Inférence statistique

Séance 8

Analyse de la variance (ANOVA)

Frédéric Sur

Mines Nancy

https://members.loria.fr/FSur/

Plan

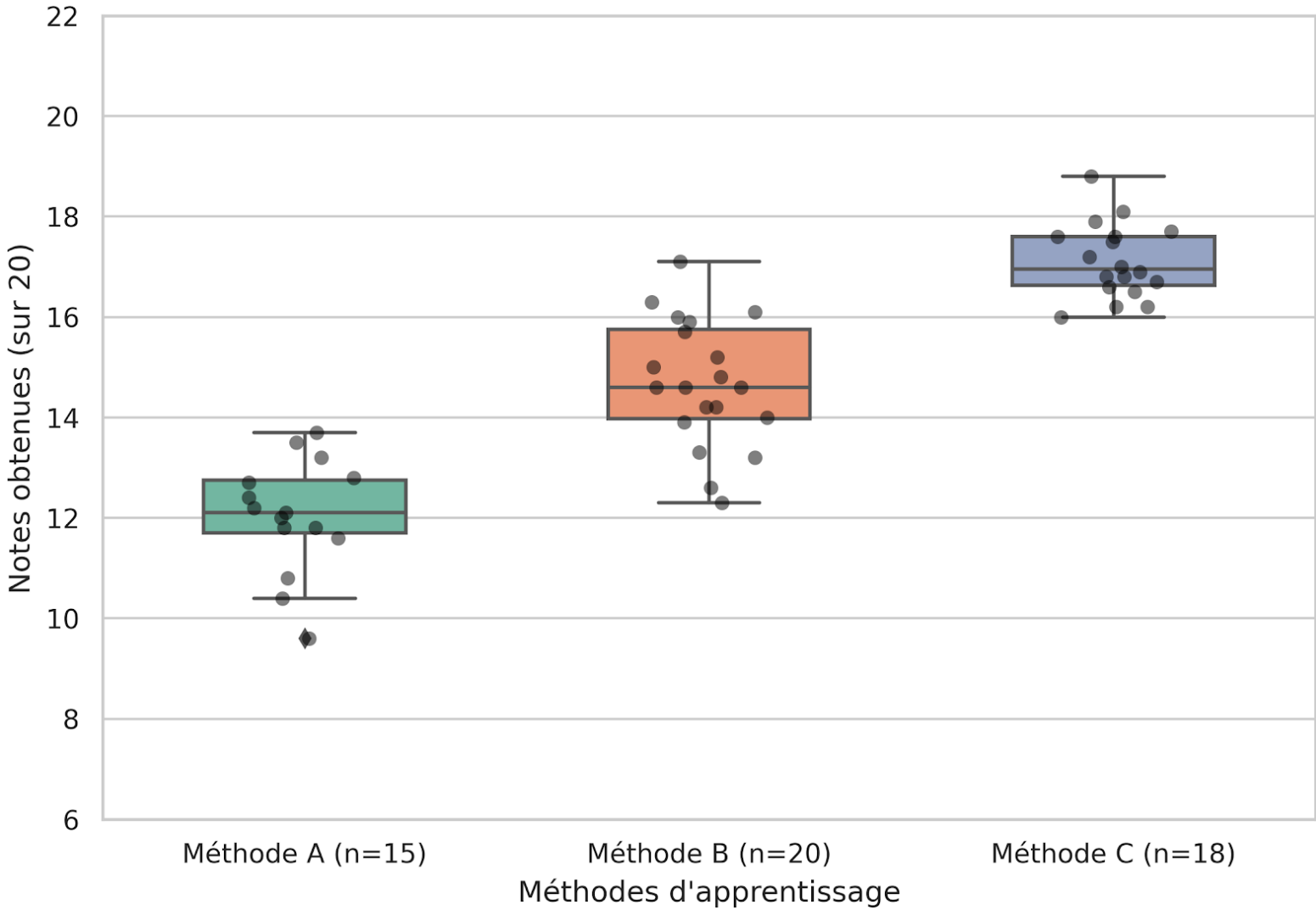
Qu'est-ce que l'ANOVA ?

ANOVA : ANalysis Of VAriance.

