Le modèle linéaire général : retour sur l'analyse de variance

J.N. Bacro
Master ESDB, Université Montpellier

2023/2024

Retour sur l'ANOVA 1 facteur

On reprend l'exemple des 3 correcteurs pour lesquels on a choisi au hasard 5 copies et on cherche à caractériser un éventuel *"effet correcteur"*

1. Modèles à considérer ?
2. soit θ le vecteur des paramètres. On sait estimer σ^2 et, **éventuellement sous contraintes**, θ.

   D'une façon générale, θ n'est pas toujours estimable mais on sait par exemple que $X\theta$ est une combinaison linéaire de θ qui est toujours estimable (car projection de $\mathbf{Y}$) . . . il existe donc des combinaisons linéaires de θ toujours estimables (i.e. estimation unique ne dépendant pas des contraintes).

   Parmi celles-ci, les contrastes sont de tout premier intérêt.

Définition :
On appelle contraste toute combinaison linéaire des paramètres $\sum_i c_i\theta_i$ vérifiant $\sum_i c_i = 0$.

Intérêt d'un contraste ?

ANOVA 1 facteur et contrastes

Dans l'approche ANOVA, on peut considérer deux écritures
(1) $y_{ij} = \mu_j + e_{ij}$ et (2) $y_{ij} = \mu + \alpha_j + e_{ij}$; l' écriture (1) ne nécessite pas de contraintes pour l'estimation . . . mais (2) si ! Soit L un contraste.

Pour (1), $L = \sum_j c_j \mu_j$ avec $\sum_j c_j = 0$ et L peut être estimé par $\hat{L} = \sum_j c_j \hat{\mu}_j$; comme $\hat{\mu}_j = y_{.j}$, et via les postulats, $y_{.j} \sim N(\mu_j, \frac{\sigma^2}{n_j})$ d'où

$$\hat{L} \sim N\left(\sum_j c_j \mu_j, \sum_j \frac{c_j^2 \sigma^2}{n_j}\right)$$

et on peut construire pour L des intervalles de confiance, faire des test etc (où σ^2 estimée par . . .).

Pour (2) ?
$\mathcal{L} = \sum_j c_j \alpha_j$ avec $\sum_j c_j = 0$ et α_j non-estimables . . . mais
$\mathcal{L} = \sum_j c_j \alpha_j + \mu \sum_j c_j = \sum_j c_j(\mu + \alpha_j) = \sum_j c_j \mu_j = L$ et $\hat{\mathcal{L}} = \hat{L}$, *ce qui permet là aussi de faire de l'inférence sur $\mathcal{L}$ . . . donc des tests associés aux paramètres considérés !*

Exemple : pour tester si les effets des correcteurs 1 et 3 sont significativement différents, quel contraste pourrait-on considérer ?

Les comparaisons multiples

Lorsque le modèle M_1 est meilleur que M_0 (i.e. effet significatif du facteur considéré), on est tenté de chercher où sont les différences, quelles espérances diffèrent de quelles autres ... faire des comparaisons multiples pour caractériser les différences

Problèmes :

1. contrôle du risque de 1ère espèce relatif à l'ensemble des tests effectués *(notion de risque global α_g)* ;
2. *si possible*, ne pas (trop) interférer dans les procédures de tests en choisissant des hypothèses a posteriori (hypothèses guidées par les résultats d'expérience)

Exemple : 3 espérances ; tester H_0 $\mu_1 = \mu_2$ et $\mu_2 = \mu_3$ et $\mu_1 = \mu_3$ contre à chaque fois H_1 différence. Pour un **test individuel**, confiance $1 - \alpha$: $P(\textit{garder } H_0 \mid H_0 \textit{ vraie}) = 1 - \alpha$ ... *quid d'une confiance sur* **ensemble des tests** $1 - \alpha_g$, *i.e. d'une confiance globale sur l'ensemble des tests ?*

$$1-\alpha_g = P(\textit{garder } H_0 \textit{ test 1 et } \ldots \textit{ et garder } H_0 \textit{ test 3} \mid \textit{les } H_0 \textit{ vraies}) \approx (1-\alpha)^3$$

Pour $\alpha = 0.05$, $(1 - \alpha)^3 = 0.857$ d'où $\alpha_g \approx 15\%$! ... $\alpha_g \approx 1 - (1 - \alpha)^{nc}$, où nc =nombre comparaisons.

