Introduction au méthodes statistiques

Cours 1

Michal W. Urdanivia*

*UGA, Faculté d'Économie, GAEL,
e-mail : michal.wong-urdanivia@univ-grenoble-alpes.fr

12 septembre 2022

Plan

1. Présentation

Objectifs

- Une introduction formelle et solide à la méthodologie statistique.
- Une discussion de la pertinence/adéquation des méthodes selon les/aux applications.

Divers

- **Prérequis :** cours de probabilités, analyse, algèbre(calcul vectoriel, matriciel, orthogonalité).
- **Références :**
  - Le matériel donné pour le cours d'abord.
  - Pour une référence "encyclopédique"(même si le titre indique que c'est "concis") vous pouvez consulter **?**.
  - Des références supplémentaires pourront être indiquées au moment de traiter chaque thème.

Divers

- **Travail :**
  1. Exercices(analytiques) pour s'approprier les méthodes et/ou traiter de points non vus dans le cours.
  2. Applications/illustrations : elle seront données/corrigées sous forme de notebooks Python, mais vous pouvez utiliser le langage de votre préférence(R, Matlab, . . .)
  3. Le travail sera donné environ toutes les deux semaines. Les corrections seront mises en ligne et ne se feront pas en séance(au moins dans la totalité)

Divers

- Site/dépôt du cours :
  `https://github.com/MWUrda/Statistiques_UGA_M1Eco_S1/tree/main`

Plan

2. Introduction

Heuristique

- On considère **l'expérience statistique suivante :**.
  - On lance une pièce de monnaie n fois, et à chaque lancer est associé son résultat qui vaut soit "pile", soit "face".
  - On veut à partir de cette expérience mesurer la probabilité du résultat "pile", que nous notons p.
  - Pour cela, **on estime** p par le pourcentage de résultats de valeur "pile" par rapport au nombre total de lancers.
- **Question :** comment garantir la validité de cette procédure ?

Heuristique

- Formellement, la procédure consiste en :
  - Pour $i = 1, 2, \ldots, n$, on définit $P_i = 1$ si le i-ème lancé donne "pile" et $P_i = 0$ sinon.
  - L'estimateur de p est la moyenne empirique :

$$\bar{P}_n = \frac{1}{n}\sum_{i=1}^{n} P_i.$$

- **<u>Question :</u>** quelle est la précision de cet estimateur ?
- Afin de répondre à ces question on propose un modèle statistique qui fournit une description/approximation de l'expérience, et qui sera considérée comme suffisamment bonne (par rapport à certains critères) pour être utile.

Heuristique

- Le modèle consiste en un ensemble d'hypothèses sur les observations $P_i, i = 1, 2, \ldots, n$.
- A partir de ces hypothèse on obtient des conclusions/résultats statistiques.
- Les hypothèses sont les suivantes :

  1. P_i est une variable aléatoire(v.a.).
  2. La loi de P_i est une loi de Bernouilli de paramètre p.
  3. Les $P_1, P_2, \ldots, P_n$ sont mutuellement indépendantes.

Heuristique

- **Discussion :**
  1. Modéliser les observations comme de v.a. est une façon de modéliser le fait que le modélisateur n'a qu'une information imparfaite sur les conditions de l'expérience. Si cette information était parfaite on pourrait prédire les résultats de l'expérience sans erreur.
  2. Comme $P_i \in \{0, 1\}$, la loi de P_i est une loi de Bernouilli de paramètre p dès lors que la pièce est équilibrée.

     . . . Pour une discussion intéressante vous pouvez consulter ici
  3. Supposer l'indépendance est approprié lorsque les conditions demeurent identiques entre les lancers. Par exemple, il n'y a pas de processus d'apprentissage entre les lancers.

Deux résultats/outils importants

- Soit $X, X_1, X_2, \ldots, X_n$ des v.a. i.i.d., avec $\mu := \mathrm{E}(X)$, $\sigma^2 := \mathrm{V}(X)$.
  1. **Lois des grands nombres(LGN) :**
     - **Loi forte :**

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_n \xrightarrow{p.s.} \mu.$$

     - **Loi faible :**

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_n \xrightarrow{p} \mu.$$

Deux résultats/outils importants

2. **Théorème central limite(TCL) :**

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

ou de manière équivalente

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Conséquences

- Les LGN appliquées à $\bar{P}_n$ impliquent :

$$\bar{P}_n \xrightarrow{p} p, \quad \text{ou} \quad \bar{P}_n \xrightarrow{p.s.} p.$$

- Ainsi lorsque $n \to \infty$, $\bar{P}_n$ est un "bon" estimateur(*estimateur consistant*) de p.
- Le TCL affine ce résultat en quantifiant la précision de l'estimateur.

