Chapitre 03 : L'analyse de la variance à un critère de classe

ANOVA (ANalysis Of VAriance) :

Introduction:

L'analyse de la variance **(ANOVA)** a pour objectif d'étudier l'influence d'un ou plusieurs facteurs sur une variable quantitative. Nous nous intéresserons ici au cas où les niveaux, ou modalités, des facteurs sont fixés par l'expérimentateur. On parle alors de modèle *fixe*.
C'est la comparaison de moyennes pour plusieurs groupes (> 2). Il s'agit de comparer la **variance intergroupe (entre les différents groupes : écart des moyennes des groupes à la moyenne totale)** à **la variance intragroupe (somme des fluctuations dans chaque groupe).** S'il n'y a pas de différence entre les groupes, ces deux variances sont (à peu près) égales. Sinon, la variance intergroupe est nécessairement la plus grande.
L'ANOVA se résume à une comparaison multiple de moyennes de différents échantillons constitués par les différentes modalités des facteurs. Les conditions d'application du test paramétrique de comparaison de moyennes s'appliquent donc à nouveau.
Elle peut être vue comme une généralisation du test de Student.
On souhaite tester les effets de **k** traitements qui ont été administrés respectivement à $\mathbf{n_1,\ldots\ldots n_k}$ individus. En analyse de variance, le paramètre susceptible d'influer sur les données étudiées s'appelle un *facteur*, et ses valeurs sont les *modalités*.
Dans le modèle probabiliste, chaque modalité correspond à un échantillon. Pour h = 1,….k, on note :

$$(X_1, X_2, \ldots, X_{\mathrm{nk}})$$

On cherche à savoir si la variabilité observée dans les données est uniquement due au hasard, ou s'il existe effectivement des différences significatives entre les classes, imputables au facteur.
Pour cela, on va comparer les variances empiriques de chaque échantillon, à la variance de l'échantillon global, de taille $\mathbf{n_1+\ldots+n_k=n}$.
La moyenne des variances (pondérée par les effectifs) résume la variabilité à l'intérieur des classes, d'où le nom de variance *intra-classes (intragroupes)*, ou variance *résiduelle*.
La variance des moyennes décrit les différences entre classes qui peuvent être dues au traitement, d'où le nom de variance *inter-classes (intra-groupes)*, ou variance *expliquée*.
On note : $\bar{X}^h$ la moyenne empirique de la h-ième classe,
V^h la variance empirique de la h-ième classe,

$\overline{X}$ la moyenne de l'échantillon global,
La moyenne des variances (variance intra-classes), Vintra
La variance des moyennes (variance inter-classes), Vinter.
S^2 : la variance de l'échantillon global.
Alors :

$$\boldsymbol{S^2} = \textbf{Vintra} + \textbf{Vinter}.$$

Principe de l'analyse de variance

- Hypothèses
  - Echantillons (groupes) indépendants
  - Distribution normale du critère au sein des groupes
  - Variances identiques d'un groupe à l'autre

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (k groupes)

H_1 : au moins l'une des moyennes diffère des autres

Notation :

Facteur	Groupe 1	Groupe 2	…	Groupe j
Effectif	n_1	n_2	…	n_j
Mesure	x_{11}	x_{12}	…	x_{1j}
Mesure	x_{21}	x_{22}	…	x_{2j}
Mesure	…	…	…	…
Mesure	x_{i1}	x_{i2}	…	x_{ij}
Moyennes	$\overline{x}_1$	$\overline{x}_2$	…	$\overline{x}_j$

x : variable à laquelle on s'intéresse
k : nombre de groupes
n_j : taille du groupe j
X_{ij} : $i^{ème}$ observation du groupe j

Décomposition de la variabilité des observations

- Mesure de la dispersion totale : SCE_T
  - Somme des carrés des écarts à la moyenne générale : $\sum (x_{ij} - \bar{x})^2$
- Mesure de la dispersion intra-groupe : SCE_R
  - Somme des carrés des écarts à la moyenne d'un groupe : $\sum (x_{ij} - \bar{x}_j)^2$
- Mesure de la dispersion inter-groupe SCE_A
  - Somme des carrés des écarts de la moyenne d'un groupe à la moyenne générale : $\sum n_j (\bar{x}_j - \bar{x})^2$

☞ $SCE_T = SCE_R + SCE_A$

Estimation de la variance inter-groupe SCE_A

- Elle ne dépend que de la dispersion des moyennes des groupes comparés

  ⇔ Somme des carrés des écart due au facteur étudié
- SCE_A a k-1 degrés de liberté
- Sa variance σ^2_A est estimée par :

$$S_A^2 = \frac{SCE_A}{k-1} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2}{k-1}$$

- Pour les calculs, on montre que SCE_A s'écrit :

$$SCE_A = \sum_j \frac{T_j^2}{n_j} - \frac{T_G^2}{n}$$

  - T_j = total des valeurs de x du groupe j (somme des valeurs x du groupe j)
  - T_G = total général (somme globale des valeurs x)

Estimation de la variance intra-groupe SCE_R

- Elle ne dépend que de la dispersion des valeurs x_{ij} au sein de chaque groupe

