

STATISTIQUES

Comparaison d'échantillons

Y-a-t 'il une différence significative entre k échantillons?

COMPARAISON DE DEUX ÉCHANTILLONS

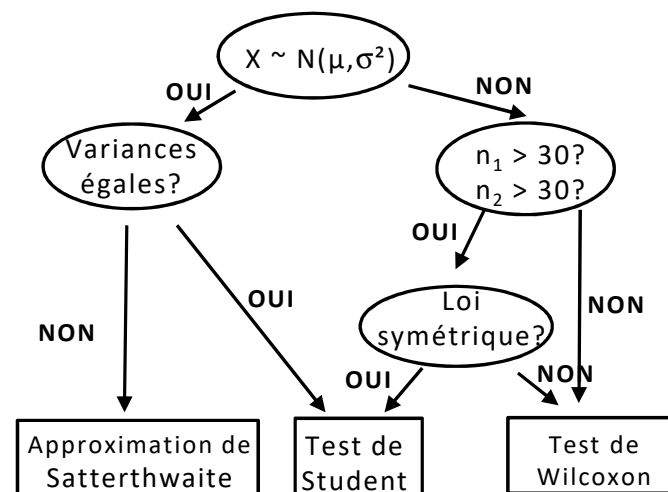
Il s'agit ici de répondre à la question:

« L'écart entre deux échantillons est-il dû au hasard ou est-il significatif? »

Exemple : Une étude menée sur 9 garçons et 12 filles montre que le QI moyen des garçons est 107 et celui des filles est 112. On sait par ailleurs que le QI suit une loi normale d'écart-type 15. Avec un test à 5%, on peut dire qu'il n'y a pas de différence significative entre le QI des filles et celui des garçons.

En pratique, on se contente de comparer les échantillons au travers de valeurs remarquables telles que la moyenne, variance,

- Le **test de Student** permet de comparer les moyennes des deux échantillons
- Le **test de Fisher** permet de comparer leurs variances



Si les échantillons présentent des valeurs extrêmes, on préfère utiliser le test de **Wilcoxon-Mann-Whitney**. Il est basé sur les positions relatives des observations et non sur la distribution des valeurs. On dit que c'est un **test non paramétrique**. En général, les tests non paramétriques sont moins puissants que les tests paramétriques.

TEST DE STUDENT

Le **test de Student** consiste à faire la différence entre les deux moyennes (proportions) et de tester si la différence est nulle

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \longleftrightarrow \begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

- Si les échantillons sont de grandes tailles, la statistique du test est

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sous l'hypothèse H_0 , elle suit une loi $N(0,1)$.

- Si les échantillons sont petits mais **gaussiens**, la statistique du test est

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{où} \quad S = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

Sous l'hypothèse H_0 , elle suit une loi de Student à $n_1 + n_2 - 2$ d.d.l.

Plus la différence s'éloigne de 0 et moins l'hypothèse H_0 est probable. La région critique est donc de la forme $W = \{T > C\}$

TEST DE WILCOXON-MANN-WHITNEY

Dans le cas du **test de Wilcoxon-Mann-Whitney**, on teste si la probabilité qu'une observation de la population X_1 soit supérieure à une observation de la population X_2 est égale à la probabilité qu'une observation de la population X_2 soit supérieure à une observation de la population X_1

$$\begin{cases} H_0 : P(X_1 > X_2) = P(X_2 > X_1) \\ H_1 : P(X_1 > X_2) \neq P(X_2 > X_1) \end{cases}$$

Pour déterminer la statistique du test on réunit les deux séries observées que l'on classe par ordre croissant. A chaque valeur classée, on associe son rang, puis on calcule la somme des rangs de la série X_1 .

$$K = \sum_{i=1}^p R_i^2 - n_i(n_i + 1)/2$$

où n_i est la taille de l'échantillon i et R_i la somme des rangs pour l'échantillon i ,

On constate que la statistique du test n'est pas basée sur les valeurs des séries mais uniquement de leur position respective. Elle n'est donc pas perturbée par une distribution non symétrique ou par des valeurs atypiques.