Conséquences

- Notons :
  - $\Phi(x)$: fonction de répartition d'une v.a. de loi $\mathcal{N}(0,1)$.
  - $\Phi_n(x)$: fonction de répartition de $\sqrt{n}\frac{\bar{P}_n - p}{\sqrt{p(1-p)}}$.
- Par le TCL $\Phi_n(x) \approx \Phi(x)$ lorsque $n \to \infty$. Par conséquent, pour tout $x > 0$,

$$\mathrm{P}\left(\left|\bar{P}_n - p\right| \geq x\right) \approx 2\left(1 - \Phi\left(\frac{x\sqrt{n}}{\sqrt{p(1-p)}}\right)\right).$$

Conséquences

- **Conséquences :**
  - Une approximation sur le degré de concentration de $\bar{P}_n$ autour de μ ;
  - Pour $\alpha \in (0,1)$, si q_α est le quantile d'ordre $(1-\alpha/2)$ de la loi $\mathcal{N}(0,1)$, alors avec une probabilité $\approx 1-\alpha$(pour $n \to \infty$) :

$$\bar{P}_n \in \left[p - \frac{q_\alpha \sqrt{p(1-p)}}{\sqrt{n}} ; p + \frac{q_\alpha \sqrt{p(1-p)}}{\sqrt{n}} \right].$$

Conséquences

- On remarque que quelque soit la valeur(inconnue) de p,

$$p(1-p) \leq 1/4.$$

- Ce faisant on a grossièrement avec une probabilité d'au moins $1-\alpha$:

$$\bar{P}_n \in \left[p - \frac{q_\alpha}{2\sqrt{n}} ; p + \frac{q_\alpha}{2\sqrt{n}} \right].$$

- Autrement dit lorsque $n \to \infty$, l'intervalle $\left[\bar{P}_n - \frac{q_\alpha}{2\sqrt{n}} ; \bar{P}_n + \frac{q_\alpha}{2\sqrt{n}} \right]$ contient p avec une probabilité $\geq 1-\alpha$.
- Cet intervalle est appelé *intervalle de confiance asymptotique* pour p.
- **<u>Question :</u>** qu'en est-il pour n pas "suffisamment" grand ?

Un autre résultat/outil : inégalité de Hoeffding

- On considère le cas i.i.d.
  - Soit n un entier positif et $X, X_1, X_2, \ldots, X_n$ des v.a. i.i.d. telles que $X \in [a, b]$ p.s.($a < b$ sont deux nombres donnés). Soit $\mu := \mathrm{E}(X)$.
  - Alors pour tout $\epsilon > 0$:

$$\mathrm{P}\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

- **Conséquence :**
  - Pour $\alpha \in (0, 1)$, avec une probabilité $(1 - \alpha)$:

$$\bar{P}_n - \sqrt{\frac{\log(2/\alpha)}{2n}} \leq p \leq \bar{P}_n + \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

  - Ce résultat est valable pour n "pas très grand".

Plan

3. Retour sur les différents types de convergence

Types de convergence

- Soit $(T_n)_{n\geq 1}$ une suite de v.a. et T une v.a.(T peut être déterministe).
  - **Convergence presque sûre(p.s.) :**

$$T_n \overset{p.s.}{\to} T \quad \text{ssi} \quad \mathrm{P}\left[\left\{\omega : T_n(\omega) \underset{n\to\infty}{\longrightarrow} T(\omega)\right\}\right] = 1.$$

  - **Convergence en probabilité :**

$$T_n \overset{p}{\to} T \quad \text{ssi} \quad \mathrm{P}\left(|T_n - T| \geq \epsilon\right) \underset{n\to\infty}{\longrightarrow} 0, \quad \forall \epsilon > 0.$$

Types de convergence

- **Convergence en $L^p (p \geq 1)$:**

$$T_n \xrightarrow{L^p} T \quad \text{ssi} \quad \mathrm{E}\left(|T_n - T|^p\right) \underset{n\to\infty}{\longrightarrow} 0.$$

- **Convergence en loi/distribution) :**

$$T_n \xrightarrow{d} T \quad \text{ssi} \quad \mathrm{P}(T_n \leq x) \underset{n\to\infty}{\longrightarrow} \mathrm{P}(T \leq x),$$

pour tout $x \in \mathbb{R}$ où la fonction de répartition/distribution de T est continue.

