Inférence statistique

Cours 5 : Tests d'Hypothèses Paramétriques

Michal W. Urdanivia*

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

19 septembre 2022

Plan

1. Introduction

2. Formulation Statistique

3. Paradigme de Neyman-Pearson

Plan

1. Introduction

Heuristique

- **<u>Idée clé :</u> décider de rejeter ou d'accepter des hypothèses.**
- **Exemples :**
  - Après avoir lancé une pièce plusieurs fois, on veut décider(tester) à partir de ces résultats si la pièce peut être considérée comme équilibrée.
  - Après avoir prélevé le sang d'un patients et mesuré la concentration d'anticorps dans l'échantillon on veut décider(tester) si le patient peut être considéré comme porteur d'un virus.
- Formellement ceci peut être écrit ainsi :
  - décider si $p = 1/2$ à partir d'un échantillon de n v.a. de Bernouilli i.i.d. ;
  - décider si $c > c_0$ à partir d'un échantillon de n v.a. i.i.d. de loi normale $\mathcal{N}(c, \sigma^2)$.

Heuristique

- **Exemple 1 :** une pièce est lancée 80 fois, et le résultat est pile 54 fois. Peut-on conclure *significativement* que la pièce est déséquilibrée ?
  - $n = 80$, $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{B}er(p)$,
  - $\bar{X}_n = 54/80 = 0.68$,
  - Si cela était vrai que $p = 0.5$, par le TCL et le théorème de Slutsky,

$$\sqrt{n}\frac{\bar{X}_n - 0.5}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \approx \mathcal{N}(0, 1),$$

  - $\sqrt{n}\frac{\bar{X}_n - 0.5}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \approx 3.45$,
  - conclusion : il paraît assez raisonnable de rejeter l'hypothèse "$p = 0.5$".

Heuristique

- **Exemple 2 :** une pièce est lancée 30 fois, et le résultat est pile 13 fois. Peut-on conclure *significativement* que la pièce est déséquilibrée ?
  - $n = 80$, $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{B}er(p)$,
  - $\bar{X}_n = 54/80 = 0.68$,
  - Si cela était vrai que $p = 0.5$, par le TCL et le théorème de Slutsky,
$$\sqrt{n}\frac{\bar{X}_n - 0.5}{\sqrt{\bar{X}_n}(1 - \bar{X}_n)} \approx \mathcal{N}(0, 1),$$
  - $\sqrt{n}\frac{\bar{X}_n - 0.5}{\sqrt{\bar{X}_n}(1-\bar{X}_n)} \approx -0.77$,
  - conclusion : il paraît impossible de rejeter l'hypothèse "$p = 0.5$" de manière significative.

Plan

2. Formulation Statistique

Définition du problème

- Soit un échantillon de n v.a. i.i.d. $X_1, X_2, \ldots, X_n$ et un modèle statistique $(\mathcal{E}, \mathcal{F}, (\mathrm{P}_\theta)_{\theta\in\Theta})$.
- Soit Θ_0 et Θ_1 des sous-ensembles disjoints de Θ.
- Considérons deux hypothèses : $\begin{cases} H_0 & : \quad \theta \in \Theta_0 \\ H_1 & : \quad \theta \in \Theta_1 \end{cases}$
- H_0 est l'**hypothèse nulle** et H_1 est l'**hypothèse alternative**.
- Si nous pensons que le vrai paramètre θ est soit dans Θ_0, soit dans Θ_1, alors on pourrait tester H_0 vs. H_1.
- On veut décider si oui ou non il faut rejeter H_0(on cherche de l'information dans les données à l'encontre de H_0).

Définition du problème

- H_0 et H_1 ne sont pas symétriques. Par exemple si H_0 : "le patient est malade"($c > c_0$) vs. H_1 : "le patient est en bonne santé"($c \leq c_0$).
- Un **test** est une statistique $\delta \in \{0, 1\}$ telle que :
  - Si $\delta = 0$, H_0 n'est pas rejetée ;
  - Si $\delta = 1$, H_0 est rejetée.
- Dans l'exemple du lancer de pièce :
  - H_0 : "$p = 1/2$" vs. H_1 : "$p \neq 1/2$",
  - et $\delta = \mathbf{1}_{\left|\sqrt{n}\frac{\bar{X}_n - 0.5}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right| > C}$ pour un seuil $C > 0$.
  - **Question :** comment choisir C ?