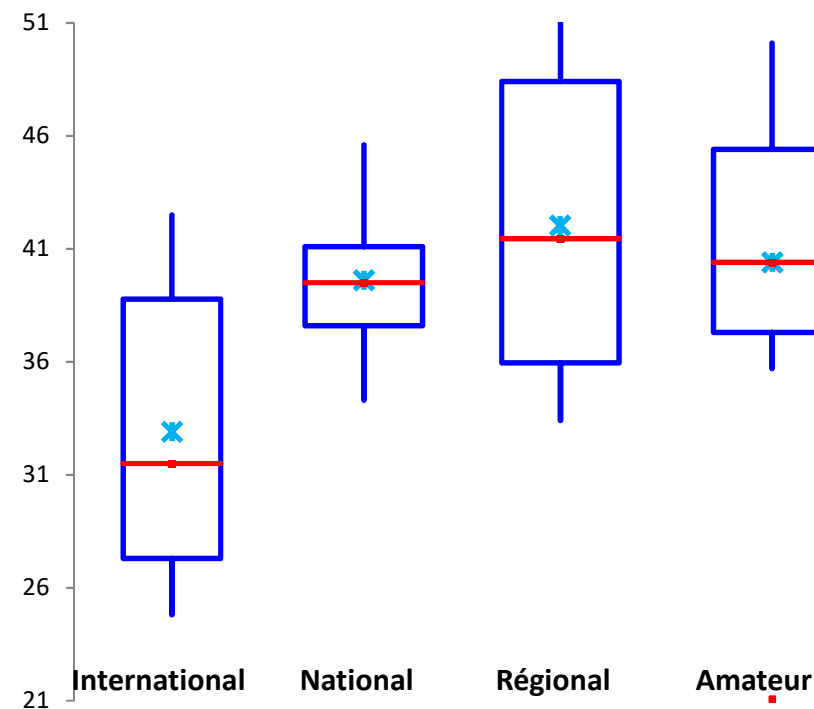
COMPARAISON DE K>2 ÉCHANTILLONS

Il s'agit ici de répondre à la question:

« Est-ce que les modalités d'une variable qualitative ont un impact significatif sur les valeurs d'une variable quantitative »

Exemple : Y-a-t'il un lien entre le niveau d'anxiété d'un joueur de tennis et le niveau de la compétition?

International	National	Régional	Amateur
24,8	45,6	33,4	21,1
26,7	41,1	34,6	35,7
27,5	34,3	36,4	37,3
30,6	37,6	39,1	39,4
...
38,2		47,9	44,5
40,5		49,9	45,4
42,5		51,2	49,8
			50,1



COMPARAISON DE $K > 2$ ÉCHANTILLONS

Les deux hypothèses testées sont :

$$\begin{cases} H_0 : \text{Les } k \text{ échantillons sont issus de la même population (pas d'impact)} \\ H_1 : \text{il existe au moins un échantillon différent des autres (impact significatif)} \end{cases}$$

Dans le cas où l'hypothèse H_1 est adoptée, on ne sait pas quel échantillon est différent, ni s'il y en a plusieurs.

Et pourquoi pas des tests 2 à 2? Considérons 7 groupes d'observations tirées indépendamment d'une même population statistique (7 échantillons). Il faudrait réaliser $7(7-1)/2 = 21$ tests pour comparer toutes les paires de groupes. Chaque test étant réalisé au niveau $\alpha = 0,05$, on a, dans chaque cas, 5 chances sur 100 de rejeter H_0 même si H_0 est vraie. La probabilité de rejeter H_0 au moins une fois à tort au cours de 21 tests est 0,66 ! **On a donc tout intérêt à utiliser un test adapté.**

- Dans le cas d'échantillons gaussiens de même variance, on utilise le test de l'analyse de la variance : **ANOVA (Analyse Of Variance)**
- Si la distribution des échantillons présentent des valeurs extrêmes, on utilise le test non paramétrique de **Kruskal-Wallis**. Tout comme le test de Man-Whitney, il repose que la façon dont les valeurs des échantillons se positionnent les uns par rapport aux autres.

ANOVA

Notons A la variable qualitative et A_i ses modalités (compétition : 4 niveaux)

On partitionne la population en sous-populations :
une sous-population pour chaque modalité du facteur


On suppose que les observations pour la modalités A_i sont les réalisations d'une variable aléatoire X_i , telle que


$$X_i \sim N(\mu_i, \sigma^2)$$


C'est-à-dire que chaque observation s'écrit sous la forme


$$x_i^k = \underbrace{\mu_i}_{\text{Effet moyen sur } A_i} + \underbrace{\varepsilon_i^k}_{\text{Erreur}}$$

Modalités			
A_1	A_2		A_p
x_1^1	x_2^1		x_p^1
		...	
$x_1^{n_1}$	$x_2^{n_2}$		$x_p^{n_p}$
Moyennes	\bar{x}_1	\bar{x}_2	\bar{x}_p
Effectifs	n_1	n_2	n_p