$\rightsquigarrow$ contrôler α_g !

Les tests multiples

Beaucoup de méthodes pour contrôler α_g ... mais aucune optimale !

Dans le cadre des tests multiples, sur les m hypothèses nulles testées, m_0 sont vraies et m_1 sont fausses. Les tests réalisés concluent à R rejets et W non rejets des hypothèses H_0 testées.

Scénario type :

H_0	Gardées	Rejetées	Total
Vraies	U	V	m_0
Fausses	T	S	m_1
Total	W	R	m

La généralisation du risque α individuel au cadre tests multiples présente différents aspects : PCER *(Per Comparison Error Rate)* ; PFER *(Per Family Error Rate)* ; FWER *(Familywise Error Rate)* ; gFWER *(Generalised Familywise Error Rate)* ; FDR *(False Discovery Rate)* ...

Les tests multiples

Scénario type :

H_0	Gardées	Rejetées	Total
Vraies	U	V	m_0
Fausses	T	S	m_1
Total	W	R	m

• Espérance du taux d'erreur de Type I : $PCER = \frac{\mathbb{E}(V)}{m}$
Si chaque test est de risque α, $PCER = \frac{m_0 \alpha}{m} \leq \alpha$ . . . *mais ne prend pas en compte la multiplicité des tests donc peu d'intérêt en pratique.*

• Espérance du nombre d'erreur de Type I : $PFER = \mathbb{E}(V)$. *Ce n'est pas une probabilité* et $PFER = mPCER$.

• Probabilité de commettre au moins une erreur de Type I :
$FWER = P(V > 0)$. Prise en compte de la multiplicité des tests
. . . historiquement, l'approche la plus utilisée !

• Probabilité de commettre au moins q erreurs de Type I :
$gFWER = P(V > q)$; *en pratique on préfère souvent utiliser FWER . . .*

Les tests multiples

Scénario type :

H_0	Gardées	Rejetées	Total
Vraies	U	V	m_0
Fausses	T	S	m_1
Total	W	R	m

- Espérance de la proportion d'erreur de type I **parmi les rejetées** :

$$\begin{aligned} FDR &= \mathbb{E}\left(\frac{V}{R}\right) = \mathbb{E}\left(\frac{V}{R} \mid R > 0\right)\mathbb{P}(R > 0) + 0 \times \mathbb{P}(R = 0) \\ &= \mathbb{E}\left(\frac{V}{R} \mid R > 0\right)\mathbb{P}(R > 0) \end{aligned}$$

Approche plus récente *souvent utilisée pour les analyses en grande dimension (-omiques).*

On peut montrer que

$$PCER \leq FDR \leq FWER \leq PFER$$

et en pratique seules les approches *FDR* et *FWER* sont couramment utilisées. Le plus souvent, *FDR* pour un cadre exploratoire (avec grand nombre d'hypothèses à tester) puis *FWER* pour étude "confirmatoire" qui suppose un contrôle plus strict du nombre de rejets à tort.

Les tests multiples

Principe général :
1) recherche d'une *p*-value *ajustée* pour chaque test *j* en relation avec α_g fixé, $1 \leq j \leq m$, généralisant alors la notion de *p*-value attachée à un test individuel.

2) pour α_g fixé, rejet de $H_0^{(j)}$ si la *j*ème *p*-value ajustée est inférieure à un α individuel choisi pour assurer un risque global α_g.

Le *FWER* coincide dans ce cas avec α_g.

Le plus souvent :

• Approche *FWER* : la plupart des méthodes sont fondées sur une approche de comparaison d'espérance de type Student et peuvent se résumer à déterminer une valeur critique de la différence *d* des moyennes égale à :

$$quantile(tabulé) * s_d$$

• Approche *FDR* : classement ordonné des *p*-value obtenues pour les différents tests et décision de rejets pour les H_0 correspondant aux plus faibles . . .

Les comparaisons multiples

Beaucoup de méthodes pour contrôler α_g ... mais aucune optimale !

