Décision et Prévision Statistiques

Thierry Verdel (2007)

Chapitre 8

L'expérimentation statistique

*Imaginons le cas suivant : un fabricant d'ampoules électriques ayant le choix entre 4 types de filaments se propose d'étudier l'influence de la nature du filament sur la durée de vie des ampoules fabriquées. Pour ce faire, il va faire fabriquer 4 échantillons d'ampoules identiques, sauf en ce qui concerne le filament, faire brûler les ampoules jusqu'à extinction, puis comparer les résultats obtenus. La technique statistique permettant cette comparaison est appelée l'**analyse de la variance**. Elle se présente comme une technique d'analyse de l'influence d'une variable qualitative appelée **facteur** (ici, le facteur « filament ») sur une variable quantitative (ici, la durée de vie des lampes). L'objectif du chapitre est de présenter cette technique dans le cas de l'influence d' un facteur, puis de plusieurs facteurs.*

1. Analyse de la variance à un facteur

1.1. Recherche de l'influence d'un facteur

Nous noterons A le facteur et appellerons A_1, ... , A_j, ... A_p ses p modalités. Le problème est l'étude de l'influence du facteur A sur la variable quantitative Y. L'expérimentation disponible a consisté à réaliser, pour chaque modalité A_j du facteur, un certain nombre n_j de mesures de la variable Y étudiée : $y_{1\,j}$, ..., $y_{i\,j}$, ..., $y_{n_{i\,j}}$.

A_1	...	A_j	...	A_p
y_{11}		$y_{1\,j}$		$y_{1\,p}$
$\vdots$		$\vdots$		$\vdots$
$\vdots$		y_{ij}		$\vdots$
$y_{n_1\,1}$		$\vdots$		$\vdots$
		$\vdots$		$y_{n_p\,p}$
		$y_{n_j\,j}$		
$\overline{y_1}$	...	$\overline{y_j}$	...	$\overline{y_p}$

1.2. La relation d'analyse de la variance

Appelons $\overline{y_1}$, ..., $\overline{y_j}$, ..., $\overline{y_p}$ les moyennes des colonnes $\overline{y_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i\,j}$, et appelons $\overline{y}$ la moyenne générale des mesures $\overline{y} = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} y_{i\,j}$ avec $n = \sum_{j=1}^{p} n_j$.

Effectuons alors la décomposition :

$$(y_{i\,j} - \overline{y}) = (\overline{y_j} - \overline{y}) + (y_{i\,j} - \overline{y_j}).$$

En élevant au carré et en sommant, le double produit est nul. En effet :

$$2 \sum_{j=1}^{p} \sum_{i=1}^{n_j} (\overline{y_j} - \overline{y}) + (y_{i\,j} - \overline{y_j}) = 2 \sum_{j=1}^{p} (\overline{y_j} - \overline{y}) \sum_{i=1}^{n_j} (y_{i\,j} - \overline{y_j}) = 0$$

par définition des moyennes $\overline{y_j}$.

On obtient donc :

$$\sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{i\,j} - \overline{y})^2 = \sum_{j=1}^{p} n_j (\overline{y_j} - \overline{y})^2 + \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{i\,j} - \overline{y_j})^2$$

relation appelée d'analyse de la variance, qui décompose la somme des carrés totale :

$$\text{SCT} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{i\,j} - \overline{y})^2$$

en une somme des carrés mesurant la variabilité *intercolonnes* (c'est-à-dire l'influence du facteur) :

$$\text{SCA} = \sum_{j=1}^{p} n_j (\overline{y_j} - \overline{y})^2$$

et une somme des carrés mesurant la variabilité *intracolonnes* (somme des carrés résiduelle) :

$\text{SCR} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{i\,j} - \overline{y_j})^2.$

Notons la grande généralité de cette relation puisqu'elle a été établie sans faire aucune hypothèse sur les données. Cependant, la structure de la relation de base : $(y_{i\,j} - \overline{y}) = (\overline{y_j} - \overline{y}) + (y_{i\,j} - \overline{y_j})$ revient à admettre implicitement l'*additivité* de l'influence du facteur $(\overline{y_j} - \overline{y})$ et d'un résidu $(y_{i\,j} - \overline{y_j})$.

