombres : méthodes de Monte-Carlo

Notons X_i la variable de Bernoulli $\mathcal{B}(p)$ telle que $X_i = 1$ si le i^e boulet atterrit dans l'eau et $X_i = 0$ sinon.

La méthode précédente revient alors à approximer p (la proportion réelle $\frac{\text{superficie}_{\text{lac}}}{\text{superficie}_{\text{terrain}}}$), en sommant les X_i

Application de la loi des grands nombres : méthodes de Monte-Carlo

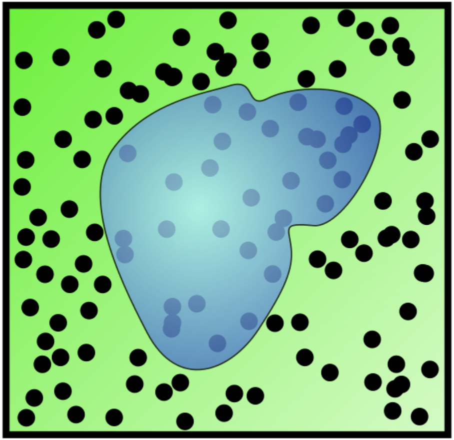
Notons X_i la variable de Bernoulli $\mathcal{B}(p)$ telle que $X_i = 1$ si le i^e boulet atterrit dans l'eau et $X_i = 0$ sinon.

La méthode précédente revient alors à approximer p (la proportion réelle $\frac{\text{superficie}_{\text{lac}}}{\text{superficie}_{\text{terrain}}}$), en sommant les X_i

En effet, la loi des grands nombres nous assure que :

$$\frac{X_1 + \cdots + X_n}{n} \to p$$

Application de la loi des grands nombres : méthodes de Monte-Carlo



Le théorème central limite (TCL)

Soient $X_1, \ldots, X_n$ des v.a.r indépendantes, de même loi, et admettant une variance.

On note $\mu = E(X_1)$ et $\sigma^2 = Var(X_1)$.

Alors,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Formellement, il s'agit d'une convergence en loi, c'est à dire que la fonction de répartition de la moyenne converge vers la fonction de répartition d'une loi normale.

Illustration du TCL

Soit la variable aléatoire suivante

$$X = \begin{cases} 1 & \text{avec probabilité } 1/3 \\ 2 & \text{avec probabilité } 1/3 \\ 3 & \text{avec probabilité } 1/3 \end{cases}$$

Illustration du TCL

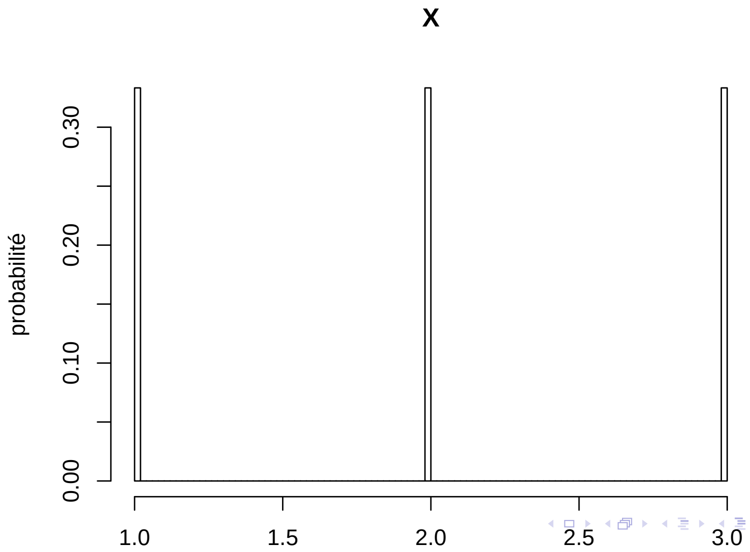


Illustration du TCL

On considère maintenant la somme de deux copies indépendantes de X, X_1 et X_2 :

$$X_1 + X_2 = \begin{cases} 1+1=2 & \text{avec probabilité } 1/9 \\ 1+2=3 & \text{avec probabilité } 1/9 \\ 1+3=4 & \text{avec probabilité } 1/9 \\ 2+1=3 & \text{avec probabilité } 1/9 \\ 2+2=4 & \text{avec probabilité } 1/9 \\ 2+3=5 & \text{avec probabilité } 1/9 \\ 3+1=4 & \text{avec probabilité } 1/9 \\ 3+2=5 & \text{avec probabilité } 1/9 \\ 3+3=6 & \text{avec probabilité } 1/9 \end{cases} = \begin{cases} 2 & \text{avec probabilité } 1/9 \\ 3 & \text{avec probabilité } 2/9 \\ 4 & \text{avec probabilité } 3/9 \\ 5 & \text{avec probabilité } 2/9 \\ 6 & \text{avec probabilité } 1/9 \end{cases}$$

