

Module 8.2 - Tests d'hypothèses à deux échantillons

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Introduction

Objectif

Avec UN échantillon :

Tester si un paramètre prend une valeur précise ($H_0 : \mu = 200$)

Avec DEUX échantillons :

Tester si le paramètre prend la même valeur dans deux populations

($H_0 : \mu_1 = \mu_2$).

Exemples :

- La déformation moyenne d'un matériau sous contrainte est-elle la même à 20 °C et à 80 °C ?
- La concentration moyenne de particules fines mesurée en zone urbaine est-elle supérieure à celle mesurée en zone rurale ?
- Les vibrations mesurées sur deux types de moteurs présentent-elles la même variabilité ?

Méthode d'analyse

Les mêmes étapes de construction du test d'hypothèses qu'avec une population :

- 1) Formuler H_0 et H_1 .
- 2) Établir une règle pour le rejet de H_0 basée sur un seuil α et une loi échantillonnale appropriée.
- 3) Utiliser les données pour calculer la statistique observée.
- 4) Décider de rejeter ou non H_0 .
- 5) Énoncer la conclusion dans les termes de l'étude.

5 cas étudiés

$H_0 : \sigma_1^2 = \sigma_2^2$ Éch. indép.	1) $X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$ $X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$
$H_0 : \mu_1 = \mu_2$ Échantillons indépendants	2) $X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$ $X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$ } σ_1^2 et σ_2^2 connues
	3) $X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \sigma^2)$ $X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \sigma^2)$ } σ_1^2 et σ_2^2 inconnues mais égales
	4) $X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$ $X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$ } σ_1^2 et σ_2^2 inconnues mais inégales
$H_0 : \mu_1 = \mu_2$ Éch. appariés	5) $X_{11}, \dots, X_{1n} \sim N(\mu_1, \sigma_1^2)$ $X_{21}, \dots, X_{2n} \sim N(\mu_2, \sigma_2^2)$ } σ_1^2 et σ_2^2 inconnues, dépend. par paires

Comparaison de deux variances

Cas 1 : Comparaison de deux variances

Deux échantillons **indépendants** :

$$X_{11}, \dots, X_{1n_1} \stackrel{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$$

$$X_{21}, \dots, X_{2n_2} \stackrel{i.i.d.}{\sim} N(\mu_2, \sigma_2^2),$$

où μ_1, μ_2, σ_1^2 et σ_2^2 sont inconnus.

On désire tester **$H_0 : \sigma_1^2 = \sigma_2^2$** . Si H_0 est vraie, alors

$$F_0 = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}.$$

Règles de décision

Au seuil α ,

- $H_1 : \sigma_1^2 \neq \sigma_2^2$: on rejette H_0 si $f_0 > F_{\alpha/2, n_1-1, n_2-1}$ ou
si $f_0 < F_{1-\alpha/2, n_1-1, n_2-1}$
- $H_1 : \sigma_1^2 > \sigma_2^2$: on rejette H_0 si $f_0 > F_{\alpha, n_1-1, n_2-1}$
- $H_1 : \sigma_1^2 < \sigma_2^2$: on rejette H_0 si $f_0 < F_{1-\alpha, n_1-1, n_2-1}$

Exemple 1

Un ingénieur en contrôle qualité souhaite comparer la performance de deux fournisseurs de tiges métalliques utilisées dans la fabrication de structures d'appui. Pour cela, il prélève 5 tiges au hasard de chaque fournisseur et mesure la résistance à la rupture (en kN) de chaque tige dans un test de traction standardisé.

Fournisseur	Résistances (en kN)					$\bar{x}$	s^2
A	12,1	11,8	12,0	12,2	11,9	12,00	0,025
B	12,3	12,1	12,2	12,4	12,4	12,28	0,017

Exemple 1 (suite)

En supposant que la loi normale est appropriée pour les résistances, peut-on conclure au seuil $\alpha = 0,05$ que la variabilité des résistances des deux fournisseurs est différente ?

Comparaison de deux moyennes

Rappel

- On pige un échantillon de taille $n_1 = 10$ dans une population où $X \sim N(30, 16)$.
- On pige un échantillon de taille $n_2 = 25$ dans une population où $Y \sim N(10, 25)$.