1.3. Le modèle

Pour permettre l'inférence statistique, il est nécessaire de poser un certain nombre d'hypothèses. Le modèle de base de l'analyse de la variance s'écrit :

$y_{i\,j} = \mu + \alpha_j + \varepsilon_{i\,j}.$

Les α_j sont des quantités inconnues, mais certaines, qui mesurent l'influence du facteur A. Pour lever leur indétermination à une constante près, on a l'habitude de poser :

$\sum_{j=1}^{p} n_j\, \alpha_j = 0.$

Les $\varepsilon_{i\,j}$ représentent les fluctuations aléatoires correspondant aux erreurs de mesure ou à l'influence des facteurs non contôlés. Nous poserons qu'il n'y a pas d'erreur systématique, ou qu'elle est contenue dans μ, donc que $E(\varepsilon_{i\,j}) = 0$.

Les hypothèses suivantes stipulent que les $\varepsilon_{i\,j}$:

- sont indépendants : $\sigma(\varepsilon_{i\,j},\ \varepsilon_{i'\,j'}) = 0$ pour $(i,\ j) \neq (i',\ j')$,

- ont même variance : $\sigma^2(\varepsilon_{i\,j}) = \sigma^2$,

- suivent des lois normales.

La plus restrictive, parmi ces hypothèses, est certainement la seconde d'après laquelle l'erreur sur la variable Y est indépendante de la valeur prise par Y, c'est-à-dire notamment, qu'elle n'est pas de type multiplicatif. Pour vérifier si elle est légitime, on dispose de plusieurs tests dont le plus connu est celui de *Bartlett*.

1.4. Test d'analyse de la variance

Il s'agit de tester si l'influence du facteur A sur la variable Y, est significative. Si le facteur A n'a pas d'influence, c'est que : $\alpha_1 = \ldots = \alpha_j = \ldots = \alpha_p = 0$. Et, en faisant l'hypothèse qu'il en est ainsi, on peut montrer que $\frac{\text{SCA}}{\sigma^2}$ suit une loi du χ^2 à $(p - 1)$ degrés de liberté. Comme, d'autre part, la quantité $\frac{\text{SCR}}{\sigma^2}$ suit une loi du χ^2 à $(n - p)$ degrés de liberté, il en résulte que la quantité :

$$f = \frac{\frac{\text{SCA}}{p-1}}{\frac{\text{SCR}}{n-p}}$$

suit une loi de Snedecor à $(p - 1)$ et $(n - p)$ degrés de liberté.

Si la valeur f calculée est supérieure au seuil f_α lu dans la table de Snedecor, on concluera à l'influence du facteur A. Si elle est inférieure, l'information disponible ne permet pas de conclure à une influence du facteur A. Il importera d'effectuer un *test à droite*. En effet, les faibles valeurs de f correspondent à des différences faibles entre les moyennes $\overline{y_j}$ des colonnes, alors que le test vise à mettre en évidence des différences fortes.

1.5. Calcul pratique

On calcule :

$$\text{SCT} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{i\,j} - \overline{y})^2 = \sum_{j=1}^{p} \sum_{i=1}^{n_j} {y_{i\,j}}^2 - n\,\overline{y}^2$$

$$\text{SCA} = \sum_{j=1}^{p} n_j(\overline{y_j} - \overline{y})^2 = \sum_{j=1}^{p} n_j\,\overline{y_j}^2 - n\,\overline{y}^2$$

et enfin par différence :

$$\text{SCR} = \text{SCT} - \text{SCA}.$$

On constitue alors le tableau suivant :

Variation	SC	Degrés de liberté	f calculé	F Snédécor
Facteur	SCA	$p-1$	$\frac{\text{SCA}/(p-1)}{\text{SCR}/(n-p)}$	F_α
Résiduelle	SCR	$n-p$		
Totale	SCT	$n-1$		

1.6. Test de linéarité d'une régression

Ce test concerne les problèmes de régression qui ont fait l'objet du chapitre 7, mais il est plus facile de le présenter si les résultats de l'analyse de la variance sont connus.

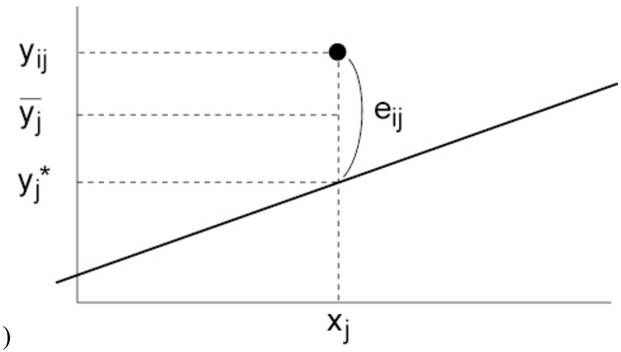
Nous avons supposé dans ce chapitre que la ligne de régression $E[Y(x)] = f(x)$ était une droite.