Objectif : comparer les moyennes de **plusieurs groupes simultanément**

Principe : analyser la variance totale des données pour déterminer si les différences des moyennes entre les groupes sont significatives **dans leur ensemble** ou dues au hasard

Exemple


22
20
18
16
14
12
10
8
6
Notes obtenues (sur 20)
Méthode A (n=15)
Méthode B (n=20)
Méthode C (n=18)
Méthodes d'apprentissage

Le problème des tests multiples

Soient 4 groupes (A, B, C, D) : on dispose pour chacun d'échantillons distribués selon $\mathcal{N}(\mu_i, \sigma^2)$ avec $i \in \{A, B, C, D\}$.

Pour comparer les 4 groupes, il faudrait faire 6 tests t de Student : AB, AC, AD, BC, BD, CD
(voir *comparaison de moyennes d'échantillons gaussiens de variances égales - séance « Test 2 » sec. 1.1*)

Problème : inflation du risque d'erreur α
Si chaque test a un seuil de signification $\alpha = 0,05$, la probabilité de commettre au moins une erreur de type I « explose » :

$$P(\text{Erreur Type I}) = 1 - (1 - 0,05)^6 \approx 0,26$$

(en supposant les tests *indépendants*)

$\rightarrow$ l'ANOVA permet de contrôler ce risque *globalement*

Plan

Hypothèses de l'ANOVA

On étudie l'effet d'une variable qualitative (k modalités), le facteur, sur une variable quantitative y continue.
$\rightarrow$ pour $1 \leqslant i \leqslant k$, on dispose d'un n_i-échantillon $(y_{ij})_{1 \leqslant j \leqslant n_i}$
nombre total d'observations : $\sum_{i=1}^{k} n_i = N$

Exemple : $k = 3$ méthodes d'apprentissage (facteur), $n_1 = 15$, $n_2 = 20$, $n_3 = 18$; effet de la méthode sur la note (variable à expliquer) ?

- **Hypothèse nulle ($\mathcal{H}_0$) :** toutes les moyennes des k groupes / modalités sont égales.

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

- **Hypothèse alternative ($\mathcal{H}_1$) :** au moins une des moyennes diffère des autres.

$$\mathcal{H}_1 : \exists (i, j) \text{ tels que } \mu_i \neq \mu_j$$

Conditions fondamentales d'application

ANOVA : test paramétrique reposant sur trois conditions fondamentales :

1. **indépendance :** les observations sont indépendantes.
2. **normalité :** les observations dans chaque modalité suivent une loi normale $\mathcal{N}(\mu_i, \sigma_i^2)$, $1 \leqslant i \leqslant k$.
3. **homoscédasticité :** les variances des différentes modalités sont égales ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$).

$\rightarrow$ dans la suite, on suppose les trois conditions vérifiées.

Modèle mathématique : le modèle linéaire

Chaque observation y_{ij} (individu j de la modalité i) se décompose en :

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- μ : moyenne générale de la population
- α_i : effet de la modalité i (avec $\sum_{i=1}^{k} n_i \alpha_i = 0$)
- ϵ_{ij} : erreur aléatoire (résidu), supposée suivre une loi $\mathcal{N}(0, \sigma^2)$

Ainsi : $y_{ij} \sim \mathcal{N}(\mu + \alpha_i, \sigma^2)$
soit : $\mu_i = \mu + \alpha_i$
$\mathcal{H}_0$ s'écrit aussi : $\forall i, \alpha_i = 0$.

Décomposition des Sommes de Carrés

Relation mathématique fondamentale :

