Chapitre 7 Analyse de la variance (ANOVA)

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.

L'analyse de variance (analysis of variance ou ANOVA) peut être vue comme une généralisation du test de Student.

On souhaite tester les effets de k traitements qui ont été administrés respectivement à n_1,……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* (ici les différents traitements).

Dans le modèle probabiliste, chaque modalité correspond à un échantillon. Pour h = 1,…..k, on note :

$$(X_1^{(h)}, \ldots, X_{n_h}^{(h)}),$$

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 n1+…+nk=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 (intra-groupes)*, 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 :

- $\overline{X}^{(h)}$ 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), V_{intra}

- La variance des moyennes (variance inter-classes), V_{inter}

S^2 la variance de l'échantillon global.

Alors :

$$S^2 = V_{intra} + V_{inter} .$$

Test d'homogénéité des variances

Pour beaucoup de tests paramétriques (ANOVA, régression), l'homogénéité des variances est une condition nécessaire.

Homogénéité des variances = homoscédasticité

Plusieurs méthodes existent pour tester l'homogénéité des variances dans plusieurs groupes qui n'ont pas nécessairement le même nombre d'objets. Un test très utilisé est le test de Bartlett, détaillé ici. Ce test est valide si les distributions des objets sont Normales (*Le test de Bartlett estime si les différentes sous-catégories d'une variable de distribution normale ont la même variance). Le test donne un résultat global et ne permet pas d'estimer les différences de variances des sous-catégories deux à deux. Il est très sensible à la non-normalité.*

.

Exemple

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**

Avant de tester l'effet du milieu (forêt) sur la densité de sapins par une ANOVA, il faut vérifier l'homogénéité des variances.

Question : à un niveau de risque de 5 %, les variances de ces trois groupes sont-elles homogènes?

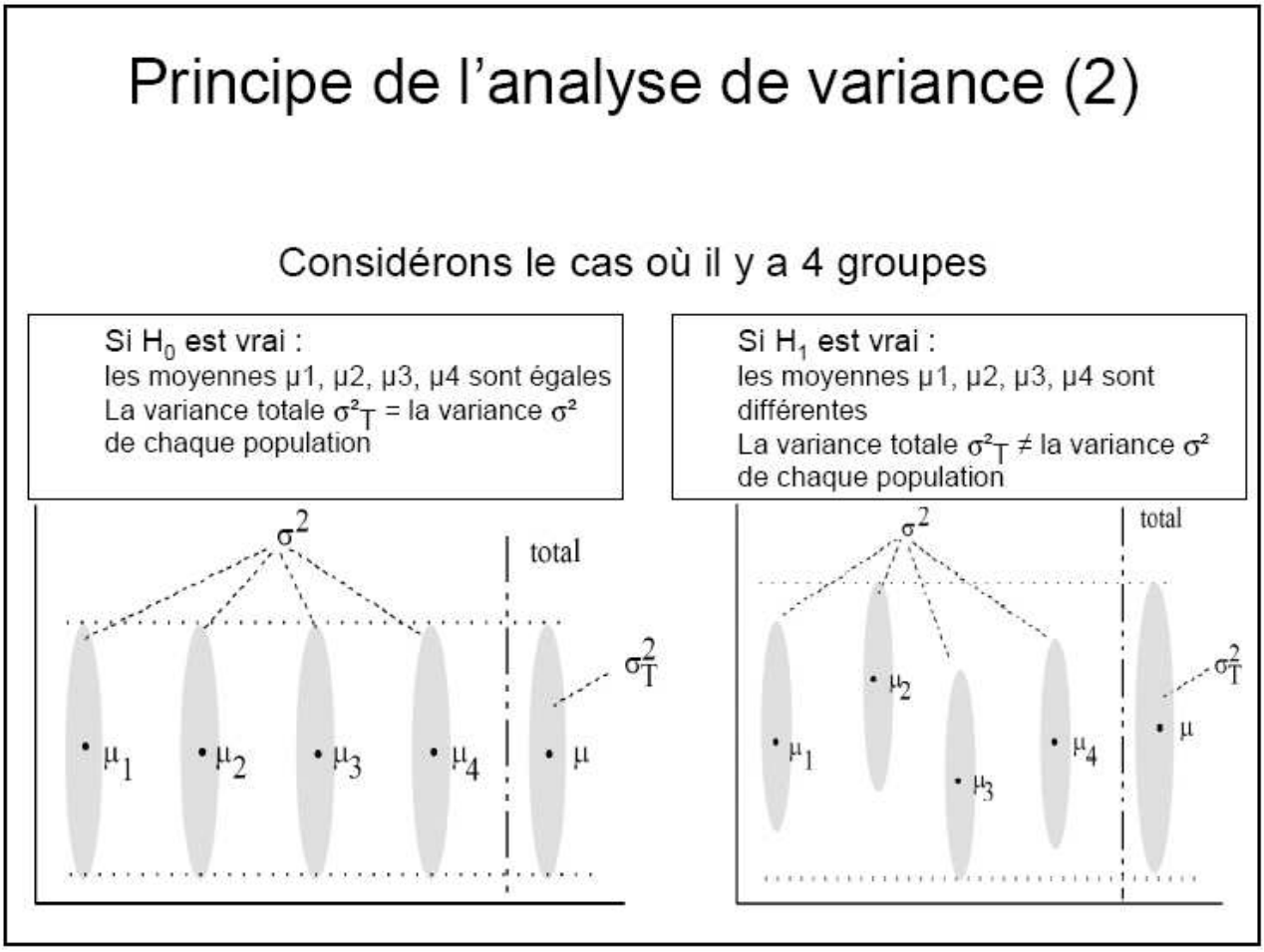
Hypothèses :
H0 : toutes les variances sont égales
H1 : au moins une des variances est différente des autres
Test :
$Sp^2 = \Sigma((ni - 1)si^2)/\Sigma(ni - 1)$
$B = (\Sigma(ni - 1))(\ln Sp^2) - \Sigma((ni - 1)\ln si^2)$

