TESTS DE COMPARAISON DE DEUX MOYENNES

A) *Cas de deux échantillons indépendants*

Le problème :
On étudie une grandeur X sur deux échantillons indépendants, issus a priori de deux populations différentes Ω_1, Ω_2.
Exemple : X est la hauteur de tige d'une plante, que l'on mesure sur deux échantillons : l'un issu d'une population Ω_1 de plantes cultivées sur un substrat azoté, et l'autre issu d'une population Ω_2 de plantes de la même espèce cultivées sur un substrat pauvre en azote.
On se place du point de vue probabiliste, on considère le choix d'un échantillon comme une expérience aléatoire.

On note :
– m_1 (resp. m_2) la moyenne théorique (ou espérance) de X sur Ω_1 (resp. Ω_2).
m_1 et m_2 sont des réels dont les valeurs ne sont pas connues.

– n_1 (resp. n_2) l'effectif de l'échantillon n°1 (resp. n°2).
– $\overline{X}_1$ (resp. $\overline{X}_2$) la moyenne observée de X sur l'échantillon n°1 (resp. n°2), ce sont des variables aléatoires qui, sur les deux échantillons dont on dispose, prennent les valeurs $\overline{x}_1$ et $\overline{x}_2$ (réels).

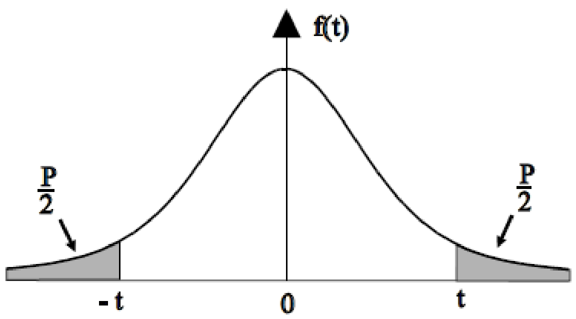
$\overline{x}_1$ et $\overline{x}_2$ peuvent être différents, mais cette différence peut être due au hasard de l'échantillonnage, ou au fait que les moyennes théoriques m_1, m_2 sont très différentes.
On s'appuie sur l'observation de $\overline{x}_1$ et $\overline{x}_2$ pour décider si on peut considérer que $m_1 \neq m_2$ (test bilatéral) ou $m_1 > m_2$ (test unilatéral); si ce n'est pas le cas, on considère qu'on ne peut pas rejeter l'hypothèse que la différence observée entre $\overline{x}_1$ et $\overline{x}_2$ soit simplement due au hasard de l'échantillonnage.
Dans l'exemple, si le test permet de décider que $m_1 > m_2$, cela peut laisser penser que la présence d'azote favorise la croissance des tiges.

Principe du test

On souhaite tester à l'aide des observations l'hypothèse $m_1 = m_2$ (hypothèse nulle, notée $\mathcal{H}_0$) contre une autre hypothèse $\mathcal{H}_1$, appelée hypothèse alternative.
$\mathcal{H}_1$ peut prendre plusieurs formes :
– soit $m_1 \neq m_2$: test bilatéral
– soit par exemple $m_1 > m_2$ (si par ailleurs on a de bonnes raisons de penser que la présence d'azote a tendance à augmenter la hauteur des tiges) : test unilatéral

1) Pour cela on se fixe un **niveau de risque**, ou niveau de signification $\alpha \in [0, 1]$: en général $\alpha = 5\%$:
c'est **le risque de 1ère espèce, ou encore la probabilité de rejeter $\mathcal{H}_0$ alors qu'elle est vraie, ou encore la probabilité de détecter une différence entre les moyennes alors qu'il n'y en a pas.**

On sait que lorsque $\mathcal{H}_0$ est vraie, une certaine statistique T (variable aléatoire calculée à partir des observations, dépendant du test utilisé, appelée statistique de test, qui rend compte de l'écart entre les observations sur les deux échantillons) suit une certaine loi connue (souvent donnée par des tables, ou, aujourd'hui, calculée par les ordinateurs) et dépendant du type de test choisi .
Pour tout réel t, on sait donc calculer $P_{\mathcal{H}_0}(T \geq t)$, ou $P_{\mathcal{H}_0}(|T| \geq t)$.
Dans les deux cas couramment rencontrés (loi normale ou loi de Student), ces grandeurs sont représentées par les aires de domaines situés sous la courbe de la densité de T dont la forme est la suivante :



La fonction $t \mapsto P_{\mathcal{H}_0}(T \geq t)$ est décroissante : $P_{\mathcal{H}_0}(T \geq t)$ diminue quand t augmente. (Idem pour $P_{\mathcal{H}_0}(|T| \geq t)$)

De manière générale, si $\mathcal{H}_0$ est vraie, T et $|T|$ ont peu de chances de prendre de grandes valeurs.

