

Module 9 - Tests d'hypothèses, partie 2

STT-1000 Probabilités et statistique

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 concentration moyenne de phosphate dans le sol est-elle la même dans la région A que dans la région B ?
- La variabilité du débit de la rivière C est-elle inférieure à celle du débit de la rivière D ?

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

$H_0 : \mu_1 = \mu_2$ Éch. indép.	1) $\left. \begin{array}{l} X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \sigma_1^2) \\ X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \sigma_2^2) \end{array} \right\} \begin{array}{l} \sigma_1^2 \text{ et } \sigma_2^2 \\ \text{connues} \end{array}$
	2) $\left. \begin{array}{l} X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \sigma^2) \\ X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \sigma^2) \end{array} \right\} \begin{array}{l} \sigma_1^2 \text{ et } \sigma_2^2 \text{ inconnues} \\ \text{mais égales} \end{array}$
	3) $\left. \begin{array}{l} X_{11}, \dots, X_{1n_1} \sim N(\mu_1, \sigma_1^2) \\ X_{21}, \dots, X_{2n_2} \sim N(\mu_2, \sigma_2^2) \end{array} \right\} \begin{array}{l} \sigma_1^2 \text{ et } \sigma_2^2 \text{ inconnues} \\ \text{mais inégales} \end{array}$
$H_0 : \mu_1 = \mu_2$ Éch. appariés	4) $\left. \begin{array}{l} X_{11}, \dots, X_{1n} \sim N(\mu_1, \sigma_1^2) \\ X_{21}, \dots, X_{2n} \sim N(\mu_2, \sigma_2^2) \end{array} \right\} \begin{array}{l} \sigma_1^2 \text{ et } \sigma_2^2 \text{ inconnues,} \\ \text{dépend. par paires} \end{array}$
$H_0 : p_1 = p_2$ Éch. indép.	5) $\begin{array}{l} X_{11}, \dots, X_{1n_1} \sim \text{Bernoulli}(p_1) \\ X_{21}, \dots, X_{2n_2} \sim \text{Bernoulli}(p_2) \end{array}$

Comparaison de deux moyennes

QUIZ!

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

Cas 1 : 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 1 : 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 2 et 3 : 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 2 et 3 : 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 2 et 3 : 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 (pas dans ce cours).

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

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

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

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

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

v est ensuite arrondi à l'entier le plus proche (pas nécessairement vers le bas) pour trouver la valeur critique dans la table de Student.

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

Exemple 1

Des échantillons d'eau sont prélevés de 2 rivières en plusieurs lieux, toujours à la même profondeur.

On mesure la concentration de zinc, en $\mu\text{g}/l$, considérée comme une variable aléatoire normale.

Lieu	1	2	3	4	5	6	$\bar{x}$	s^2
Rivière 1	6,51	5,43	4,32	5,73	3,90	4,11	5,00	1,092
Rivière 2	6,03	4,92	3,60	5,13	2,58		4,45	1,850

Exemple 1 (suite)

Au seuil de 5%, peut-on dire que la concentration moyenne est plus élevée dans la rivière 1 que dans la rivière 2?

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 A (on note la distance).

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

On note 40 observations indépendantes.

Scénario 2 :

20 golfeurs frappent une balle avec chacun des bâtons A et B.

On note 20 paires d'observations.

Lequel des scénarios est préférable ?

Notion d'appariement

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

$Var(\bar{X}_A - \bar{X}_B)$ 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.

Notion d'appariement

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

$Var(\bar{X}_A - \bar{X}_B)$ 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.

Problème : les X_{Aj} ne sont pas indépendants des X_{Bj} .

On ne peut pas utiliser les tests vus aux cas 1, 2 et 3.

Cas 4 : Populations normales, échantillons appariés

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

$$D_1 = X_{A1} - X_{B1}$$

$$D_2 = X_{A2} - X_{B2}$$

...

$$D_n = X_{An} - X_{Bn}$$

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.

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).

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).
- 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 ?

Comparaisons de deux proportions

Cas 5 : Comparaison de deux proportions

Dans le cas de proportion, on a $X_{11}, \dots, X_{1n_1} \stackrel{i.i.d.}{\sim} \text{Bernoulli}(p_1)$, et $X_{21}, \dots, X_{2n_2} \stackrel{i.i.d.}{\sim} \text{Bernoulli}(p_2)$.

Alors, si $n_1 > 30$ et $n_2 > 30$, par le TCL

$$\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} \approx N \left(p_1, \frac{p_1(1-p_1)}{n_1} \right)$$

$$\hat{p}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i} \approx N \left(p_2, \frac{p_2(1-p_2)}{n_2} \right)$$

Sous l'hypothèse nulle, $p_1 = p_2 = p$ et on a donc

$$\hat{p}_1 - \hat{p}_2 \approx N \left(0, p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

Cas 5 : Comparaison de deux proportions

La statistique du test est alors

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ où } \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Un test de niveau approximativement α rejette H_0 si

- $H_1 : p_1 \neq p_2$: on rejette H_0 si $|z_0| > z_{\alpha/2}$
- $H_1 : p_1 > p_2$: on rejette H_0 si $z_0 > z_\alpha$
- $H_1 : p_1 < p_2$: on rejette H_0 si $z_0 < -z_\alpha$

Réponses des exemples

Réponses des exemples

- **Exemple 1 (zinc, rivières 1 et 2) : test-t à deux échantillons indépendants à variances égales, $S_p^2 \approx 1,429$ et $t_0 = (5,00 - 4,45) / \sqrt{S_p^2(1/6 + 1/5)} \approx 0,76$; comme $t_0 = 0,76 < t_{0,05,9} = 1,833$, on ne rejette pas H_0 : on ne peut pas conclure, au seuil de 5 %, que la concentration moyenne est plus élevée dans la rivière 1.**
- **Exemple 2 (examens de juin et septembre, échantillons appariés) : différences $D_j = X_{1j} - X_{2j}$ de moyenne $\bar{D} = 4,5$ et d'écart-type $S_D \approx 1,690$; $t_0 = \bar{D} / (S_D / \sqrt{8}) \approx 7,53$, largement supérieur à $t_{0,025,7} = 2,365$: on rejette H_0 et on conclut, au seuil de 5 %, que les étudiants sont significativement moins forts en septembre (baisse moyenne de 4,5 points).**

Réponses des quiz

Réponses des quiz

- **Quiz (distributions de $\bar{X}$ et $\bar{X} - \bar{Y}$, avec $n_1 = 10$, $X \sim N(30, 16)$ et $n_2 = 25$, $Y \sim N(10, 25)$) :**
 $\bar{X} \sim N(30; 1,6)$; $\bar{X} - \bar{Y} \sim N(20; 2,6)$.
- **Quiz (exemple 2 analysé à tort comme deux échantillons indépendants à variances égales) :**
règle de rejet bilatérale avec $df = n_1 + n_2 - 2 = 14$: on rejette H_0 si $|t_0| > t_{0,025,14} = 2,145$; comme $t_0 = 1,28 < 2,145$, on ne rejette pas H_0 — conclusion contraire à celle du test apparié (correct) de l'exemple 2, ce qui illustre la perte de puissance lorsqu'on ignore l'appariement.

Bibliographie

Ce document est rédigé avec Beamer ainsi qu'une classe .cls fournie par Jérôme Soucy. Les diapositives sont adaptées des diapositives créées par Thierry Duchesne et Emmanuelle Reny-Nolin pour le cours STT-1900. Les graphiques proviennent de ces mêmes diapositives.

Pour signaler une erreur dans ce document, veuillez écrire à l'adresse : julien.miron@mat.ulaval.ca.