$C = 1 + 1/(3(k - 1))[\Sigma 1/(ni - 1) - 1/(\Sigma(ni - 1))]$
$BC = B/C$. Sous H_0, B_C suit une loi du Khi-carré (χ^2) à $(k - 1)$ ddl (ν).

Condition : distributions Normales des populations d'origine.

Règle de décision : H_0 est rejetée si $B_C > \chi^2_{0,05;2}$, soit 5,99.
Calcul du test :
$Sp^2 = ((7)42,214 + (5)131,867 + (6)86,476)/(7 + 5 + 6) = 81,872$
$B = (7 + 5 + 6)\ln 81,872 - (7 \ln 42,214 + 5 \ln 131,867 + 6 \ln 86,476) = 1,925$
$C = 1 + (1/6)[(1/7 + 1/5 + 1/6) - (1/(7 + 5 + 6))] = 1,076$
$B_C = 1,925/1,076 = 1,789$
Décision : $B_C < 5,99$, H_0 est acceptée : Les trois variances sont homogènes.

Explication de ANOVA à un critère (ou facteur)



Principe de l'analyse de variance (3)

- La dispersion totale σ^2_T a 2 composantes
  - Fluctuations individuelles : σ^2 qui est la variance interne à chaque groupe (variance intra-groupe)
  - Fluctuations entre les groupes : la variation entre les μ_i qui correspond à la variabilité entre les groupes (variance inter-groupe)
- Si la variabilité inter-groupe > la variabilité intra-groupe $\Rightarrow$ 2 moyennes au moins diffèrent
- Principe général :
  - ☞ Décomposer σ^2_T en ses 2 parties
  - ☞ Tester si σ^2_T est différent de σ^2

Principe de l'analyse de variance (4)

- Hypothèses
  - Echantillons (groupes) indépendants
  - Distribution normale du critère au sein des groupes
  - Variances identiques d'un groupe à l'autre
- L'ANOVA est un test robuste (résultats assez peu affectés par de légers écarts à ces hypothèses)

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

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

Conventions de notations

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	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_j$

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 : $\sum n_j (\bar{x}_j - \bar{x})^2$

☞ $SCE_T = SCE_R + SCE_A$

ANOVA : méthode de calcul (1)

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)

ANOVA : méthode de calcul (2)

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)

ANOVA : méthode de calcul (3)

- 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 $\nu_1 = k-1$ et $\nu_2 = n-k$ 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 $F_0 > 1$

ANOVA : Execution du test (1)

$H_o : \sigma^2_A = \sigma^2_R$ $\quad$ $H_1 : \sigma^2_A > \sigma^2_R$

1. Calculer $F_o = \frac{S^2_A}{S^2_R}$ 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

ANOVA : Execution du test (2)

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	$s_A^2 = \frac{SCE_A}{k-1}$	$F_0 = \frac{s_A^2}{s_R^2}$
Résiduelle	SCE_R	n-k	$s_R^2 = \frac{SCE_R}{n-k}$	
Total	$SCE_T = SCE_A + SCE_R$	n-1		

Avec :

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

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

Exemple
Mêmes données que précédemment, mais la question devient : la densité moyenne de sapin est-elle la même dans les 3 forêts ?