$$SCT = SCF + SCR$$

- **SCT (Total)** : Variation totale des observations

$$SCT(= T) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$

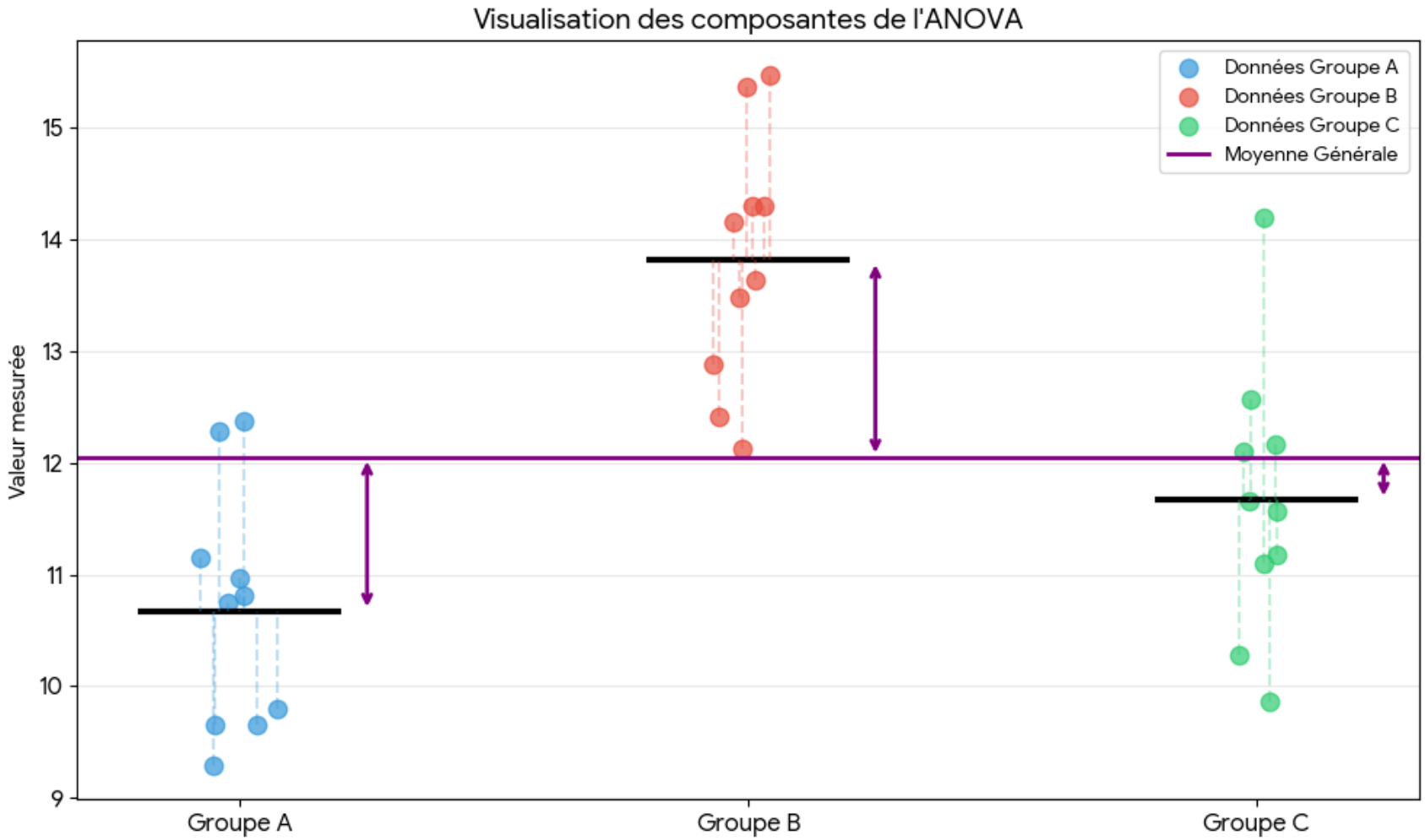
- **SCF (facteur)** : variation expliquée par le facteur

$$SCF(= A) = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

- **SCR (résiduelle)** : variation non expliquée, ou résiduelle.

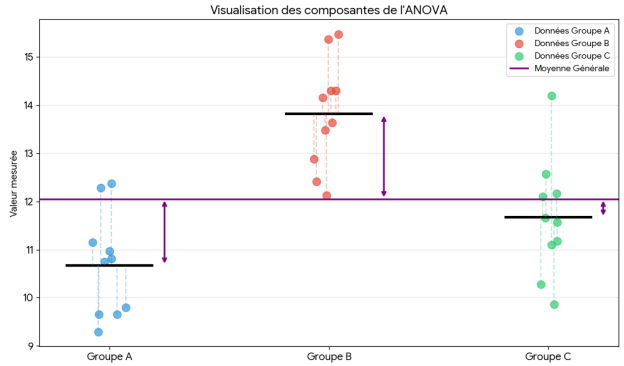
$$SCR(= R) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

Illustration : $SCT = SCF + SCR$



SCF ? SCR ? Estimation des μ_i (α_i) ?

