Statistiques inférentielles
4 - Tests de comparaison d'échantillons

A. BOURHATTAS

CY-Tech ING2-GSI

Année universitaire 2023-2024

Comparaison de deux échantillons

Section 1

Comparaison de deux échantillons

Introduction :

- Nous partons de 2 échantillons indépendants $X_1, X_2, .., X_{n_x}$ et $Y_1, Y_2, .., Y_{n_y}$, représentant les v.a. X et Y.
- $E(X) = \mu_x$ et $Var(X) = \sigma_x^2$.
  $E(Y) = \mu_y$ et $Var(Y) = \sigma_y^2$.
- Nous calculons les moyennes empiriques $\overline{X}$ et $\overline{Y}$. Ainsi que les variances empiriques corrigées S_x^{*2} et S_y^{*2}.
- Nous nous demandons : Est ce que ces deux échantillons sont issus de la même population ? Ont-ils les mêmes caractéristiques ?
- Nous nous limterons aux tests d'égalité des moyennes, et d'égalité des variances.

Cas des grands échantillons 1 :

- Dans ce cas, $n_x > 30$ et $n_y > 30$.
- Les hypothèses à tester sont :
$$\begin{cases} (H_0) & \mu_x = \mu_y \\ (H_1) & \mu_x \neq \mu_y \end{cases} \iff \begin{cases} (H_0) & \mu_x - \mu_y = 0 \\ (H_1) & \mu_x - \mu_y \neq 0 \end{cases}.$$
- La variable de décision est $D = \overline{X} - \overline{Y}$.
Sa variance est $\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} \simeq \frac{s_x^{*2}}{n_x} + \frac{s_y^{*2}}{n_y}$.
- La statistique utile est la suivante : sous (H_0),
$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^{*2}}{n_x} + \frac{S_y^{*2}}{n_y}}} \sim \mathcal{N}(0,1).$$

Cas des grands échantillons 2 :

- $D = \overline{X} - \overline{Y}$, et sous (H_0), $\quad Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^{*2}}{n_x} + \frac{S_y^{*2}}{n_y}}} \sim \mathcal{N}(0,1)$
- La région critique est donc de la forme $W = \{|D| > C\} = \{|Z| > k\}$
- La table de la loi $\mathcal{N}(0,1)$ donnera $k = z_{1-\frac{\alpha}{2}}$, c à d $F_Z(k) = 1 - \dfrac{\alpha}{2}$.
- On en déduit $C = k\sqrt{\dfrac{s_x^{*2}}{n_x} + \dfrac{s_y^{*2}}{n_y}}$, et les règles de décision :
- Si $|D| < C$, on valide (H_0), les moyennes sont égales.
- Si $|D| > C$, on valide (H_1), les moyennes sont différentes

Petits échantillons gaussiens, même variance 1 :

- $X \sim \mathcal{N}(\mu_x, \sigma^2)$ et $Y \sim \mathcal{N}(\mu_y, \sigma^2)$
- On teste toujours $\begin{cases} (H_0) & \mu_x - \mu_y = 0 \\ (H_1) & \mu_x - \mu_y \neq 0 \end{cases}$
- La variable de décision est toujours : $D = \overline{X} - \overline{Y}$
- On introduit la quantité $S^2 = \dfrac{(n_x - 1)s_x^{*2} + (n_y - 1)s_y^{*2}}{n_x + n_y - 2}$ meilleure estimation de σ^2.
- La statistique utile est : $T = \dfrac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$ qui suit la loi de Student à $(n_x + n_y - 2)$ d.d.l.

Petits échantillons gaussiens, même variance 2 :

- L'obtention de la région critique, et les règles de décision s'obtiennent de la même manière que précédemment à partir de la table de la loi de Student.
- $C = k\ s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$, avec k tel que : $F_T(k) = 1 - \frac{\alpha}{2}$.
- Si $|D| < C$, on valide (H_0), les moyennes sont égales.
- Si $|D| > C$, on valide (H_1), les moyennes sont différentes.

