Analyse de la variance
ANOVA

Exemple.

21 candidats, 3 examinateurs (resp. 6,8 et 7 étudiants)

Examinateur	**A**	**B**	**C**
Notes	10,11,11 12,13,15	8,11,11,13 14,15,16,16	10,13,14,14 15,16,16
Effectif	6	8	7
Moyenne	12	13	14

Exemple.

21 candidats, 3 examinateurs (resp. 6,8 et 7 étudiants)

Examinateur	**A**	**B**	**C**
Notes	10,11,11 12,13,15	8,11,11,13 14,15,16,16	10,13,14,14 15,16,16
Effectif	6	8	7
Moyenne	12	13	14

"effet d'examinateur"?

Exemple.

21 candidats, 3 examinateurs (resp. 6,8 et 7 étudiants)

Examinateur	**A**	**B**	**C**
Notes	10,11,11 12,13,15	8,11,11,13 14,15,16,16	10,13,14,14 15,16,16
Effectif	6	8	7
Moyenne	12	13	14

"effet d'examinateur"?

ANOVA : pour étudier l'effet des variables qualitatives sur une variable quantitative

- *facteur (variable qualitative)* : prend un nombre fini de valeurs, une valeur = une classe.

- *facteur (variable qualitative)* : prend un nombre fini de valeurs, une valeur = une classe. Exemple : facteur "examinateur"

- *facteur (variable qualitative)* : prend un nombre fini de valeurs, une valeur = une classe. Exemple : facteur "examinateur"
- *niveau (ou population)* : les différentes valeurs prises par un facteur.

- *facteur (variable qualitative)* : prend un nombre fini de valeurs, une valeur = une classe. Exemple : facteur "examinateur"
- *niveau (ou population)* : les différentes valeurs prises par un facteur. Ex : niveaux A, B, C

- *facteur (variable qualitative)* : prend un nombre fini de valeurs, une valeur = une classe. Exemple : facteur "examinateur"
- *niveau (ou population)* : les différentes valeurs prises par un facteur. Ex : niveaux A, B, C
- *test de l'effet d'un facteur* : tester si les moyennes des populations sont égales.

- *facteur (variable qualitative)* : prend un nombre fini de valeurs, une valeur = une classe. Exemple : facteur "examinateur"
- *niveau (ou population)* : les différentes valeurs prises par un facteur. Ex : niveaux A, B, C
- *test de l'effet d'un facteur* : tester si les moyennes des populations sont égales.

La variable étudiée : Y, à valeurs numériques

- *facteur (variable qualitative)* : prend un nombre fini de valeurs, une valeur = une classe. Exemple : facteur "examinateur"
- *niveau (ou population)* : les différentes valeurs prises par un facteur. Ex : niveaux A, B, C
- *test de l'effet d'un facteur* : tester si les moyennes des populations sont égales.

La variable étudiée : Y, à valeurs numériques (note).

1. Analyse de variance à un facteur
   - Introduction
   - Terminologie
   - Données
   - Modèles statistiques
   - Estimation des paramètres
2. Tests d'hypothèses
3. Analyse de variance à deux facteurs

Un seul facteur F

k niveaux

k échantillons de tailles respectives $n_1, ..., n_k$

Un seul facteur F

k niveaux

k échantillons de tailles respectives $n_1, \ldots, n_k$

Effectif total

$$n = \sum_{i=1}^{k} n_i$$

Un seul facteur F

k niveaux

k échantillons de tailles respectives $n_1, ..., n_k$

Effectif total

$$n = \sum_{i=1}^{k} n_i$$

À chaque expérience, on mesure la valeur de la variable Y.

Un seul facteur F

k niveaux

k échantillons de tailles respectives $n_1, ..., n_k$

Effectif total

$$n = \sum_{i=1}^{k} n_i$$

À chaque expérience, on mesure la valeur de la variable Y.

Données

Niveau (population)	**Nb. obs.**	**Valeurs de Y**
1	n_1	$y_{11}, y_{12}, \ldots, y_{1n_1}$
2	n_2	$y_{21}, y_{22}, \ldots, y_{2n_2}$
$\vdots$	$\vdots$	....
k	n_k	$y_{k1}, y_{k2}, ..., y_{kn_k}$

Sommes et moyennes empiriques

Sommes et moyennes empiriques

Pour le niveau i :

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$$

$$\overline{Y_{i\bullet}} = \frac{1}{n_i} Y_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

Sommes et moyennes empiriques

Pour le niveau i :

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$$

$$\overline{Y_{i\bullet}} = \frac{1}{n_i} Y_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

Globalement

$$Y_{\bullet\bullet} = \sum_{i=1}^{k} Y_{i\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

et

$$\overline{Y_{\bullet\bullet}} = \frac{1}{n} Y_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{k} Y_{i\bullet} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

Hypothèse :

les k échantillons sont indépendants et de loi Normale.

Hypothèse :

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$

Hypothèse :

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et Y_{ij}, $Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Hypothèse :

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et Y_{ij}, $Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Autrement dit, pour chaque i, $(y_{ij})_{j \leq n_i}, \ldots, y_{in_i}$ est un échantillon standard.

Hypothèse :

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et Y_{ij}, $Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Autrement dit, pour chaque i, $(y_{ij})_{j \leq n_i}, \ldots, y_{in_i}$ est un échantillon standard.

L'écart-type (théorique) est le même pour tous les niveaux.

Hypothèse :

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et Y_{ij}, $Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Autrement dit, pour chaque i, $(y_{ij})_{j \leq n_i}, \ldots, y_{in_i}$ est un échantillon standard.

L'écart-type (théorique) est le même pour tous les niveaux. La moyenne (théorique) peut varier avec le niveau.

Hypothèse :

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et Y_{ij}, $Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Autrement dit, pour chaque i, $(y_{ij})_{j \leq n_i}, \dots, y_{in_i}$ est un échantillon standard.

L'écart-type (théorique) est le même pour tous les niveaux. La moyenne (théorique) peut varier avec le niveau.

On veut savoir si les moyennes m_i sont toutes égales ou non.

$Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$

$$Y_{ij} = m_i + \varepsilon_{ij} \qquad i = 1, ..., k \qquad j = 1, ..., n_i$$

$Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$

$$\begin{aligned} Y_{ij} &= m_i + \varepsilon_{ij} \qquad i = 1, ..., k \qquad j = 1, ..., n_i \\ &= \mu + \alpha_i + \varepsilon_{ij} \end{aligned}$$

avec $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.
Avec
- $\mu =$ « effet moyen » ;
- $\alpha_i =$ effet du niveau i du facteur F.

$$Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$$

$$\begin{aligned} Y_{ij} &= m_i + \varepsilon_{ij} \qquad i = 1, ..., k \qquad j = 1, ..., n_i \\ &= \mu + \alpha_i + \varepsilon_{ij} \end{aligned}$$

avec $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.
Avec
- $\mu =$ « effet moyen » ;
- $\alpha_i =$ effet du niveau i du facteur F.
Contrainte : $\sum_{i=1}^{k} n_i \alpha_i = 0$

$Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$

$$\begin{aligned} Y_{ij} &= m_i + \varepsilon_{ij} \qquad i = 1, ..., k \qquad j = 1, ..., n_i \\ &= \mu + \alpha_i + \varepsilon_{ij} \end{aligned}$$

avec $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.
Avec
- $\mu =$ « effet moyen » ;
- $\alpha_i =$ effet du niveau i du facteur F.

