

INFERENCE STATISTIQUE II:

COMPARAISON DE MOYENNES

Test de Student pour un seul échantillon

Exemple 1. Selon les courbes de croissance, il est raisonnable de supposer que la taille moyenne des *Suisesses* à l'âge de 18 ans est de 164 cm (v. TP I question 4.2). Sur la base de l'échantillon d'étudiants (premier chapitre) nous allons tester l'hypothèse \mathcal{H}_0 que taille moyenne de la *population d'étudiantes* de laquelle l'échantillon a été pris est de 164 cm.

En général:

- On considère une population \mathcal{P} .
- On s'intéresse à une **variable quantitative** X pour les individus de \mathcal{P} .
- Soit μ la moyenne de population (inconnue) de X
- Soit μ_0 **une valeur hypothétique moyenne** de X . On veut tester

$$\mathcal{H}_0 : \mu = \mu_0$$

contre une des alternatives suivantes:

$$\mathcal{H}_1 : \mu \neq \mu_0 \quad \text{ou bien} \quad \mathcal{H}_1 : \mu > \mu_0 \quad \text{ou bien} \quad \mathcal{H}_1 : \mu < \mu_0.$$

- On utilisera un échantillon x_1, \dots, x_n de taille n .

Procédé de test approximatif

- on calcule
la moyenne $m(X)$
l'écart type $s(X)$
l'**écart standardisé** entre $m(X)$ et μ_0 :

$$t = \frac{m(X) - \mu_0}{s(X)/\sqrt{n}}$$

- on s'appuie sur le résultat théorique suivant.

Distribution d'échantillonnage de t : Sous \mathcal{H}_0 , et si n est "grand" les valeurs possibles de t ont approximativement une distribution de Gauss standard.

Que signifie n "grand" ?

Malheureusement, la réponse dépend de la distribution de X !

Remarque: $s(X)/\sqrt{n}$ est l'écart type des valeurs possibles de $m(X)$.
Il diminue si n augmente.

Test approximatif bilatéral

Hypothèses: $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu \neq \mu_0$

Règle de décision: fixer un *niveau* α (p.ex. 10%) et

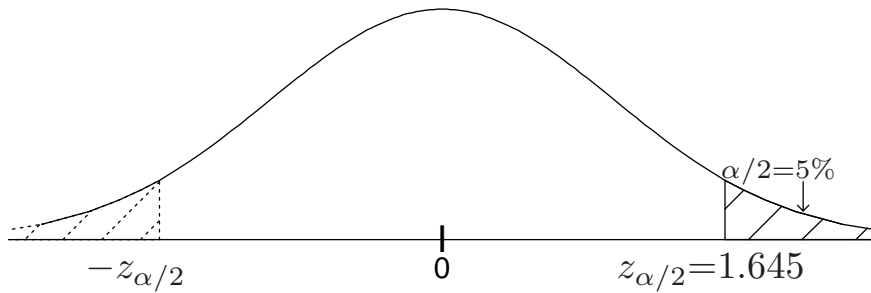
rejeter \mathcal{H}_0 si $t < -z_{\alpha/2}$ ou si $t > z_{\alpha/2}$.

La probabilité de rejeter \mathcal{H}_0 si elle est vraie est

$$\text{Prob}(t < -z_{\alpha/2} \text{ ou } t > z_{\alpha/2}) \approx \alpha$$

ayant défini $z_{\alpha/2}$ de la façon suivante:

α	10%	5%	1%
$z_{\alpha/2}$	1.645	1.960	2.576



Exemple 1, continuation

Soit $\mathcal{H}_0 : \mu = 164\text{cm}$ et $\mathcal{H}_1 : \mu \neq 164\text{cm}$, $\alpha = 10\%$.

En utilisant la formule $s^2(X) = \sum(x_i - m(X))^2 / (n - 1)$ on obtient:

$$m(X) = 164.2\text{cm}, \quad s(X) = 4.09\text{cm}, \quad t = 0.19.$$

$\Rightarrow \mathcal{H}_0$ ne peut pas être rejetée.

Test approximatif *unilatéral*

Hypothèses: $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu > \mu_0$

Règle de décision: rejeter \mathcal{H}_0 si $t > z_\alpha$

La probabilité de rejeter \mathcal{H}_0 si elle est vraie est

$$\text{Prob}(t > z_\alpha) \approx \alpha$$

ayant défini z_α de la façon suivante:

α	10%	5%	1%
z_α	1.282	1.645	2.326

Procédé usuel: test de Student ou t-test pour un seul échantillon

- On s'appuie sur le résultat théorique suivant.

