

Cours N°9 : tests de comparaison portant sur deux échantillons.

[Dans ce cours, je ne traite pas les nombreux exemples utilisés par la prof pour vous faire comprendre les buts des tests en vous donnant des cas concrets. ALLEZ EN COURS, c'est important pour vous !]

I. Introduction

Pour savoir si des paramètres diffèrent dans les deux populations, on s'intéresse à leur différence

$$\Delta = \mu_A - \mu_B.$$

On observe alors la différence entre les valeurs expérimentales $d = m_A - m_B$.

Si d est très petit, on n'aura pas envie de conclure que Δ est différent de 0.

Si d est très grand, on pourra penser que Δ est différent de 0.

Soit X_A et X_B deux variables aléatoires indépendantes avec $E(X_A) = \mu_A$ et $E(X_B) = \mu_B$.

On donne les moyennes expérimentales m_A et m_B et la différence des moyennes $d = m_A - m_B$.

Les fluctuations d'échantillonnage de d dépendent des lois de M_A et M_B avec $D = M_A - M_B$.

$$D = M_A - M_B$$

$$E(D) = E(M_A - M_B) = \mu_A - \mu_B$$

$$\text{Var}(D) = \text{Var}(M_A - M_B) = \text{Var}(M_A) + \text{Var}(-M_B) = \text{Var}(M_A) + \text{Var}(M_B) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$$

- Si X_A et X_B suivent des lois normales, D suit la loi normale quels que soient n_A et n_B .
- Si n_A et n_B sont supérieurs ou égaux à 30, D suit une loi normale quelles que soient les lois de X_A et X_B .

$$D \sim N\left(\mu_A - \mu_B; \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)$$

II. Test de comparaison de deux moyennes

A. Hypothèses

On réalise un **test de comparaison de deux moyennes**. On s'intéresse pour cela à une **variable X continue**. On dispose de **deux échantillons A et B indépendants** de $n_A \geq 30$ et $n_B \geq 30$

observations, avec **m_A moyenne expérimentale sur A** et **m_B moyenne expérimentale sur B**.

X suit une loi quelconque de moyenne μ_A dans la population A et μ_B dans la population B.

On pose comme hypothèses :

- $H_0 : \mu_A = \mu_B$
- $H_1 : \mu_A \neq \mu_B$

B. Construction du test

Sous $H_0 : \mu_A = \mu_B$.

On construit un test basé sur la distribution de $m_A - m_B$ sous H_0 .

$$D = M_A - M_B$$

$$E(D) = \mu_A - \mu_B$$

$$\text{Var}(D) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} \quad (\text{on estime } \sigma_A^2 \text{ par } s^2 \text{ et } \sigma_B^2 \text{ par } s^2)$$

On travaille sur de **grands échantillons** $n_A \geq 30$ et $n_B \geq 30$

Soit :

$$Z = \frac{M_A - M_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Sous H_0 :

- $E(Z) = 0$ (car $M_A - M_B = 0$)

- $\text{Var}(Z) = \frac{\text{Var}(D)}{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = 1$

- Si on dispose de **grands échantillons** ($n_A \geq 30$ et $n_B \geq 30$)

Donc $Z \sim N(0 ; 1)$

On forme la quantité :

$$z = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

La **zone de rejet de H_0** est toujours la même, lorsque $|z| > 1,96$

Il ne faut pas oublier les **conditions de validité** du test ! ($n_A \geq 30$ et $n_B \geq 30$)

C. Calcul de la puissance

La **puissance** est la **probabilité de détecter que les moyennes de A et de B sont différentes si elles le sont** ($P(\text{rejet } H_0 | H_1 \text{ vraie})$).

C'est la probabilité que $|z| > 1,96$ quand A et B sont différents ($\mu_A \neq \mu_B$).

La puissance dépend d'une **hypothèse alternative spécifique** qui dépend de la valeur (non nulle) pour $\Delta = \mu_A - \mu_B$.