Définition du problème

- **Région de rejet** d'un test δ :

$$\mathcal{R}_\delta = \{x \in \mathcal{E}^n : \delta(x) = 1\}.$$

- **Érreur du(de) premier(ère) type(espèce)** d'un test δ(rejeter H_0 quand cette hypothèse est vraie) :

$$\begin{aligned}\alpha_\delta : \Theta_0 &\to \mathbb{R}\\ \theta &\mapsto \mathrm{P}_\theta(\delta = 1).\end{aligned}$$

- **Érreur du(de) deuxième type(espèce)** d'un test δ(ne pas rejeter H_0 quand H_1 vraie) :

$$\begin{aligned}\beta_\delta : \Theta_1 &\to \mathbb{R}\\ \theta &\mapsto \mathrm{P}_\theta(\delta = 0).\end{aligned}$$

- **Puissance** du test δ :

$$\pi_\delta = \inf_{\theta \in \Theta_1} \left(1 - \beta_\delta(\theta)\right).$$

Définition du problème

- Le **niveau** d'un test δ est α si,

$$\alpha_\delta(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- Le **niveau asymptotique** d'un test δ est α si,

$$\lim_{n\to\infty} \alpha_\delta(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

- Typiquement la forme d'un test est donnée par

$$\delta = \mathbf{1}_{T_n > c},$$

pour une statistique T_n et un seuil $c \in \mathbb{R}$.

- T_n est alors appelée **statistique de test** et la région $\mathcal{R}_\delta$ est,

$$\mathcal{R}_\delta := \{T_n > c\}.$$

Un exemple

- $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{B}er(p)$, pour $p \in (0, 1)$.
- Test :

$$H_0 : "p = 1/2" \text{vs.} H_1 : "p \neq 1/2",$$

  avec un niveau asymptotique $\alpha \in (0, 1)$.
- Statistique :

$$T_n = \left| \sqrt{n} \frac{\hat{p}_n - 0.5}{\sqrt{\hat{p}_n(1 - \hat{p}_n)}} \right|,$$

  où $\hat{p}_n$ est l'estimateur du MV.
- Si H_0 est vraie, alors par le TCL et le théorème de Slutsky,

$$\mathrm{P}\left(T_n > q_{1-\alpha/2}\right) \underset{n \to \infty}{\to} 0.05.$$

- Soit $\delta_\alpha = \mathbf{1}_{T_n > q_{1-\alpha/2}}$.

Un exemple

- Revenons cas du lancer de pièce et fixons $\alpha = 5\%$, d'où $q_{1-\alpha/2} = 1.96$, et alors :
  - Dans l'exemple 1, H_0 est rejeté au niveau asymptotique de 5% par le test $\delta_{5\%}$.
  - Dans l'exemple 2, H_0 n'est pas rejeté au niveau asymptotique de 5% par le test $\delta_{5\%}$
- **Question** : dans l'exemple 1, pour quel niveau α, le test δ_α ne rejetterait H_0 ? et dans l'exemple 2 pour quel niveau α le test δ_α rejetterait H_0 ?

P-value

- **<u>Définition</u>** : la p-value(asymptotique) d'un test δ_α est le plus petit niveau(asymptotique) α pour lequel δ_α conduit à rejeter H_0. Elle est aléatoire car elle dépend de l'échantillon.
- **<u>Régle d'or :</u>**
  - p-value $< \alpha \Leftrightarrow H_0$ est rejeté par δ_α pour le niveau(asymptotique)α.
  - Plus la p-value est petite, et plus on est confiant dans le rejet de H_0.
- Exemple 1 : p-value $= \mathrm{P}\left(|Z| > 3.45\right) \ll 0.01$.
- Exemple 2 : p-value $= \mathrm{P}\left(|Z| > 0.77\right) \approx 0.44$.

Plan

3. Paradigme de Neyman-Pearson

Idée de Neyman-Pearson

- Pour une hypothèse donnée, parmi tous les tests de niveau(ou de niveau asymptotique) α, est-il possible de trouver celui qui a la puissance maximale ?
- **Exemple :** le test trivial $\delta = 0$ qui ne rejette jamais H_0 a un niveau parfait($\alpha = 0$) mais une puissance médiocre($\pi_\delta = 0$).
- La théorie de Neyman-Pearson offre des test qui sont(le plus) puissant pour un niveau donné.
- Nous allons seulement considérer certains cas.

