Statistique : Intervalles de confiance et tests

Joseph Salmon

Septembre 2014

Plan du cours

Intervalle de confiance

- Contexte : on a une estimation $\hat{g}(y_1, \dots, y_n)$ d'une grandeur $g(\theta)$. On veut un intervalle $\hat{I}$ autour de $\hat{g}$ qui contient g avec une grande probabilité.
- On construit $\hat{I} = [A, B]$ à partir des observations $(y_1, \dots, y_n)$: l'intervalle est une variable aléatoire

$$\mathbb{P}(\hat{I} \text{ contient } g) = \mathbb{P}(A \leq g \text{ et } B \geq g) = 95\%$$

Intervalle de confiance de niveau α

Intervalle de confiance

Un intervalle de confiance de niveau α pour la grandeur $g = g(\theta)$ est une fonction de l'échantillon

$$\hat{I} : (y_1, \ldots, y_n) \mapsto \hat{I} = [A(y_1, \ldots, y_n), B(y_1, \ldots, y_n)]$$

telle que, quelle que soit le paramètre $\theta \in \Theta$,

$$\mathbb{P}\left[g(\theta) \in \hat{I}(y_1, \ldots, y_n)\right] \geq 1 - \alpha \qquad \text{lorsque } y_i \sim \mathbb{P}_\theta$$

Rem: des choix classiques sont $\alpha = 5\%, 1\%, 0.1\%$, etc.

Exemple : sondage

- Sondage d'une élection à deux candidats : A et B. Le choix du i-ème sondé suit une loi de Bernoulli de paramètre p, $y_i = 1$ s'il vote A, 0 sinon. Ainsi,

$$\Theta = [0, 1] \text{ et } \theta = p.$$

- but : estimer $g(\theta) = p$.
- échantillon de taille n : un estimtateur raisonnable est alors

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}_n$$

intervalle de confiance pour p ?

Sondage : intervalle de confiance

- Chercher un intervalle $\hat{I} = [\hat{p} - \delta, \hat{p} + \delta]$ tel que $\mathbb{P}(p \in \hat{I}) \geq 0.95 \Leftrightarrow$ chercher δ tel que $\mathbb{P}\left[\,|\hat{p} - p| > \delta\right] \leq 0.05$
- Ingrédient : inégalité de **Tchebyschev** (si $\mathbb{E}(X^2) < +\infty$)

$$\boxed{\forall \delta > 0, \quad \mathbb{P}(|X - \mathbb{E}(X)| > \delta) \leq \frac{\mathrm{Var}(X)}{\delta^2}}$$

Pour $X = \hat{p} = \frac{1}{n}\sum_{i=1}^n Y_i$ on a $\mathbb{E}(\hat{p}) = p$ et $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$:

$$\forall p \in (0,1), \forall \delta > 0, \quad \mathbb{P}(|\hat{p} - p| > \delta) \leq \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

Application numérique : pour un intervalle de confiance à 95%, on peut choisir δ tel que $\frac{1}{4n\delta^2} = 0.05$, *i.e.*, $\delta = \sqrt{\frac{1}{4 \times 0.05 \times n}}$.
Si $n = 1000$ et $\hat{p} = 55\%$, on obtient

$$\delta = 0.07 \; ; \quad \hat{I} = [0.48, 0.62]$$

Théorème central limite

- $y_1, y_2, \ldots,$ des variables aléatoires *i.i.d.* de carré intégrable.
- μ et σ leur espérance et écart-type théoriques.

Théorème central limite (TCL)

La loi de la moyenne empirique renormalisée

$$\sqrt{n}\left(\frac{\bar{y}_n - \mu}{\sigma}\right)$$

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

- Reformulation : La moyenne empirique se comporte approximativement comme une loi normale $\mathcal{N}(\mu, \sigma^2/n)$

Intervalles de confiance asymptotiques

- Exemple du sondage : $\hat{p} = 0.55$, $n = 1000$
- On suppose que n est suffisamment grand pour que

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} y_n - p}{\sqrt{p(1-p)}} \sim \mathcal{N}(0,1) \qquad \text{rappel : } p(1-p) = \mathrm{Var}(Y)$$

- On connaît les quantiles de la loi normale (numériquement)
- D'après le TCL, et l'approximation des quantiles gaussiens

$$\mathbb{P}\left[-1.96 < \sqrt{n}\,\frac{0.55 - p}{\sqrt{p(1-p)}} < 1.96\right] \approx 0.95$$

On résout en p (équations de degré deux) :

$$\mathbb{P}\left[0.52 < p < 0.58\right] = 0.95$$

nouvel intervalle de confiance : $\hat{I} = [0.52, 0.58]$: meilleur !

