Tests sur la moyenne

Mohamed LEMDANI

MISO
Université de Lille

29 Septembre 2021

Comparaison d'une moyenne à une moyenne théorique (test à un échantillon)

Variable observée X : quantitative continue.

Paramètre étudié : moyenne de X sur la population (μ).

Objectif : Comparer ce paramètre μ à une valeur théorique (connue) μ_0.

Hypothèses : $H_0 : \{\mu = \mu_0\}$ contre $H_1 : \{\mu \neq \mu_0\}$ (cas bilatéral) ou $\{\mu <> \mu_0\}$ (unilatéral).

Données : n observations (sur l'échantillon) de $X \Longrightarrow x_1, x_2, \ldots, x_n$.

Calcul de la moyenne $\bar{x} = \dfrac{\sum x_i}{n}$ et de la variance $s^2 = \dfrac{n}{n-1}\left(\dfrac{\sum x_i^2}{n} - \bar{x}^2\right)$.

Variable de décision : $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim St_\nu$ sous H_0, avec $\nu = n - 1$.

Conditions d'utilisation :

- $X \sim \mathcal{N}(\mu, \sigma)$ ou
- n "grand" ($n \geqslant 30$).

Préciser le risque α et construire les zones d'acceptation/rejet à partir de la table de Student.

Exemple 3

Dans le cadre d'un contrôle de qualité, pour lequel le poids moyen d'un comprimé doit être égal à 3g, on prélève et pèse 10 comprimés. On note : $\sum x_i = 32.0$g et $\sum x_i^2 = 103.84$ g^2. Peut-on conclure à un écart par rapport aux exigences de production, au seuil de 5 % ?

Variable observée : $X =$ poids d'un comprimé, observée sur un échantillon de taille $n = 10$ (représentant la population du lot de comprimés testé).

$\mu =$ moyenne de X sur le lot et $\mu_0 = 3$ g.

$H_0 : \{\mu = \mu_0\}$ *versus* $H_1 : \{\mu \neq \mu_0\}$.

Variable de décision : $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim St_9$ sous H_0.

Conditions : $n = 10$ (petit échantillon) $\Longrightarrow$ condition de normalité de X nécessaire : $X \sim \mathcal{N}(\mu, \sigma)$ (admise).

Calculs : $\bar{x} = 32.0/10 = 3.2$ g et $s^2 = \dfrac{10}{9}\left(\dfrac{103.84}{10} - 3.2^2\right) = 0.36 \Longrightarrow s = \sqrt{0.36} = 0.6$

$\Longrightarrow t_c = \dfrac{3.2 - 3}{0.6\sqrt{10}} \approx 1.054.$

Zone de non-rejet : $t_c \in [-2.262, 2.262] \Longrightarrow$ non rejet de H_0 au seuil de 5% (au risque de 5%, on ne peut pas conclure à un écart par rapport aux exigences de production).

Comparaison de deux moyennes (variances) pour deux échantillons indépendants

Variable observée X : quantitative continue.

Paramètres étudiés : moyennes (et variances) de X sur la population : μ_1, μ_2 (et σ_1^2, σ_2^2).

Objectif : Comparer μ_1 à μ_2 (et σ_1^2 à σ_2^2).

Hypothèses : $H_0 : \{\mu_1 = \mu_2\}$ *versus* $H_1 : \{\mu_1 \neq \mu_2\}$ (ou $\{\mu_1 < \mu_2\}$ ou $\{\mu_1 > \mu_2\}$).

Test paramétrique : test de Student (ou test t).

Variable de décision : $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}}$ ou $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{{s_1}^2}{n_1} + \dfrac{{s_2}^2}{n_2}}}$.

Conditions d'utilisation :

- $X \sim \mathcal{N}(\mu_1, \sigma_1)$ sur Pop 1, $X \sim \mathcal{N}(\mu_2, \sigma_2)$ sur Pop 2 **ou**
- $n_1, n_2 \geqslant 30$.

Le test de Student ne peut pas être utilisé si

- n_1 ou $n_2 < 30$ **et**
- $X \nsim \mathcal{N}$ sur cet échantillon.

Dans ce cas, utiliser un test non paramétrique (ici celui de Mann et Whitney).

Mise en œuvre du test de Student

Avant de comparer les moyennes μ_1 et μ_2, d'abord comparer les variances σ_1^2 et σ_2^2.

Hypothèses : $K_0 : \{\sigma_1^2 = \sigma_2^2\}$ *versus* $K_1 : \{\sigma_1^2 \neq \sigma_2^2\}$.

Si non-rejet de K_0 : utiliser $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$ pour tester H_0 contre H_1.

Si rejet de K_0 : utiliser $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ pour tester H_0 contre H_1.

Comparaison des variances le test de Fisher.