Principe de l'ANOVA



L'ANOVA compare deux sources de variabilité :

1. **variabilité inter-groupes (Facteur) :** différences entre les moyennes des groupes et la moyenne générale.
2. **variabilité intra-groupe (Erreur) :** dispersion des observations autour de la moyenne de leur propre groupe.

Principe de décision : si la variation expliquée par le facteur (SCF) est *significativement* supérieure à la variation résiduelle (SCR), on rejette $\mathcal{H}_0$ et on conclut que l'effet du facteur est significatif.

La statistique de Fisher (F)

On calcule le rapport des carrés moyens :

$$F = \frac{SCF/(k-1)}{SCR/(N-k)}$$

- si $\mathcal{H}_0$ est vraie, $F \simeq 1$ car $E(F|\mathcal{H}_0) = (N-k)/(N-k-2)$
- si $\mathcal{H}_0$ est fausse (effet significatif du facteur), $F > 1$

Propriété : sous les trois conditions fondamentales de l'ANOVA, la statistique F est distribuée selon la loi de Fisher-Snedecor $F(k-1, N-k)$

$\rightarrow$ test unilatéral à droite : on compare F_{obs} à la valeur critique F_{crit} de la table de Fisher, ou on regarde la p-valeur.

Le tableau d'ANOVA à un facteur

Source	Somme des carrés	ddl	Carré moyen	F
Facteur	SCF	$k-1$	$\frac{SCF}{k-1}$	$\frac{SCF/(k-1)}{SCR/(N-k)}$
Erreur	SCR	$N-k$	$\frac{SCR}{N-k}$	
Total	SCT	$N-1$		

Dans les logiciels, colonne supplémentaire avec la *p*-valeur

$\rightarrow$ à comparer avec le risque α fixé *a priori*
(*p*-valeur $< 5\%$: rejet de $\mathcal{H}_0$)

Et après ? Les tests *post-hoc*

Question : si l'ANOVA permet de rejeter $\mathcal{H}_0$, on sait qu'une modalité diffère, mais **laquelle** ?

On ne peut pas faire de tests de Student indépendants (à cause de l'*inflation de l'erreur* α).

$\rightarrow$ on utilise des **tests de comparaisons multiples** (post-hoc) qui ajustent le seuil de significativité.

approche *post-hoc* la plus simple :

tests de Student sur les paires de modalités/groupes avec correction de Bonferroni (on divise α par le nombre de tests.)
Inconvénient : le critère de rejet devient très difficile à atteindre, alors que les tests ne sont généralement pas indépendants.

Test HSD de Tukey (hors-programme)

Test post-hoc le plus courant

$\rightarrow$ plutôt que comparer indépendamment *les écarts entre toutes les paires de moyennes*, on compare *l'écart entre la plus grande et la plus petite moyenne* parmi les k groupes (condition plus stricte).

Statistique de test : $q = |\bar{x}_{\text{max}} - \bar{x}_{\text{min}}| / \sqrt{SCR/N}$

La distribution de q sous $\mathcal{H}_0$ est connue, dépend de k, $df = N - k$

$\rightarrow$ à un risque α (5%), on calcule q_{crit} (qui s'adapte $\ll$ automatiquement $\gg$ à k)

(5% de chances pour que $\bar{x}_{\text{max}} - \bar{x}_{\text{min}} > q_{\text{crit}} \sqrt{SCR/N}$ "par hasard")

HSD$= q_{\text{crit}} \sqrt{SCR/N}$: *honestly significant difference*

$\rightarrow$ différence minimale qui doit exister entre les moyennes de deux groupes pour que l'on puisse affirmer qu'ils sont statistiquement différents, en tenant compte de la variabilité globale de toutes les données

Règle de décision :

si $\bar{x}_i - \bar{x}_j > HSD$, alors différence significative entre groupes i et j.

Une question de méthodologie...

Pourquoi faire une ANOVA si HSD (ou d'autres tests post-hoc) contrôle le risque d'erreur global ?

Il se peut que l'ANOVA ne permette pas de rejeter $\mathcal{H}_0$ (moyennes de groupes identiques), mais que des tests post-hoc détectent des paires de moyennes significativement différentes...
$\rightarrow$ ANOVA fait une analyse globale, des différences entre groupes peuvent se moyenner. Un test post-hoc se focalise sur des différences deux-à-deux.