En reprenant la statistique du test :

$$Z = \frac{M_A - M_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Soit $s_D^2 = \frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}$

Alors $Z = \frac{M_A - M_B}{s_D}$

Sous $H_1 : \Delta \neq 0$: Z n'a plus une distribution normale centrée. On utilise alors Z' .

$$E(Z') = \frac{1}{s_D} (E[M_A] - E[M_B]) = \frac{\mu_A - \mu_B}{s_D} = \frac{\Delta}{s_D}$$

$$\text{Var}(Z') = \frac{1}{s_D^2} \text{Var}(M_A - M_B) = \frac{s_D^2}{s_D^2} = 1$$

$$\text{Donc } Z' \sim N\left(\frac{\Delta}{s_D}; 1\right)$$

$$P(|Z'| > 1,96 / \mu_A - \mu_B = \Delta) = p_1 + p_2$$

$$p_1 = P(Z' < -1,96 / Z' \sim N\left[\frac{\Delta}{s_D}; 1\right])$$

$$p_2 = P(Z' > 1,96 / Z' \sim N\left[\frac{\Delta}{s_D}; 1\right])$$

Très souvent, p_1 ou p_2 est très faible, donc **on en néglige une des deux**.

Pour **augmenter la puissance**, on peut :

- On augmente n_A et n_B .
- On fait diminuer s_D .

III. Test de comparaison de deux proportions

A. Hypothèses

On réalise un **test de comparaison de deux proportions**. On s'intéresse pour cela à une **variable X dichotomique**. On dispose de **deux échantillons A et B indépendants** de n_A et n_B observations, avec **p_A proportion de $x = 1$ sur A** et **p_B proportion de $x = 1$ sur B**.

X suit une loi de Bernoulli de paramètre π_A dans la population A et π_B dans la population B.

On pose comme hypothèses :

- $H_0 : \pi_A = \pi_B$
- $H_1 : \pi_A \neq \pi_B$

B. Construction du test

Sous $H_0 : \pi_A = \pi_B = \pi$

On construit un test basé sur la distribution de $P_A - P_B$ sous H_0 .

$$D = P_A - P_B$$

$$E(D) = \pi_A - \pi_B$$

$$Var(P_A) = \frac{\pi_A(1-\pi_A)}{n_A}$$

$$Var(P_B) = \frac{\pi_B(1-\pi_B)}{n_B}$$

$$Var(D) = Var(P_A) + Var(P_B)$$

On ne connaît pas π_A et π_B . En fait, on estime $Var(D)$ sous H_0 , en utilisant la proportion commune p estimée sur l'ensemble de l'échantillon ($n_A + n_B$).

On calcule p pour estimer π , la proportion totale de $x = 1$ sur les deux échantillons :

$$p = \frac{n_A \times p_A + n_B \times p_B}{n_A + n_B}$$

C'est la moyenne pondérée des proportions observées.

$$\text{Si } n_A = n_B; \quad p = \frac{p_A + p_B}{2}$$

Sous H_0 :

$$E(D) = 0$$

$$Var(D) \approx \frac{p(1-p)}{n_A} + \frac{p(1-p)}{n_B}$$

$$Var(D) \approx p(1-p) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)$$

$$\text{Si } n_A = n_B; \text{ alors } Var(D) \approx \frac{2p(1-p)}{n}$$

Soit :

$$Z = \frac{P_A - P_B}{\sqrt{p(1-p) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}}$$

Sous H_0 :

$$E(Z) = 0$$

$$Var(Z) = \frac{Var(D)}{p(1-p) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)} = 1$$

Si on travaille sur de grands échantillons :

- $n_A \times p \text{ ET } n_A \times (1-p) \geq 5$
- $n_B \times p \text{ ET } n_B \times (1-p) \geq 5$

Donc $Z \sim N(0 ; 1)$

On forme la quantité :

$$z = \frac{p_A - p_B}{\sqrt{p(1-p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

La **zone de rejet de H_0** est toujours la même, lorsque $|z| > 1,96$

Il faut **vérifier les conditions de validité !**

- $n_A \times p$ ET $n_A \times (1-p) \geq 5$
- $n_B \times p$ ET $n_B \times (1-p) \geq 5$

IV. Nombre de sujets nécessaires

A. Introduction et principes

On veut faire un essai qui ait une bonne chance de conclure s'il y a une différence entre les traitements. Souvent, une « **bonne puissance** » correspond à **80%** ($\beta = 20\%$).

On peut calculer le **nombre de sujets nécessaires** pour des tests de comparaison de **deux moyennes** observées ou de **deux proportions** observées sur des échantillons.