Si l'expérimentation a été menée de telle sorte que, pour chaque valeur de la variable explicative X, on dispose de q mesures de la variable expliquée Y, il est possible de tester la linéarité de la régression.

En fait, il n'est pas nécessaire qu'il y ait le même nombre q de mesures pour chaque valeur de X, mais seulement qu'il y en ait plusieurs. Nous nous placerons toutefois ici dans ce cas particulier. On dispose donc du tableau des observations ci-dessous, qui a la même structure que celui d'une analyse de la variance.

x_1	...	x_j	...	x_p
y_{11}		$y_{1\,j}$		$y_{1\,\text{p}}$
$\vdots$		$\vdots$		$\vdots$
$\vdots$		$y_{i\,j}$		$\vdots$
$\vdots$		$\vdots$		$\vdots$
y_{q1}		y_{qj}		y_{qp}
$\overline{y_1}$	...	$\overline{y_j}$	...	$\overline{y_p}$

Le principe du test de linéarité consiste à s'assurer que les moyennes $\overline{y_j}$ ne sont pas « trop éloignées » de la droite de régression.



(8.1)

Le déroulement en est le suivant. Soit :

$$y_j^* = a x_j + b$$

le point d'abcisse x_j de la droite des moindres carrés et soit :

$$e_{ij} = (y_{ij} - y_j^*)$$

le résidu correspondant à l'observation y_{ij}. On peut décomposer e_{ij} en :

$$e_{ij} = (\overline{y_j} - y_j^*) + (y_{ij} - \overline{y_j}).$$

En élevant au carré et en sommant sur i et j, le double produit est nul et on obtient :

$$\sum_{j=1}^{p} \sum_{i=1}^{q} (y_{ij} - y_j^*)^2 = q \sum_{j=1}^{p} (\overline{y_j} - y_j^*)^2 + \sum_{j=1}^{p} \sum_{i=1}^{q} (y_{ij} - \overline{y_j})^2$$

Soit ${S_1}^2$ le premier terme de la décomposition :

$${S_1}^2 = q \sum_{j=1}^{p} (\overline{y_j} - y_j^*)^2$$

qui est appelé le *défaut d'ajustement*.

Et soit ${S_2}^2$ le second terme :

$${S_2}^2 = \sum_{j=1}^{p} \sum_{i=1}^{q} (y_{ij} - \overline{y_j})^2$$

Si on a bien affaire à une *régression linéaire*, $\frac{{S_1}^2}{\sigma^2}$ suit une loi du χ^2 à $(p - 2)$ degrés de liberté. Comme d'autre part, $\frac{{S_2}^2}{\sigma^2}$ suit une loi du χ^2 à $q(p - 1)$ degrés de liberté, et que ${S_1}^2$ et ${S_2}^2$ sont indépendantes (puisque chaque $\overline{y_j}$ est indépendant de $\sum_{i=1}^{q} (y_{ij} - \overline{y_j})^2$), il en résulte que le quotient :

$$f = \frac{\frac{{S_1}^2}{p-2}}{\frac{{S_2}^2}{q(p-1)}}$$

suit une loi de Snedecor à $(p - 2)$ et $q\,(p - 1)$ degrés de liberté. Cette propriété permet de tester la linéarité. Ici encore, c'est un test à droite qu'il faut faire, puisque ce que l'on veut éventuellement montrer c'est une valeur élevée du défaut d'ajustement.

2. Etude de l'influence de deux facteurs

Imaginons que le fabricant d'ampoules évoqué plus haut, se préoccupe d'étudier l'influence, sur la durée de vie des ampoules, non seulement du type de filament utilisé, mais également de la nature du gaz de remplissage.

Il pourrait évidemment faire, d'une part, une première étude " filament " en utilisant l'analyse de la variance à un facteur, puis procéder, d'autre part, à une étude " gaz " en tous points analogue. Cela fait, il lui resterait à rapprocher les résultats de ces deux études pour se faire une idée de l' influence des deux facteurs étudiés. Mais en procédant de la sorte, il postulera implicitement l' additivité des influences " filament " et " gaz ", ce qui n'est pas acquis.