Rappel sur la loi du χ^2

- **Définition :** pour un entier positif d, la loi du χ^2 à d degrés de liberté est la loi de la v.a. $Z_1^2 + Z_2^2 + \cdots + Z_d^2$, où $Z_1, Z_2, \ldots, Z_d \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$.
- Exemples :
  - Si $Z \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$, alors $\|Z\|_2^2 \sim \chi_d^2$.
  - Le théorème de Cochran énonce que pour $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, si S_n est la variance empirique, alors,
  $$\frac{nS_n}{\sigma^2} \sim \chi_{n-1}^2.$$
  - $\chi_2^2 = \mathcal{E}xp(1/2)$.

Rappel sur la loi de Student

- **Définition :** pour un entier positif d, la loi de Student à d degrés de liberté (notée t_d) est la loi de la v.a. $\frac{U}{\sqrt{V/d}}$, où $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2_d$, et $U \perp\!\!\!\perp V$.
- Exemple : le théorème de Cochran énonce que pour $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, si S_n est la variance empirique, alors,

$$\sqrt{n-1}\left(\frac{\bar{X}_n - \mu}{\sqrt{S_n}}\right) \sim t_{n-1}.$$

Test de Wald

- Soit un échantillon de v.a. i.i.d., $X1, X_2, \ldots, X_n$ avec le modèle statistique $(\mathbf{E}, \mathbf{F}, (\mathrm{P}_\theta)_{\theta\in\Theta})$, où $\Theta \subseteq \mathbb{R}^d$(pour $d \geq 1$), et soit $\theta_0 \in \Theta$ une valeur donnée de θ.
- Considérons les hypothèses suivantes :

$$\begin{cases} H_0 : & \theta & = \theta_0, \\ H_1 : & \theta & \neq \theta_0. \end{cases}$$

- Soit $\hat{\theta}^{MV}$ l'estimateur du MV, et supposons que les conditions techniques pour cet estimateur soient satisfaites.
- Si H_0 est vraie, alors,

$$\sqrt{n}\mathcal{I}(\hat{\theta}^{MV})^{1/2}\left(\hat{\theta}_n^{MV} - \theta_0\right) \xrightarrow{d} \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d), \quad \text{par rapport à } \mathrm{P}_{\theta_0}.$$

Test de Wald

- D'où,

$$\underbrace{n\left(\hat{\theta}_n^{MV} - \theta_0\right)^\top \mathcal{I}(\hat{\theta}^{MV})\left(\hat{\theta}_n^{MV} - \theta_0\right)}_{T_n} \xrightarrow{d} \chi^2_d, \quad \text{par rapport à } \mathrm{P}_{\theta_0} .$$

- Test de Wald de niveau asymptotique $\alpha \in (0, 1)$:

$$\delta = \mathbf{1}_{T_n > q_{1-\alpha}},$$

  où $q_{1-\alpha}$ est le quantile $(1 - \alpha)$ du χ^2_d(voir tables).
- Remarque : le test de Wald est aussi valide si H_1 présente la forme "$\theta > \theta_0$" ou "$\theta < \theta_0$" où "$\theta = \theta_1$", ...

Test du rapport de vraisemblance

- Soit un échantillon de v.a. i.i.d., $X1, X_2, \ldots, X_n$ avec le modèle statistique $(\mathbf{E}, \mathbf{F}, (\mathrm{P}_\theta)_{\theta\in\Theta})$, où $\Theta \subseteq \mathbb{R}^d$(pour $d \geq 1$).
- On considère que l'hypothèse nulle présente la forme,

$$H_0 : (\theta_{r+1}, \theta_{r+2}, \ldots, \theta_d) = (\theta^0_{r+1}, \theta^0_{r+2}, \ldots, \theta^0_d),$$

  pour des valeurs données $\theta^0_{r+1}, \theta^0_{r+2}, \ldots, \theta^0_d$.
- Soit,

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \ell_n(\theta), \quad \text{(estimateur du MV)},$$

  et,

$$\hat{\theta}^c_n = \arg\max_{\theta\in\Theta_0} \ell_n(\theta), \quad \text{(estimateur du MV contraint)}.$$