1. Quelle est la loi de $\bar{X}$?

2. Quelle est la loi de $\bar{X} - \bar{Y}$?

Cas 2 : Populations normales, variances connues, échantillons indépendants

On recueille **deux échantillons indépendants** dans des populations où la variable X est distribuée selon des lois normales de moyennes μ_1 et μ_2 et de variances σ_1^2 et σ_2^2 **connues**.

$$X_{11}, \dots, X_{1n_1} \stackrel{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$$

$$X_{21}, \dots, X_{2n_2} \stackrel{i.i.d.}{\sim} N(\mu_2, \sigma_2^2).$$

On connaît la loi des moyennes :

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

$$\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

On déduit la loi de la différence des moyennes :

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Cas 2 : Statistique de test

Sous l'hypothèse nulle $H_0 : \mu_1 = \mu_2$, on a que

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

Au seuil α ,

- $H_1 : \mu_1 \neq \mu_2$: on rejette H_0 si $|z_0| > z_{\alpha/2}$
- $H_1 : \mu_1 > \mu_2$: on rejette H_0 si $z_0 > z_\alpha$
- $H_1 : \mu_1 < \mu_2$: on rejette H_0 si $z_0 < -z_\alpha$

Cas 3 et 4 : Et si les variances sont inconnues ?

Si σ_1^2 et σ_2^2 sont inconnues, on doit utiliser un estimateur des variances des deux distributions :

- si les variances sont égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), on peut **combiner les estimations de variance des deux échantillons pour estimer cette variance**;

Cas 3 et 4 : Et si les variances sont inconnues ?

Si σ_1^2 et σ_2^2 sont inconnues, on doit utiliser un estimateur des variances des deux distributions :

- si les variances sont égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), on peut **combiner les estimations de variance des deux échantillons pour estimer cette variance** ;
- si les variances ne sont pas égales ($\sigma_1^2 \neq \sigma_2^2$), on doit **estimer σ_1^2 et σ_2^2 séparément à partir de chaque échantillon**.

Cas 3 et 4 : Et si les variances sont inconnues ?

Si σ_1^2 et σ_2^2 sont inconnues, on doit utiliser un estimateur des variances des deux distributions :

- si les variances sont égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), on peut **combiner les estimations de variance des deux échantillons pour estimer cette variance** ;
- si les variances ne sont pas égales ($\sigma_1^2 \neq \sigma_2^2$), on doit **estimer σ_1^2 et σ_2^2 séparément à partir de chaque échantillon**.

Il faut faire un test de comparaison des variances (test F)
pour distinguer ces deux situations, comme on vient de le voir au cas 1.

Cas 3 : Lois normales, $\sigma_1^2 = \sigma_2^2$ inconnues

Si le test F ne rejette pas l'hypothèse nulle d'égalité des variances, on combine ces deux estimateurs :

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2$$

pour obtenir une estimation globale (plus précise) de σ^2 :

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Cas 3 : Lois normales, $\sigma_1^2 = \sigma_2^2$ inconnues

Si le test F ne rejette pas l'hypothèse nulle d'égalité des variances, on combine ces deux estimateurs :

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2$$

pour obtenir une estimation globale (plus précise) de σ^2 :

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Sous $H_0 : \mu_1 = \mu_2$,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right]\right)$$

Statistique de test

Si H_0 est vraie,

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}.$$

Au seuil α ,

- $H_1 : \mu_1 \neq \mu_2$: on rejette H_0 si $|t_0| > t_{\alpha/2, n_1+n_2-2}$
- $H_1 : \mu_1 > \mu_2$: on rejette H_0 si $t_0 > t_{\alpha, n_1+n_2-2}$
- $H_1 : \mu_1 < \mu_2$: on rejette H_0 si $t_0 < -t_{\alpha, n_1+n_2-2}$

Cas 4 : Lois normales, $\sigma_1^2 \neq \sigma_2^2$ inconnues

Si le test F rejette l'hypothèse nulle d'égalité des variances, on doit estimer séparément les variances σ_1^2 et σ_2^2 . Il n'y a pas de statistique T exacte pour tester $H_0 : \mu_1 = \mu_2$. Par contre

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_v \quad \text{où} \quad v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Au seuil α ,

- $H_1 : \mu_1 \neq \mu_2$: on rejette H_0 si $|t_0^*| > t_{\alpha/2,v}$
- $H_1 : \mu_1 > \mu_2$: on rejette H_0 si $t_0^* > t_{\alpha,v}$
- $H_1 : \mu_1 < \mu_2$: on rejette H_0 si $t_0^* < -t_{\alpha,v}$

Nomenclature

- Cas 3 : ce test pour variances égales est souvent appelé **test-t à deux échantillons** (*two-sample t-test*).
- Cas 4 : ce test approximatif pour variances inégales est souvent appelé **test de Welch** (*Welch's test*).