L'analyse de la variance à deux facteurs va permettre de traiter globalement le problème, et de mettre en évidence, éventuellement, ce qu'il est convenu d'appeler les *interactions* des facteurs étudiés.

2.1. Plan factoriel

Soit, d'une façon générale, A et B les deux facteurs dont on se propose d'étudier l'influence sur une variable quantitative Y. Nous appellerons $A_1, \ldots, A_i, \ldots, A_p$ les p modalités du facteur A, et $B_1, \ldots, B_j, \ldots, B_q$ les q modalités du facteur B. La mise en oeuvre de l'analyse de la variance à deux facteurs nécessite de disposer d'au moins une mesure de Y pour toute combinaison (A_i, B_j) des modalités des facteurs.

Nous admettrons que l'expérimentation a permis de réaliser r répétitions, c'est-à-dire r mesures pour chacune des $p\,q$ combinaisons des modalités des facteurs. Le cas où il n'y a pas de répétitions ($r = 1$) fera l'objet d'un paragraphe particulier.

Les essais sont donc menés de façon à obtenir le tableau de mesures ci-dessous, une des difficultés de l'expérimentation étant d'éviter les mesures manquantes.

	A_1	$\cdots$	A_i	$\cdots$	A_p
B_1	y_{111} $\vdots$ y_{11r}	$\cdots$	y_{i11} $\vdots$ y_{i1r}	$\cdots$	y_{p11} $\vdots$ y_{p1r}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_j	y_{1j1} $\vdots$ y_{1jr}	$\cdots$	y_{ij1} $\vdots$ y_{ijk} $\vdots$ y_{ijr}	$\cdots$	y_{pj1} $\vdots$ y_{pjr}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_q	y_{1q1} $\vdots$ y_{1qr}	$\cdots$	y_{iq1} $\vdots$ y_{iqr}	$\cdots$	y_{pq1} $\vdots$ y_{pqr}

Le plan d'expérience ainsi réalisé est appelé *plan factoriel*. Il est dit *équilibré* parce qu'il y a le même nombre de mesures dans chaque ligne et dans chaque colonne. Il existe d'autres plans d'expérience équilibrés qui évitent le principal inconvénient du plan factoriel, qui est d'être très coûteux du point de vue du nombre de mesures à effectuer.

2.2. Modèle additif et modèle avec interaction

Le modèle le plus général, en admettant l'additivité des erreurs ε_{ijk}, est le suivant :

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}.$$

En explicitant μ_{ij}, un modèle couramment utilisé est le modèle additif :

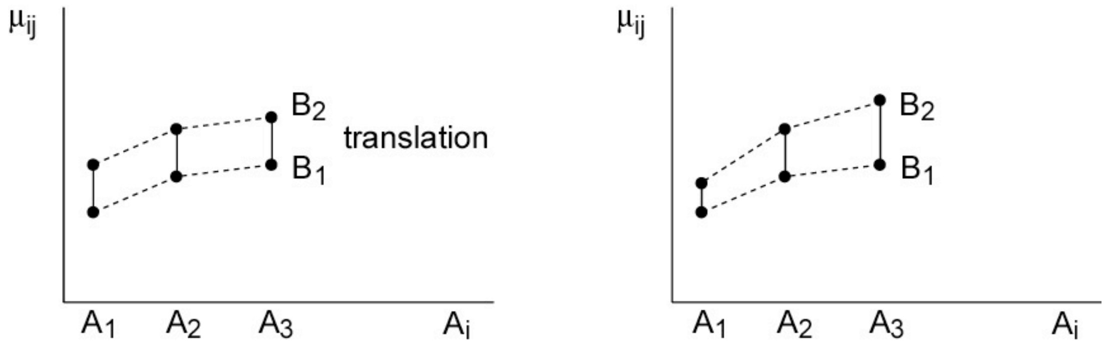
$$\mu_{ij} = \mu + \alpha_i + \beta_j.$$

On suppose ainsi qu'il y a *additivité* des effets : l'action conjuguée des modalités A_i et B_j est la somme des actions isolées de A_i d'une part et de B_j d'autre part.

Si l'on ne suppose pas réalisée cette hypothèse restrictive d'additivité, on adopte le modèle avec *interaction* :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

Il n'y a plus additivité des effets car, aux actions directes de A_i et B_j, s'ajoute le terme γ_{ij} qui traduit un effet supplémentaire dû à la conjonction des modalités A_i et B_j.