 X_1


 X_2


 \dots


 X_p

L'analyse de la variance consiste à effectuer le test d'hypothèses suivant

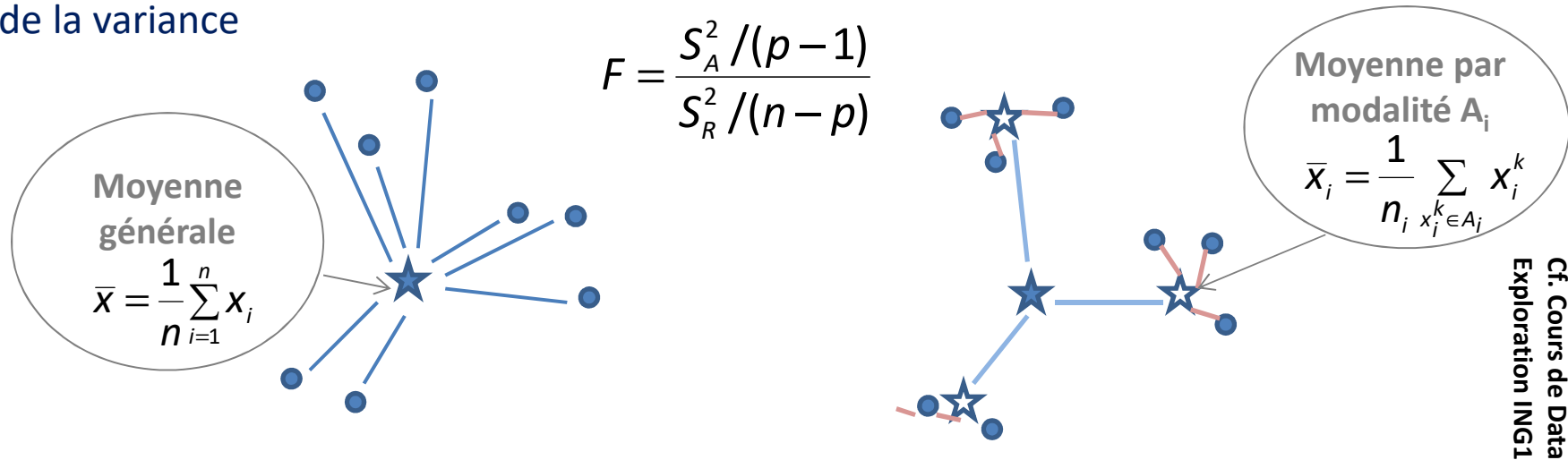
$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_p \\ H_1 : \exists i \neq j \mu_i \neq \mu_j \end{cases}$$

Trois hypothèses :

- A influe linéairement sur la moyenne: $\mu_i = \mu + \alpha_i$ où α_i est l'effet de la modalité A_i
- A n'influe pas sur la variance (variance constante) : $\varepsilon_i \sim N(0, \sigma^2)$ (on pourra utiliser le test de Fisher)
- Les échantillons (pour chaque modalité) sont gaussiens

ANOVA

La statistique du test de l'ANOVA est construite à partir de la formule de la décomposition de la variance



$$F = \frac{S_A^2 / (p-1)}{S_R^2 / (n-p)}$$

Variance totale = Variance expliquée par A + Variance résiduelle

$$S^2 = \sum_i \sum_j (X_i^j - \bar{X})^2 \quad S_A^2 = \sum_{i=1}^p n_i (\bar{X}_i - \bar{X})^2 \quad S_R^2 = \sum_i \sum_j (X_i^j - \bar{X}_i)^2$$

Sous l'hypothèse H_0 , la statistique suit une loi de Fisher-Snedecor $F_{\alpha}(p-1, n-p)$ et la région critique est de la forme $W = \{F > C\}$.

Exemple : niveau d'anxiété d'un joueur de tennis et le niveau de la compétition

Source de variation	Somme des carrées	Degré de liberté	Carré moyen
Expliquée	384,18	$p-1=3$	128,06
Résiduelle	1377,96	$n-p=26$	53
Totale	1762,14	$n-1=29$	

$$\Rightarrow F = \frac{128,06}{53} = 2,42$$

Avec un risque $\alpha=5\%$, on obtient dans la table de $F(3,26)$ un seuil $C=3$.

$F=2,42 < 3$ donc garde donc H_0 . Le niveau de compétition n'a donc pas d'impact sur l'anxiété.

TEST DE KRUSKAL-WALLIS

Si on désigne par M_i le paramètre de position l'échantillon i , les hypothèses nulle H_0 et alternative H_1 du test de Kruskal-Wallis sont les suivantes :

$$\begin{cases} H_0 : M_1 = M_2 = \dots = M_p \\ H_1 : \text{il existe au moins un couple } (i, j) \text{ tel que } M_i \neq M_j \end{cases}$$

Le calcul de la statistique du test de Kruskal-Wallis fait intervenir comme pour le test de Wilcoxon-Mann-Whitney le rang des observations, une fois les p échantillons mélangés.

$$K = \frac{12}{n(n+1)} \sum_{i=1}^p \frac{R_i^2}{n_i} - 3(n+1)$$

où n_i est la taille de l'échantillon i , n la somme des n_i , et R_i la somme des rangs pour l'échantillon i parmi l'ensemble des échantillons.

Lorsque $k=2$ le test de Kruskal-Wallis est équivalent au test de Wilcoxon-Mann-Whitney

COMPLÉMENT : TEST DE FISHER

Le **test de Fisher** consiste à faire comparer les deux variances

$$\left\{ \begin{array}{l} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{array} \right.$$

Le test n'est valable que pour des échantillons gaussiens.

La statistique du test est

$$F = \frac{S_1^2}{S_2^2}$$

Sous l'hypothèse H_0 , elle suit une loi de Fisher $F_{n_1-1; n_2-1}$.

La région critique est donc de la forme $W=\{F>C\}$

Exemple : Un courtier rapporte que le taux de rendement moyen pour un échantillon d'actions de 10 compagnies pétrolières est de 12.6% avec un écart-type de 3.9%. Le taux de rendement moyen des actions de 8 compagnies de service est de 10.9% avec un écart-type de 3.5%. Avec un risque de 5% peut-on conclure que les actions des compagnies pétrolières sont plus volatiles que celles de compagnies de service?