Test d'égalité de proportions 1 :

- Dans ce cas-ci, on a deux Bernouilli :
  $X \sim \mathcal{B}(p_x)$ et $Y \sim \mathcal{B}(p_y)$
- $n_x > 30$ et $n_y > 30$.
- On a les deux fréquences empiriques F_x et F_y.
- On teste $\begin{cases} (H_0) & p_x - p_y = 0 \\ (H_1) & p_x - p_y \neq 0 \end{cases}$
- La variable de décision est : $D = F_x - F_y$
- On introduit la quantité $f_0 = \dfrac{n_x f_x + n_y f_y}{n_x + n_y}$ meilleure estimation de p.
- La statistique utile est :
$$Z = \frac{F_x - F_y}{\sqrt{f_0(1-f_0)\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim \mathcal{N}(0,1).$$

Test d'égalité de proportions 2 :

- L'obtention de la région critique, et les règles de décision s'obtiennent de la même manière à partir de la table de la loi normale.
- $C = k\sqrt{f_0(1-f_0)\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}$, avec k tel que : $F_Z(k) = 1 - \frac{\alpha}{2}$.
- Si $|D| < C$, on valide (H_0), les proportions sont égales.
- Si $|D| > C$, on valide (H_1), les proportions sont différentes.

Egalité de moyennes, échantillons appariés 1 :

- Il s'agit dans ce cas de deux mesures ou deux expériences faites sur les mêmes individus.
- Les deux échantillons ne sont pas indépendants.
- On définit $D_i = X_i - Y_i$, $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ et

$$S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$$

- La variable de décision est : $\overline{D}$
- La statistique utile est : $T = \frac{\overline{D}}{S_d}\sqrt{n}$ qui suit la loi de Student à $(n-1)$ d.d.l.

Egalité de moyennes, échantillons appariés 2 :

- L'obtention de la région critique, et les règles de décision s'obtiennent de la même manière à partir de la table de la loi de Student.
- $C = k \dfrac{S_d}{\sqrt{n}}$, avec k tel que : $F_T(k) = 1 - \dfrac{\alpha}{2}$.
- Si $|\overline{D}| < C$, on valide (H_0), les moyennes sont égales.
- Si $|\overline{D}| > C$, on valide (H_1), les moyennes sont différentes.

Test d'égalité de variances 1 :

- $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ et $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$
- On teste $\begin{cases} (H_0) & \sigma_x - \sigma_y = 0 \\ (H_1) & \sigma_x - \sigma_y \neq 0 \end{cases}$
- La variable de décision est : $K = \dfrac{S_x^{*2}}{S_y^{*2}}$.
- Elle suit la loi de Fisher-Snedecor à $(n_x - 1)$ et $(n_y - 1)$ dégrés de liberté.
- La table correspondant à $\alpha = 5$ % ressemble à :

Loi de Fisher F

$$P(F_{\nu_1,\nu_2} < \mathsf{f}_{\nu_1,\nu_2,\alpha}) = \alpha$$

α = 0,975

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100
1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1008	1013
2	38,5	39,0	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,2	14,1	14,0	14,0
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,46	8,38	8,32
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,23	6,14	6,08

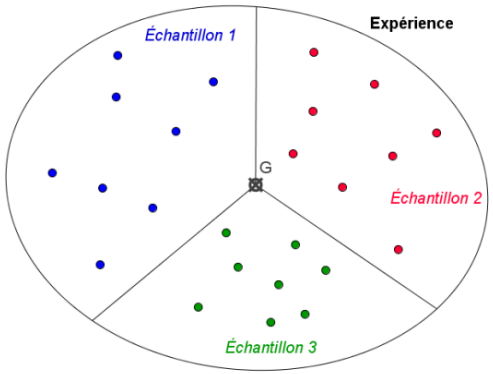
Test d'égalité de variances 2 :

- Il s'agit, normalement, d'un test bilatéral. Mais si on choisit de mettre au numérateur la plus grande des variances empiriques, on a une région critique de la forme $W = \{K > C\}$.
- C est obtenu par la table de la loi de Fisher avec la relation : $F_K(C) = 1 - \frac{\alpha}{2}$.
- Si $K < C$, on valide (H_0), les variances sont égales.
- Si $K > C$, on valide (H_1), les variances sont différentes.