$\mu_{ij} = \mu + \alpha_i + \beta_j$ $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$

(8.2)

On dit que $(\alpha_1, ..., \alpha_p)$ et $(\beta_1, ..., \beta_q)$ sont les actions des facteurs A et B, tandis que $(\gamma_{11}, ..., \gamma_{pq})$ sont les interactions du couple (A, B). On peut encore dire que le modèle avec interaction traduit le fait que l'action du facteur A, par exemple, dépend des modalités du facteur B, comme l'illustrent les figures ci-dessus.

Pour lever l'indétermination de μ, on pose les relations suivantes :

$$\sum_{i=1}^{p} \alpha_i = \sum_{j=1}^{q} \beta_j = \sum_{i=1}^{p} \gamma_{ij} = \sum_{j=1}^{q} \gamma_{ij} = 0.$$

2.3. Relation d'analyse de la variance

Appelons $\overline{y_i}$ la moyenne d'une colonne du tableau des mesures : $\overline{y_i} = \frac{1}{qr} \sum_{jk} y_{ijk}$.

Appelons $\overline{y_j}$ la moyenne d'une ligne du tableau : $\overline{y_j} = \frac{1}{pr} \sum_{ik} y_{ijk}$.

Appelons $\overline{y_{ij}}$ la moyenne d'une case du tableau : $\overline{y_{ij}} = \frac{1}{r} \sum_{k} y_{ijk}$.

Appelons enfin $\overline{y}$ la moyenne générale des mesures : $\overline{y} = \frac{1}{pqr} \sum_{ijk} y_{ijk}$.

Effectuons alors la décomposition :

$$(y_{ijk} - \overline{y}) = (\overline{y_i} - \overline{y}) + (\overline{y_j} - \overline{y}) + [(\overline{y_{ij}} - \overline{y}) - (\overline{y_i} - \overline{y}) - (\overline{y_j} - \overline{y})] + (y_{ijk} - \overline{y_{ij}}).$$

En élevant au carré et en sommant, les doubles produits s'annulent par définition des différentes moyennes, à la condition stricte que le tableau soit complet, c'est-à-dire qu'il n'y ait aucune mesure manquante. On obtient par conséquent :

$$\sum_{ijk} (y_{ijk} - \overline{y})^2 =$$
$$q\,r \sum_i (\overline{y_i} - \overline{y})^2 + p\,r \sum_j (\overline{y_j} - \overline{y})^2 + r \sum_{ij} [(\overline{y_{ij}} - \overline{y}) - (\overline{y_i} - \overline{y}) - (\overline{y_j} - \overline{y})]^2 + \sum_{ijk} (y_{ijk} - \overline{y_{ij}})^2$$

que nous noterons symboliquement :

$$\text{SCT} = \text{SCA} + \text{SCB} + \text{SCAB} + \text{SCR}.$$

C'est la relation d'analyse de la variance. Elle permet de décomposer la somme des carrés totale en quatre sommes. Les deux premières correspondent respectivement aux *actions* de A et de B. La troisième correspond à l'*interaction* de A et B. La dernière est la somme des carrés *résiduelle*.

2.4. Les tests d'analyse de la variance

Admettons, comme dans le cas d'un seul facteur, que les $\varepsilon_{i\,j\,k}$ sont des variables aléatoires centrées, de même variance σ^2, indépendantes, et qu'elles suivent des lois normales. Il est alors possible d'effectuer une inférence statistique à partir des observations, et de tester :

- la présence d'une interaction,

- l'influence d'un facteur.

2.4.1. Test de l'interaction

Faisons l'hypothèse qu'il n'y a pas d'interaction des facteurs A et B, c'est-à-dire que :

$$\forall\ i,\ j\ :\ \gamma_{i\,j} = 0.$$

On montre que, s'il en est ainsi, la quantité $\frac{\text{SCAB}}{\sigma^2}$ suit une loi du χ^2 à $(p-1)\,(q-1)$ degrés de liberté.

Comme d'autre part, la quantité $\frac{\text{SCR}}{\sigma^2}$ suit une loi du χ^2 à $(n - p\,q) = p\,q\,(r-1)$ degrés de liberté, il en résulte que le quotient :

$$f_{A\,B} = \frac{\frac{\text{SCAB}}{(p-1)\,(q-1)}}{\frac{\text{SCR}}{p\,q(r-1)}}$$

suit une loi de Snedecor à $(p-1)\,(q-1)$ et $p\,q\,(r-1)$ degrés de liberté, s'il n'y a pas d'interaction.