Test du rapport de vraisemblance

- Statistique de test :

$$T_n = 2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^c)\right).$$

- **<u>Théorème :</u>** supposons que H_0 soit vraie et que les conditions techniques de l'estimateur du MV soient satisfaites. Alors,

$$T_n \xrightarrow{d} \chi^2_{d-r}, \quad \text{par rapport à } \mathrm{P}_\theta\,.$$

- Test du rapport de vraisemblance de niveau asymptotique $\alpha \in (0, 1)$:

$$\delta = \mathbf{1}_{T_n > q_{1-\alpha}},$$

où $q_{1-\alpha}$ est le quantile $(1-\alpha)$ du χ^2_{d-r}(voir tables).

Tests d'hypothèses implicites

- Soit $X_1, X_2, \ldots, X_n$ des v.a. i.i.d., et soit $\theta \in \mathbb{R}^d$ un paramètre associé à la loi de X(une v.a. "générique"). Cela peut être par exemple, un moment, un paramètre du modèle statistique, etc.
- Soit $g : \mathbb{R}^d \to \mathbb{R}^k$ une application continûment dérivable(avec $k < d$).
- Considérons les hypothèses suivantes :

$$\begin{array}{lll} H_0 : & g(\theta) & = 0 \\ H_1 : & g(\theta) & \neq 0. \end{array}$$

- Par exemple : $g(\theta) = (\theta_1, \theta_2)$(ici $k = 2$), ou $g(\theta) = \theta_1 - \theta_2$(ici $k = 1$), ...

Tests d'hypothèses implicites

- Supposons qu'on dispose d'un estimateur asymptotiquement normal $\hat{\theta}_n$:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}_d\left(\mathbf{0}, \Sigma(\theta)\right).$$

- Méthode Delta :

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Gamma(\theta)),$$

où $\Gamma(\theta) = \nabla g(\theta)^\top \Sigma(\theta) \nabla g(\theta) \in \mathbb{R}^{k \times k}$.

- Supposons que $\Sigma(\theta)$ soit inversible, et $\nabla g(\theta)$ de rang k, de sorte que $\Gamma(\theta)$ soit aussi inversible et que,

$$\sqrt{n}\Gamma(\theta)^{-1/2}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k).$$

Tests d'hypothèses implicites

- Par le théorème de Slutsky si $\Gamma(\theta)$ est continue en θ,

$$\sqrt{n}\Gamma(\hat{\theta}_n)^{-1/2}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k).$$

- En conséquence, si H_0 est vraie, i.e., $g(\theta) = 0$,

$$\underbrace{ng(\hat{\theta}_n)^\top \Gamma^{-1}(\hat{\theta}_n)g(\hat{\theta}_n)}_{T_n} \xrightarrow{d} \chi^2_k.$$

- Test de niveau asymptotique $\alpha \in (0, 1)$:

$$\delta = \mathbf{1}_{T_n > q_{1-\alpha}},$$

  où $q_{1-\alpha}$ est le quantile $(1 - \alpha)$ du χ^2_k(voir tables).

Le cas multinomial : test du χ^2

- Soit $\mathcal{E} = \{a_1, a_2, \dots, a_k\}$ un ensemble dénombrable et $(\mathrm{P}_{\mathbf{p}})_{\mathbf{p}\in\Delta_k}$ la famille de toutes les lois de probabilité sur $\mathcal{E}$:

$$\Delta_k = \{\mathbf{p} = (\mathrm{p}_1, \mathrm{p}_2, \dots, \mathrm{p}_k) \in (0,1)^K : \sum_{j=1}^{k} \mathrm{p}_j = 1\}.$$

- Pour $\mathbf{p} \in \Delta_k$ et $X \sim \mathrm{P}_{\mathbf{p}}$,

$$\mathrm{P}_{\mathbf{p}}(X = a_j) = \mathrm{p}_j, \quad j = 1, 2, \dots, k.$$

Le cas multinomial : test du χ^2

- Soit $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathrm{P}_{\mathbf{p}}$, pour un $\mathbf{p} \in \Delta_k$ inconnu, $\mathbf{p}^0 \in \Delta^K$ donné.
- On souhaite tester : "$H_0 : \mathbf{p} = \mathbf{p}^0$" vs., "$H_1 : \mathbf{p} \neq \mathbf{p}^0$" au niveau asymptotique $\alpha \in (0, 1)$.
- Exemple : si $\mathbf{p}^0 = (1/k, 1/k, \ldots, 1 : k)$, on teste si $\mathrm{P}_{\mathbf{p}}$ est la loi uniforme sur $\mathcal{E}$.