Est-on vraiment obligé de faire une ANOVA ?
$\rightarrow$ cela dépend de la question expérimentale...
Si on s'intéresse aux différences entre paires de groupes en suspectant qu'un groupe est « différent », on peut passer directement à HSD.
Néanmoins, HSD nécessite SCR calculé par ANOVA, donc tant qu'on y est...

Une autre « bonne » raison : ANOVA puis post-hoc semble une tradition bien ancrée dans la littérature scientifique.

Plan

ANOVA à deux facteurs

On étudie l'effet de deux facteurs (variables qualitatives) simultanément sur une variable continue.

Intérêt :

- on utilise le même échantillon pour deux questions
- permet d'étudier l'*interaction* entre les facteurs

Effets principaux et interaction

Une ANOVA à deux facteurs revient à tester trois hypothèses nulles distinctes :

1. **Effet principal A :** les moyennes des p modalités du facteur A sont-elles égales ?
2. **Effet principal B :** les moyennes des q modalités du facteur B sont-elles égales ?
3. **Effet d'interaction (AxB) :** l'effet du facteur A dépend-il du niveau (modalité) du facteur B ?

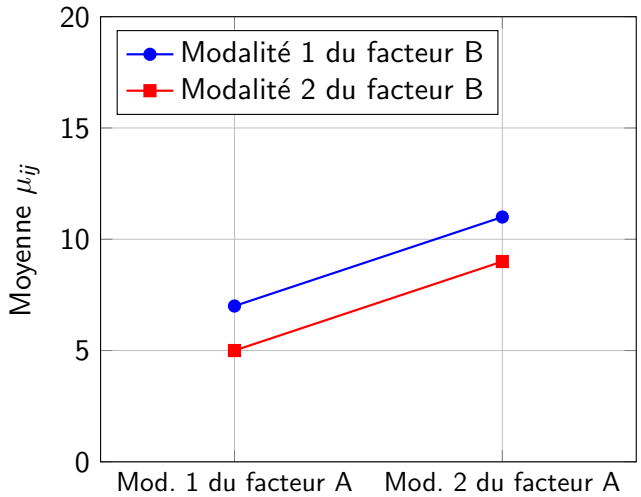
Remarque : voir la discussion sur $\mathcal{H}_0$ dans les notes de cours.

Comprendre l'interaction

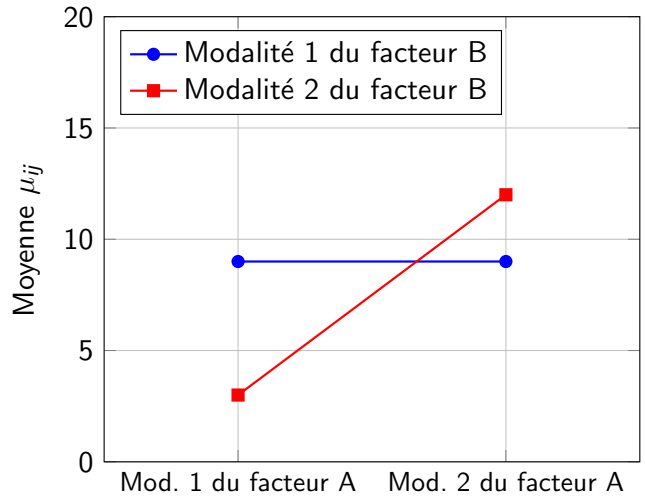
Il y a interaction si l'effet d'un facteur n'est pas le même selon la modalité de l'autre facteur.

Exemple : certains médicaments font baisser la pression artérielle chez les hommes, mais l'augmentent chez les femmes. L'effet sur la tension (variable quantitative à expliquer) du médicament (facteur A) dépend du sexe (facteur B).

En traçant les moyennes de chaque sous-groupe (paire de modalités) :



pas d'interaction : les segments reliant les moyennes sont parallèles



interaction : les segments se croisent ou divergent

Modèle linéaire à deux facteurs

L'équation du modèle complet devient :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

- μ : moyenne générale
- α_i : effet principal de la modalité i du facteur A $(\sum_i \alpha_i = 0)$
- β_j : effet principal de la modalité j du facteur B $(\sum_j \beta_j = 0)$
- γ_{ij} : effet d'interaction entre la modalité i du facteur A et la modalité j du facteur B $(\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0)$
- ϵ_{ijk} : erreur résiduelle $\sim \mathcal{N}(0, \sigma^2)$

y_{ijk} : k-ème observation de la modalité i du facteur A et de la modalité j du facteur B.