1. Approche Bonferroni
   Si J groupes, on fait $k = \frac{J(J-1)}{2}$ comparaisons ; $\alpha_g = 1 - (1-\alpha)^k \leq k\alpha$ si α petit ... *prendre* $\alpha = \frac{\alpha_g}{k}$.
   Exemple : si 4 groupes, pour avoir $\alpha_g = 5\%$, chaque test à $\alpha = 0.8\%$!
2. Approche Sidak
   même raisonnement sans approximation α petit : $\alpha = 1 - (1-\alpha_g)^{1/k}$;
3. Approche Holm-Bonferroni
   On ordonne les p-value individuelles des m tests faits :
   $p_{1,m} \leq p_{2,m} \leq \ldots \leq p_{m-1,m} \leq p_{m,m}$. Pour α fixé, soit i_0 le plus petit i tel que $p_{i,m} > \frac{\alpha}{m-i+1}$. On rejette les $H_0^{(j,m)}$ pour $1 \leq j \leq i_0 - 1$.

   Principe :

   $$p_{1,K} \leq \ldots \leq \overbrace{p_{m-i+1,m} \leq \ldots \leq p_{m,m}}^{i\ p\text{-value} \Rightarrow i\ \text{tests}} ;$$ si $p_{m-i+1,m} \leq \frac{\alpha}{i}$, les i tests considérés sont significatifs via Bonferroni, c'est-à-dire pour les i tels que

   $$p_{i,m} \leq \frac{\alpha}{m-i+1}$$

   Dès que les p-value dépassent cette valeur, les test sont non-significatifs, d'où $i_0 - 1$ tests significatifs.

Les comparaisons multiples

Beaucoup de méthodes pour contrôler α_g ... mais aucune optimale !

1. Approche Dunnett
   Comparaison de J traitements à un traitement de référence (placébo). Test fondé sur la statistique $\max\{T_1, \ldots, T_m\}$ où T_i représente la statistique de Student pour la comparaison de deux espérances. Rejet de H_0^j si $T_j \geq d_{1-\alpha}$ avec $d_{1-\alpha}$ quantile $(1-\alpha)$ de la distribution du $\max\{T_1, \ldots, T_m\}$.
2. Approche Newman-Keuls
   Comparaison, dans un ordre déterminé, entre la plus grande et la plus petite moyenne d'un groupe de ℓ moyennes à des valeurs tabulées représentant la différence maximale attendue pour la comparaison de ℓ moyennes issues d'une loi normale. Si rejet H_0, on sépare en 2 groupes distincts de $\ell - 1$ moyennes en excluant respectivement le min et le max du groupe de ℓ moyennes et processus itéré jusqu'à non-rejet.
3. Approche Scheffé
   Fondée sur la construction d'I.C. simultané valide pour tous les contrastes. On rejette l'égalité à 0 d'un contraste L si
   $$|\hat{L}| > \hat{\sigma}\sqrt{(J-1)F_{1-\alpha_g}(J-1, N-r)}\sqrt{\sum_{j=1}^{J}\frac{c_j^2}{n_j}}$$

Les comparaisons multiples

- Approche FDR

H_0	Gardées	Rejetées	Total
Vraies	U	V	m_0
Fausses	T	S	m_1
Total	W	R	m

Rmq : si chaque test est fait au risque $\frac{\alpha_g}{m}$ (Bonferroni), on a

$$P(V > 0) = m\frac{\alpha_g}{m} = \alpha_g.$$

On ordonne les m p-value $p_{1,m} \leq p_{2,m} \leq \ldots \leq p_{m-1,m} \leq p_{m,m}$ et on décide de rejeter les H_0 correspondantes aux k plus faibles p-value $p_{1,m}, \ldots, p_{k,m}$.
On a alors :

$$FDR = \frac{m_0}{k} p_{k,m} \leq \frac{m}{k} p_{k,m}.$$

$\rightsquigarrow$ chercher le plus grand k tel que $\frac{m}{k} p_{k,m} \leq \alpha_g$

$\hookrightarrow$ au risque global α_g, rejeter les k H_0 associées aux tests de p-value les plus faibles vérifiant :

$$p_{k,m} \leq \frac{k}{m}\alpha_g.$$

ANOVA 1 facteur : exemple

Des embryons de ver à soie ont été soumis pendant leurs développements à des expositions continues à différentes doses de rayons gamma (6 doses). Pour chaque dose, on dispose de 5 mesures. On donne le temps moyen de développement en fonction de la dose reçue : $y_{\cdot 1} = 28.82$; $y_{\cdot 2} = 23.98$; $y_{\cdot 3} = 14.64$; $y_{\cdot 4} = 19.92$; $y_{\cdot 5} = 13.26$; $y_{\cdot 6} = 18.70$. On donne : $\sum_{i,j}(y_{ij} - y_{\cdot\cdot})^2 = 1129.97$.