  ⇔ Somme des carrés des écart intra-classe ou résiduelle
- SCE_R a n-k degrés de liberté
- Sa variance σ^2_R est estimée par :

$$S_R^2 = \frac{SCE_R}{n-k} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n-k}$$

- Pour les calculs, on montre que SCE_A s'écrit :

$$SCE_R = \sum_{ij} x_{ij}^2 - \sum_j \frac{T_j^2}{n_j}$$

avec T_j = total des valeurs de x du groupe j (somme des valeurs x du groupe j)

- Après avoir décomposé la variance totale, le principe consiste à comparer S^2_A/S^2_R

☞ Tester si le rapport des 2 variances S^2_A/S^2_R est proche de 1

☞ Statistique de test distribuée selon une loi dite de Fisher à ν_1= k-1 et ν_2 = n-k degrés de liberté (ddl)

- $F_0 = S^2_A/S^2_R$
- Test unilatéral dans tous les cas
- si H_0 vraie : $S^2_A \approx S^2_R$ et donc $F_0 \approx 1$
- si H_1 vraie : $S^2_A > S^2_R$ et donc $F_0 > 1$

$H_o : \sigma_A^2 = \sigma_R^2$ $\quad$ $H_1 : \sigma_A^2 > \sigma_R^2$

1. Calculer $F_o = \frac{s_A^2}{s_R^2}$ à partir des observations sur l'échantillon

2. Comparer F_o à la valeur seuil de F^{k-1}_{n-k} :

=> règle de décision

$F_o \geq F^{k-1}_{n-k\,(\alpha)}$: rejet de H_o (au risque α) **d'indépendance**

$F_o < F^{k-1}_{n-k\,(\alpha)}$: non rejet de H_o

Tableau d' "analyse de la variance"

Source de variation	Somme des carrés des écarts	ddl	Carré moyen (ou variance)	F
Entre groupes (facteur A)	SCE_A	k-1	$s_A^2 = \frac{SCE_A}{k-1}$	$F_o = \frac{s_A^2}{s_R^2}$
Résiduelle	SCE_R	n-k	$s_R^2 = \frac{SCE_R}{n-k}$	
Total	$SCE_T = SCE_A + SCE_R$	n-1		

Avec :

$$S_A^2 = \frac{SCE_A}{k-1} = \frac{\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2}{k-1}$$

$$S_R^2 = \frac{SCE_R}{n-k} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n-k}$$

Exercice 01:

Nombre/km2 (densité) de sapins poussant dans 3 (= k) forêts différentes (groupes) :

Groupe 1	Groupe 2	Groupe 3
45	78	354
34	69	338
35	86	351
29	58	332
42	57	341
37	64	358
44		347
28		
Variance	**42,214**	**131,867**	**86,476**

Question : à un niveau de risque de 5 %, la densité moyenne de sapin est-elle la même dans les 3 forêts ?

Solution :

On veut tester :

H_0 : toutes les densités moyennes sont identiques / le type de foret n'influence pas sur la densité des sapins

H_1 : Il y a au moins une densité moyenne différente des autres / le type de foret influence (a un effet) sur la densité des sapins

$n_1 = 8$; $n_2 = 6$; $n_3 = 7$ donc : n=8+6+7= 21= le nombre d'observations ;

Total$_1$ =$\sum x_{i1}$ =294, Total$_2$ = $\sum x_{i2}$= 412, Total$_3$= $\sum x_{i3}$= 2421 ;

$$T = \sum x_{ij} = \sum x_{i1} + \sum x_{i2} + \sum x_{i3} = 294 + 412 + 2421 = 3127$$

$$\sum\sum {x_{ij}}^2 = (45)^2 + (34)^2 + \cdots + (347)^2 = 877889,$$

k (Nombre de groupes) = **3**

$$SCE_R = \sum\sum {x_{ij}}^2 - \sum \frac{T_j^2}{n_j}$$

$$= 877889 - \left(\frac{294^2}{8} + \frac{412^2}{6} + \frac{2421^2}{7}\right) = 1473,69$$

$$SCE_A = \sum \frac{T_j^2}{n_j} - \frac{T^2}{n} = \left(\frac{294^2}{8} + \frac{412^2}{6} + \frac{2421^2}{7}\right) - \frac{(3127)^2}{21} = 410790,119$$

$$S_R^2 = \frac{SCE_R}{(n-k)} = \frac{1473,69}{(21-3)} = 81,872$$

$$S_A^2 = \frac{SCE_A}{(k-1)} = \frac{410790,119}{(3-1)} = 205395,060.$$

Test statistique :

La statistique $F_{Obs} = \frac{S_A^2}{S_R^2} = \frac{205395,060}{81,872} = 2508,743$

F_{Obs} est comparée à une valeur tabulée F_{Tab} à (k-1=3 – 1 = 2) et (n-k=21 – 3 = 18) degrés de liberté

Donc:

$F_{Tabulée}$ = F(5 %, 2; 18) = **3,555** à.

Décision :

$F_{calculé}$ =**2508,743**> $F_{Tabulée} = \mathbf{3,555}$: on rejette l'hypothèse nulle.

Donc, les densités moyennes de sapins ne sont pas les mêmes = le facteur « Forêt » a un effet sur la densité des sapins (il y a un effet du milieu (forêt) sur la densité de sapins) = Il y a une relation de dépendance.