Comparaison de plusieurs échantillons

Cadre général :

- Nous partons de $k \geq 3$ échantillons indépendants.



- Hypothèses à tester :

$$\begin{cases} (H_0) & \text{Les } k \text{ échantillons sont issus de la même population} \\ (H_1) & \text{Il y a au moins un échantillon différent des autres} \end{cases}$$

- Dans le cas où les échantillons sont gaussiens de même variance, nous utiliserons le test d'analyse de variance ANOVA.

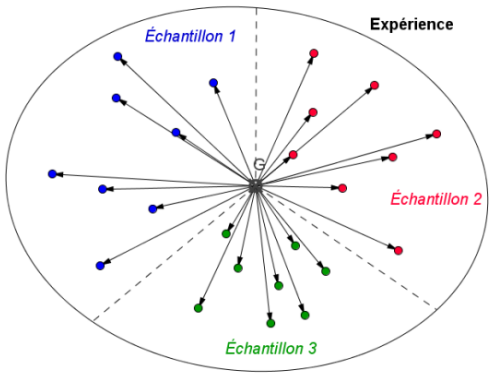
Principe de l'ANOVA :

- On a k échantillons de taille $n_1, \ldots, n_k$.
- j-ème échantillon $(X_1^{(j)}, \ldots, X_{n_j}^{(j)})$ associé à la v.a. $X^{(j)}$.
- Echantillons gaussiens, indépendants et de même variance :
  $X^{(j)} \sim \mathcal{N}(\mu_j, \sigma^2)$ pour tout j.
- On va donc tester : $\begin{cases} (H_0) & \mu_1 = \mu_2 = ... = \mu_k \\ (H_1) & \exists (i,j) \quad \mu_i \neq \mu_j \end{cases}$
- Ce sont les différentes mesures de variance qui vont nous permettre de décider.

Quelques définitions :

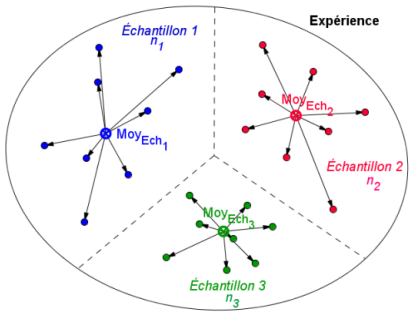
On note :

- $\overline{X}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_i^{(j)}$: moyenne empirique de $X^{(j)}$,
- $V^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_i^{(j)} - \overline{X^{(j)}})^2$: var empirique de $X^{(j)}$,
- $\overline{X}$: la moyenne de l'échantillon global,

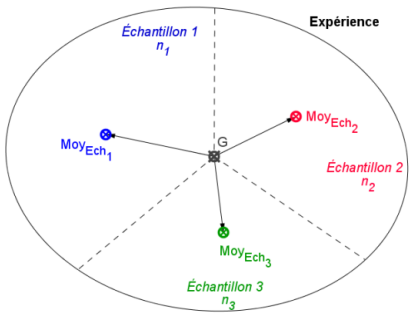


Quelques définitions :

- $V_{intra} = \sum_{j=1}^{k} \frac{n_j}{n} V^{(j)}$: Variance intra-classes, ou résiduelle,



- $V_{inter} = \sum_{j=1}^{k} \frac{n_j}{n} (\overline{X}^{(j)} - \overline{X})^2$: Variance inter-classes, entre échantillons.



Résultat principal :

Propriété :

1. La var totale est égale à la somme des var inter et intra : $V_{tot} = V_{inter} + V_{intra}$.
2. La variable de décision est : $D = \frac{\frac{V_{inter}}{k-1}}{\frac{V_{intra}}{n-k}}$
3. Sous l'hypothèse (H_0), d'égalité des moyennes, D suit la loi de Fisher à $(k-1)$ et $(n-k)$ degrés de liberté.

$$D \sim \mathcal{F}(k-1, n-k)$$

Test ANOVA :