1. Ecrire les modèles à comparer pour répondre à la question : y a t il un effet de l'exposition ?
2. Sachant que $\hat{\sigma^2} = 11.79$, reconstruire la table d'ANOVA correspondante.
3. Si l'on veut comparer les 5 dernières espérances avec un risque global de 10%, à quel risque doit-on effectuer les tests individuels ?
4. La table de NK suivante donne les valeurs d'étendues critiques pour la comparaion d'un groupe de ℓ moyennes :

nbre de moyennes ℓ	2	3	4	5	6
étendue critique	4.48	5.42	5.99	6.40	6.71

Que peut-on conclure ici pour la comparaison des espérances de développement en fonction de l'exposition ?

Cas de deux facteurs : l'ANOVA à deux facteurs croisés

On considère maintenant l'effet éventuel de deux facteurs A et B, par exemple ayant respectivement deux modalités $A1$, $A2$ et $B1$, $B2$, avec 2 répétitions par traitement.

	B1	B2
A1	12 ; 14	18 ; 20
A2	19 ; 23	21 ; 33

Quels modèles envisageables ?

ANOVA à deux facteurs croisés

Exemple :
on considère deux facteurs A et B, respectivement à deux modalités $A1$, $A2$ et $B1$, $B2$, avec 2 répétitions par traitement.

	B1	B2
A1	12 ; 14	18 ; 20
A2	19 ; 23	21 ; 33

Quels modèles envisageables ?

$1 \leq i \leq 2$; $1 \leq j \leq 2$; $1 \leq k \leq 2$
M_0 $y_{ijk} = \mu + e_{ijk}$;
M_1 $y_{ijk} = \mu_i + e_{ijk}$; $y_{ijk} = \mu + \alpha_i + e_{ijk}$
M_2 $y_{ijk} = \mu_j + e_{ijk}$; $y_{ijk} = \mu + \beta_j + e_{ijk}$
M_3 $y_{ijk} = \mu_{ij} + e_{ijk}$; $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$

M_3 est un modèle *additif* : on voit que
$\forall i$, $\mu_{ij} - \mu_{ij'} = (\mu + \alpha_i + \beta_j) - (\mu + \alpha_i + \beta_{j'}) = \beta_j - \beta_{j'}$
$\forall j$, $\mu_{ij} - \mu_{i'j} = (\mu + \alpha_i + \beta_j) - (\mu + \alpha_{i'} + \beta_j) = \alpha_i - \alpha_{i'}$ les effets des facteurs se combinent en s'additionnant simplement . . . *l'effet de A est le même quelque soit le niveau de B considéré (et reciproquement).*

Peut on visualiser graphiquement les effets potentiels considérés ?

ANOVA à deux facteurs croisés

	B1	B2
A1	12 ; 14	18 ; 20
A2	19 ; 23	21 ; 33

graphe représentant les variations des moyennes ?

ANOVA à deux facteurs croisés

Et pour :

	B1	B2
A1	17 ; 19	13 ; 15
A2	15 ; 17	21 ; 23

??

ANOVA à deux facteurs croisés

Et pour :

	B1	B2
A1	17 ; 19	13 ; 15
A2	15 ; 17	21 ; 23

??

On voit que l'effet de A varie selon le niveau de B considéré !

$\hookrightarrow$ dans un tel cas, **tester directement un effet de A ou de B (on parle alors d'effet principal) n'est pas possible . . .**

ANOVA à deux facteurs croisés

Modèle avec interaction :
M4 $y_{ijk} = \mu_{ij} + e_{ijk}$; $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$

Ici, la différence entre μ_{ij} et $\mu_{ij'}$ va dépendre de la modalité i de A :

$$\mu_{ij} - \mu_{ij'} = (\mu + \alpha_i + \beta_j + \gamma_{ij}) - (\mu + \alpha_i + \beta_{j'} + \gamma_{ij'}) = \beta_j - \beta_{j'} + \gamma_{ij} - \gamma_{ij'}$$

$\rightsquigarrow$ dépend du niveau i de A : $\gamma_{ij} - \gamma_{ij'}$ module la différence $\beta_j - \beta_{j'}$ caractéristique de l'effet de A !