Contrainte : $\sum_{i=1}^{k} n_i \alpha_i = 0$

On veut tester si les α_i sont tous nuls.

Vectoriellement, le modèle s'écrit

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ \vdots \\ m_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

$$Y = X\beta + \varepsilon$$

Donc, l'analyse de variance est un modèle linéaire.

(m_i)

Il faut trouver ainsi les valeurs des m_i qui minimisent la fonction :

$$T((m_i)) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - m_i)^2$$

(m_i)

Il faut trouver ainsi les valeurs des m_i qui minimisent la fonction :

$$T((m_i)) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - m_i)^2$$

On obtient que : $\hat{m}_i = \bar{Y}_{i\bullet}$.

(m_i)

Il faut trouver ainsi les valeurs des m_i qui minimisent la fonction :

$$T((m_i)) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - m_i)^2$$

On obtient que : $\hat{m}_i = \bar{Y}_{i\bullet}$.
Sous l'hypothèse de normalité et d'indépendance des échantillons, $\bar{Y}_{i\bullet}$ est un estimateur sans biais de m_i et

$$\hat{m}_i = \bar{Y}_{i\bullet} \sim \mathcal{N}\left(m_i, \frac{\sigma^2}{n_i}\right)$$

$\mu, \ \alpha_i$

$$\varepsilon_{ij} = \bar{\varepsilon}_{\bullet\bullet} + (\bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet}) + (\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet})$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{\varepsilon_{\bullet\bullet}^2} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{\varepsilon_{i\bullet}} - \overline{\varepsilon_{\bullet\bullet}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\varepsilon_{ij} - \overline{\varepsilon_{i\bullet}})^2$$

$\mu, \ \alpha_i$

$$\varepsilon_{ij} = \bar{\varepsilon}_{\bullet\bullet} + (\bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet}) + (\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet})$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \overline{\varepsilon_{\bullet\bullet}^2} + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\overline{\varepsilon_{i\bullet}} - \overline{\varepsilon_{\bullet\bullet}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\varepsilon_{ij} - \overline{\varepsilon_{i\bullet}})^2$$

On a

$$\varepsilon_{ij} = Y_{ij} - \mu - \alpha_i, \qquad \varepsilon_{i\bullet} = Y_{i\bullet} - n_i\mu - n_i\alpha_i, \qquad \overline{\varepsilon_{i\bullet}} = \overline{Y_{i\bullet}} - \mu - \alpha_i$$

$$\varepsilon_{\bullet\bullet} = Y_{\bullet\bullet} - \sum_{i=1}^{k} n_i\mu - \sum_{i=1}^{k} n_i\alpha_i, \qquad \varepsilon_{\bullet\bullet} = Y_{\bullet\bullet} - n\mu, \qquad \overline{\varepsilon_{\bullet\bullet}} = \overline{Y_{\bullet\bullet}} - \mu$$

$\mu, \ \alpha_i$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{\bullet\bullet}}-\mu)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i\bullet}}-\overline{Y_{\bullet\bullet}}-\alpha_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{i\bullet}})^2$$

$\mu, \ \alpha_i$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{\bullet\bullet}}-\mu)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i\bullet}}-\overline{Y_{\bullet\bullet}}-\alpha_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{i\bullet}})^2$$

Le membre de droite est minimisé pour :

$$\hat{\mu} = \overline{Y_{\bullet\bullet}}, \qquad \hat{\alpha}_i = \overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}}$$

$\mu, \ \alpha_i$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{\bullet\bullet}}-\mu)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i\bullet}}-\overline{Y_{\bullet\bullet}}-\alpha_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{i\bullet}})^2$$

Le membre de droite est minimisé pour :

$$\hat{\mu} = \overline{Y_{\bullet\bullet}}, \qquad \hat{\alpha}_i = \overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}}$$

On a bien $\sum_{i=1}^{k} n_i\hat{\alpha}_i = 0$

$\mu,\ \alpha_i$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{\bullet\bullet}}-\mu)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i\bullet}}-\overline{Y_{\bullet\bullet}}-\alpha_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{i\bullet}})^2$$

Le membre de droite est minimisé pour :

$$\hat{\mu}=\overline{Y_{\bullet\bullet}},\qquad \hat{\alpha}_i=\overline{Y_{i\bullet}}-\overline{Y_{\bullet\bullet}}$$

On a bien $\sum_{i=1}^{k} n_i\hat{\alpha}_i=0$:

$$\sum_{i=1}^{k} n_i\hat{\alpha}_i=\sum_{i=1}^{k} n_i\overline{Y_{i\bullet}}-\sum_{i=1}^{k} n_i\overline{Y_{\bullet\bullet}}=\sum_{i=1}^{k} Y_{i\bullet}-n\overline{Y_{\bullet\bullet}}=0$$

σ^2

L'estimateur de σ^2 est :

$$S_n^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

1. Analyse de variance à un facteur
2. Tests d'hypothèses
   - Tableau d'analyse de variance
   - Test d'égalité des k effets
   - Comparaison de moyennes
3. Analyse de variance à deux facteurs

Niveaux

1 seul niveau ou k niveaux ?

Niveaux

1 seul niveau ou k niveaux ?

$H_0 : m_1 = m_2 = \ldots = m_k = m$

contre

Niveaux

1 seul niveau ou k niveaux ?

$H_0 : m_1 = m_2 = \ldots = m_k = m$

contre $H_1 : \exists i, j \in \{1, \ldots, k\}$ tels que $m_i \neq m_j$.

Niveaux

1 seul niveau ou k niveaux ?
$H_0 : m_1 = m_2 = \ldots = m_k = m$
contre $H_1 : \exists i, j \in \{1, \ldots, k\}$ tels que $m_i \neq m_j$.

Ou : $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
contre

Niveaux

1 seul niveau ou k niveaux ?
$H_0 : m_1 = m_2 = ... = m_k = m$
contre $H_1 : \exists i, j \in \{1, ..., k\}$ tels que $m_i \neq m_j$.

Ou : $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$
contre $H_1 : \exists i \in \{1, ..., k\}$ tel que $\alpha_i \neq 0$

Niveaux

1 seul niveau ou k niveaux ?
$H_0 : m_1 = m_2 = ... = m_k = m$
contre $H_1 : \exists i, j \in \{1, ..., k\}$ tels que $m_i \neq m_j$.