2.4.2. Test de l'influence d'un facteur

Faisons l'hypothèse que le facteur A, par exemple, n'a pas d'influence sur la variable Y. On montre alors que la quantité $\frac{\text{SCA}}{\sigma^2}$ suit une loi du χ^2 à $(p - 1)$ degrés de liberté.

Par conséquent, la quantité :

$$f_A = \frac{\frac{\text{SCA}}{(p-1)}}{\frac{\text{SCR}}{p\,q(r-1)}}$$

suit une loi de Snedecor à $(p-1)$ et $p\,q\,(r-1)$ degrés de liberté.

2.4.3. Exécution des calculs

On calcule SCA, SCB, SCAB et SCR par les formules suivantes :

$$\text{SCA} = q\,r \sum_i \overline{y_i}^2 - p\,q\,r\,\overline{y}^2,$$

$$\text{SCB} = p\,r \sum_j \overline{y_j}^2 - p\,q\,r\,\overline{y}^2,$$

$$\text{SCAB} = r \sum_{i\,j} \overline{y_{i\,j}}^2 - p\,q\,r\,\overline{y}^2 - \text{SCA} - \text{SCB},$$

$$\text{SCT} = \sum_{i\,j\,k} {y_{i\,j\,k}}^2 - p\,q\,r\,\overline{y}^2.$$

Puis SCR s'obtient par différence :

$$\text{SCR} = \text{SCT} - \text{SCA} - \text{SCB} - \text{SCAB}.$$

On dresse enfin le tableau :

SC	DL	f calculé	F Snédécor
SCA	$p-1$	$f_A = \frac{\text{SCA}/(p-1)}{\text{SCR}/p\,q\,(\text{r}-1)}$	F_A
SCB	$q-1$	$f_B = \frac{\text{SCB}/(q-1)}{\text{SCR}/p\,q\,(\text{r}-1)}$	F_B
SCAB	$(p-1)(q-1)$	$f_{AB} = \frac{\text{SCAB}/(p-1)(q-1)}{\text{SCR}/p\,q\,(\text{r}-1)}$	F_{AB}
SCR	$p\,q\,(r-1)$		
SCT	$p\,q\,r-1$		

2.5. Analyse de la variance sans répétitions

Supposons qu'on n'ait réalisé qu'une seule mesure $y_{i\,j}$ pour chaque couple de modalités (A_i, B_j), conformément au tableau ci-dessous.

	A_1	$\cdots$	A_i	$\cdots$	A_p
B_1	y_{11}	$\cdots$	y_{i1}	$\cdots$	y_{p1}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_j	$y_{1\,j}$	$\cdots$	$y_{i\,j}$	$\cdots$	$y_{p\,j}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_q	$y_{1\,q}$	$\cdots$	$y_{i\,q}$	$\cdots$	$y_{p\,q}$

L'équation d'analyse de la variance s'écrit alors :

$$\sum_{i\,j} (y_{i\,j} - \overline{y})^2 = q\,r \sum_i (\overline{y_i} - \overline{y})^2 + p\,r \sum_j (\overline{y_j} - \overline{y})^2$$
$$+r \sum_{i\,j} [(y_{i\,j} - \overline{y}) - (\overline{y_i} - \overline{y}) - (\overline{y_j} - \overline{y})]^2$$

soit, avec les notations habituelles :

$$\text{SCT} = \text{SCA} + \text{SCB} + \text{SCAB}.$$

Il devient impossible de tester l'interaction, puisqu'on ne dispose plus d'une quantité telle que SCR permettant, par division, d'éliminer σ^2 et d'obtenir une loi de Snedecor. Il est donc nécessaire, dans ce cas de faire l'hypothèse (impossible à vérifier) qu'il n'y a pas d'interaction. On doit donc adopter le modèle additif :

$$y_{i\,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i\,j}\,.$$

Sous cette condition, et quelles que soient les actions des facteurs A et B, on montre, comme dans le cas général, que $\frac{\text{SCAB}}{\sigma^2}$ suit une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté.

Dès lors, pour tester l'influence de A, par exemple, faisons l'hypothèse que les α_i sont tous nuls. Elle entraine que $\frac{\text{SCA}}{\sigma^2}$ suit une loi du χ^2 à $(p-1)$ degrés de liberté et, par conséquent, que la quantité :

$f_A = \frac{\frac{\text{SCA}}{(p-1)}}{\frac{\text{SCAB}}{(p-1)(q-1)}}$ suit une loi de Snedecor à $(p-1)$ et $(p-1)(q-1)$ degrés de liberté.