Si la condition

C : la distribution de population de X est Gaussienne

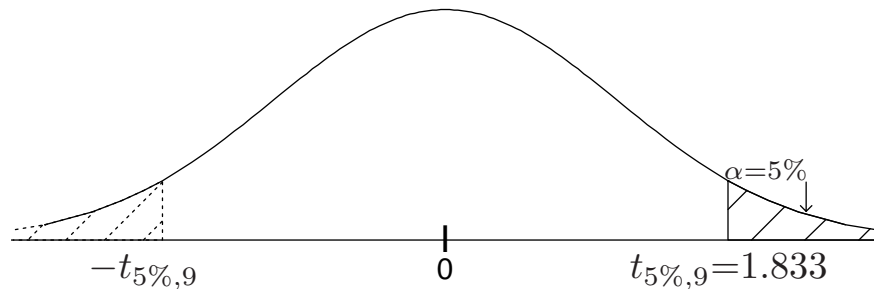
alors, sous \mathcal{H}_0 , la distribution d'échantillonnage de t est une distribution de Student à $n - 1$ degrés de liberté.

- Les distributions de Student sont symétriques (similaires à une Gaussienne).
- Elles dépendent d'un paramètre appelé degrés de liberté (d.d.l.)
- Si le d.d.l est fixé, la distribution est complètement spécifiée (\rightarrow table).
- Si d.d.l. > 30 , Student \approx Gauss standard.

Soit $t_{\alpha, n-1}$ le percentile $1 - \alpha$ de cette distribution.

D'après la table on a par exemple:

$$\begin{aligned} t_{5\%, 9} &= 1.833, & t_{2.5\%, 9} &= 2.262, \\ t_{5\%, 14} &= 1.761, & t_{2.5\%, 14} &= 2.145, \\ t_{5\%, \infty} &= 1.645, & t_{2.5\%, \infty} &= 1.960. \end{aligned}$$



- En pratique, les pas suivants sont nécessaires.

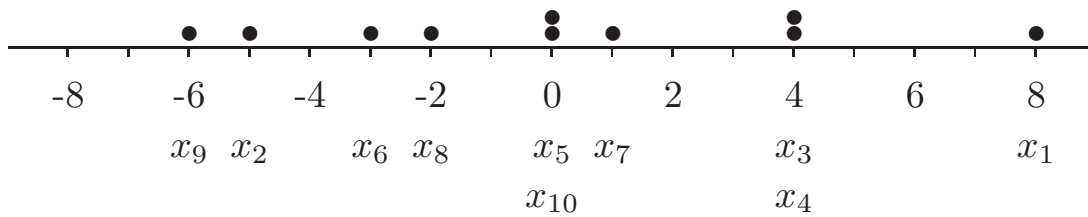
Premier pas: analyse graphique

Représenter les données graphiquement dans le but de s'assurer que la condition C ne soit pas clairement violée.

Exemple 2. Supposons que les valeurs observées x_i sont:

$$x_1 = 8, x_2 = -5, x_3 = 4, x_4 = 4, x_5 = 0, x_6 = -3,$$

$$x_7 = 1, x_8 = -2, x_9 = -6, x_{10} = 0.$$



L'analyse graphique de ces données ne suggère pas de violations de C .

Il ne serait pas approprié d'appliquer le test de Student aux données suivantes:



Deuxième pas: calcul

Calculer t

Troisième pas: décision pour le test au niveau α

- Si $\mathcal{H}_1 : \mu \neq \mu_0$: rejeter \mathcal{H}_0 si $t > t_{\alpha/2, n-1}$ ou $t < -t_{\alpha/2, n-1}$.
- Si $\mathcal{H}_1 : \mu > \mu_0$: rejeter \mathcal{H}_0 si $t > t_{\alpha, n-1}$.
- Si $\mathcal{H}_1 : \mu < \mu_0$: rejeter \mathcal{H}_0 si $t < -t_{\alpha, n-1}$.

Exemple 2, continuation.

Soit $\mathcal{H}_0 : \mu_0 = 0$, $\mathcal{H}_1 : \mu_0 \neq 0$, $\alpha = 10\%$.

On obtient: $n = 10$, $m(X) = 0.1$, $s(X) = 4.35$, et $t = 0.072$.

Donc, $-1.833 < t < 1.833$ et on ne peut pas rejeter \mathcal{H}_0 au niveau 10%.

Intervalle de confiance pour μ

L'intervalle de confiance 95% (bilatéral) pour μ est l'ensemble de toutes les valeurs hypothétiques μ_0 que seraient acceptées par un test (bilatéral) de niveau 5%.