On ne peut conclure à un effet de B qu'à niveau fixé de A.

Visualisation ?

Les estimations des γ_{ij} supposent aussi des contraintes pour assurer l'identifiabilité.

ANOVA à deux facteurs croisés

Remarques :

• quand on écrit $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, si l'on choisit les contraintes $\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$ on a

$$y_{ij\cdot} = y_{\cdots} + (y_{i\cdot\cdot} - y_{\cdots}) + (y_{\cdot j\cdot} - y_{\cdots}) + \hat{\gamma}_{ij}$$

d'où

$$\hat{\gamma}_{ij} = y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + y_{\cdots}$$

On a donc
$\sum_{j=1}^{J} \hat{\gamma}_{ij} = \sum_{j=1}^{J} y_{ij\cdot} - Jy_{i\cdot\cdot} - \sum_{j=1}^{J} y_{\cdot j\cdot} + Jy_{\cdots} = Jy_{i\cdot\cdot} - Jy_{i\cdot\cdot} - Jy_{\cdots} + Jy_{\cdots} = 0$;
idem pour $\sum_{i=1}^{I} \hat{\gamma}_{ij}$

$\rightsquigarrow$ contraintes sur les $\hat{\gamma}_{ij}$!

• Combien de γ_{ij} indépendants ? du modèle estimable : IJ paramètres, et de $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ on déduit : $IJ - I - J + 1 = (I-1)(J-1)$

• pour pouvoir estimer les γ_{ij} il faut des répétitions des traitements !
si pas de répétitions,

$$\hat{\gamma}_{ij} = y_{ij} - y_{i\cdot} - y_{\cdot j} + y_{\cdot\cdot} = y_{ij} - [y_{\cdot\cdot} + (y_{i\cdot} - y_{\cdot\cdot}) + (y_{\cdot j} - y_{\cdot\cdot})] \equiv \hat{e}_{ij}$$

où (e_{ij}) résidus du modèle $\mu + \alpha_i + \beta_j$.
L'interaction γ_{ij} ne peut alors être dissociée du résidu du modèle . . . *donc inestimable !*

ANOVA à deux facteurs croisés : estimation des paramètres

Le vecteur des paramètres θ est estimé via l'inverse généralisée de la matrice ${}^tXX$; on a montré que

$$\boxed{\hat{\theta} = ({}^tXX)^- \, {}^tX\mathbf{Y}}$$

Exercice :

1. Ecrire la matrice X correspondant à une anova de deux facteurs croisés à 2 niveaux et 2 répétitions par niveau ;
2. Vérifier que l'estimation nécessite des contraintes ; combien ?

$$X = \begin{pmatrix} 1&1&0&1&0&1&0&0&0 \\ 1&1&0&1&0&1&0&0&0 \\ 1&1&0&0&1&0&1&0&0 \\ 1&1&0&0&1&0&1&0&0 \\ 1&0&1&1&0&0&0&1&0 \\ 1&0&1&1&0&0&0&1&0 \\ 1&0&1&0&1&0&0&0&1 \\ 1&0&1&0&1&0&0&0&1 \end{pmatrix}$$

$$X'X = \begin{pmatrix} 8&4&4&4&4&2&2&2&2 \\ 4&4&0&2&2&2&2&0&0 \\ 4&0&4&2&2&0&0&2&2 \\ 4&2&2&4&0&2&0&2&0 \\ 4&2&2&0&4&0&2&0&2 \\ 2&2&0&2&0&2&0&0&0 \\ 2&0&2&2&0&0&0&2&0 \\ 2&0&2&0&2&0&0&0&2 \end{pmatrix}$$

$$X'X\theta = X'X \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{pmatrix}$$