Ou : $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$
contre $H_1 : \exists i \in \{1, ..., k\}$ tel que $\alpha_i \neq 0$

Sous H_0, le modèle a la forme :

$$M_{reduit} : Y_{ij} = \mu + \varepsilon_{ij}$$

L'estimation pour μ est $\hat{\mu} = \bar{Y}_{\bullet\bullet}$ et la prévision de Y_{ij} est $\hat{Y}_{ij} = \hat{\mu}$.
Alors, le résidu, sous H_0 est :
$Y_{ij} - \bar{Y}_{\bullet\bullet}$.
La variabilité totale est :

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{\bullet\bullet}})^2$$

On peut écrire : $Y_{ij} - \overline{Y_{\bullet\bullet}} = (Y_{ij} - \overline{Y_{i\bullet}}) + (\overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}})$

L'estimation pour μ est $\hat{\mu} = \bar{Y}_{\bullet\bullet}$ et la prévision de Y_{ij} est $\hat{Y}_{ij} = \hat{\mu}$.
Alors, le résidu, sous H_0 est :
$Y_{ij} - \bar{Y}_{\bullet\bullet}$.
La variabilité totale est :

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{\bullet\bullet}})^2$$

On peut écrire : $Y_{ij} - \overline{Y_{\bullet\bullet}} = (Y_{ij} - \overline{Y_{i\bullet}}) + (\overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}})$
et on obtient :

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{\bullet\bullet}})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i\bullet}})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{Y_{i\bullet}} - \overline{Y_{\bullet\bullet}})^2$$

(variabilité totale=variabilité résiduelle + variabilité due au modèle) : ST=SR+SM.

Tableau d'analyse de variance

Source de variation	ddl	S.C.	Carré moyen
Régression	$k-1$	SM	$SM/(k-1)$
Résiduelle	$n-k$	SR	$SR/(n-k)$
Totale	$n-1$	ST	

1. Analyse de variance à un facteur
2. Tests d'hypothèses
   - Tableau d'analyse de variance
   - Test d'égalité des k effets
   - Comparaison de moyennes
3. Analyse de variance à deux facteurs

Pour tester l'hypothèse H_0 on utilise la statistique :

$$Z = \frac{SM/(k-1)}{SR/(n-k)} \sim F(k-1, n-k) \qquad \text{(sous } H_0\text{)}$$

Pour un risque α fixé, la zone d'acceptation est :
$ZA_{H_0,\alpha} = [0\ ;\ f_{k-1,n-k;1-\alpha}]$

Exemple : Notes

21 candidats, 3 examinateurs (resp. 6,8 et 7 étudiants)

Examinateur	**A**	**B**	**C**
Notes	10,11,11 12,13,15	8,11,11,13 14,15,16,16	10,13,14,14 15,16,16
Effectif	6	8	7
Moyenne	12	13	14

$$Y_{ij} = m_i + \varepsilon_{ij}$$
$$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \;\; n_2 = 8 \;\; n_3 = 7$$

$Y_{ij} = m_i + \varepsilon_{ij}$
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

ou encore $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

$Y_{ij} = m_i + \varepsilon_{ij}$
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

ou encore $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

Les estimations des paramètres : $\hat{\mu} = \bar{y}_{\cdot\cdot} = 13.05$,
$\hat{\alpha}_1 = \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} = 12 - 13.05 = -1.05$,
$\hat{\alpha}_2 = \bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot} = 13 - 13.05 = -0.05$,
$\hat{\alpha}_3 = \bar{y}_{3\cdot} - \bar{y}_{\cdot\cdot} = 14 - 13.05 = 0.95$.

$Y_{ij} = m_i + \varepsilon_{ij}$
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

ou encore $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

Les estimations des paramètres : $\hat{\mu} = \bar{y}_{\cdot\cdot} = 13.05$,
$\hat{\alpha}_1 = \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} = 12 - 13.05 = -1.05$,
$\hat{\alpha}_2 = \bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot} = 13 - 13.05 = -0.05$,
$\hat{\alpha}_3 = \bar{y}_{3\cdot} - \bar{y}_{\cdot\cdot} = 14 - 13.05 = 0.95$.

H_0 : pas d'effet examinateur sur la notation

$Y_{ij} = m_i + \varepsilon_{ij}$
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

ou encore $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,
$i = 1, 2, 3, \qquad j = 1, \ldots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7$

Les estimations des paramètres : $\hat{\mu} = \bar{y}_{\mathbf{..}} = 13.05$,
$\hat{\alpha}_1 = \bar{y}_{1\mathbf{.}} - \bar{y}_{\mathbf{..}} = 12 - 13.05 = -1.05$,
$\hat{\alpha}_2 = \bar{y}_{2\mathbf{.}} - \bar{y}_{\mathbf{..}} = 13 - 13.05 = -0.05$,
$\hat{\alpha}_3 = \bar{y}_{3\mathbf{.}} - \bar{y}_{\mathbf{..}} = 14 - 13.05 = 0.95$.

H_0 : pas d'effet examinateur sur la notation
$H_0 : m_1 = m_2 = m_3 = m$ contre $H_1 : \exists i \neq j$ tel que $m_i \neq m_j$
$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ contre $H_1 : \exists i \neq j$ tel que $\alpha_i \neq 0$

On obtient : SM=12.95, SR=98

On obtient : SM=12.95, SR=98 donc
$z = (SM/(3-1))/(SR/(21-3)) = 1.19$.

On obtient : SM=12.95, SR=98 donc
$z = (SM/(3-1))/(SR/(21-3)) = 1.19$.
La zone d'acceptation est $ZA_{H_0;1-\alpha} = [0\,;f_{2,18;0.95}] = [0\,;3.55]$.

On obtient : SM=12.95, SR=98 donc
$z = (SM/(3-1))/(SR/(21-3)) = 1.19$.
La zone d'acceptation est $ZA_{H_0;1-\alpha} = [0\ ; f_{2,18;0.95}] = [0\ ; 3.55]$.
Donc, H_0 est acceptée : les examinateurs ont le même système de notation.

Le rejet de l'hypothèse d'égalité des moyennes ne signifie pas que tous les m_i sont différents entre eux. On cherche souvent à tester l'égalité entre deux moyennes :
$H_0 : m_h = m_j$ contre $H_1 : m_h \neq m_j$ pour $h \neq j$.
On utilise la statistique de test :

$$Z = \frac{\bar{Y}_{h\bullet} - \bar{Y}_{j\bullet}}{\sqrt{\frac{SR}{n-k}}\sqrt{\frac{1}{n_h} + \frac{1}{n_j}}} \sim t_{n-k}$$

(t_{n-k} : loi de Student à $n-k$ degrés de liberté.)

Le rejet de l'hypothèse d'égalité des moyennes ne signifie pas que tous les m_i sont différents entre eux. On cherche souvent à tester l'égalité entre deux moyennes :
$H_0 : m_h = m_j$ contre $H_1 : m_h \neq m_j$ pour $h \neq j$.
On utilise la statistique de test :

$$Z = \frac{\bar{Y}_{h.} - \bar{Y}_{j.}}{\sqrt{\frac{SR}{n-k}}\sqrt{\frac{1}{n_h} + \frac{1}{n_j}}} \sim t_{n-k}$$

(t_{n-k} : loi de Student à $n - k$ degrés de liberté.)
La zone d'acceptation $ZA_{H_0,1-\alpha} = [-t_{n-k;1-\alpha/2} \,;\, t_{n-k;1-\alpha/2}]$.

