Test de Wald pour l'égalité des moyennes

Maxence Caucheteux, Hassen Kallala, Armand De Cacqueray, Sabri Bouafia

École des ponts et chaussées

8 janvier 2025

Introduction

Objectifs

- But : déterminer si des moyennes sont égales ou non.
- Formellement, on dispose de deux échantillons $\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,n_1})$ et $\mathbf{X}_2 = (X_{2,1}, \ldots, X_{2,n_2})$.
- On pose $\mu_1 = \mathbb{E}(X_{1,1})$ et $\mu_2 = \mathbb{E}(X_{2,1})$
- A-t-on $\mu_1 = \mu_2$?

Hypothèses

Choix des hypothèses

$$H_0 = \{\mu_1 = \mu_2\}, \quad H_1 = \{\mu_1 \neq \mu_2\}$$

Remarques :

- Test bilatéral
- De base, on suppose que les moyennes sont égales.
- On veut éviter de dire que les moyennes sont différentes alors que ce n'est pas le cas.
- On aurait également pu effectuer un test unilatéral via le choix de $H_0 = \{\mu_1 \leq \mu_2\}$ et $H_1 = \{\mu_1 > \mu_2\}$ (la pertinence du choix dépend de la situation).

Construction de la statistique de test

Pour $j \in \{1, 2\}$, on pose :

- $\mu_j = \mathbb{E}(X_{j,1})$
- $\sigma_j^2 = \mathbb{V}(X_{j,1})$ (non connus)
- $Z_{n_1,n_2} = \overline{X}_{1,n_1} - \overline{X}_{2,n_2}$ (différence empirique des moyennes des deux échantillons)

Normalité asymptotique

On peut montrer que :

$$\frac{Z_{n_1,n_2} - (\mu_1 - \mu_2)}{\sqrt{\mathbb{V}(Z_{n_1,n_2})}} \xrightarrow[n_1,n_2 \to \infty]{\text{loi}} \mathcal{N}(0, 1)$$

Mais on ne connaît pas σ_1 et σ_2.

Construction de la statistique de test

C'est pourquoi on introduit les variances empiriques :

$$\forall j \in \{1,2\}, \quad \hat{\sigma}^2_{j,n_j} = \frac{1}{n_j}\sum_{k=1}^{n_j}(X_{j,k} - \overline{X}_{j,n_j})^2$$

Forte consistance

On a :

$$\frac{\sqrt{\frac{\hat{\sigma}^2_{1,n_1}}{n_1} + \frac{\hat{\sigma}^2_{2,n_2}}{n_2}}}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} \xrightarrow[n_1,n_2\to\infty]{\text{p.s.}} 1$$

Construction de la statistique de test

Région de rejet

Avec la région de rejet

$$W_n = \left\{ Y_n := \frac{|Z_{n_1,n_2}|}{\sqrt{\frac{\hat{\sigma}^2_{1,n_1}}{n_1} + \frac{\hat{\sigma}^2_{2,n_2}}{n_2}}} \geq \phi_{1-\alpha/2} \right\},$$

le test est consistant et de niveau asymptotique α.

Rappel : $H_0 = \{\mu_1 = \mu_2\}$ et $H_1 = \{\mu_1 \neq \mu_2\}$

Calcul de la p-valeur

p-valeur

Avec $\mu = \mu_1 = \mu_2$, on a :

$$\begin{aligned} \text{p-valeur} &= \mathbb{P}_\mu(Y_n \geq y_n^{\text{obs}}) \\ &\simeq 2(1 - F_G(y_n^{\text{obs}})) \end{aligned}$$

où F_G est la fonction de répartition de la gaussienne centrée réduite.

Similarités avec le test de Wald à un échantillon

- Le test de Wald à deux échantillons généralise le test de Wald à un échantillon.
- Pour le test de Wald à un échantillon : $H_0 = \{\mu = \mu_0\}$ et $H_1 = \{\mu \neq \mu_0\}$ où μ_0 est fixé et connu.
- En prenant pour deuxième échantillon des variables aléatoires déterministes égales à μ_0 dans le test de Wald à deux échantillons, on retrouve le test de Wald à un échantillon.