$$\begin{array}{ll} (1) & 8\mu + 4\alpha_1 + 4\alpha_2 + 4\beta_1 + 4\beta_2 + 2\gamma_{11} + 2\gamma_{12} + 2\gamma_{21} + 2\gamma_{22} \\ (2) & 4\mu + 4\alpha_1 + 2\beta_1 + 2\beta_2 + 2\gamma_{11} + 2\gamma_{12} \\ (3) & 4\mu + 4\alpha_2 + 2\beta_1 + 2\beta_2 + 2\gamma_{21} + 2\gamma_{22} \\ (4) & 4\mu + 2\alpha_1 + 2\alpha_2 + 4\beta_1 + 2\gamma_{11} + 2\gamma_{21} \\ (5) & 4\mu + 2\alpha_1 + 2\alpha_2 + 4\beta_2 + 2\gamma_{12} + 2\gamma_{22} \\ (6) & 2\mu + 2\alpha_1 + 2\beta_1 + 2\gamma_{11} \\ (7) & 2\mu + 2\alpha_2 + 2\beta_1 + 2\gamma_{12} \\ (8) & 2\mu + 2\alpha_2 + 2\beta_2 + 2\gamma_{22} \end{array}$$

(1) = (2) + (3)

(1) = (4) + (5)

(3) = (7) + (8)

(4) = (6) + (7)

(5) donc

(1)
(2)
(3)
(4)
(5)
(6)
(7)

4 équations pour 9 paramètres
donc on doit rajouter 5
equations (contraints)

ANOVA à deux facteurs croisés : tests des effets des facteurs

Les tests des effets de facteur se font d'une manière générale par comparaisons de deux modèles (modèle ayant plus de paramètres est-il significativement meilleur que le modèle en présentant moins ?).

On peut envisager au moins deux appproches :

• une approche pas à pas où l'on regarde *de proche en proche* la significativité des facteurs considérés en comparant les modèles associés (approche de type *SSI*). On est alors amené à considérer **des comparaisons deux à deux de modèles qui ne diffèrent que par la présence ou non d'un facteur** (qui fait donc l'objet du test) ;

• une approche où l'on **compare les modèles d'intérêt à un même modèle** qui est alors le modèle "maximal" au sens où il fait intervenir tous les facteurs (et éventuellement leurs interactions) (approche type *SSIII*)

On verra que dans un cadre particulier d'expérience (*expérience équilibrée, ou encore orthogonales*) ces deux approches sont strictement équivalentes.

Comparaisons de modèles emboités

L'approche vue pour comparer les modèles M_0 et M_1 de l'ANOVA à un facteur se généralise à la comparaison de deux modèles *dits emboités*.

Définition 1 : Soient M et M' deux modèles ayant respectivement k et $k+u$ paramètres. On dit que M est emboité dans M' si lorsque l'on égalise à 0 certains paramètres de M' on retrouve M.

Rmq : on donnera une seconde définition plus générale dans la suite du cours.

Exemple de modèles emboités ?

Proposition :
Sous les postulats du modèle linéaire, pour comparer les modèles emboités M et M', on peut considérer la statistique :

$$F = \frac{\frac{SCR_M - SCR_{M'}}{ddl_{SCR_M} - ddl_{SCR_{M'}}}}{\frac{SCR_{M'}}{ddl_{SCR_{M'}}}} \sim Fish(ddl_{SCR_M} - ddl_{SCR_{M'}}, ddl_{SCR_{M'}})$$

ANOVA à deux facteurs croisés : tests des effets des facteurs

Anova deux facteurs avec interaction, les différents modèles :

M_0 $\mathbb{E}(y_{ijk}) = \mu$

M_1 $\mathbb{E}(y_{ijk}) = \mu + \alpha_i$

M_1' $\mathbb{E}(y_{ijk}) = \mu + \beta_j$

M_2 $\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j$

M_3 $\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$

Quels modèles emboités ?

Quels tests ?

M_3/M_0 revient à tester

$$H_0 : \alpha_1 = \ldots = \alpha_I = \beta_1 = \ldots = \beta_J = \gamma_{11} = \ldots = \gamma_{IJ} = 0$$

contre

$$H_1 : \text{"au moins un non nul"}.$$

Statistique de test ?

ANOVA à deux facteurs croisés : tests des effets des facteurs

$$F = \frac{\frac{SCR_{M_0} - SCR_{M_3}}{ddl_{SCR_{M_0}} - ddl_{SCR_{M_3}}}}{\frac{SCR_{M_3}}{ddl_{SCR_{M_3}}}}$$

ddl SCR_{M_0} : N-1 ;
ddl SCR_{M_3} : N-IJ ; en effet, on a IJ paramètres irréductibles (écriture μ_{ij}) ; on peut aussi compter les paramètres irréductibles dans l'écriture de M_3 :
$1 + (I-1) + (J-1) + x$ où x ddl des γ_{ij}. On a vu que ddl des γ_{ij}=$(I-1)(J-1)$ (on peut obtenir x par différence :
$x = IJ - 1 - (I-1) - (J-1) = (I-1)(J-1))$ et on en déduit que le nombre de paramètres irréductibles dans l'écriture de M_3 est
$1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$

Rmq : on peut aussi noter que les contraintes $\sum_i \gamma_{ij} = 0$ et $\sum_j \gamma_{ij} = 0$ impliquent $I + J - 1$ dépendances et donc le ddl des γ_{ij} est de
$IJ - (I + J - 1) = (I-1)(J-1)$.

