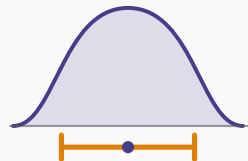


Statistique inférentielle

Chapitre 18 : L'analyse de la variance



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Pourquoi un outil de plus

La décomposition

L'application

Ce qu'il faut retenir

Pourquoi un outil de plus

Trois groupes, trois comparaisons

Bac contre Bac+2, Bac+2 contre Bac+5, Bac contre Bac+5. Trois tests à 5 % chacun.

$$P(\text{au moins un rejet à tort}) = 1 - (0,95)^3 = 0,143$$

Le risque réel n'est plus 5 % mais 14 % – et il monte à 40 % avec cinq groupes.

La solution

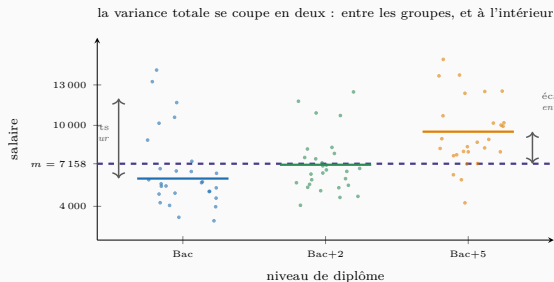
Un **test unique**, portant sur les trois moyennes à la fois :

$$H_0 : m_1 = m_2 = m_3 \quad H_1 : \text{au moins une diffère}$$

Le risque annoncé redevient le risque réel.

La décomposition

Ce que l'ANOVA regarde



La décomposition, en une ligne

$$\underbrace{SCT}_{3\,375\,061} = \underbrace{SCE}_{869\,848} + \underbrace{SCR}_{2\,505\,213}$$

(en milliers)

$$\eta^2 = \frac{869\,848}{3\,375\,061} = 0,258$$

Le diplôme explique 25,8 % de la dispersion des salaires. Les trois quarts restants viennent d'ailleurs.

C'est déjà le R^2 de la régression
— la même quantité, sous un autre nom, et le chapitre 20 le montrera.

L'ANOVA n'est pas une méthode de plus : c'est la décomposition de la variance du cours de descriptive, transformée en test.

La décomposition, déjà vue en descriptive

$$\underbrace{\text{SCT}}_{\text{variation totale}} = \underbrace{\text{SCE}}_{\text{entre les groupes}} + \underbrace{\text{SCR}}_{\text{à l'intérieur}}$$

$$\text{SCT} = \sum_{i,j} (x_{ij} - \bar{x})^2 \quad \text{SCE} = \sum_i n_i (\bar{x}_i - \bar{x})^2 \quad \text{SCR} = \sum_{i,j} (x_{ij} - \bar{x}_i)^2$$

Le nom du test est trompeur

On appelle cela « analyse de la variance », et l'on y compare pourtant des **moyennes**.

La raison est dans la figure : si les moyennes de groupe s'écartaient peu, la dispersion entre elles serait faible devant la dispersion interne. **C'est en comparant deux dispersions qu'on juge des moyennes.**

La statistique de test

$$F = \frac{SCE/(k-1)}{SCR/(n-k)} \sim \text{Fisher à } (k-1) \text{ et } (n-k) \text{ degrés de liberté}$$

où k est le nombre de groupes et n l'effectif total.

Pourquoi cette forme

On divise chaque somme de carrés par son nombre de degrés de liberté — la règle du chapitre 4, appliquée deux fois.

Sous H_0 , les deux quantités estiment la même variance résiduelle : leur rapport vaut 1 en moyenne. Un F nettement supérieur à 1 signale que les moyennes diffèrent.

L'application

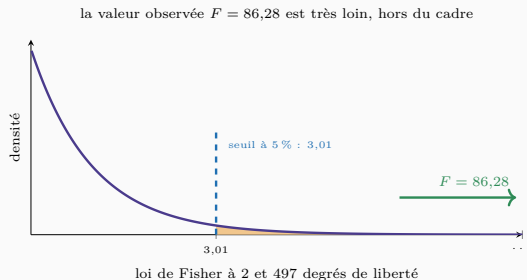
Le diplôme explique-t-il le salaire

Les trois groupes

Diplôme	n_i	Moyenne
Bac	219	6 058,90
Bac+2	174	7 080,75
Bac+5	107	9 534,39
Ensemble	500	7 158,26

Le tableau d'analyse de la variance

Source	Somme de carrés	ddl	Carré moyen
Entre (SCE)	869 847 811	2	434 923 906
Résiduelle (SCR)	2 505 213 175	497	5 040 671
Totale	3 375 060 986	499	



Lire la statistique F

$$F = \frac{SCE/(k-1)}{SCR/(n-k)} = \frac{434\,924}{5\,041} = 86,28$$

Numérateur : la dispersion *entre* les diplômes. Dénominateur : la dispersion *résiduelle*.

Sous H_0 , les trois moyennes sont égales et le rapport vaut 1 en moyenne. Ici il vaut 86.

$p < 0,0001$. Le diplôme a un effet, et le doute n'est pas permis.

Mais l'ANOVA s'arrête là : elle dit que *les moyennes diffèrent*, pas *lesquelles*. Il faut ensuite des comparaisons deux à deux.

Le résultat

$$F = \frac{434\,923\,906}{5\,040\,671} = 86,28 \quad \text{contre un seuil de } 3,01$$

$p < 0,0001$: on rejette H_0 . Les salaires moyens diffèrent selon le diplôme.

Le rapport de corrélation, et ce qu'il ajoute

$$\eta^2 = \frac{\text{SCE}}{\text{SCT}} = \frac{869\,848}{3\,375\,061} = 0,258$$

Le diplôme explique 25,8 % de la dispersion des salaires. **Les trois quarts viennent d'ailleurs** — ancienneté, poste, performance, hasard.

Le test dit que l'effet est réel; η^2 dit qu'il est important sans être déterminant. Les deux sont nécessaires.

Elle rejette en bloc

« Au moins une moyenne diffère » : elle ne dit pas laquelle, ni de combien.

Ici la lecture des trois moyennes suffit — l'écart Bac / Bac+5 dépasse trois mille dirhams. Sur des cas moins nets, il faut des **comparaisons multiples** avec un seuil corrigé.

Les conditions de validité

1. Échantillons **indépendants**.
2. Population **normale** dans chaque groupe — ou effectifs suffisants.
3. Variances **comparables** entre groupes.

La troisième est la plus fragile; l'ANOVA la supporte assez bien si les effectifs sont équilibrés, mal s'ils ne le sont pas.

Ce qu'il faut retenir

Six points

1. Comparer k groupes deux à deux **gonfle le risque réel** : 14 % pour trois groupes, 40 % pour cinq.
2. L'ANOVA teste toutes les moyennes en **un seul test**.
3. $SCT = SCE + SCR$: la décomposition du cours de statistique descriptive.
4. F compare deux dispersions pour juger des **moyennes**; sous H_0 il vaut 1 en moyenne.
5. Sur nos données : $F = 86,28$, $p < 0,0001$, et $\eta^2 = 0,258$.
6. L'ANOVA rejette **en bloc** : elle ne désigne pas les groupes responsables.

La suite

Il reste un type de données non traité : les **effectifs de tableau**, quand les deux variables sont qualitatives.

Le chapitre 19 y consacre le dernier test du cours — celui qui ferme la boucle avec le tableau de contingence de la statistique descriptive.