Interprétations: intervalle des valeurs "plausibles" de μ

On démontre que

$$\left(m(X) - t_{\alpha/2, n-1} \frac{s(X)}{\sqrt{n}}, \quad m(X) + t_{\alpha/2, n-1} \frac{s(X)}{\sqrt{n}} \right)$$

est un intervalle de confiance $1 - \alpha$ pour μ .

Exemple 2, continuation.

L'intervalle $(-3.02, 3.22)$ est un intervalle de confiance 95% pour μ .

Remarque. En principe, on peut construire des intervalles de confiance pour une proportion ou pour toute autre caractéristique moyenne de population.

Données appariées et non appariées

Considérons le problème de comparer deux ensembles de données.
Selon leur structure, il faut distinguer deux cas.

Données non-appariées

Exemple 3. Est-ce que la concentration lipidique chez des sujets avec un problème circulatoire et chez des sujets sains est la même ?

Sains:	4.90	5.40	5.60	5.90	6.20	6.75			
Malades:	5.40	6.00	6.25	6.50	6.60	6.75	7.40	7.90	

Une mesure est effectuée pour chaque sujet.

Données appariées

Exemple 4. Est-ce que la concentration lipidique se modifie si le sang est conservé pendant un certain temps ? Les échantillons de sang de 10 sujets d'une certaine population ont été analysés immédiatement après la prise de sang et 8 mois après.

Avant:	74	80	75	136	104	102	90	100	95	84
Après:	66	85	71	132	104	105	89	102	101	84

Deux mesures sont effectuées pour chaque unité d'observation.

On se demande si les deux mesures de chaque échantillon sont suffisamment éloignées pour qu'on puisse décider qu'il y a un effet de la conservation.

Test de Student pour données appariées

On considère les différences d_i entre les données appariées et on réduit l'analyse des données originales à l'analyse des différences.

Soit D la variable qui représente la différence et μ sa moyenne de population.

On teste $\mathcal{H}_0 : \mu = 0$ en utilisant le test de Student pour un échantillon.

La condition d'application est

$$C : D \sim \mathcal{N}(\mu, \sigma^2),$$

c'est-à-dire, la distribution de D est Normale de moyenne μ et variance σ^2 .

Exemple 4, continuation

Les différences d_i entre les paires de mesures sont

$$\begin{aligned}d_1 = 8, \quad d_2 = -5, \quad d_3 = 4, \quad d_4 = 4, \quad d_5 = 0, \quad d_6 = -3, \\d_7 = 1, \quad d_8 = -2, \quad d_9 = -6, \quad d_{10} = 0.\end{aligned}$$

Ces différences coïncident avec les données analysées dans l'Exemple 2.

Le test de Student pour données non-appariées

- On considère deux populations \mathcal{P}_1 et \mathcal{P}_2
- Soit X une variable quantitative pour les individus de \mathcal{P}_1 et \mathcal{P}_2
- Soit μ_1 la moyenne de X dans \mathcal{P}_1 et μ_2 la moyenne de X dans \mathcal{P}_2
- Tester $\mathcal{H}_0 : \mu_1 = \mu_2$ contre une des alternatives suivantes
 $\mathcal{H}_1 : \mu_1 \neq \mu_2$ ou bien $\mathcal{H}_1 : \mu_1 < \mu_2$ ou bien $\mathcal{H}_1 : \mu_1 > \mu_2$.
- On s'appuie sur un échantillon x_1, \dots, x_m de X dans \mathcal{P}_1
et un échantillon x'_1, \dots, x'_n de X dans \mathcal{P}_2

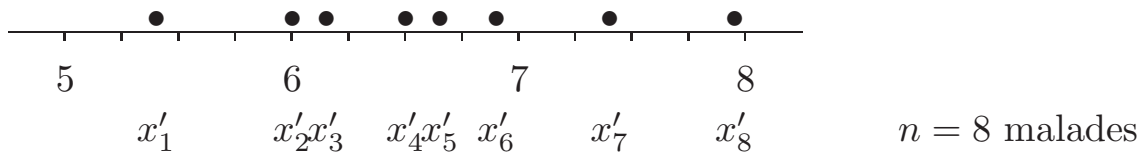
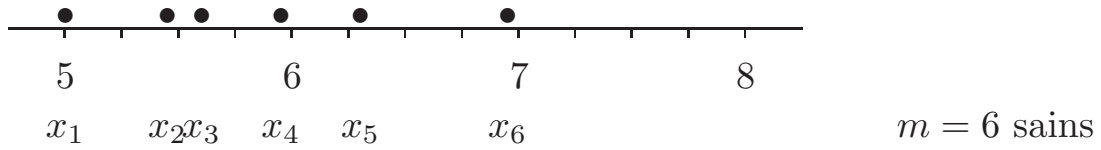
Condition d'application

C : Les deux échantillons proviennent de deux populations Gaussiennes avec la même variance.