$\hookrightarrow$ loi de la statistique de test F : $Fish(IJ-1, N-IJ)$

ANOVA à deux facteurs croisés : tests des effets des facteurs

Tester les effets des facteurs ?

On utilise les statistiques de comparaisons de modèles correspondantes !

ddl mis en cause ?

SCR_{M_0} : $N-1$;
SCR_{M_1} : $N-I$;
$SCR_{M_{1'}}$: $N-J$;
SCR_{M_2} : $N-(I+J-1)$;
SCR_{M_3} : $N-IJ$;

Rmq : à chaque fois on retrouve $N-$(nombre de paramètres irréductibles) !

Ex : tester l'effet du facteur ligne (i.e. H_0 $\alpha_1 = \ldots = \alpha_I = 0$ contre H_1)

$$F = \frac{\frac{SCR_{M_0} - SCR_{M_1}}{I-1}}{\frac{SCR_{M_3}}{N-IJ}}$$

ou bien

$$F = \frac{\frac{SCR_{M_{1'}} - SCR_{M_2}}{I-1}}{\frac{SCR_{M_3}}{N-IJ}}$$

ANOVA à deux facteurs croisés : tests des effets des facteurs

Rmq :
dans les expressions précédentes, par exemple la seconde, on aurait du prendre

$$F = \frac{\frac{SCR_{M_{1'}} - SCR_{M_2}}{I-1}}{\frac{SCR_{M_2}}{N-(I+J-1)}}$$

car comparaison $M_{1'}$ et M_2 !

Mais l'idée est que l'on va faire des comparaisons de modèle conditionnellement à un modèle "le plus large" fixé (par exemple M_3) ... du coup, on cherche à utiliser l'estimation de σ^2 qui a priori est la plus "fine" et dans ce cas c'est celle sous le modèle le plus large qui est considérée, c'est-à-dire ici $\frac{SCR_{M_3}}{N-IJ}$ ...

Ex :
donner les statistiques de tests pour les effets

- colonne
- d'interaction ; que remarquez-vous dans ce dernier cas ?

ANOVA à deux facteurs croisés : tests des effets des facteurs

Pour tester les effets des facteurs ligne et colonne deux approches sont donc possibles et se distinguent par les comparaisons de modèles mis en jeu.

Question : les deux approches donnent - elles le même résultat ? . . . en d'autres termes,

la décomposition de la somme des carrés totale corrigée (SCR_{M_0}) est-elle toujours unique ?

D'une façon générale, la réponse est non !

Il existe des cas où les approches donnent le même résultat (expérience équilibrée ou orthogonale) . . .

Une façon de visualiser ce problème est de considérer les projections sur le plan $\mathcal{P}$ engendré par les colonnes de la matrice design X et des différents sous-espaces de $\mathcal{P}$ mis en jeu par les colonnes de X relatives aux facteurs et à l'interaction des facteurs . . .

$\rightsquigarrow$ *illustration géométrique du modèle linéaire en général et des tests relatifs à l'ANOVA à deux facteurs avec interaction en particulier.*

ANOVA à deux facteurs croisés : tests des effets des facteurs

ANOVA à deux facteurs croisés : décomposition de la SCT_c

Les différences entre les sommes de carrés résiduelles des modèles emboités permettent de mettre en évidence les sommes de carrés imputables au facteur considéré (du fait de l'égalité fondamentale de l'anova).