Variable de décision : $F = \dfrac{s_1^2}{s_2^2} \sim F_{n_2-1}^{n_1-1}$ sous H_0 ou $F = \dfrac{s_2^2}{s_1^2} \sim F_{n_1-1}^{n_2-1}$ sous H_0.

Conditions d'utilisation : identiques à celles du test t.

Choix de la formule de F ($\alpha = 5\%$) : $F = \dfrac{s_{grand}^2}{s_{petit}^2}$.

Si non-rejet de K_0, $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \sim St_{n_1+n_2-2}$ sous H_0.

Si rejet de K_0, $t \sim St_\nu$ sous H_0, avec ν "compliqué".

Exemple 4

On souhaite comparer les durées de vie moyennes de 2 types d'ampoules électriques. Pour cela on observe la durée de vie X de 10 ampoules de type "ancien" et de 12 ampoules de type "nouveau". On note, pour le type "ancien", $\sum x_i = 18\,000$ h et $\sum x_i^2 = 32\,760\,000$ h^2, pour le type "nouveau", $\sum x_i = 24\,000$ h et $\sum x_i^2 = 48\,550\,000$ h^2.
Peut-on affirmer que les ampoules de type nouveau sont plus efficaces au seuil de 5% ?
On admettra la normalité de la durée de vie d'une ampoule, pour chacun des types.
Variable observée : $X =$ 'Durée de vie d'une ampoule' (h) observée sur deux échantillons de tailles $n_1 = 10$ et $n_2 = 12$ (1= 'Ancien' et 2 = 'Nouveau').

$H_0 : \{\mu_1 = \mu_2\}$ *versus* $H_1 : \{\mu_1 < \mu_2\}$.

$K_0 : \{\sigma_1 = \sigma_2\}$ *versus* $K_1 : \{\sigma_1 \neq \sigma_2\}$.

Calculs : $\bar{x}_1 = 18\,000/10 = 1\,800$ h et $\bar{x}_2 = 24\,000/12 = 2\,000$ h.

$$s_1^2 = \frac{10}{9}\left(\frac{32\,760\,000}{10} - 1\,800^2\right) = 40\,000 \text{ h}^2.$$

$$s_2^2 = \frac{12}{11}\left(\frac{48\,550\,000}{12} - 2\,000^2\right) = 50\,000 \text{ h}^2.$$

Exemple 4 (suite)

$X =$ 'Durée de vie d'une ampoule' (h).

$H_0 : \{\mu_1 = \mu_2\}$ *versus* $H_1 : \{\mu_1 < \mu_2\}$ et $K_0 : \{\sigma_1 = \sigma_2\}$ *versus* $K_1 : \{\sigma_1 \neq \sigma_2\}$.

$n_1 = 10, n_2 = 12, \bar{x}_1 = 1\,800$ h, $\bar{x}_2 = 2\,000$ h, $s_1^2 = 40\,000$ h^2 et $s_2^2 = 50\,000$ h^2.

Conditions : $X \sim \mathcal{N}(\mu_1, \sigma_1)$ sur Pop 1, $X \sim \mathcal{N}(\mu_2, \sigma_2)$ sur Pop 2 (énoncé).

Test K_0/K_1 : $F = s_2^2/s_1^2 \sim \mathcal{F}_9^{11}$, sous K_0.

Zone de rejet de K_0 (5%) : $F_c \notin [3.9, +\infty[\Longrightarrow$ non-rejet de K_0 au seuil de 5%.

$F_c = 50\,000/40\,000 = 1.25$.

Variance commune : $s^2 = \dfrac{9 \times 40\,000 + 11 \times 500\,000}{10 + 12 - 2} = 45\,500$ h^2.

$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \sim St_{20}$ sous H_0.

Zone de rejet unilatérale ($\alpha = 5\%$) : $t_c \in \,]-\infty, -1.725]$.

Rejet $\Rightarrow$ choix de $H_1 \Rightarrow \mu_1 < \mu_2 \Rightarrow \bar{x}_1 < \bar{x}_2 \Rightarrow t < 0$: rejet côté $-\infty$, α(table) $= 10\%$.

Calcul : $t_c = \dfrac{1\,800 - 2\,000}{\sqrt{\frac{45\,500}{10} + \frac{45\,500}{12}}} \approx -2.190 \Longrightarrow$ Rejet de H_0 au seuil de 5% (au risque 5%, on peut dire qu'en moyenne les ampoules de type "Récent" ont une durée de vie plus longue).

Problématique

X : variable quantitative continue, définie sur deux populations $\Longrightarrow$ comparer les moyennes μ_1 et μ_2 de X $\Longrightarrow$ comparaison à partir d'échantillons indépendants.
Exemple : comparer les tailles moyennes d'arbres de deux espèces différentes.

Cas de "traitements" pour une même population : comparer deux traitements (Population = Patients), deux méthodes de dosages (Population = Flacons), ...
Possibilité de prendre un échantillon pour chaque traitement (affectation par tirage au sort de deux groupes indépendants) ou de constituer un seul échantillon où chaque individu subit les deux traitements (séries appariées).