Illustration du TCL

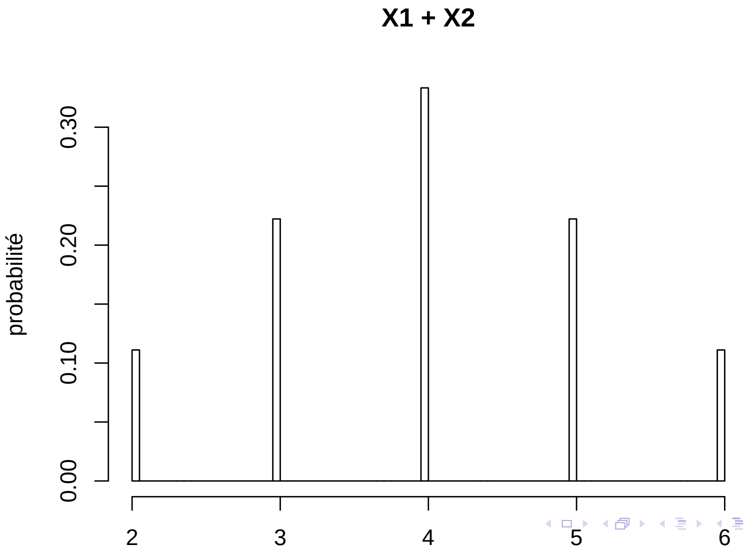


Illustration du TCL

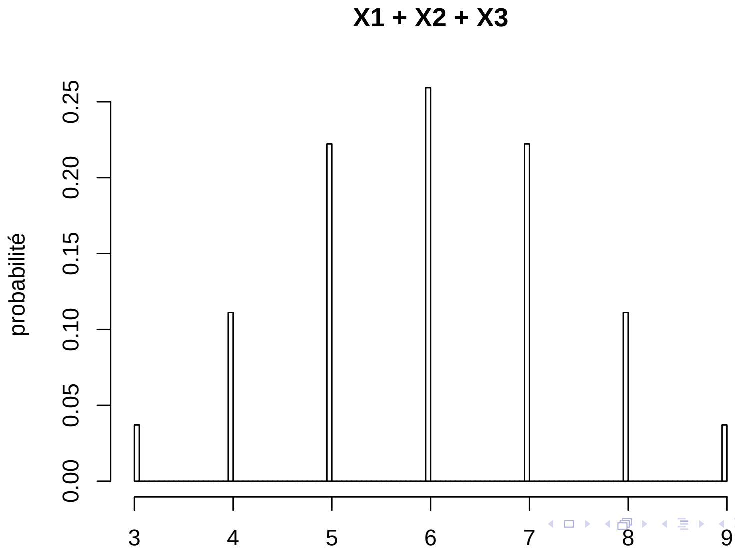


Illustration du TCL

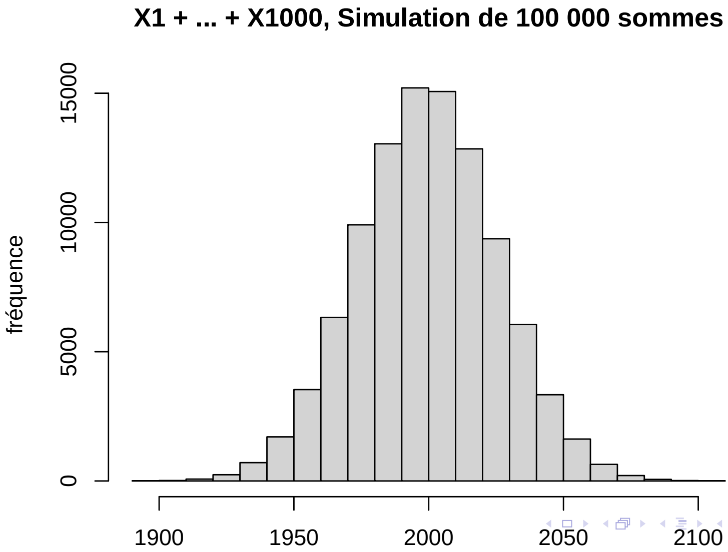


Illustration du TCL

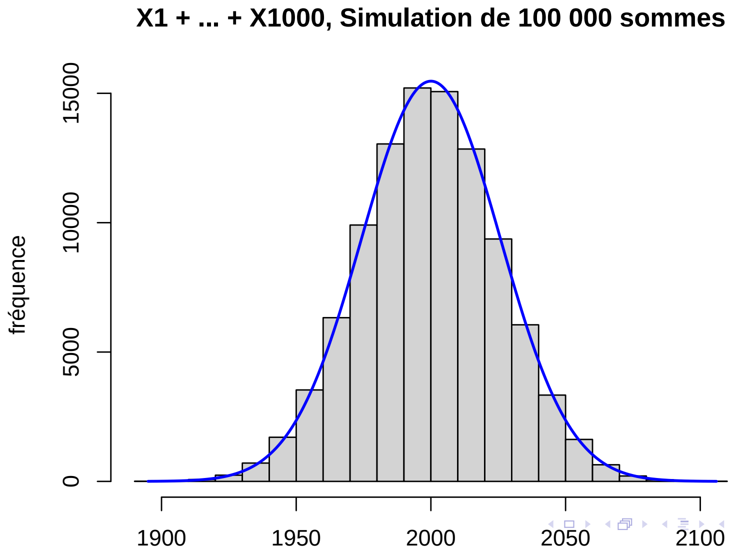


Planche de Galton

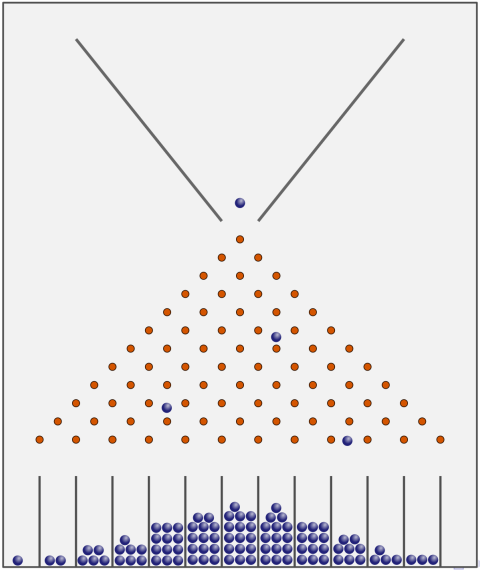


Planche de Galton

Sur la i^e ligne, il y a i clou(s).

Sur chaque ligne, la probabilité qu'une bille passe à gauche ou à droite du clou est la même et vaut $p = 0.5$.

Planche de Galton

Sur la i^e ligne, il y a i clou(s).

Sur chaque ligne, la probabilité qu'une bille passe à gauche ou à droite du clou est la même et vaut $p = 0.5$.

On considère qu'il y a n lignes et donc $n + 1$ boîtes en bas de la planche

Le nombre de chemins possibles amenant vers la k^e boîte (celle tout à gauche est la 0^e) est de C_n^k

Au final, la probabilité qu'une bille atterrisse dans la k^e boîte est

$$C_n^k p^k (1 - p)^{n-k}$$

Planche de Galton

Sur la i^e ligne, il y a i clou(s).

Sur chaque ligne, la probabilité qu'une bille passe à gauche ou à droite du clou est la même et vaut $p = 0.5$.

On considère qu'il y a n lignes et donc $n+1$ boîtes en bas de la planche

Le nombre de chemins possibles amenant vers la k^e boîte (celle tout à gauche est la 0^e) est de C_n^k

Au final, la probabilité qu'une bille atterrisse dans la k^e boîte est

$$C_n^k p^k (1-p)^{n-k}$$

$\Rightarrow$ loi binomiale !

Planche de Galton

La planche de Galton a été inventée par Francis Galton (1822 - 1911)

Elle démontre la convergence de la loi binomiale $\mathcal{B}(n, p)$ vers la loi normale $\mathcal{N}(np, np(1-p))$ lorsque $n \to \infty$

Intérêt du TCL

On le verra par la suite, le théorème central limite est un outil très puissant qui justifiera la **construction d'intervalles de confiance** à partir de la loi normale.