Hypothèse : même effectifs (nombre de répétitions) r pour chaque paire de modalité (i, j).
$\rightarrow$ nombre total d'observations : $N = pqr$

Décomposition de la variance

La somme des carrés totale est divisée en **quatre** parties :

$$SCT = SCA + SCB + SCI + SCR$$

- SCA : variation expliquée par le facteur A

$$\frac{SCA}{N} \simeq \frac{1}{p} \sum_i \alpha_i^2$$

- SCB : variation expliquée par le facteur B

$$\frac{SCB}{N} \simeq \frac{1}{q} \sum_j \beta_j^2$$

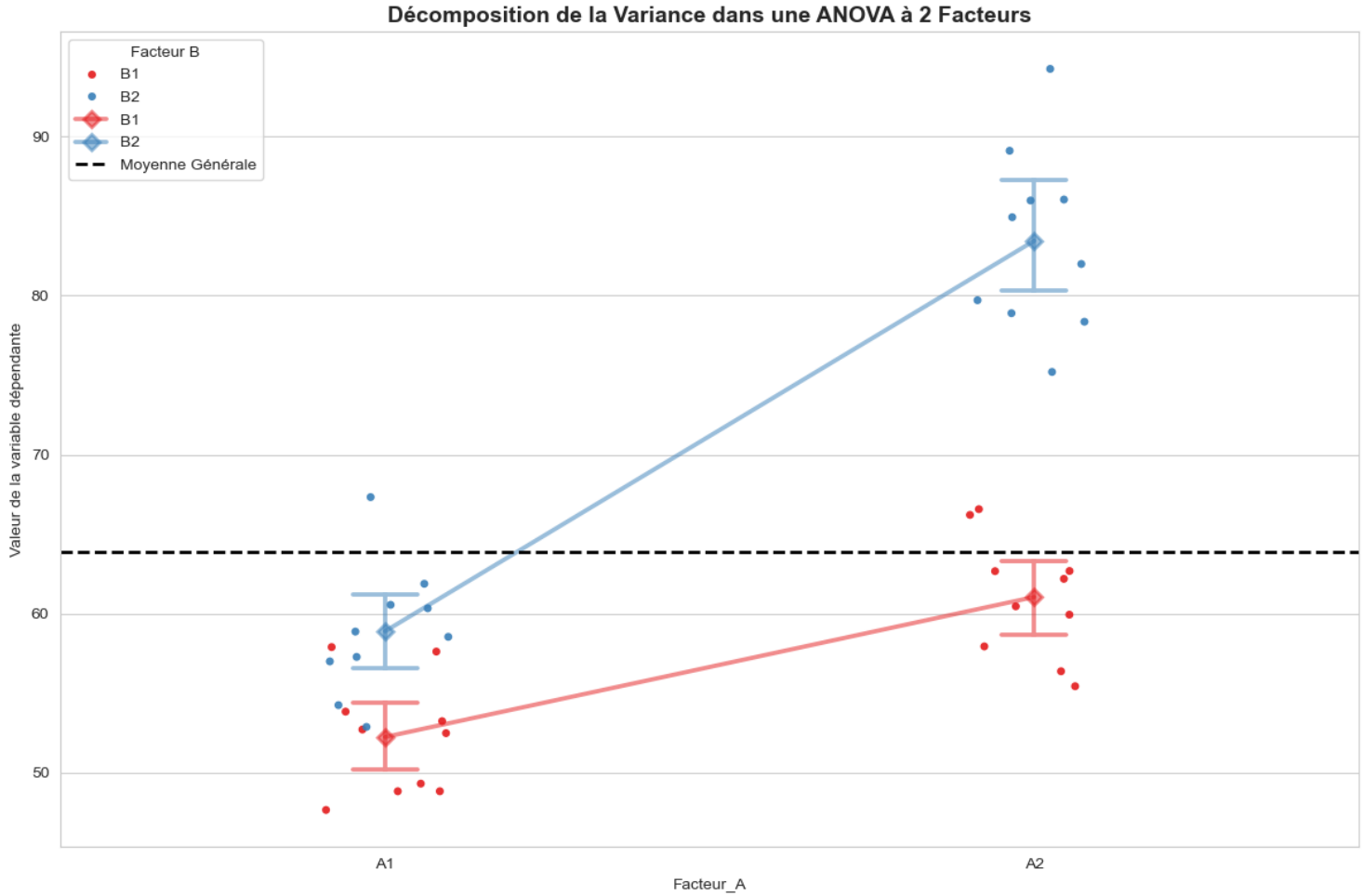
- SCI : variation expliquée par l'interaction