- **Remarque :** les définition précédentes peuvent s'étendre aux vecteurs de v.a.(i.e., v.a. dans $\mathbb{R}^d$, pour $d \geq 2$).

Caractérisations de la convergence en distribution

- Les propositions suivantes sont équivalentes :
  (i) $T_n \xrightarrow{d} T$;
  (ii) $\mathrm{E}\left[f(T_n)\right] \underset{n\to\infty}{\longrightarrow} \mathrm{E}\left[f(T)\right]$, pour toutes les fonctions continues et bornées f ;
  (iii) $\mathrm{E}\left(e^{ixT_n}\right) \underset{n\to\infty}{\longrightarrow} \mathrm{E}\left(e^{ixT}\right)$, pour tout $x \in \mathbb{R}$.

Propriétés importantes

- Si $(T_n)_{n\geq 1}$ converge p.s., alors elle converge aussi en probabilité, et les deux limites sont égales p.s.
- Si $(T_n)_{n\geq 1}$ converge en L^p alors elle converge aussi en L^q pour tout $q \leq p$ et en probabilité, et les deux limites sont égales p.s.
- Si f est une fonction continue :

$$T_n \stackrel{p.s./p/d}{\longrightarrow} T \Rightarrow f(T_n) \stackrel{p.s./p/d}{\longrightarrow} f(T).$$

Limites et opérations

- On peut additionner, multiplier, ..., les limites presque sûrement et en probabilité. Si $U_n \xrightarrow{p.s./p} U$, et $V_n \xrightarrow{p.s./p} V$, alors :
  - $U_n + V_n \xrightarrow{p.s./p} U + V$,
  - $U_n V_n \xrightarrow{p.s./p} UV$,
  - Si en plus, $V \neq 0$ p.s., alors $\frac{U_n}{V_n} \xrightarrow{p.s./p} \frac{U}{V}$.
- **Remarque :** ces règles ne s'appliquent pas à la convergence en distribution sauf lorsque la pair (U_n, V_n) converge vers en distribution vers (U, V).

Un autre exemple

- On observe les temps entre les arrivées de personnes dans une file d'attente(par exemple d'un centre d'appel) : $T_1, T_2, \ldots, T_n$.
- On suppose que ces temps sont :
  - Mutuellement indépendants.
  - Représentés par des v.a de loi exponentielle de paramètre commun $\lambda > 0$.
- On veut estimer λ à partir des observations.

Un autre exemple

- **Discussion des hypothèses :**
  - Indépendance mutuelle : les personnes sont supposées ne pas être liées entre elles de sorte qu'elles ne décident pas leur arrivée en fonction de celle des autres.
  - $T_1, T_2, \ldots, T_n$ sont de v.a de loi exponentielle : absence de mémoire de la loi exponentielle ;

$$\mathrm{P}(T_1 > t + s) = \mathrm{P}(T_1 > s), \quad \forall s, t \geq 0.$$

  - Les lois exponentielles de $T_1, T_2, \ldots$ ont le même paramètre λ : comportement homogène dans la population.

Un autre exemple

- Densité de T_1 :

$$f(t) = \lambda e^{-\lambda t}, \quad \forall t \geq 0.$$

- $\mathrm{E}(T_1) = \frac{1}{\lambda}$.
- Et un estimateur naturel de $\frac{1}{\lambda}$ est :

$$\bar{T}_n := \frac{1}{n} \sum_{i=1}^{n} T_i.$$

- Un estimateur naturel de λ est :

$$\hat{\lambda} := \frac{1}{\bar{T}_n}.$$

Un autre exemple

- Par les LGNs,

$$\bar{T}_n \xrightarrow{p.s./p} \frac{1}{\lambda}.$$

- Et donc :

$$\hat{\lambda} \xrightarrow{p.s./p} \lambda.$$

- Par le TCL :

$$\sqrt{n}\left(\bar{T}_n - \frac{1}{\lambda}\right) \xrightarrow{d} \mathcal{N}(0, \lambda^{-2}).$$

- **Question :** comment le résultat sur le TCL ci-dessus se traduit/transfère t-il sur $\hat{\lambda}$? Comment obtenir un intervalle de confiance asymptotique pour λ ?

La méthode Delta

- Soit $(Z_n)_{n\geq 1}$ une suite de v.a. qui vérifient :

$$\sqrt{n}\,(Z_n - \vartheta) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

pour $\vartheta \in \mathbb{R}$ et $\sigma^2 > 0$(la suite $(Z_n)_{n\geq 1}$ est dite *asymptotiquement normale autour de* ϑ).