Pour calculer une puissance, on fixe un **risque de première espèce α** et un **risque de seconde espèce β** . On choisit une **différence Δ entre les deux moyennes ou les deux proportions** qu'on juge intéressante.

Le nombre de sujets doit être tel que **si la différence entre les traitements vaut Δ (H_1)**, la **probabilité de tomber dans la zone de rejet est $1 - \beta$** . Ce nombre repose donc sur le **calcul de la puissance**.

B. Nombre de sujets nécessaires pour comparer deux moyennes

Sous H_1 , Z ne suit plus une loi normale d'espérance 0 et de variance 1. $Z' \sim N\left(\frac{\Delta}{s_D}; 1\right)$

Le **calcul de la puissance** donne donc :

$$P(|Z'| > \varepsilon_{\alpha}/Z' \sim N\left[\frac{\Delta}{s_D}; 1\right]) = p_1 + p_2$$

$$P_1 = P(Z' < -\varepsilon_{\alpha}/Z' \sim N\left[\frac{\Delta}{s_D}; 1\right])$$

$$P_2 = P(Z' > \varepsilon_{\alpha}/Z' \sim N\left[\frac{\Delta}{s_D}; 1\right])$$

On néglige soit P_1 soit P_2 , donc $1 - \beta = P_1$ ou P_2 .

$$1 - \beta = P\left(Z' > \frac{\varepsilon_\alpha - \Delta}{s_d}\right)$$

Si la puissance est supérieure à 50%, $\frac{\varepsilon_\alpha - \Delta}{s_d} < 0$

$$P(Z > x_\beta) = \beta \quad \text{donc} \quad P(Z > -x_\beta) = 1 - \beta$$

$$\frac{\varepsilon_\alpha - \Delta}{s_d} = -x_\beta$$

$$\text{Or } s_d^2 = \frac{2\sigma^2}{n}$$

$$\text{Donc } x_\beta = -\varepsilon_\alpha + \Delta \sqrt{\frac{n}{2\sigma^2}}$$

$$n \geq (x_\beta + \varepsilon_\alpha)^2 \left(\frac{2\sigma^2}{\Delta^2}\right)$$

Souvent $\alpha = 5\%$, $\beta = 20\%$, $\varepsilon_\alpha = 1,96$ et $x_\beta = 0,842 \approx 0,85$.

[Dans le formulaire, page 9, vous avez divers valeurs d' ε_α et x_β et les calculs associés $(a + b)^2$]

C. Nombre de sujets nécessaires pour comparer deux proportions

On suppose qu'on a deux échantillons A et B de même taille n et qu'on travaille sur de grands échantillons.

$$H_0 : \pi_A = \pi_B$$

La statistique du test est :

$$z = \frac{p_A - p_B}{\sqrt{p(1-p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$$\text{avec } p = \frac{n_A \times p_A + n_B \times p_B}{n_A + n_B}$$

On ne connaît pas p et on se donne une valeur à priori π pour le traitement « de référence ».

Sous H_1 , Z ne suit plus une loi normale d'espérance 0 et de variance 1. $Z' \sim N\left(\frac{\Delta}{s_D}; 1\right)$

$$\text{avec } s_D = \sqrt{\frac{2\pi(1-\pi)}{n}}$$

Le calcul de la puissance est le même que pour la comparaison de moyennes en remplaçant σ^2 par $\pi(1-\pi)$.

$$1 - \beta = P\left(Z > \frac{\varepsilon_\alpha - \Delta}{s_d}\right)$$

$$\frac{\varepsilon_\alpha - \Delta}{s_d} = -x_\beta$$

$$\text{Or } s_d^2 = \frac{2\pi(1-\pi)}{n}$$

Donc :

$$n \geq (x_\beta + \varepsilon_\alpha)^2 \left(\frac{2\pi(1-\pi)}{\Delta^2} \right)$$

On prend le π de référence pour le calcul des sujets nécessaires.

On vérifie tout d'abord les conditions de validité ! Il faut que :

- $n\pi \geq 5$
- $n(1-\pi) \geq 5$

On réalise le calcul du nombre de sujets nécessaires pour avoir une **puissance suffisante de montrer une différence cliniquement pertinente**.

Ce document, ainsi que tous les cours P1, sont disponibles gratuitement sur <http://coursplbichat-larib.weebly.com>