On a vu comment comparer les populations d'un même facteur. Supposons maintenant qu'un expérimentateur souhaite comparer l'influence de trois régimes alimentaires et de deux exploitations sur la production laitière. Les résultats expérimentaux sont dans le tableau suivant.

Expl $\downarrow$ R.alim $\rightarrow$	**A**	**B**	**C**	**Total**	**Moyenne**
1	7	36	2	45	15
2	13	44	18	75	215
Total	20	80	20	120	
Moyenne	10	40	10		20

1. Analyse de variance à un facteur
2. Tests d'hypothèses
3. Analyse de variance à deux facteurs
   - Introduction
   - Données
   - Modèle sans interaction (additif) : $r = 1$
     - Estimation des paramètres
     - Tableau d'analyse de variance
     - Test d'hypothèse
     - Test d'un facteur.
   - Modèle avec interaction
   - Modèle hiérarchique

Deux facteurs (variables) F1 et F2.
p niveaux pour F1, q niveaux pour F2
Pour chaque couple (i, j) de niveaux, on a $r(\geq 1)$ observations de la variable dépendante Y.

Deux facteurs (variables) F1 et F2.
p niveaux pour F1, q niveaux pour F2
Pour chaque couple (i, j) de niveaux, on a $r(\geq 1)$ observations de la variable dépendante Y.

F1 / F2	**1**	$\dots$	**i**	$\dots$	**p**
1	$y_{111}, \dots, y_{11r}$	$\dots$	$y_{i11}, \dots, y_{i1r}$	$\dots$	$y_{p11}, \dots, y_{p1r}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$y_{1j1}, \dots, y_{1jr}$	$\dots$	$y_{ij1}, \dots, y_{ijr}$	$\dots$	$y_{pj1}, \dots, y_{pjr}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
q	$y_{1q1}, \dots, y_{1jq}$	$\dots$	$y_{iq1}, \dots, y_{iqr}$	$\dots$	$y_{pq1}, \dots, y_{pqr}$

Dans la cellule (i, j) : les valeurs (observations) y_{ijk} :
i : niveau (population) du facteur $F1$,
j : niveau de $F2$
k : la k-ième répétition pour un couple (i, j).

Notations :

$$\begin{cases} y_{ij\bullet} = \sum_{k=1}^{r} y_{ijk} & \bar{y}_{ij\bullet} = \frac{1}{r} y_{ij\bullet} \\ y_{i\bullet\bullet} = \sum_{j=1}^{q} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{i\bullet\bullet} = \frac{1}{qr} y_{\bullet\bullet} \\ y_{\bullet j\bullet} = \sum_{i=1}^{p} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{\bullet j\bullet} = \frac{1}{pr} y_{\bullet j\bullet} \\ y_{\bullet\bullet\bullet} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{\bullet\bullet\bullet} = \frac{1}{pqr} y_{\bullet\bullet\bullet} \end{cases}$$

Notations :

$$\left\{\begin{array}{ll} y_{ij\bullet} = \sum_{k=1}^{r} y_{ijk} & \bar{y}_{ij\bullet} = \frac{1}{r} y_{ij\bullet} \\ y_{i\bullet\bullet} = \sum_{j=1}^{q} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{i\bullet\bullet} = \frac{1}{qr} y_{\bullet\bullet} \\ y_{\bullet j\bullet} = \sum_{i=1}^{p} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{\bullet j\bullet} = \frac{1}{pr} y_{\bullet j\bullet} \\ y_{\bullet\bullet\bullet} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{\bullet\bullet\bullet} = \frac{1}{pqr} y_{\bullet\bullet\bullet} \end{array}\right.$$

Les observations y_{ijk} sont des réalisations de la v.a. Y_{ijk} sur laquelle on fait les hypothèses :

$$\left\{\begin{array}{ll} Y_{ijk} \sim \mathcal{N}(m_{ij}, \sigma^2) & \forall k = 1, \ldots, r \\ Y_{ijk}, Y_{i'j'k'} & \text{indépendantes} \end{array}\right.$$

En ce qui concerne le nombre r de répétitions on a 2 situations :

- $r > 1$
- $r = 1$. Il n'y a pas de répétition et on va noter $Y_{ij\bullet}$ par Y_{ij}.

Alors, les modèles statistiques considérés seront fonction de ces 2 situations.

En ce qui concerne le nombre r de répétitions on a 2 situations :

- $r > 1$
- $r = 1$. Il n'y a pas de répétition et on va noter $Y_{ij\bullet}$ par Y_{ij}.

Alors, les modèles statistiques considérés seront fonction de ces 2 situations.
Même nombre de répétitions de l'expérience pour chaque couple de facteurs.

Les problèmes étudiés sont les mêmes que pour un seul facteur :

- écrire un modèle statistique de Y fonction des facteurs ;

Les problèmes étudiés sont les mêmes que pour un seul facteur :

- écrire un modèle statistique de Y fonction des facteurs ;
- estimer les effets des niveaux des deux facteurs ;

Les problèmes étudiés sont les mêmes que pour un seul facteur :

- écrire un modèle statistique de Y fonction des facteurs ;
- estimer les effets des niveaux des deux facteurs ;
- test d'hypothèse

Le modèle le plus simple est d'additionner les effets du facteur $F1$ avec les effets du facteur $F2$:

$$m_{ij} = \mu + \alpha_i + \beta_j$$

Le modèle le plus simple est d'additionner les effets du facteur $F1$ avec les effets du facteur $F2$:

$$m_{ij} = \mu + \alpha_i + \beta_j$$

où :

- μ est l'effet moyen
- α_i est l'effet dû au niveau i du facteur $F1$;
- β_j est l'effet dû au niveau j du facteur $F2$;

Le modèle le plus simple est d'additionner les effets du facteur $F1$ avec les effets du facteur $F2$:

$$m_{ij} = \mu + \alpha_i + \beta_j$$

où :
- μ est l'effet moyen
- α_i est l'effet dû au niveau i du facteur $F1$;
- β_j est l'effet dû au niveau j du facteur $F2$;

Puisque $Y_{ij} \sim \mathcal{N}(m_{ij}, \sigma^2)$ on peut considérer un modèle :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Avec comme contraintes

$$\sum_{i=1}^{p} \alpha_i = 0 \qquad \sum_{j=1}^{q} \beta_j = 0$$

Il faut trouver les valeurs de m_{ij} (ou de μ, α_i, β_j) qui minimisent la fonction :

$$T((m_{ij})_{ij}) = \sum_{i=1}^{p}\sum_{j=1}^{q} \varepsilon_{ij}^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}(Y_{ij} - m_{ij})^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}(Y_{ij} - \mu - \alpha_i - \beta_j)^2$$

On utilise la même technique que pour l'analyse de variance à un facteur, et on obtient :

$$\hat{\alpha}_i = \overline{Y_{i.}} - \overline{Y_{..}} \qquad \hat{\beta}_i = \overline{Y_{.j}} - \overline{Y_{..}} \qquad \hat{\mu} = \overline{Y_{..}}$$

La valeur prédite pour Y_{ij} est :

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y_{i.}} + \overline{Y_{.j}} - \overline{Y_{..}}$$

Exemple.