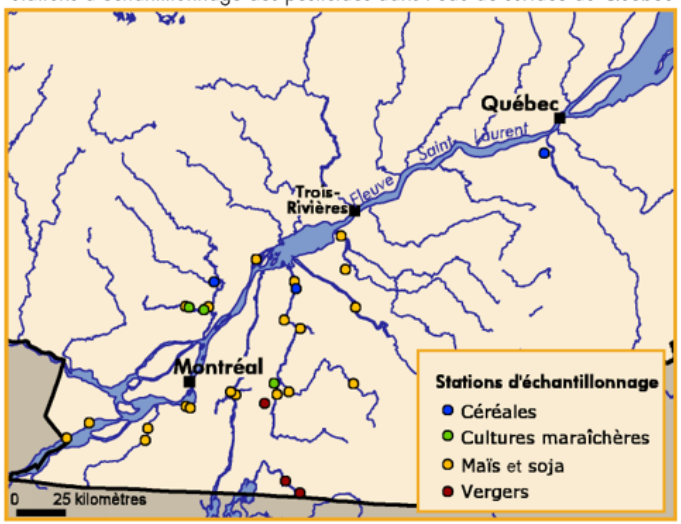
La région critique est de la forme : $W = \{D > C\}$
avec $C =$ fractile d'ordre $1 - \alpha$ pour $\mathcal{F}(k-1, n-k)$
La règle de décision est donc :

- Si $D < C$, on valide (H_0), les moyennes sont égales. Les échantillons sont issus de la même population.
- Si $D > C$, on valide (H_1). Il y a au moins un échantillon différent des autres.

Exemple 1 :

On veut savoir si la quantité de nitrates, prélevées le long d'une rivière, varie d'une station à l'autre. Une différence significative pourrait indiquer des déversements de nitrates entre ces stations.

Stations d'échantillonnage des pesticides dans l'eau de surface au Québec



Ministère de l'Environnement, 2004

Pour cela, on dispose des résultats de 10 prélévements effectués dans 3 stations différentes ($k = 3$).

Exemple 2 :

Station 1	Station 2	Station 3
50,00	162,00	120,00
52,00	350,00	120,00
123,00	125,00	122,00
100,00	320,00	221,00
200,00	112,00	253,00
250,00	200,00	141,00
220,00	40,00	182,00
220,00	162,00	175,00
300,00	160,00	160,00
220,00	250,00	214,00

Exemple 3 :

- Le test ANOVA, effectué à l'aide de Libre Office ou excel donne :

ANOVA - facteur unique

Alpha	0,05

Groupes	Compter	Somme	Moyenne	Variance
1 colonne	10	1735	173,5	7445,6111
2 colonne	10	1881	188,1	9048,9889
3 colonne	10	1708	170,8	2203,7333

Source de la variation	SS	df	MS	F	Valeur P	Critique F
Entre les groupes	1732,4667	2	866,2333	0,1390	0,8709	3,3541
À travers les groupes	168285,0000	27	6232,7778			
Total	170017,4667	29				

- $D = F = 0.139$, et le seuil est $C =$ Critique F $= 3.3541$.
- $D < C$, on valide (H_0). il n'y a pas de différence significative entre échantillons.
- Ou bien : La p-valeur, Valeur P$= 0.87$ est très supérieure à α.

Remarque :

- Le test de comparaison d'échantillons que l'on vient de voir, peut être vu comme un test d'indépendance entre la variable X concernant tous les échantillons et une variable qualitative dont les modalités sont associées à chaque échantillon.
- On associe à chaque échantillon une modalité A_j d'une variable qualitative A.
- L'hypothèse (H_0) devient : les échantillons sont identiques, donc A n'a pas d'effet sur X : variables indépendantes.
- L'hypothèse (H_1) devient : il y a un échantillon différent des autres, les modalités de A ont un effet sur X : variables liées.

Cas d'échantillons compliqués :

- Lorsque l'hypothèse de normalité n'est pas possible,
- ou lorsque les échantillons contiennent des valeurs extrêmes (aberrantes),
  on fait appel à des tests non paramétriques :
- test de Wilcoxon-Mann-Whitney, dans le cas de 2 échantillons, ou
- test de Kruskal-Wallis dans le cas de plus de 3 échantillons.