Présentation du problème

- Jeu de données : une liste d'élèves du collège dont on sait s'ils ont Internet chez eux ou pas et dont on connaît le nombre d'échecs aux examens passés.
- Le nombre d'échecs à des examens mesure en quelque sorte les performances de l'élève.
- Question : Les élèves sans accès à Internet réussissent-ils autant à l'école que les élèves qui ont accès à Internet ?

Résultats

Échantillon	Élèves avec Internet	Élèves sans Internet
n	$n_1 = 827$	$n_2 = 217$
μ	$\mu_1 = 0.24$	$\mu_2 = 0.36$
σ	$\sigma_1 = 0.62$	$\sigma_2 = 0.77$

Table – Tableau des statistiques des échantillons

Calcul de la p-valeur

On a :

$$\text{p-valeur} = 0.034 \text{ i.e. } 3.4\%$$

On rejette H_0.

Rappel : $\alpha = 5\%$

Pouvoir statistique du test

- On s'intéresse au pouvoir statistique du test
- On cherche à calculer $\mathbb{P}_{\mu_1,\mu_2}(W_n)$
- On fixe les variances des deux échantillons et on trace le pouvoir en fonction de $\Delta\mu = \mu_1 - \mu_2$

Expression du pouvoir statistique

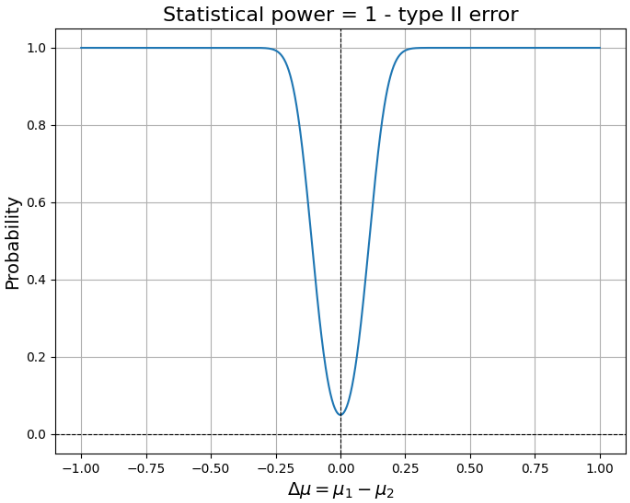
Expression du pouvoir statistique

Le pouvoir statistique de ce test est donné par la formule suivante :

$$\text{Statistical Power} = 1 - F_G\left(\phi_{1-\alpha/2} - \frac{\Delta\mu}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) + F_G\left(-\phi_{1-\alpha/2} - \frac{\Delta\mu}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

où F_G est la fonction de répartition de la normale centrée réduite.

Pouvoir statistique du test



Conclusion à tirer du test

- On rejette l'hypothèse selon laquelle les élèves avec et sans Internet réussissent autant à l'école.
- Les élèves qui n'ont pas internet réussissent moins bien.

Autre application : comparaison de proportions

- Il s'agit du cas particulier du modèle de Bernoulli.
- Les deux échantillons $\mathbf{X_1}$ et $\mathbf{X_2}$ suivent des lois de Bernoulli $\mathcal{B}(p_1)$ et $\mathcal{B}(p_2)$.
- Question du type : est-ce que $p_1 = p_2$? C'est-à-dire, est-ce que les deux groupes réussissent l'épreuve tout autant ?
- Exemples d'application :
  - Effets cliniques : comparer l'efficacité de deux traitements.
  - Marketing : évaluer le taux de clics entre deux campagnes.
  - Éducation : comparer le taux de réussite de deux méthodes pédagogiques.
  - Production : tester la qualité de deux lots industriels.
  - Environnement : comparer le taux de survie de deux espèces en milieu contrôlé.

Autre application : comparaison de proportions

- Exemple du cours : Neymar a réussi 70 pénaltys sur 85 (82%) et Mbappé en a réussi 20 sur 25 (80%)
- Question : Neymar est-il meilleur que Mbappé au pénalty ?
- p_1 (resp. p_2) : probabilité que Neymar marque un penalty (resp. p_2)
- Avec les hypothèses $H_0 = \{p_1 \leq p_2\}$ et $H_1 = \{p_1 > p_2\}$, H_0 n'est pas rejeté à 5%

Fin de l'exposé