F1 : le régime alimentaire, prend 3 valeurs (A, B, C), donc $p = 3$.
F2 : l'exploitation, prend 2 valeurs (1 et 2), donc $q = 2$.

Exemple.

F1 : le régime alimentaire, prend 3 valeurs (A, B, C), donc $p = 3$.
F2 : l'exploitation, prend 2 valeurs (1 et 2), donc $q = 2$.
Modèle statistique :

$$Y_{ij} = \mu + \alpha_i + \beta_j \qquad i = 1, 2, 3 \quad j = 1, 2$$

où : α_i est l'effet de l'exploitation n°i sur Y, β_1 est l'effet du régime A sur la production laitière...

Estimations

$$\hat{\mu} = \bar{y}_{\bullet\bullet} = 20,$$

Estimations

$$\hat{\mu} = \bar{y}_{..} = 20,$$
$$\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..} = 10 - 20 = -10,$$
$$\hat{\alpha}_2 = 20,$$
$$\hat{\alpha}_3 = -10,$$

Estimations

$$
\begin{aligned}
&\hat{\mu} = \bar{y}_{\cdot\cdot} = 20,\\
&\hat{\alpha}_1 = \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} = 10 - 20 = -10,\\
&\hat{\alpha}_2 = 20,\\
&\hat{\alpha}_3 = -10,\\
&\hat{\beta}_1 = \bar{y}_{\cdot 1} - \bar{y}_{\cdot\cdot} = 15 - 20 = -5,\\
&\hat{\beta}_2 = 5.
\end{aligned}
$$

Estimations

$\hat{\mu} = \bar{y}_{..} = 20,$
$\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..} = 10 - 20 = -10,$
$\hat{\alpha}_2 = 20,$
$\hat{\alpha}_3 = -10,$
$\hat{\beta}_1 = \bar{y}_{.1} - \bar{y}_{..} = 15 - 20 = -5,$
$\hat{\beta}_2 = 5.$
La prévision de Y_{11} (exploitation 1 et régime alimentaire A) :
$\hat{Y}_{11} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1 = 20 - 10 - 5 = 5.$

Tableau d'analyse de variance

En partant de l'identité :

$$Y_{ij} - \overline{Y_{..}} = (Y_{ij} - \overline{Y_{i.}} - \overline{Y_{.j}} + \overline{Y_{..}}) + (\overline{Y_{i.}} - \overline{Y_{..}}) + (\overline{Y_{.j}} - \overline{Y_{..}})$$

Tableau d'analyse de variance

En partant de l'identité :

$$Y_{ij} - \overline{Y_{..}} = (Y_{ij} - \overline{Y_{i.}} - \overline{Y_{.j}} + \overline{Y_{..}}) + (\overline{Y_{i.}} - \overline{Y_{..}}) + (\overline{Y_{.j}} - \overline{Y_{..}})$$

On obtient :

$$\sum_{i,j}(Y_{ij} - \overline{Y_{..}})^2 = \sum_{i,j}(Y_{ij} - \overline{Y_{i.}} - \overline{Y_{.j}} + \overline{Y_{..}})^2 + q\sum_{i=1}^{p}(\overline{Y_{i.}} - \overline{Y_{..}})^2 + p\sum_{j=1}^{q}(\overline{Y_{.j}} - \overline{Y_{..}})$$

ou encore $ST = SR + S_{F1} + S_{F2}$.

Tableau d'analyse de variance

En partant de l'identité :

$$Y_{ij} - \overline{Y_{..}} = (Y_{ij} - \overline{Y_{i.}} - \overline{Y_{.j}} + \overline{Y_{..}}) + (\overline{Y_{i.}} - \overline{Y_{..}}) + (\overline{Y_{.j}} - \overline{Y_{..}})$$

On obtient :

$$\sum_{i,j}(Y_{ij}-\overline{Y_{..}})^2 = \sum_{i,j}(Y_{ij}-\overline{Y_{i.}}-\overline{Y_{.j}}+\overline{Y_{..}})^2+q\sum_{i=1}^{p}(\overline{Y_{i.}}-\overline{Y_{..}})^2+p\sum_{j=1}^{q}(\overline{Y_{.j}}-\overline{Y_{..}})$$

ou encore $ST = SR + S_{F1} + S_{F2}$.

Source de variation	ddl	S.C.	Carré moyen
F1	$p-1$	S_{F1}	$S_{F1}/(p-1)$
F2	$q-1$	S_{F2}	$S_{F2}/(q-1)$
Résidu	$(p-1)(q-1)$	SR	$SR/(p-1)(q-1)$
Totale	$pq-1$	ST	

Tests

Deux types d'hypothèse : si le modèle significatif

Tests

Deux types d'hypothèse : si le modèle significatif ou effet de chaque facteur.

Tests

Deux types d'hypothèse : si le modèle significatif ou effet de chaque facteur.

Modèle significatif ?

Le modèle n'est pas significatif si aucun des deux facteurs n'influence Y :

Tests

Deux types d'hypothèse : si le modèle significatif ou effet de chaque facteur.

Modèle significatif ?

Le modèle n'est pas significatif si aucun des deux facteurs n'influence Y :

$H_0 : \alpha_1 = \cdots = \alpha_p = \beta_1 = \cdots = \beta_q = 0$

contre :

$H_1 : \exists i \in \{1, \ldots, p\}$ ou $\exists j \in \{1, \ldots, q\}$ t.q. $\alpha_i \neq 0$ ou $\beta_i \neq 0$.

Tests

Deux types d'hypothèse : si le modèle significatif ou effet de chaque facteur.

Modèle significatif ?

Le modèle n'est pas significatif si aucun des deux facteurs n'influence Y :

$H_0 : \alpha_1 = \cdots = \alpha_p = \beta_1 = \cdots = \beta_q = 0$

contre :

$H_1 : \exists i \in \{1, \ldots, p\}$ ou $\exists j \in \{1, \ldots, q\}$ t.q. $\alpha_i \neq 0$ ou $\beta_i \neq 0$.

Le modèle réduit est : $Y_{ij} = \mu + \varepsilon_{ij}$.

Tests

Deux types d'hypothèse : si le modèle significatif ou effet de chaque facteur.

Modèle significatif ?

Le modèle n'est pas significatif si aucun des deux facteurs n'influence Y :

$H_0 : \alpha_1 = \cdots = \alpha_p = \beta_1 = \cdots = \beta_q = 0$

contre :

$H_1 : \exists i \in \{1, \ldots, p\}$ ou $\exists j \in \{1, \ldots, q\}$ t.q. $\alpha_i \neq 0$ ou $\beta_i \neq 0$.

Le modèle réduit est : $Y_{ij} = \mu + \varepsilon_{ij}$.