Pour $t \geq 0$, $P_{\mathcal{H}_0}(|T| \geq t) = P_{\mathcal{H}_0}([T \geq t] \cup [T \leq -t]) = P_{\mathcal{H}_0}(T \geq t) + P_{\mathcal{H}_0}(T \leq -t)$

Pour les statistiques T couramment utilisées $P_{\mathcal{H}_0}(T \leq -t) = P_{\mathcal{H}_0}(T \geq t)$ donc $\boxed{P_{\mathcal{H}_0}(|T| \geq t) = 2\ P_{\mathcal{H}_0}(T \geq t).}$

<u>**Idée**</u> : On calcule la valeur observée de T sur les deux échantillons, notée T_{obs}.
Si T_{obs} est " grand" (en valeur absolue pour un test bilatéral), comme on avait peu de chances d'observer ce qu'on observe sous l'hypothèse $\mathcal{H}_0$, on est amené à rejeter cette hypothèse, donc à penser que $\mathcal{H}_1$ est vraie .

2) <u>**Choix de la statistique de test T:**</u>
<u>**Cas de deux échantillons de grande taille ($n_1 \geq 30$, $n_2 \geq 30$) : test Z**</u>

La statistique T est plutôt notée Z : $Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\dfrac{S'^2_1}{n_1} + \dfrac{S'^2_2}{n_2}\right)}}$ où $S'^2_i = \dfrac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(X_{k,i} - \overline{X_i}\right)^2$

Si $\mathcal{H}_0$ est vraie, Z suit la loi normale centrée réduite.

$Z_{obs} = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\left(\dfrac{s'^2_1}{n_1} + \dfrac{s'^2_2}{n_2}\right)}}$ où s'^2_i est la réalisation de S'^2_i sur l'échantillon n°i.

<u>**Cas de deux échantillons de petite taille. ($n_1 < 30$, $n_2 < 30$) : test T**</u>

Sous l'hypothèse que les variances de X sur Ω_1 et Ω_2 sont égales, on prend la <u>statistique T de Student</u> :

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S'^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \text{ où } S'^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{k=1}^{n_1}(X_{k,1} - \overline{X}_1)^2 + \sum_{k=1}^{n_2}(X_{k,2} - \overline{X}_2)^2\right).$$

Si $\mathcal{H}_0$ est vraie, T suit généralement la loi de Student-Fisher à $n_1 + n_2 - 2$ degrés de liberté.

$$T_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s'^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \quad \text{avec} \quad s'^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{k=1}^{n_1}(x_{k,1} - \overline{x}_1)^2 + \sum_{i=1}^{n_2}(x_{k,2} - \overline{x}_2)^2\right)$$

3) <u>**Réalisation pratique du test**</u> :

- <u>**Méthode "à l'ancienne":**</u>
  On obtient à l'aide d'une table de la loi sachant $\mathcal{H}_0$ de T ou Z , ou d'un logiciel, un réel t_α (ou t_α') appelé <u>valeur critique</u> tel que
  - $P_{\mathcal{H}_0}(|T| \geq t_\alpha) \approx \alpha$ pour un test bilatéral
  - $P_{\mathcal{H}_0}(T \geq t_\alpha') \approx \alpha$ pour un test unilatéral

  (avec α= 5%, , sous l'hypothèse $\mathcal{H}_0$, $|T|$ a au plus 5% de chances de dépasser t_α et T a au plus 5% de chances de dépasser t'_α)
  - Si $|T_{obs}| \geq t_\alpha$ (test bilatéral) ou $T_{obs} \geq t'_\alpha$ (test unilatéral) on rejette alors $\mathcal{H}_0$: on admet que $\mathcal{H}_1$ est vraie, c'est-à-dire que la différence entre les moyennes observées n'est pas due qu'au hasard : dans l'exemple, le facteur azote semble avoir une influence sur la taille de la tige; si on espérait mettre en évidence cette différence le résultat du test est dit <u>significatif</u>.
  - Si $|T_{obs}| < t_\alpha$ (ou $T_{obs} < t_\alpha'$) on ne peut pas rejeter $\mathcal{H}_0$, donc le test ne met pas en évidence une différence entre m_1 et m_2 ; on dit alors que le résultat du test n'est <u>pas significatif</u> (mais il ne *prouve* pas que $m_1 = m_2$)