Exercices du chapitre 8

Exercice 1

Les données suivantes représentent l'effet du temps T (en heures), sur la perte H (en ppm) de l'hydrogène contenu dans des échantillons d'acier à 20 degrés centigrades.

T	H		$\log(T)$
1	7.7	8.5	0
2	7.5	8.1	0.69
6	6.2	6.8	1.79
17	5.7	5.3	2.83
30	4.2	4.6	3.4

a) Le facteur temps a-t-il une influence sur la perte en hydrogène ?

b) Peut-on admettre que la relation entre H et T est de la forme $H = \alpha + \beta \log(T)$?

Avec $y = H$ et $x = \log(T)$, on donne :

$2\,\Sigma_i\, x_i = 17.422$ $\Sigma_i\, x_i^2 = 46.4982$

$\Sigma_{i,j}\, y_{i\,j} = 64.6$ $\Sigma_{i,j}\, y_{i\,j}^2 = 437.46$

$2\,\Sigma_i\, \overline{y_i}^2 = 436.62$ $\Sigma_{i,j}\, x_i\, y_{i\,j} = 95.084$

Exercice 2

Un laboratoire utilise 4 thermomètres de façon interchangeable pour faire des mesures de température. Pour étudier si les résultats diffèrent suivant les thermomètres, ces derniers ont été placés dans un récipient maintenu à température constante. Trois lectures ont été faites avec chaque thermomètres. Les résultats en degrés centigrades ont été les suivants :

Thermo 1	Thermo 2	Thermo 3	Thermo 4
0.9	0.3	−0.6	0
1.2	−0.2	−1.0	0.4
0.8	0.1	−0.7	0.5

Que peut-on en conclure ? On donne :

$\Sigma_{i,j}\, y_{i\,j} = 1.7$ $\Sigma_{i,j}\, y_{i\,j}^2 = 5.29$ $3\,\Sigma_i\, \overline{y_i}^2 = 4.85$

Exercice 3

Le modèle $y_{i\,j} = \mu + \alpha_i + \varepsilon_{i\,j}$, où les α_i sont des quantités inconnues mais déterminées, est appelé modèle à *effets fixes*. Un autre modèle est le celui qui est appelé à *effets aléatoires* et qui s'écrit $y_{i\,j} = \mu + a_i + b_{i\,j}$ où a_i est une réalisation d'une variable aléatoire A avec $E(A) = 0$ et $b_{i\,j}$ est une réalisation d'une variable aléatoire B avec $E(B) = 0$. Les variables A et B sont caractérisées par leurs variances σ_A^2 et σ_B^2 et sont indépendantes. L'exemple suivant est une application de ce dernier modèle.

Un ciment est caractérisé par sa résistance à la compression (mesurée sur des prismes de béton fabriqués à partir d'un échantillon du ciment). Pour étudier la variabilité des productions journalières d'un four à ciment, on a réalisé deux mesures sur des échantillons prélevés chaque jour pendant une période de 10 jours. On a obtenu les résultats ci-après.

Jour	1	2	3	4	5	6	7	8	9	10
Mesure 1	290	294	281	273	271	270	274	266	264	272
Mesure 2	298	273	272	285	318	311	290	315	259	300

$\Sigma_{i,j}\, y_{i\,j} = 5676$ $\qquad \Sigma_{i,j}\, {y_{i\,j}}^2 = 1616672$ $\qquad \Sigma_i\, \overline{y_i}^2 = 806314$

a) Estimer la variance des mesures et la variance des productions.

b) Sont-elles différentes ? Conclure.

c) On donne $\sum_i (y_{i\,1} - y_{i\,2})^2 = 9126$. Rapprocher ce résultat de ceux de l'analyse de la variance.

Exercice 4

Les produits appelés " cru ", à l'entrée d'un four de cimenterie, sont ajustés en tenant compte du titre en carbonates. On veut savoir avec quelle précision est connu ce titre dans le cadre de l'usine. Le laboratoire dispose de cinq chimistes. Il y a deux batteries de dosage.