Exemple 1 (suite)

Toujours avec les données de résistance à la rupture et en se fiant au résultat du test du cas 1, a-t-on que les moyennes sont égales, au seuil $\alpha = 5\%$?

Fournisseur	Résistances (en kN)					$\bar{x}$	s^2
A	12,1	11,8	12,0	12,2	11,9	12,00	0,025
B	12,3	12,1	12,2	12,4	12,4	12,28	0,017

Exemple 1 (suite)

Exemple 1 (suite)

On aimerait maintenant comparer le fournisseur A à un nouveau fournisseur, qu'on nomme C.

Fournisseur	Résistances (en kN)					$\bar{x}$	s^2
A	12,1	11,8	12,0	12,2	11,9	12,00	0,025
C	11,7	12,9	11,8	12,7	12,1	12,24	0,288

Sachant que nous rejetons H_0 au test d'égalité des variances, a-t-on que les moyennes sont égales au seuil $\alpha = 5\%$?

Exemple 1 (suite)

Deux scénarios d'échantillonnage

On veut comparer deux types de bâtons de golf. On collecte 40 observations.

Scénario 1 :

20 golfeurs frappent une balle avec le bâton 1 (on note la distance).

20 autres golfeurs frappent avec le bâton 2 (on note la distance).

On note donc 40 observations indépendantes.

Scénario 2 :

20 golfeurs frappent une balle avec chacun des bâtons 1 et 2 dans un ordre aléatoire.

On note 20 paires d'observations.

Notion d'appariement

- Scénario 1 : Comment distinguer si la différence observée est due aux golfeurs ou aux bâtons ?

$V(\bar{X}_1 - \bar{X}_2)$ risque d'être trop élevée pour conclure qu'un bâton est meilleur que l'autre.

- Scénario 2 : On contrôle la variabilité entre les golfeurs en les faisant frapper avec les deux bâtons.

S'il y a une différence significative, elle sera due aux bâtons et non à la variabilité entre golfeurs, pour autant que l'ordre de frappe (1 avant 2 ou vice-versa) soit aléatoire pour chaque golfeur (pour contrer les effets de la fatigue, par exemple).

Notion d'appariement

- Scénario 1 : Comment distinguer si la différence observée est due aux golfeurs ou aux bâtons ?

$V(\bar{X}_1 - \bar{X}_2)$ risque d'être trop élevée pour conclure qu'un bâton est meilleur que l'autre.

- Scénario 2 : On contrôle la variabilité entre les golfeurs en les faisant frapper avec les deux bâtons.

S'il y a une différence significative, elle sera due aux bâtons et non à la variabilité entre golfeurs, pour autant que l'ordre de frappe (1 avant 2 ou vice-versa) soit aléatoire pour chaque golfeur (pour contrer les effets de la fatigue, par exemple).

Problème : Dans le scénario 2, les X_{1j} ne sont pas indépendants des X_{2j} .
On ne peut pas utiliser les tests vus aux cas 2, 3 et 4 dans le scénario 2.

Cas 5 : Populations normales, échantillons appariés

Considérons plutôt les différences D_j qui, elles, **sont indépendantes** :

$$D_1 = X_{11} - X_{21}$$

$$D_2 = X_{12} - X_{22}$$

...

$$D_n = X_{1n} - X_{2n}$$

Un échantillon de n différences

Dans le cas de deux échantillons appariés, un test de $H_0 : \mu_1 = \mu_2$ est équivalent à tester $H_0 : \mu_D = 0$ dans le cas d'un seul échantillon $D_1, \dots, D_n$ d'une loi normale avec variance inconnue.

La statistique du test est donc

$$T_0 = \frac{\bar{D}}{S_D / \sqrt{n}} \sim t_{n-1} \quad \text{sous } H_0,$$

où

$$\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j = \frac{1}{n} \sum_{j=1}^n (X_{1j} - X_{2j}) \text{ et} \quad S_D^2 = \frac{1}{n-1} \sum_{j=1}^n (D_j - \bar{D})^2.$$

Un échantillon de n différences

Dans le cas de deux échantillons appariés, un test de $H_0 : \mu_1 = \mu_2$ est équivalent à tester $H_0 : \mu_D = 0$ dans le cas d'un seul échantillon $D_1, \dots, D_n$ d'une loi normale avec variance inconnue.