Statistique de test :

$$Z = \frac{(S_{F1} + S_{F2})/(p+q-2)}{SR/(p-1)(q-1)} \sim F(p+q-2, (p-1)(q-1)) \qquad \text{sous } H_0$$

Test d'un facteur

Supposons que l'on veut tester l'effet de F1.
H_0 : F1 n'influe pas Y sachant que F2 est dans le modèle.
$H_0 : \alpha_1 = \cdots = \alpha_p = 0$ contre $H_1 : \exists i \in \{1, \ldots, p\}$ t.q. $\alpha_i \neq 0$.
Le modèle réduit est :
$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$.

Test d'un facteur

Supposons que l'on veut tester l'effet de F1.
H_0 : F1 n'influe pas Y sachant que F2 est dans le modèle.
$H_0 : \alpha_1 = \cdots = \alpha_p = 0$ contre $H_1 : \exists i \in \{1, \ldots, p\}$ t.q. $\alpha_i \neq 0$.
Le modèle réduit est :
$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$.(modèle à un facteur)

Test d'un facteur

Supposons que l'on veut tester l'effet de F1.
H_0 : F1 n'influe pas Y sachant que F2 est dans le modèle.
$H_0 : \alpha_1 = \cdots = \alpha_p = 0$ contre $H_1 : \exists i \in \{1, \ldots, p\}$ t.q. $\alpha_i \neq 0$.
Le modèle réduit est :
$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$.(modèle à un facteur)
H_0 : la moyenne m_{ij} ne dépend pas de i.

Test d'un facteur

Supposons que l'on veut tester l'effet de F1.
H_0 : F1 n'influe pas Y sachant que F2 est dans le modèle.
$H_0 : \alpha_1 = \cdots = \alpha_p = 0$ contre $H_1 : \exists i \in \{1, \ldots, p\}$ t.q. $\alpha_i \neq 0$.
Le modèle réduit est :
$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$.(modèle à un facteur)
H_0 : la moyenne m_{ij} ne dépend pas de i.
Statistique de test :

$$Z = \frac{(S_{F1})/(p-1)}{SR/(p-1)(q-1)} \sim F(p-1, (p-1)(q-1)) \qquad \text{sous } H_0$$

Exemple.

Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Exemple.

Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Significativité du modèle : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = 0$:

$$Z =$$

Exemple.

Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Significativité du modèle : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = 0$:

$$Z = \frac{(S_{F1} + S_{F2})/(3+2-2)}{SR/2} \sim F(3,2) \qquad \text{sous } H_0$$

Exemple.

Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Significativité du modèle : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = 0$:

$$Z = \frac{(S_{F1} + S_{F2})/(3+2-2)}{SR/2} \sim F(3,2) \qquad \text{sous } H_0$$

$$ZA = [0\ ;\ f_{3,2:0.95}] = [0\ ;\ 19.2],$$

Exemple.

Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Significativité du modèle : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = 0$:

$$Z = \frac{(S_{F1} + S_{F2})/(3+2-2)}{SR/2} \sim F(3,2) \qquad \text{sous } H_0$$

$$ZA = [0\,;f_{3,2:0.95}] = [0\,;19.2],\ z_{\text{obs}} = \frac{1350/3}{14} = 32.1 \notin ZA.$$

Exemple.

Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Significativité du modèle : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = 0$:

$$Z = \frac{(S_{F1} + S_{F2})/(3+2-2)}{SR/2} \sim F(3,2) \qquad \text{sous } H_0$$

$ZA = [0\ ;f_{3,2:0.95}] = [0\ ;19.2]$, $z_{obs} = \frac{1350/3}{14} = 32.1 \notin ZA$.
Donc H_0 est rejetée et le modèle est significatif.

Régime alimentaire influent ?

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle.

Régime alimentaire influent ?

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle.
$H_1 : \exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$.

Régime alimentaire influent ?

H_0 : $\alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle.
H_1 : $\exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$.
Sous H_0 : $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, 3, \; j = 1, 2.$

Régime alimentaire influent ?

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle.
$H_1 : \exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$.
Sous H_0 : $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, 3,\ j = 1, 2.$
La statistique de test $Z = \frac{S_{F1}/2}{SR/2}$ suit la loi $F(2, 2)$ sous H_0.

Régime alimentaire influent ?

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle.
$H_1 : \exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$.
Sous H_0 : $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, 3, \; j = 1, 2.$
La statistique de test $Z = \frac{S_{F1}/2}{SR/2}$ suit la loi $F(2, 2)$ sous H_0.
$ZA = [0 \, ; f_{2,2:0.95}] = [0 \, ; 19.0]$,

Régime alimentaire influent ?

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle.
$H_1 : \exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$.
Sous H_0 : $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, 3, \; j = 1, 2.$
La statistique de test $Z = \frac{S_{F1}/2}{SR/2}$ suit la loi $F(2, 2)$ sous H_0.
$ZA = [0 \,; f_{2,2:0.95}] = [0 \,; 19.0]$, $z = \frac{600}{14} = 42.86 \notin ZA$.

Régime alimentaire influent ?

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle.
$H_1 : \exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$.
Sous H_0 : $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, 3,\ j = 1, 2.$
La statistique de test $Z = \frac{S_{F1}/2}{SR/2}$ suit la loi $F(2, 2)$ sous H_0.
$ZA = [0\,; f_{2,2:0.95}] = [0\,; 19.0]$, $z = \frac{600}{14} = 42.86 \notin ZA$.
Donc H_0 est rejetée, le régime alimentaire est un facteur influent pour la production laitière.

Par exemple, si on utilise deux engrais simultanément, on espère que l'action des engrais se complète et que les plantes se développent mieux du fait de cette concomitance.

Par exemple, si on utilise deux engrais simultanément, on espère que l'action des engrais se complète et que les plantes se développent mieux du fait de cette concomitance.
On peut ainsi mesurer l'influence de divers dosages de chacun des engrais.

Par exemple, si on utilise deux engrais simultanément, on espère que l'action des engrais se complète et que les plantes se développent mieux du fait de cette concomitance.
On peut ainsi mesurer l'influence de divers dosages de chacun des engrais.
L'interaction peut être bénéfique (synergie) ou néfaste (antagonisme).

Par exemple, si on utilise deux engrais simultanément, on espère que l'action des engrais se complète et que les plantes se développent mieux du fait de cette concomitance.
On peut ainsi mesurer l'influence de divers dosages de chacun des engrais.
L'interaction peut être bénéfique (synergie) ou néfaste (antagonisme).
p valeurs pour le facteur $F1$, q pour le facteur $F2$, r observations pour chaque couple de facteur.

Le modèle est

$$Y_{ijk} = m_{ij} + \gamma_{ij} + \epsilon_{ijk}$$
$$\text{que l'on écrit : } Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Les ϵ_{ijk} sont des v.a. Normales, centrées et de même écart-type σ.

Le modèle est

$$Y_{ijk} = m_{ij} + \gamma_{ij} + \epsilon_{ijk}$$

$$\text{que l'on écrit : } Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Les ϵ_{ijk} sont des v.a. Normales, centrées et de même écart-type σ.