Le cas multinomial : test du χ^2

- Vraisemblance du modèle :

$$\mathcal{L}_n(X_1, X_2, \ldots, X_n, \mathbf{p}) = \mathrm{p}_1^{N_1}\,\mathrm{p}_2^{N_2}\ldots\mathrm{p}_k^{N_k},$$

où $N_j = \#\left\{i = 1, 2, \ldots, n : X_i = a_j\right\}$.

- Soit $\hat{\mathbf{p}}$ l'estimateur du MV :

$$\hat{\mathrm{p}} = \frac{N_j}{n}, \quad j = 1, 2, \ldots, k.$$

- **Remarque :** $\hat{\mathbf{p}}$ maximise $\mathcal{L}_n(X_1, X_2, \ldots, X_n, \mathbf{p})$ **sous la contrainte**,

$$\sum_{j=1}^{k} \mathrm{p}_j = 1.$$

Le cas multinomial : test du χ^2

- Si H_0 est vraie, alors $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}^0)$ est asymptotiquement normale, et le résultat suivant est établi,
- **Théorème :**

$$\underbrace{n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j^0)^2}{p_j^0}}_{T_n} \xrightarrow{d} \chi^2_{k-1}.$$

- Test du χ^2 de niveau asymptotique α : $\delta_\alpha = \mathbf{1}_{T_n > q_{1-\alpha}}$, où $q_{1-\alpha}$ est quantile d'ordre $1-\alpha$ du χ^2_{k-1}.
- La p-value asymptotique de ce test est : $\mathrm{p-value} = \mathrm{P}(Z > T_n | T_n)$, où $Z \sim \chi^2_{k-1}$ et $Z \perp\!\!\!\perp T_n$.

Cas Gaussien : test de Student

- Soit $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ pour $\mu \in \mathbb{R}$, $\sigma^2 > 0$, et soit aussi un $\mu_0 \in \mathbb{R}$ donné.
- On veut tester :

$$H_0 : "\mu = \mu_0" \text{vs.} " H_1 : \mu \neq \mu_0,$$

  au niveau asymptotique $\alpha \in (0, 1)$.
- **1er cas** : σ^2 est connu.
  - Soit $T_n = \sqrt{n}\left(\frac{\bar{X}_n - \mu_0}{\sigma}\right)$.
  - Alors $T_n \sim \mathcal{N}(0, 1)$ et,

$$\delta_\alpha = \mathbf{1}_{|T_n| > q_{1-\alpha/2}}$$

  est un test au niveau(non asymptotique) α.

Cas Gaussien : test de Student

- **2ème cas** : σ^2 est inconnu.
  - Soit $\tilde{T}_n = \sqrt{n-1}\left(\frac{\bar{X}_n-\mu_0}{\sqrt{S_n}}\right)$, où S_n est la variance empirique.
  - Théorème de Cochran :
    - $\bar{X}_n \perp\!\!\!\perp S_n$;
    - $\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}$.
  - Et par conséquent, $\tilde{T}_n \sim t_{n-1}$: loi de Student à $n-1$ degrés de liberté.

Cas Gaussien : test de Student

- Test de Student de niveau(non asymptotique) $\alpha \in (0, 1)$:

$$\delta_\alpha = \mathbf{1}_{|\tilde{T}_n| > q_{1-\alpha/2}},$$

  où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de t_{n-1}.
- Si H_1 est "$\mu > \mu_0$", le test de Student de niveau $\alpha \in (0, 1)$ est :

$$\delta'_\alpha = \mathbf{1}_{\tilde{T}_n > q_{1-\alpha}},$$

  où $q_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de t_{n-1}.

- Avantage du test de Student :
  - Non asymptotique.
  - Peut être réalisé sur des échantillons petits.
- Inconvénients du test de Student : suppose que l'échantillon est gaussien.