L'expérimentation a été menée dans le but de mettre en évidence une influence possible de l'opérateur ou de la batterie sur le résultat d'un dosage. Le tableau suivant indique les résultats pour 10 crus, chacun d'eux ayant fait l'objet d'une analyse par chaque opérateur et sur chaque batterie. Dans chaque case du tableau, le premier nombre correspond à la batterie n°1 et le second à la batterie n°2.

Opérateur	Cru									
	1	2	3	4	5	6	7	8	9	10
B	77.25 77.10	79.25 79.45	78.50 78.55	87.50 87.30	82.75 82.90	84.25 84.05	80.80 81.05	77.65 77.70	78.80 78.55	85.90 86.05
C	77.35 77.30	79.60 79.70	78.80 78.75	87.40 87.45	83.05 83.05	84.25 84.20	81.10 81.00	78.00 78.00	78.70 78.75	86.20 86.00
G	77.15 77.40	79.40 79.75	78.60 78.75	87.50 87.55	83.15 83.15	84.45 84.45	81.20 81.20	78.00 78.05	78.60 78.70	86.10 86.45
N	77.05 77.15	79.10 79.40	78.40 78.45	87.10 87.05	82.95 83.15	84.25 84.15	80.90 81.05	77.85 77.90	78.40 79.05	85.90 86.30
P	77.15 77.25	79.40 79.55	78.60 78.45	87.20 87.30	82.95 82.90	84.15 84.05	81.10 81.00	77.95 77.85	78.90 78.75	86.20 86.05

Analyser les résultats. On donne les sommes de carrés correspondants aux actions des facteurs et à leurs interactions doubles et triple :

SCT	1166.5085	SCAB	0.5725
SCA	1164.4534	SCBC	0.1957
SCB	0.7013	SCAC	0.1059
SCC	0.0552	SCABC	0.4245

A : Cru
B : Opérateur
C : Batérie

Exercice 5

On veut connaître l'influence des facteurs de fabrication sur la qualité d'un ciment et, plus particulièrement, l'influence sur la cuisson des trois facteurs : qualité de l'alumine, température de cuisson, temps de palier de cuisson.

Les critères de cuisson sont : la teneur en alumine libre et la teneur en chaux libre des produits à la sortie du four.

Lors des essais effectués, six crus ont été testés (cru industriel témoin, cru AH_3, cru Guilini, cru A6000, cru A8000, cru A9000), à trois températures différentes (1350°, 1400°, 1450°). A chaque température, les temps de palier étaient de 10, 20 et 30 mn. On a obtenu les résultats suivants :

Températures		1350 °		1400 °		1450 °	
Temps	Crus	Al_2O_3	Ca0 libre	$Al_2 0_3$	Ca0 libre	$Al_2 0_3$	Ca0 libre
10 mn	Témoin	11.20	1.10	6.10	0.40	3.55	0.10
	AH_3	10.75	0.75	4.30	0.20	3.65	0.05
	Guilini	8.60	0.25	5.45	0.10	4.50	0.05
	A6000	10.25	1.55	5.45	0.25	1.75	0.05
	A8000	7.40	2.55	3.20	0.85	0.85	0.05
	A9000	4.70	0.95	2.20	0.20	1.70	0.07
20 mn	Témoin	9.50	0.65	5.00	0.20	1.55	0.10
	AH_3	14.10	1.20	3.55	0.20	2.35	0.15
	Guilini	7.60	0.15	5.30	0.10	4.00	0.15
	A6000	9.40	1.30	5.60	0.40	2.45	0.10
	A8000	5.70	1.50	2.05	0.55	0.95	0.07
	A9000	3.20	0.45	1.55	0.10	0.50	0.08
30 mn	Témoin	8.25	0.70	5.15	0.15	2.65	0.00
	AH_3	15.55	1.80	2.20	0.10	2.75	0.08
	Guilini	7.80	0.20	4.65	0.05	3.00	0.06
	A6000	7.75	0.65	4.70	0.30	2.60	0.05
	A8000	3.55	1.05	2.05	0.55	0.80	0.05
	A9000	3.00	0.35	1.60	0.10	0.65	0.08

Analyser ces résultats. On donne les éléments de calcul :

Somme	Alumine	Chaux
SCA	364.96	7.66
SCB	145.55	2.45
SCC	8.00	0.28
SCAB	87.90	2.30
SCBC	6.90	0.80
SCCA	1.00	0.28
SCABC	23.40	1.20

A : Température (3 modalités)
B : Cru (6 modalités)
C : Temps de palier (3 modalités)