La statistique du test est donc

$$T_0 = \frac{\bar{D}}{S_D / \sqrt{n}} \sim t_{n-1} \quad \text{sous } H_0,$$

où

$$\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j = \frac{1}{n} \sum_{j=1}^n (X_{1j} - X_{2j}) \text{ et} \quad S_D^2 = \frac{1}{n-1} \sum_{j=1}^n (D_j - \bar{D})^2.$$

Le test bilatéral apparié rejette $H_0 : \mu_D = 0$ au seuil α si $|t_0| > t_{\alpha/2, n-1}$. On adapte la règle pour les tests unilatéraux :

Un échantillon de n différences : règles de décision

Pour tester $H_0 : \mu_1 = \mu_2 \Leftrightarrow \mu_D = 0$, au seuil α avec

- $H_1 : \mu_1 \neq \mu_2 \Leftrightarrow \mu_D \neq 0$: on rejette H_0 si $|t_0| > t_{\alpha/2, n-1}$
- $H_1 : \mu_1 > \mu_2 \Leftrightarrow \mu_D > 0$: on rejette H_0 si $t_0 > t_{\alpha, n-1}$
- $H_1 : \mu_1 < \mu_2 \Leftrightarrow \mu_D < 0$: on rejette H_0 si $t_0 < -t_{\alpha, n-1}$

Quelques remarques importantes

- C'est **le contexte** qui détermine si les données sont appariées ou non.

Quelques remarques importantes

- C'est **le contexte** qui détermine si les données sont appariées ou non.
- Si on analyse des données appariées sans tenir compte de l'appariement, on risque de diminuer grandement la significativité des données (c.-à-d. de surestimer le seuil observé). [QUIZ : Pourquoi?]

Quelques remarques importantes

- C'est **le contexte** qui détermine si les données sont appariées ou non.
- Si on analyse des données appariées sans tenir compte de l'appariement, on risque de diminuer grandement la significativité des données (c.-à-d. de surestimer le seuil observé). [QUIZ : Pourquoi?]
- Il est avantageux d'utiliser l'appariement dans une expérience **quand les unités expérimentales sont hétérogènes**.

Exemple 2

On veut savoir si les étudiants de 16 ans sont aussi forts en mathématiques en juin qu'en septembre.

On choisit 8 étudiants au hasard qui passent un examen similaire avant et après les vacances.

Voici leurs résultats.

Étudiant	1	2	3	4	5	6	7	8	Moyenne
Examen 1	77	74	82	73	87	69	66	81	76,125
Examen 2	72	68	76	68	84	68	61	76	71,625

Exemple 2 (suite)

En supposant que les notes à chaque examen suivent une loi normale, que peut-on conclure sur la différence de moyennes entre le premier et le deuxième examen, au seuil $\alpha = 0,05$?

Exemple 2 (suite)

QUIZ!

Par erreur, vous faites l'analyse de l'exemple 2 comme s'il s'agissait d'échantillons indépendants avec variances égales.

1. Quelle est votre règle de rejet de H_0 ?
2. La valeur observée de la statistique du test est 1,28. Quelle est votre conclusion ?

Réponses aux exemples

1. $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$ (fournisseurs A et B). Puisque $0,104 < f_0 = 1,47 < 9,60$ ($\alpha^* = 0,72$), on ne rejette pas l'égalité des variances.
 $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ (cas 3). Puisque $|t_0| = 3,06 > 2,306$ ($\alpha^* = 0,016$), on rejette H_0 . La résistance moyenne du fournisseur B est significativement supérieure à celle du fournisseur A.
Pour les fournisseurs A et C, $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ (cas 4). Puisque $|t_0^*| = 0,96 \ngtr 2,62$ ($\alpha^* = 0,38$), on ne rejette pas H_0 . Les résistances moyennes de A et C ne sont pas significativement différentes.
2. $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$. Puisque $|t_0| = 7,53 > 2,365$ ($\alpha^* = 0,00013$), on rejette H_0 . La moyenne à l'examen 1 est significativement supérieure à celle de l'examen 2.