Tests d'hypothèses pour le "Pile ou face"

- On veut tester une hypothèse sur le paramètre θ.
- On l'appelle hypothèse nulle $\mathcal{H}_0$
  Exemple: 'la pièce est non biaisée" : $\mathcal{H}_0 = \{p = 0.5\}$.
  Exemple: 'la pièce est peu biaisée', $\mathcal{H}_0 = \{0.45 \leq p \leq 0.55\}$
- L'hypothèse alternative $\mathcal{H}_1$ est (souvent) le contraire de $\mathcal{H}_0$.
  Exemple: $\mathcal{H}_1 = \{p \neq 0.5\}$
  Exemple: $\mathcal{H}_1 = \{p \notin [0.45, 0.55]\}$
- « Faire un test » : déterminer si les données permettent de rejeter l'hypothèse $\mathcal{H}_0$. On cherche une region R pour laquelle si $(y_1, \ldots, y_n) \in R$ on rejette l'hypothèse $\mathcal{H}_0$. R est la région de **rejet**.

Joseph Salmon

Rejet ou acceptation ?

Statistique : Intervalles de confiance et tests

Intervalles de confiance

Théorèmes limites

Tests d'hypothèses

Joseph Salmon

Présomption d'innocence en faveur de $\mathcal{H}_0$

Même si $\mathcal{H}_0$ n'est pas rejetée par le test, on ne peut en général pas conclure que $\mathcal{H}_0$ est vraie !

Rejeter $\mathcal{H}_1$ est souvent impossible car $\mathcal{H}_1$ est trop générale.
e.g., $\{p \in [0, 0.5[\cup]0.5, 1]\}$ ne peut pas être rejetée !

- $\mathcal{H}_0$ s'écrit sous la forme $\{\theta \in \Theta_0\}$, avec $\Theta_0 \subset \Theta$
- $\mathcal{H}_1$ s'écrit sous la forme $\{\theta \in \Theta_1\}$, avec $\Theta_1 \subset \Theta$

Rem: $\{\theta \in \Theta_0\}$ et $\{\theta \in \Theta_1\}$ sont disjoints.

Risques de première et de seconde espèce

	$\mathcal{H}_0$	$\mathcal{H}_1$
Non rejet de $\mathcal{H}_0$	Juste	Faux (acceptation à tort)
Rejet de $\mathcal{H}_0$	Faux (Rejet à tort)	Juste

- Risque de première espèce : probabilité de rejeter à tort

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta((y_1, \dots, y_n) \in R)$$

- Risque de seconde espèce

$$\beta = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta\left((y_1, \dots, y_n) \notin R\right)$$

Niveau/Puissance

Niveau du test

$1 - \alpha =$ probabilité d'« accepter » à raison (si $\mathcal{H}_0$ est valide)

Puissance du test

$1 - \beta =$ probabilité de rejeter $\mathcal{H}_0$ à raison (si $\mathcal{H}_1$ est valide)

En général, lorsqu'on parle de « test à 95% » on parle d'un test de niveau $1 - \alpha \geq 95\%$.

Statistique de test et région de rejet

Objectif classique : construire un test de niveau $1-\alpha$

- On cherche une fonction des données $T_n(y_1,\dots,y_n)$ dont on connaît la loi si $\mathcal{H}_0$ est vraie : T_n est appelée *statistique de test*.
- On définit une région de rejet ou région critique de niveau α, une région R telle que, sous $\mathcal{H}_0$,

$$\mathbb{P}(T_n(y_1,\dots,y_n)\in R)\leq \alpha$$

- Régle de rejet de $\mathcal{H}_0$: on rejette si $\boxed{T_n(y_1,\dots,y_n)\in R}$

Exemple gaussien : nullité de la moyenne

- Modèle : $\Theta = \mathbb{R}$, $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$.
- Hypothèse nulle : $\mathcal{H}_0 : \{\theta = 0\}$
- Sous $\mathcal{H}_0$, $\quad T_n(y_1, \dots, y_n) = \frac{1}{\sqrt{n}} \sum_i y_i \sim \mathcal{N}(0, 1)$
- Région critique pour T_n ? Quantiles gaussiens : sous H_0,

$$\mathbb{P}(T_n \in [-1.96, 1.96]) = 0.95$$

On prend $R = [-1.96, 1.96]^C =]-\infty, -1.96[\cup]1.96, +\infty[$.

- **Exemple numerique** : si $T_n = 1.5$, on ne rejette **PAS** $\mathcal{H}_0$ au niveau 95%

Statistique : Intervalles de confiance et tests

Intervalles de confiance

Théorèmes limites

Tests d'hypothèses

Joseph Salmon