Premier pas: analyse graphique

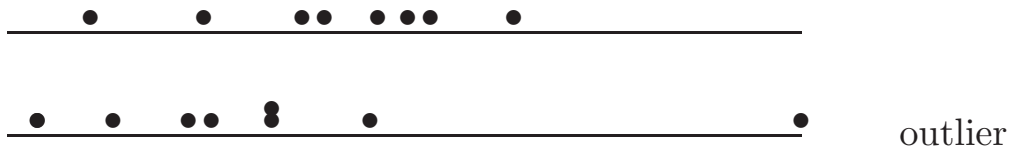
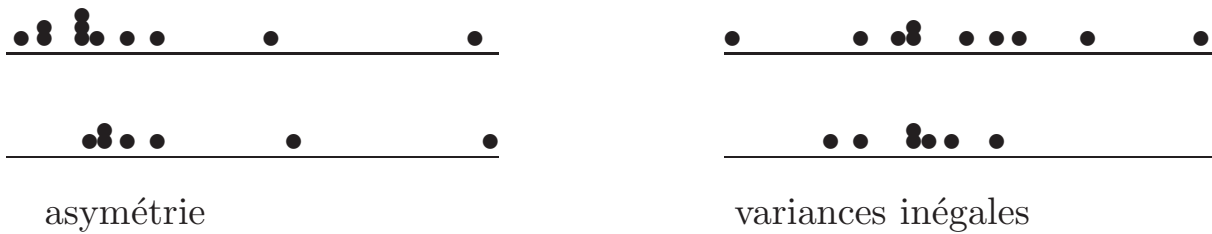
Représenter les données graphiquement dans le but de déterminer si la condition d'application est approximativement satisfaite.

Exemple 3, continuation



L'analyse graphique de ces données ne suggère pas de violations de C .

Il ne serait pas approprié d'appliquer le test aux données suivantes:



Deuxième pas: calcul

Calculer

$$m(X), \quad m(X'), \quad s(X), \quad s(X'), \quad d = m(X) - m(X'),$$

$$s(D) = \sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)s(X)^2 + (n-1)s(X')^2}{(m-1) + (n-1)}},$$

$$t = \frac{d}{s(D)}.$$

Troisième pas: décision pour le test au niveau α^*

- Si $\mathcal{H}_1 : \mu_1 \neq \mu_2$, rejeter \mathcal{H}_0 si $t > t_{\alpha/2, m+n-2}$ ou $t < -t_{\alpha/2, m+n-2}$.
- Si $\mathcal{H}_1 : \mu_1 < \mu_2$, rejeter \mathcal{H}_0 si $t < -t_{\alpha, m+n-2}$.
- Si $\mathcal{H}_1 : \mu_1 > \mu_2$, rejeter \mathcal{H}_0 si $t > t_{\alpha, m+n-2}$.

Exemple 3, continuation

On s'intéresse à $\mathcal{H}_1 : \mu_1 < \mu_2$ et à un niveau de 5%.

On a: $m(X) = 5.79$, $m(X') = 6.60$, $d = -0.81$, $s(X) = 0.782$, $s(X') = 0.645$,
 $s(D) = 0.393$, $t = -2.06$, $m + n - 2 = 12$.

A l'aide des tables, on trouve que pour une distribution de Student à 12 d.d.l.:

$$P(t < -1.782 \text{ ou } t > 1.782) = 10\% \text{ et donc } t_{5\%, 12} = 1.782.$$

Comme $-2.06 < -1.782$, On peut donc rejeter \mathcal{H}_0 .

* Résultat théorique

Sous \mathcal{H}_0 et C , la distribution d'échantillonnage de t est une distribution de Student à $(m + n - 2)$ degrés de liberté.

Complément: la “p-value”

Dans les articles, la “décision” du test est souvent laissée au lecteur. L’auteur lui fournit la p-value.

Soit w_0 la valeur observée d’une certaine statistique du test. La p-value est la probabilité sous \mathcal{H}_0 qu’un autre échantillonnage de la même population produise une statistique de test w aussi extrême ou plus extrême que w_0 .

Exemple 3, continuation

On a $w_0 = -2.06$.

A l’aide des tables on peut seulement établir que $p = P(t \leq -2.06) < 5\%$.

Un calcul exact (par ex. avec Excel) donne $p = P(t \leq -2.06) = 0.031$.

Comme $p < 5\%$ le lecteur peut rejeter \mathcal{H}_0 au niveau 5%.