Test t échantillons indépendants : $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$ = Différence/Variabilité.

$$\text{Variabilité} = \begin{cases} \text{Intergroupe} + \text{Intragroupe en échantillons indépendants,} \\ \text{Intragroupe en échantillons appariés.} \end{cases}$$

Mise en évidence plus facile d'une différence dans le cas d'échantillons appariés.

Séries appariées : essai croisé (*cross over*)

- Échantillon n de sujets suivant deux traitements A et B (l'un puis l'autre).
- Tirage au sort : groupe "A suivi de B" et groupe "B suivi de A".
- Période de *wash out* entre les deux traitements.
- Critère de jugement observé à deux reprises : X (pour A) et Y (pour B).

Présentation des données :

Sujets	1	2	...	n
X	X_1	X_2	...	X_n
Y	Y_1	Y_2	...	Y_n
D	$D_1 = X_1 - Y_1$	$D_1 = X_2 - Y_2$	...	$D_n = X_n - Y_n$

Hypothèses : $H_0 : \{\mu_X = \mu_Y\}$ *versus* $H_1 : \{\mu_X \neq \mu_Y\}$ (ou $\{\mu_X < \mu_Y\}$ ou $\{\mu_X > \mu_Y\}$).

Calcul de $D = X - Y$ (ou $D = Y - X$).

Hypothèses (réécriture) : $H_0 : \{\mu_D = 0\}$, $H_1 : \{\mu_D \neq 0\}$ (ou $\{\mu_D < 0\}$ ou $\{\mu_D > 0\}$)

$\Longrightarrow$ test de type Observé/Théorique (à un échantillon).

Test paramétrique (Student) : $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{\bar{d}}{s_D/\sqrt{n}}$.

Séries appariées : test t (suite)

$H_0 : \{\mu_D = 0\}$ *versus* $H_1 : \{\mu_D \neq 0\}$ (ou $\{\mu_D < 0\}$ ou $\{\mu_D > 0\}$).

$t = \dfrac{\bar{d}}{s_D/\sqrt{n}} \sim St_{n-1}$ sous H_0.

Conditions :

- $D \sim \mathcal{N}(\mu_D, \sigma_D)$ **ou**
- $n \geqslant 30$.

Conditions non remplies ($n < 30$ et $D \nsim \mathcal{N}$) $\Longrightarrow$ utiliser un test non paramétrique (test de Wilcoxon).

Autres types d'appariements : familial, par voisinage, . . .

Exemple 5 : On souhaite comparer les rendements moyens à l'hectare, pour une céréale, entre deux types d'engrais A et B. Pour cela, on dispose de 10 parcelles situées dans des régions différentes et dont chacune est découpée en deux : une moitié où sera utilisé l'engrais A et l'autre ou l'on utilisera l'engrais B. On note, à la fin de la saison, le rendement à l'hectare X (en tonnes). Peut-on conclure à un rendement plus élevé en moyenne pour l'engrais A au seuil de 10% ?

Exemple 5 (suite)

Parcelle	1	2	3	4	5	6	7	8	9	10
X_A	2.5	4.3	6.6	5.4	3.8	4.2	3.9	4.7	2.9	3.7
X_B	2.6	4.0	6.4	5.2	3.8	4.1	3.8	4.8	2.7	3.6
D	−0.1	0.3	0.2	0.2	0.0	0.1	0.1	−0.1	0.2	0.1

Échantillon de taille $n = 10$ avec deux observations X_A et X_B pour chaque parcelle.

$H_0 : \{\mu_A = \mu_B\}$ *versus* $H_1 : \{\mu_A > \mu_B\}$.

$D = X_A - X_B \Longrightarrow \quad H_0 : \{\mu_D = 0\}$ *versus* $H_1 : \{\mu_D > 0\}$.

Variable de décision : $t = \dfrac{\bar{d}}{s_D/\sqrt{n}} \sim St_9$ sous H_0.

Conditions : $n = 10 \Longrightarrow$ nécessité de la condition de normalité ($D \sim \mathcal{N}(\mu_D, \sigma_D)$).

Zone de rejet unilatérale à 10% : $t_c \in [1.383, +\infty[$ (rejet de H_0 au seuil de 10%).

$H_1 \Longrightarrow \mu_D > 0 \Longrightarrow t > 0$.

Calculs : $\sum d_i = 1.0$, $\sum d_i^2 = 0.26$.

$$\bar{d} = 1.0/10 = 0.1 \text{ et } s_D^2 = \frac{10}{9}\left(\frac{0.26}{10} - 0.1^2\right) = 0.16/9 = 16/(9 \times 100).$$

$$s_D = \sqrt{16/(9 \times 100)} = 4/30 \Longrightarrow t_c = \frac{0.1 \times \sqrt{10}}{4/30} \approx 2.372.$$