Tests de comparaison de moyenne

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2017-10-02

Contents

Contenu

Nous présentons ici l'une des applications les plus populaires de la statistique: le test de comparaison de moyennes.

Ce test est utilisé dans un grand nombre de contextes, nous l'appliquerons ici à deux types de données:

1. Des **données artificielles** générées en tirant des échantillons au sein de deux populations qui suivent des distributions normales, et qui, selon le cas, présentent ou non une différence de moyenne. Le but sera de comprendre la mise en application du test et l'interprétation de ses résultats dans des conditions où nous contrôlons tous les paramètres (nous savons s'il existe ou non une différence entre les moyennes de populations).
2. Des **données transcriptomiques** obtenues au moyen de biopuces. Nous testerons si un gène donné présente une différence d'expression entre deux groupes d'échantillons (par exemple des patients souffrant de deux types de cancers différents).

   Note: la technologie des biopuces a largement été remplacée par le séquençage à haut débit, et le RNA-seq a rapidement été adopté pour les études transcriptomiques. Cependant l'analyse d'expression différentielle avec le RNA-seq requiert des concepts plus avancés, qui feront partie de cours ultérieurs.

L'hypothèse à tester

Principe général:

- On observe une différence entre les moyennes d'échantillons, et on veut savoir si celle-ci résulte du hasard de l'échantillonnage ou d'une différence réelle entre les populations.
- On pose l'***hypothèse nulle*** (H_0) selon laquelle il n'existe pas de différence entre les deux populations. L'***hypothèse alternative*** (H_1) est qu'il existe une différence.
- On évalue la probabilité que les échantillons aient été générés selon cette hypothèse nulle.
- Si cette probabilité est trop faible, on rejette l'hypothèse nulle (RH_0).
- Sinon, on accepte (temporairement) l'hypothèse nulle (AH_0).

Test bilatéral (*two-tailed test*)

$$H_0 : \mu_1 = \mu_2 H_1 : \mu_1 \neq \mu_2$$

Il s'agit ici d'un **test bilatéral** (***two-tailed test***): nous désirons détecter une éventuelle différence indépendamment de son signe.

Test unilatéral (*one-tailed test*)

Modalité alternative: **test unilatéral** (***one-tailed test***), où l'on ne s'intéresse qu'à des différences allant dans une direction donnée.

Différences positives (*right-tailed test*):

$$H_0 : \mu_1 \leq \mu_2 H_1 : \mu_1 > \mu_2$$

Différences négatives (*left-tailed test*):

$$H_0 : \mu_1 \geq \mu_2 H_1 : \mu_1 < \mu_2$$

Hypothèses de travail

- Il existe différentes méthodes pour tester l'égalité entre deux moyennes.
- Le choix de la méthode dépend de la nature des données.
- Avant de réaliser le test il est crucial de se poser quelques questions afin de choisir la méthode appropriée.
- Il s'agit de **vérifier les hypothèses de travail**, c'est-à-dire des hypothèses sur lesquelles reposera l'approche envisagée.

Hypothèse de normalité

Les populations dont les échantillons sont tirés suivent-elles des distributions normales ?

- Si oui on peut réaliser des tests paramétriques (ils reposent sur une hypothèse de normalité). Dans le cas contraire il faut recourir à des tests non-paramétriques.

  Pourquoi ? Les tables de probabilités de risques des tests paramétriques ont été établies sur base de modèles mathématiques reposant sur l'hypothèse de normalité.

- En cas de non-normalité, **les échantillons sont-ils de grande taille ?** Si oui on peut se rabattre sur les méthodes paramétriques

  Pourquoi ? En vertu du **théorème central limite**, les moyennes d'échantillon tendent vers une normale même si les populations-mères ne sont pas normales.

Hypothèse d'homoscédasticité (égalité des variances)