Contraintes :

$$\sum_i \alpha_i = 0 \qquad \sum_j \beta_j = 0 \qquad \sum_{ij} \gamma_{ij} = 0$$

Le modèle est

$$Y_{ijk} = m_{ij} + \gamma_{ij} + \epsilon_{ijk}$$

$$\text{que l'on écrit : } Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Les ϵ_{ijk} sont des v.a. Normales, centrées et de même écart-type σ.

Contraintes :

$$\sum_i \alpha_i = 0 \qquad \sum_j \beta_j = 0 \qquad \sum_{ij} \gamma_{ij} = 0$$

Lien entre (m_{ij}) et (α_i), (β_j), (γ_{ij})

$$\alpha_i = \bar{m}_{i.} - \bar{m}_{..}, \qquad \beta_j = \bar{m}_{.j} - \bar{m}_{..}, \qquad \gamma_{ij} = m_{ij} - \bar{m}_{i.} - \bar{m}_{.j} + \bar{m}_{..}$$

Les hypothèses à tester peuvent être

$$H_O : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0 \text{ ou } \bar{m}_{1.} = \bar{m}_{2.} = \cdots = \bar{m}_{p.}$$

ou

$$H_O' : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ ou } \bar{m}_{.1} = \bar{m}_{.2} = \cdots = \bar{m}_{.p}$$

Les hypothèses à tester peuvent être

$$H_O : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0 \text{ ou } \bar{m}_{1.} = \bar{m}_{2.} = \cdots = \bar{m}_{p.}$$

ou

$$H'_O : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ ou } \bar{m}_{.1} = \bar{m}_{.2} = \cdots = \bar{m}_{.p}$$

ou l'absence d'interactions :

$$H''_0 : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{pq} = 0$$

On calcule les sommes des carrés des écarts

$$\begin{array}{rll}
S_{F1} & = qr\sum_i (y_{i..} - \bar{y}_{...})^2 & (p-1)\ ddl \\
S_{F2} & = pr\sum_j (y_{.j.} - \bar{y}_{...})^2 & (q-1)\ ddl \\
S_{F12} & = ST - S_{F1} - S_{F2} - SR & (p-1)(q-1)\ ddl \\
SR & = \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2 & pq(r-1)\ ddl \\
ST & = \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2 & (rpq-1)\ ddl
\end{array}$$

Tableau d'analyse de la variance

Sources	ddl	Sommes	Carrés moyens
1er facteur	$p-1$	S_{F1}	$S_{F1}/(p-1)$
2ème facteur	$q-1$	S_{F2}	$SM_{F2}/(q-1)$
Interaction	$(p-1)(q-1)$	S_{F12}	$S_{F12}/((p-1)(q-1))$
Résidus	$pq(r-1)$	SR	$SR/(pq(r-1))$
Totaux	$pqr-1$	ST	

On commence par tester l'interaction

On commence par tester l'interaction : sous H_0'',

$$f_{F12} = \frac{S_{F12}/((p-1)(q-1))}{SR/(pq(r-1))}$$

suit la loi de Fisher $((p-1)(q-1), pq(r-1))$.

On commence par tester l'interaction : sous H_0'',

$$f_{F12} = \frac{S_{F12}/((p-1)(q-1))}{SR/(pq(r-1))}$$

suit la loi de Fisher $((p-1)(q-1), pq(r-1))$.
On rejette H_0 lorsque f_{F12} dépasse le fractile d'ordre $1-\alpha$ de cette loi.
Si on accepte H_0'' (pas d'interaction), on teste l'influence de $F1$ puis de $F2$:

En l'absence d'interaction, sous H_0 (respectivement H_0')

$$f_{F1} = \frac{S_{F1}/(p-1)}{SR/pq(r-1)} \qquad \text{et} \qquad f_{F2} = \frac{S_{F2}/(q-1)}{SR/pq(r-1)}$$

suivent des lois de Fisher à $(p-1, pq(r-1))$ (respectivement $(q-1, pq(r-1)))$ degrés de liberté.
On rejette l'hypothèse H_0 (ou H_0') si la valeur f observée est supérieure au fractile d'ordre $1-\alpha$ de la loi en question.

Exemples

On sélectionne plusieurs régions (1er facteur), puis, à l'intérieur de chacune des régions, plusieurs exploitations agricoles (2ème facteur), et on mesure la quantité de lait produite annuellement par r vaches dans chacune des exploitations
Données : y_{ijk}.

Exemples

On sélectionne plusieurs régions (1er facteur), puis, à l'intérieur de chacune des régions, plusieurs exploitations agricoles (2ème facteur), et on mesure la quantité de lait produite annuellement par r vaches dans chacune des exploitations

Données : y_{ijk}.

Modèle hiérarchique

Exemples

On sélectionne plusieurs régions (1er facteur), puis, à l'intérieur de chacune des régions, plusieurs exploitations agricoles (2ème facteur), et on mesure la quantité de lait produite annuellement par r vaches dans chacune des exploitations
Données : y_{ijk}.
Modèle hiérarchique : aucune raison d'avoir un lien entre les exploitations n˚1 de chacun des régions.
$\overline{y_{i..}}$: moyenne des exploitations de la région i

Exemples

On sélectionne plusieurs régions (1er facteur), puis, à l'intérieur de chacune des régions, plusieurs exploitations agricoles (2ème facteur), et on mesure la quantité de lait produite annuellement par r vaches dans chacune des exploitations
Données : y_{ijk}.
Modèle hiérarchique : aucune raison d'avoir un lien entre les exploitations n°1 de chacun des régions.
$\overline{y_{i\bullet\bullet}}$: moyenne des exploitations de la région i : intéressant.
$\overline{y_{\bullet j\bullet}}$: moyenne des j-ièmes exploitations de chaque région

Exemples

On sélectionne plusieurs régions (1er facteur), puis, à l'intérieur de chacune des régions, plusieurs exploitations agricoles (2ème facteur), et on mesure la quantité de lait produite annuellement par r vaches dans chacune des exploitations
Données : y_{ijk}.
Modèle hiérarchique : aucune raison d'avoir un lien entre les exploitations n˚1 de chacun des régions.
$\overline{y_{i\bullet\bullet}}$: moyenne des exploitations de la région i : intéressant.
$\overline{y_{\bullet j\bullet}}$: moyenne des j-ièmes exploitations de chaque région : non pertinent.

Modèle :

$$Y_{ijk} = \mu + \alpha_i + \beta_{j|i} + \epsilon_{ijk}$$

Modèle :

$$Y_{ijk} = \mu + \alpha_i + \beta_{j|i} + \epsilon_{ijk}$$

μ : production moyenne
α_i : apport de la région i
$\beta_{i|j}$: apport de l'exploitation j dans la région i

Modèle :

$$Y_{ijk} = \mu + \alpha_i + \beta_{j|i} + \epsilon_{ijk}$$

μ : production moyenne
α_i : apport de la région i
$\beta_{i|j}$: apport de l'exploitation j dans la région i
ϵ_{ijk} : vaiid Normale centrée et de variance σ^2.