$$\frac{SCI}{N} \simeq \frac{1}{pq} \sum_{ij} \gamma_{ij}^2$$

- SCR : variation résiduelle

$$\frac{SCR}{N} \simeq \sigma^2$$

Illustration graphique



SCT : écart entre les observations et la moyenne générale
SCA : écart entre les moyennes des modalités de A et la moyenne générale
SCB : écart entre les moyennes des modalités de B et la moyenne générale
SCI : écart entre les moyennes des sous-groupes et les moyennes des modalités de A et B
SCR : écart entre les observations et la moyenne de leur sous-groupe

Le tableau de l'ANOVA à deux facteurs

Source	Somme des carrés	ddl	Carré moyen	F
Facteur A	SCA	$p-1$	$\frac{SCA}{p-1}$	$F_A = \frac{SCA/(p-1)}{SCR/pq(r-1)}$
Facteur B	SCB	$q-1$	$\frac{SCB}{q-1}$	$F_B = \frac{SCB/(q-1)}{SCR/pq(r-1)}$
Interaction AB	SCI	$(p-1)(q-1)$	$\frac{SCI}{(p-1)(q-1)}$	$F_{AB} = \frac{SCI/((p-1)(q-1))}{SCR/(pq(r-1)}$
Erreur	SCR	$pq(r-1)$	$\frac{SCR}{pq(r-1)}$	
Total	SCT	$N-1$		

où p = nombre de modalités de A, q = nombre de modalités de B
r : nombre de répétitions, et $N = pqr$: nombre d'observations

Proposition : sous hypothèse $\mathcal{H}_0$ (tous les μ_{ij} sont égaux à μ), alors les statistiques F suivent des lois de Fisher-Snedecor à ddl et $pq(r-1)$ d.d.l.

$\rightarrow$ si la variation expliquée par l'interaction, le facteur A, ou le facteur B est significativement supérieure à la variation résiduelle, on rejette $\mathcal{H}_0$.
$\rightarrow$ test unilatéral à droite

Tests d'hypothèse dans l'ANOVA à deux facteurs

On commence par le test sur l'interaction.
Si l'interaction est significative (rejet de $\mathcal{H}_0$), il devient difficile d'interpréter les effets de A et B.

Si l'interaction n'est pas significative (on ne peut pas rejeter $\mathcal{H}_0$ sur le critère F_{AB}), on teste l'effet des facteurs A et B.

Remarque : sous $\mathcal{H}_0$, $\forall i, j, \alpha_i + \beta_j + \gamma_{ij} = 0$, donc (somme sur j) $\alpha_i = 0$, (somme sur i) $\beta_j = 0$, et $\gamma_{ij} = 0$.
$\rightarrow$ c'est la raison pour laquelle certains auteurs parlent de trois hypothèses $\mathcal{H}_0$ distinctes pour F_A, F_B, et F_{AB}.

L'implication importante est que rejeter $\mathcal{H}_0$ pour F_{AB} par exemple veut dire que les γ_{ij} ne sont pas tous nuls (ce qui est plus facile à interpréter que $\ll$ il existe une différence entre deux moyennes μ_{ij} et $\mu_{i'j'}$ $\gg$)

Plan

Conclusion

L'ANOVA est un outil puissant pour comparer les moyennes de multiples groupes de modalités tout en contrôlant l'erreur de type I.

- un facteur : évalue l'impact d'une seule variable (ex1)
- deux facteurs : évalue l'impact de deux variables **et** permet de découvrir si l'une modifie le comportement de l'autre (interaction) (ex2)

→ penser à vérifier les conditions d'application (indépendance, normalité, homogénéité des variances) avant d'interpréter les résultats (ex3 sur ordinateur)

→ après ANOVA, tests post-hoc (ex3 sur ordinateur)