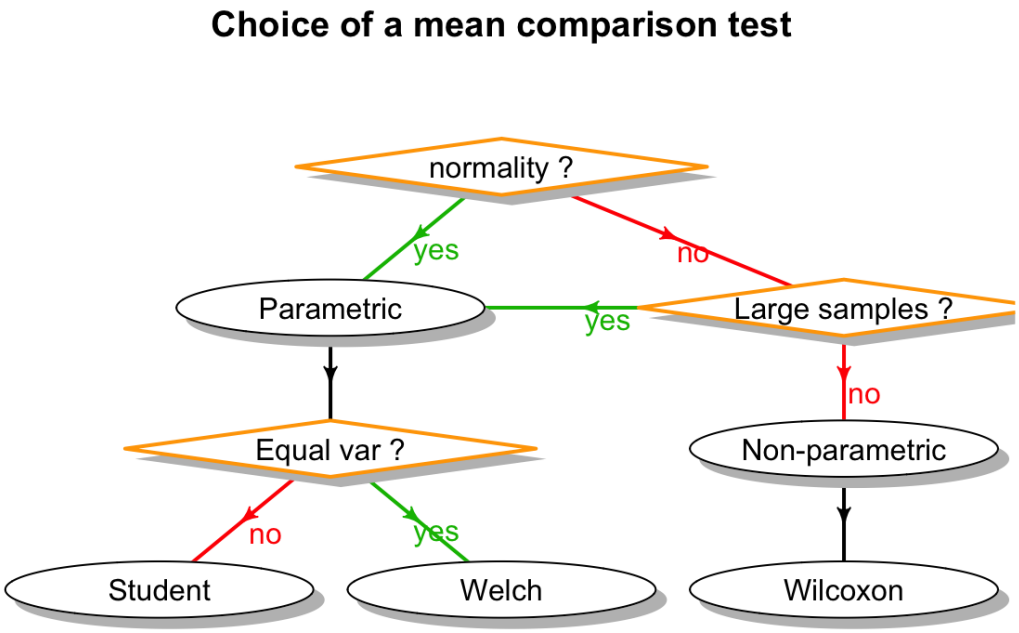
Pour les tests paramétriques, **les populations ont-elles (vraisemblablement) la même variance ?**

- Si oui on peut appliquer le test de Student.

- Dans le cas contraire, test de Welch.

  Pourquoi ? La distribution de probabilité de Student a été calculée sur base d'une hypothèse d'**homoscédasticité** (égalité des variances). Le test de Welch effectue une correction en cas d'**hétéroscédasticité** (inégalité des variances), en modifiant le nombre de degrés de liberté en fonction des différences entre variances.

Schéma du choix d'un test de comparaison de moyenne



Test de Student

Hypothèses d travail: **normalité** (ou bien grands échantillons), **homoscédasticité**.

Statistique:

$$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Sens du test	Critère de décision
Bilatéral	RH_0 if $\lvert t_S\rvert > t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}$
Unilatéral à droite	RH_0 if $t_S > t_{1-\alpha}^{n_1+n_2-2}$

Sens du test	Critère de décision
Unilatéral à gauche	$RH_0 \quad \text{if} \quad t_S < t_{alpha}^{n_1+n_2-2} = -t_{1-\alpha}^{n_1+n_2-2}$

Exercice

Un chercheur a analysé, à l'aide de biopuces, le niveau d'expression de l'ensemble des gènes à partir d'échantillons sanguins prélevés chez 50 patients (np=50) et chez 50 sujets témoins (nt=50). Il s'intéresse particulièrement à un gène qui semble montrer une différence entre les 2 groupes. Ainsi, il ré-analyse l'expression du même gène dans les mêmes échantillons en utilisant une autre technique, la qPCR. Il obtient

- pour les patients, une moyenne $m_p = 21$
- pour les contrôles, une moyenne $m_t = 10$
- des écarts-types identiques pour les 2 groupes $s_p = s_t = s = 15$

Afin de tester si la différence observée entre les moyennes est significative, le chercheur décide d'effectuer un test de Student.

a. Le choix du test de Student vous semble-t-il approprié ? Justifiez le choix du chercheur. Quelles auraient été les alternatives envisageables ?

b. Sachant qu'a priori on ne sait pas dans quel sens la maladie pourrait affecter le niveau d'expression du gène, préconisez-vous un test uni- ou bilatéral ?

c. Formulez l'hypothèse nulle et expliquez-la en une phrase.

d. Sur base de la formule ci-dessous, calculez la statistique t de Student.

e. Indiquez, en vous basant sur les tables fournies, la p-valeur correspondante.

f. Interprétez la p-valeur, et aidez le chercheur à tirer les conclusions de son étude.

Exercice

Un groupe de chercheurs a détecté l'association, avec la résistance à la bilharziose, de taux élevés d'IgE spécifiques, une classe particulière d'anticorps. D'autres chercheurs ont cherché à répliquer ce résultat dans une population indépendante exposée à la bilharziose. Les résultats obtenus sont indiqués ci-dessous.

- Pour les sujets résistants ($n_r = 32$), la moyenne $m_r = 10$.
- Pour les sujets susceptibles ($n_s = 32$), la moyenne $m_s = 7$.
- Les écarts-types des deux groupes sont égaux : $s_r = s_s = s = 2.8$.

a. Quelle méthode préconisez-vous pour tester l'égalité des moyennes (justifiez) ? Quelles sont les hypothèses de travail de ce test ?

b. En partant du principe que ces conditions sont remplies dans le cas présent, formulez l'hypothèse nulle et calculez le score t de Student (formule ci-dessous). Enfin, estimez P valeur à partir de la table fournie.

c. A l'issue du test, quelle décision prenez-vous ? Justifiez votre réponse.