Décomposition de la variance

$$y_{ijk} - \overline{y_{\dots}} = (\overline{y_{i..}} - \overline{y_{\dots}}) + (\overline{y_{ij.}} - \overline{y_{i..}}) + (y_{ijk} - \overline{y_{ij}})$$

$$\begin{aligned}
ST &= \sum_{ijk} (Y_{ijk} - \overline{Y_{\dots}})^2 \\
&= \sum_{ijk} \left[(\overline{Y_{i..}} - \overline{Y_{\dots}})^2 + (\overline{Y_{ij.}} - \overline{Y_{i..}})^2 + (Y_{ijk} - \overline{Y_{ij.}})^2 \right] \\
&= qr \sum_{i} (\overline{Y_{i..}} - \overline{Y_{\dots}})^2 + r \sum_{jk} (\overline{Y_{ij.}} - \overline{Y_{i..}})^2 + \sum_{ijk} (Y_{ijk} - \overline{Y_{ij.}})^2 \\
&= SM_a + SM_{b|a} + SR
\end{aligned}$$

Tableau d'analyse de la variance

Sources	ddl	Sommes	Carrés moyens
1er facteur	$p-1$	SM_a	$SM_a/(p-1)$
2ème facteur	$p(q-1)$	$SM_{b\|a}$	$SM_{b\|a}/(p(q-1))$
Résidus	$pq(r-1)$	SR	$SR/(pq(r-1))$
Totaux	$pqr-1$	ST	

On peut tester

- Si le 1er facteur a une influence :

$$H_0 : a_1 = \ldots = a_p = 0$$

  On calcule $(SM_a/(p-1))/(SM_{b|a}/p(q-1))$ et on le compare à $F_{1-\alpha;p-1;p(q-1)}$
- Si le 2ème facteur une influence à l'intérieur du 1er

$$H_0 : \beta_{j|i} = 0, \text{ pour tous } i \text{ et } j.$$

  On calcule $SM_{b|a}/(p(q-1))/(SR/(pq(r-1)))$ et on le compare à $F_{1-\alpha;p(q-1),pq(r-1)}$

Exemple : Rendement fourrager dans 2 types de prairies, 3 prairies pour chacun des 2 types, 5 parcelles dans chacune des 3×2 prairies.

n°	Type 1			Type 2		
	Pr. 1	Pr. 2	Pr. 3	Pr. 1	Pr. 2	Pr. 3
1	2.06	1.59	1.92	2.91	1.57	2.43
2	2.99	2.63	1.85	3.27	1.82	2.17
3	1.98	1.98	2.14	3.45	2.69	2.37
4	2.95	2.25	1.33	3.92	3.25	2.89
5	2.70	2.09	1.83	4.34	3.11	2.24

Exemple : Rendement fourrager dans 2 types de prairies, 3 prairies pour chacun des 2 types, 5 parcelles dans chacune des 3×2 prairies.

n°	Type 1			Type 2		
	Pr. 1	Pr. 2	Pr. 3	Pr. 1	Pr. 2	Pr. 3
1	2.06	1.59	1.92	2.91	1.57	2.43
2	2.99	2.63	1.85	3.27	1.82	2.17
3	1.98	1.98	2.14	3.45	2.69	2.37
4	2.95	2.25	1.33	3.92	3.25	2.89
5	2.70	2.09	1.83	4.34	3.11	2.24

On calcule les moyennes pour chacune des prairies : 2.536, 2.108, 1.814 et 3.578, 2.488, 2.420.

La moyenne globale est égale à 2.491.

On a

$$SM_a = 3.427, \quad SM_{b|a} = 5.541, \quad SR = ST - SM_a - SM_{b|a} = 5.742$$

On obtient le tableau d'analyse de la variance

Sources	ddl	Sommes	Carrés moyens
Type	$p-1=1$	$SM_a = 3.427$	$SM_a/(p-1) = 3.427$
Prairie	$p(q-1)=4$	$SM_{b\|a} = 5.541$	$SM_{b\|a}/(p(q-1)) = 1.385$
Résidus	$pq(r-1)=24$	$SR = 5.742$	$SR/(pq(r-1)) = 0.239$
Totaux	$pqr-1=29$	$ST = 14.710$	

On peut tester :

- Si le 1er facteur a une influence :

$$\frac{SM_a/(p-1)}{SM_{b|a}/(p(q-1))} = \frac{3.427}{1.385} = 2.474$$

On peut tester :

- Si le 1er facteur a une influence :

$$\frac{SM_a/(p-1)}{SM_{b|a}/(p(q-1))} = \frac{3.427}{1.385} = 2.474$$

à comparer avec $F_{0.95;1;4} = 7.71$: non significatif

On peut tester :

- Si le 1er facteur a une influence :

$$\frac{SM_a/(p-1)}{SM_{b|a}/(p(q-1))} = \frac{3.427}{1.385} = 2.474$$

  à comparer avec $F_{0.95;1;4} = 7.71$: non significatif

- Si le 2ème facteur a une influence à l'intérieur du 1er :

$$\frac{SM_{b|a}/(p(q-1))}{SR/(pq(r-1))} = \frac{1.385}{0.239} = 5.79$$

On peut tester :

- Si le 1er facteur a une influence :

$$\frac{SM_a/(p-1)}{SM_{b|a}/(p(q-1))} = \frac{3.427}{1.385} = 2.474$$

  à comparer avec $F_{0.95;1;4} = 7.71$: non significatif

- Si le 2ème facteur a une influence à l'intérieur du 1er :

$$\frac{SM_{b|a}/(p(q-1))}{SR/(pq(r-1))} = \frac{1.385}{0.239} = 5.79$$

  à comparer avec $F_{0.95;4;24} = 2.8$: significatif

On peut tester :

- Si le 1er facteur a une influence :

$$\frac{SM_a/(p-1)}{SM_{b|a}/(p(q-1))} = \frac{3.427}{1.385} = 2.474$$

  à comparer avec $F_{0.95;1;4} = 7.71$: non significatif

- Si le 2ème facteur a une influence à l'intérieur du 1er :

$$\frac{SM_{b|a}/(p(q-1))}{SR/(pq(r-1))} = \frac{1.385}{0.239} = 5.79$$

  à comparer avec $F_{0.95;4;24} = 2.8$: significatif

Cela signifie que le type de prairie n'a pas d'influence significative sur le rendement fourrager,

On peut tester :

- Si le 1er facteur a une influence :

$$\frac{SM_a/(p-1)}{SM_{b|a}/(p(q-1))} = \frac{3.427}{1.385} = 2.474$$

  à comparer avec $F_{0.95;1;4} = 7.71$: non significatif

- Si le 2ème facteur a une influence à l'intérieur du 1er :

$$\frac{SM_{b|a}/(p(q-1))}{SR/(pq(r-1))} = \frac{1.385}{0.239} = 5.79$$

  à comparer avec $F_{0.95;4;24} = 2.8$: significatif

Cela signifie que le type de prairie n'a pas d'influence significative sur le rendement fourrager,mais que les prairies ne sont pas homogènes à l'intérieur d'un type donné.