<u>Utilisation de la méthode "à l'ancienne" avec LibreOffice ou OpenOffice ou EXCEL</u> :

- **Pour calculer T_{obs} ou Z_{obs} :**

  On obtient $s'^2_i = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(x_{k,i} - \overline{x}_i\right)^2$ à l'aide de la fonction VAR, en transmettant comme argument la plage des données $x_{k,i}$ sur l'échantillon n°i .

  On obtient $\sum_{k=1}^{n_i}\left(x_{k,i} - \overline{x}_i\right)^2$ à l'aide de la fonction SOMME.CARRES.ECARTS, en transmettant comme argument la plage des données $x_{k,i}$ sur l'échantillon n°i.

- **Pour un test Z :** Pour α = 5% la table donne $P_{\mathcal{H}_0}(|Z| \geq 1{,}96) \approx 0{,}05$ et $P_{\mathcal{H}_0}(Z \geq 1{,}65) \approx 0{,}05$

- **Pour un test de Student** :
  - Pour obtenir le seuil t_α tel que $P_{\mathcal{H}_0}(|T| \geq t_\alpha) \approx \alpha$, où ddl est le nombre de degrés de liberté : utiliser la fonction LOI.STUDENT.INVERSE.BILATERALE :
    syntaxe : =LOI.STUDENT.INVERSE.BILATERALE(alpha,ddl)
  - Pour obtenir le seuil t'_α tel que $P_{\mathcal{H}_0}(T \geq t'_\alpha) \approx \alpha$, où ddl est le nombre de degrés de liberté : utiliser la même fonction, mais en remplaçant α par 2α.

- **<u>Méthode moderne avec utilisation de l'informatique</u> :**

On ne calcule pas de valeur critique, mais on calcule la <u>probabilité associée à T_{obs}</u> (appelée *p*-value en anglais): c'est $P_{\mathcal{H}_0}(|T| \geq |T_{obs}|)$ pour un test bilatéral, $P_{\mathcal{H}_0}(T > T_{obs})$ pour un test unilatéral.

Cette probabilité ne dépend pas de α, et elle diminue lorsque $|T_{obs}|$ ou T_{obs} augmente ; si elle est petite, on avait peu de chances d'observer ce qu'on observe sous l'hypothèse $\mathcal{H}_0$, ce qui amène à rejeter cette hypothèse.

On compare la probabilité associée avec le niveau de risque α :

- si $P_{\mathcal{H}_0}(|T| \geq |T_{obs}|) \leq \alpha$, (ou $P_{\mathcal{H}_0}(T > T_{obs}) \leq \alpha$ on rejette $\mathcal{H}_0$: le résultat du test <u>est significatif</u>
- sinon on ne peut pas rejeter $\mathcal{H}_0$: le résultat n'est <u>pas significatif</u>.

Avant de faire un test de Student on devrait faire un test d'égalité des variances :
avec <u>LibreOffice ou OpenOffice</u> : fonction TESTF(données1, données2)
avec <u>EXCEL</u> : F.TEST(données1, données2)
Ces fonctions renvoient une autre *p-value*
-si le résultat est inférieur à α on rejette l'égalité des variances
-sinon on l'accepte

<u>Obtention de la probabilité associée à T_{obs} avec un test T</u> de Student (p-value)

avec <u>LibreOffice ou OpenOffice</u> : : **fonction TESTSTUDENT**

pour un test bilatéral : =TESTSTUDENT(données1,données2, 2, type)
pour un test unilatéral : =TESTSTUDENT(données1,données2, 1, type), où la série de données 1 correspond à celle de plus grande moyenne (avec l'hypothèse alternative $m_1 > m_2$)
prendre type=2 si on accepte l'égalité des variances, et type = 3 sinon

<u>avec Excel</u> : Fonction T.TEST
pour un test bilatéral : =T.TEST(données1, données2, 2, type)
pour un test unilatéral : = T.TEST(données1, données2, 1, type) , où la série de données 1 correspond à celle de plus grande moyenne (avec l'hypothèse alternative $m_1 > m_2$)

prendre type=2 si on accepte l'égalité des variances, et type = 3 sinon