Hypothèses

H_0: toutes les moyennes selon le facteur sont égales
H_1: au moins une des moyennes μ_r est différente des autres

Variable dépendante : Densité en sapin (nb/km2)
Facteur : Forêt (s = 3 niveaux).
Calculs :

$Total_1 = 294$, $Total_2 = 412$, $Total_3 = 2421$; Total (T)= 3127
$\Sigma\Sigma x^2 = 877889$
$n = 21$; $n_1 = 8$; $n_2 = 6$; $n_3 = 7$
k (Nbre de groupes) = 3
$SCE_R = \Sigma\Sigma x^2 - \Sigma(Tj^2/nj) = 877889 - (294^2/8 + 412^2/6 + 2421^2/7) = 1473,69$
$SCE_A = \Sigma(Tj^2/nj) - T2/n = (294^2/8 + 412^2/6 + 2421^2/7) - 31272/21 = 410790,119$
$S_R^2 = SCE_R/(n - k) = 1473,69/(21 - 3) = 81,872$
$S_A^2 = SCE_A/(k - 1) = 410790,119/(3 - 1) = 205395,060$

Test statistique

$F = S_A^2 / S_R^2 = 205395,060/81,872 = 2508,743$

F est comparé à un F à (3 – 1 = 2) et (21 – 3 = 18) degrés de liberté
Donc:
Fcritique = F(2; 18) = 3,555 à 5 %. Attention, l'ANOVA est *toujours* un test unilatéral.

Si Fcal > F* (H0) : on rejette l'hypothèse d'indépendance
Si Fcal < F* (H1) on accepte l'hypothèse d'indépendance, on accepte H_1 (pas de relation entre les variables).

Décision

Fcalculé > Fcritique: 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.

Exemple 2 :

ANOVA - Exemple (1)

Groupes	Poids (kg) [50 - 59]	[60 - 69]	[70 - 79]	[80 - 89]	Total
Effectifs (nj)	10	10	10	10	40
Corticoïdes urinaires (mg/24h)					
$\overline{x}_i$	3,78	5,26	5,97	6,79	218 et $\overline{x}=5,45$
$n_j \cdot (\overline{x}_j - \overline{x})^2$	27,89	0,36	2,70	17,96	48,91
$\sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$	6,84	22,26	19,94	20,83	69,87

Construire le tableau d'ANOVA

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	$s_A^2 = \frac{SCE_A}{k-1}$	$F_0 = \frac{s_A^2}{s_R^2}$
Résiduelle	SCE_R	n-k	$s_R^2 = \frac{SCE_R}{n-k}$	
Total	$SCE_T = SCE_A + SCE_R$	n-1		

H_0 : lorsque le poids augmente, on n'assiste pas forcement à une augmentation des corticoïdes urinaires.

ANOVA - Exemple (2)

Source de variation	Somme des carrés	Degré de liberté (ddl)	Variance	F
Entre groupes $\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$	$SCE_A = 48{,}91$	$k-1 = 4-1 = 3$	$S_A^2 = \frac{48{,}91}{3} = 16{,}30$	$\frac{16{,}30}{1{,}94} = 8{,}40$
Résiduelle $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$	$SCE_R = 69{,}87$	$n-k = 40-4 = 36$	$S_R^2 = \frac{69{,}87}{36} = 1{,}94$	
Totale	$SCE_T = 118{,}78$	$n-1 = 40-1 = 39$		

$\Rightarrow F_3^{36}(5\%) = 2{,}90 \Rightarrow$ on rejette H0 au risque de 5%

$\Rightarrow F_3^{36}(1\%) = 4{,}60 \Rightarrow$ on rejette H0 au risque de 1%

Exercice 1

Le tableau suivant présente des mesures de la hauteur (en mm) de la plante *Saede brassica*, réalisées dans plusieurs milieux différents. Un chercheur désire comparer ces données afin de connaître l'effet du milieu sur la taille de *S. brassica* (on admet que les données suivent une distribution Normale).

	Milieu 1	**Milieu 2**	**Milieu 3**	**Milieu 4**	**Milieu 5**
	12	141	56	87	241
	15	146	67	105	264
	12	135	43	79	225
	18	147	78	123	257
	24	154	45	114	248
	32		69		258
	31				236
	15				
Tj (=Σxj)	159	723	358	508	1729

1. Quelle analyse permet d'estimer l'effet du milieu sur la hauteur des plantes ?
2. Quelles sont les conditions requises pour pouvoir réaliser cette analyse ?
3. Vérifiez ces conditions et réalisez l'analyse statistique appropriée.

Exercice 2

On veut savoir si la quantité de nitrates varie d'une station à l'autre le long d'une rivière. Pour cela, on prélève en 10 points (n=10) chaque fois une certaine quantité d'eau dans 3 stations différentes (k=3).

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