- Soit $g : \mathbb{R} \to \mathbb{R}$ continûment différentiable au point ϑ. Alors :
  - $(g(Z_n))_{n\geq 1}$ est aussi asymptotiquement normale.
  - Plus précisément :

$$\sqrt{n}\,(g(Z_n) - g(\vartheta)) \xrightarrow{d} \mathcal{N}(0, g'(\vartheta)^2\sigma^2).$$

Conséquences de la méthode Delta

- $\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda^2)$.
- Par conséquent, pour $\alpha \in (0, 1)$ et n suffisamment grand :

$$\left|\hat{\lambda} - \lambda\right| \leq \frac{q_\alpha \lambda}{\sqrt{n}}.$$

- <u>**Question :**</u> est-ce que $\left[\hat{\lambda} - \frac{q_\alpha \lambda}{\sqrt{n}}; \hat{\lambda} + \frac{q_\alpha \lambda}{\sqrt{n}}\right]$ peut être utilisé comme intervalle de confiance asymptotique pour λ ?
- La réponse est non car il dépend de λ.

Conséquences de la méthode Delta

- Deux solutions sont possibles.
  1. **Solution sous dépendance :**
     - Nous avons :

$$
\begin{aligned}
\left|\hat{\lambda} - \lambda\right| \leq \frac{q_\alpha \lambda}{\sqrt{n}} &\Leftrightarrow \lambda\left(1 - \frac{q_\alpha}{\sqrt{n}}\right) \leq \hat{\lambda} \leq \lambda\left(1 + \frac{q_\alpha}{\sqrt{n}}\right) \\
&\Leftrightarrow \hat{\lambda}\left(1 + \frac{q_\alpha}{\sqrt{n}}\right)^{-1} \leq \lambda \leq \hat{\lambda}\left(1 - \frac{q_\alpha}{\sqrt{n}}\right)^{-1}
\end{aligned}
$$

     - Et l'intervalle de confiance asymptotique pour lambda est :

$$
\left[\hat{\lambda}\left(1 + \frac{q_\alpha}{\sqrt{n}}\right)^{-1} ; \hat{\lambda}\left(1 - \frac{q_\alpha}{\sqrt{n}}\right)^{-1}\right]
$$

  2. Solution systématique par le **théorème de Slutsky.**

Théorème de Slutsky

- Soit $(X_n)_{n\geq 1}, (Y_n)_{n\geq 1}$ deux suites de v.a. telles que :

  (i) $X_n \xrightarrow{d} X$;

  (ii) $Y_n \xrightarrow{p} c$,

  où X est une v.a. et c un nombre réel donné.

- Alors :

$$(X_n, Y_n) \xrightarrow{d} (X, c).$$

- En particulier :

$$X_n + Y_n \xrightarrow{d} X + c,$$

$$X_n Y_n \xrightarrow{d} cX,$$

$$\ldots$$

Conséquences du théorème de Slutsky

- Par la méthode Delta, on sait que :

$$\sqrt{n}\frac{\hat{\lambda}-\lambda}{\lambda} \xrightarrow{d} \mathcal{N}(0,1).$$

- Par la loi faible des grands nombres :

$$\hat{\lambda} \xrightarrow{p} \lambda.$$

- Et donc par le théorème de Slutsky :

$$\sqrt{n}\frac{\hat{\lambda}-\lambda}{\hat{\lambda}} \xrightarrow{d} \mathcal{N}(0,1).$$

- Un autre intervalle de confiance asymptotique pour λ est :

$$\left[\hat{\lambda}-\frac{q_\alpha\hat{\lambda}}{\sqrt{n}};\hat{\lambda}+\frac{q_\alpha\hat{\lambda}}{\sqrt{n}}\right].$$

Conséquences du théorème de Slutsky

Remarque :

- Dans le premier example(lancers de pièce avec résultat pile/face), nous avons utilisé une solution sous dépendance : "$p(1-p) \leq 1/4$".
- Avec le théorème de Slutsky nous obtenons l'intervalle de confiance asymptotique :

$$\left[\hat{P}_n - \frac{q_\alpha \sqrt{\hat{P}_n(1-\hat{P}_n)}}{\sqrt{n}} ; \hat{P}_n + \frac{q_\alpha \sqrt{\hat{P}_n(1-\hat{P}_n)}}{\sqrt{n}}\right].$$

Références

Wasserman, Larry. 2004. *All of Statistics : A Concise Course in Statistical Inference*. Springer Texts in Statistics.