Montrer que
$\sum_{ijk}(y_{ijk} - y_{...})^2 = \sum_{ijk}(y_{ijk} - y_{ij.})^2 + \sum_i n_{i+}(y_{i..} - y_{...})^2 + \sum_j n_{+j}(y_{.j.} - y_{...})^2 +$
$\sum_{ij}(y_{ij.} - y_{i:} - y_{.j.} + y_{...})^2 - 2\sum_{ij} n_{ij}(y_{i..} - y_{...})(y_{.j.} - y_{...})$

Si l'on suppose des contraintes telles que $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_{i,j} \gamma_{ij} = 0$, la décomposition précédente correspond à :

$$SCR_{M_0} = SCR_{M_3} + SCF_L + SCF_C + SCF_{L*C} - 2\sum_{ij} n_{ij}(y_{i..} - y_{...})(y_{.j.} - y_{...})$$

On voit que la somme des carrés totale corrigée se décompose uniquement en les facteurs si

1. n_{ij}=Cte **cas équi-répété** ;
2. $n_{ij} = \frac{n_{i+}n_{+j}}{n}$ **cas équilibré** ;

Analyse de la variance : modèle

Expressions explicites des estimations sous contraintes "sommes".

Le modèle :

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Contraintes :

$$\sum \alpha_i = \sum \beta_j = 0$$

$$\forall i, j \qquad \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

Estimations :

$$\begin{aligned}
\widehat{\mu} &= y_{...} \\
\widehat{\alpha}_i &= y_{i..} - y_{...} \\
\widehat{\beta}_j &= y_{.j.} - y_{...} \\
\widehat{\gamma}_{ij} &= y_{ij.} - y_{.j.} - y_{i..} + y_{...}
\end{aligned}$$

ANOVA à deux facteurs croisés : décomposition de la SCT_c

Table d'analyse de variance du modèle avec interaction :

Source de variation	Degrés de liberté	Somme des carrés	Carrés moyens	F
Facteur L	I - 1	SCF_A	$V_L = \frac{SCF_A}{I-1}$	$\frac{V_L}{V_R}$
Facteur C	J - 1	SCF_B	$V_C = \frac{SCF_B}{J-1}$	$\frac{V_C}{V_R}$
Interaction	(I - 1) (J - 1)	SCF_{L*C}	$V_{L*C}\frac{SCF_{L*C}}{(I-1)(J-1)}$	$\frac{V_{L*C}}{V_R}$
Résiduelle	N - IJ	SCR_{M_3}	$V_R = \frac{SCR_{M_3}}{N-IJ}$	
Totale (cor)	N - 1	SCR_{M_0}		

ce qui permet de conclure quant à la significativité des effets via les p-values des tests F.

Rappel : si l'interaction est significative, les tests des effets principaux des facteurs ne peuvent être interprétés !
$\rightsquigarrow$ *étudier l'effet de l'un des facteurs à des niveaux fixés de l'autre !*

Analyse de la variance : réduction

Si les expériences ne sont pas équirépétées ou équilibrées (données manquantes, dispositif expérimental trop lourd ...)

Il n'y a plus additivité des sommes de carrés

Réduction $R(c/\mu, a, b)$: diminution de la somme de carrés résiduelle lorsque l'on passe du modèle comportant les effets a et b au modèle comportant a,b,c.

Sommes de type I, II, III

	Type I	Type II	Type III
facteur 1 α	$R(\alpha/\mu)$	$R(\alpha/\mu, \beta)$	$R(\alpha/\mu, \beta, \gamma)$
facteur 2 β	$R(\beta/\mu, \alpha)$	$R(\beta/\mu, \alpha)$	$R(\beta/\mu, \alpha, \gamma)$
interaction γ	$R(\gamma/\mu, \alpha, \beta)$	$R(\gamma/\mu, \alpha, \beta)$	$R(\gamma/\mu, \alpha, \beta)$

Remarque : Par défaut, *R* travaille avec des *SS* de type I.

Analyse de la variance : moyennes ajustées

Dans le cas non équirépété les moyennes des effets ne sont pas comparables parce que calculées sur des bases différentes.

Moyennes ajustées *(lsmeans)* :

$$\widetilde{\mu}_i = \frac{1}{J}\sum_j E(Y_{ijk}) = \mu + \alpha_i + \frac{1}{J}\sum_j \beta_j + \frac{1}{J}\sum_j \gamma_{ij}$$

$$\widehat{\widetilde{\mu}}_i = \widehat{\mu} + \widehat{\alpha}_i + \frac{1}{J}\sum_j \widehat{\beta}_j + \frac{1}{J}\sum_j \widehat{\gamma}_{ij}$$