ANOVA à trois facteurs : cas non balancé

Amavi Anna Joyce ATCHOU
Rasmane BAMOGO
Papa Abdourahmane CISSE
Moussa DIEME
Seman Giovanni GADO
Oumar Farouk MOUSSA MAHAMADOU
Cheikh Sadibou NGOM

7/01/2025

Plan de la présentation

- **Définition :**

  L'ANOVA à trois facteurs est une méthode statistique qui permet d'étudier simultanément l'effet de trois variables qualitatives (les facteurs) sur une variable quantitative.

- **Caractéristiques principales :**

  • Comparer les moyennes des groupes formés par les différentes combinaisons des modalités des trois facteurs.

  • Décomposer la variance totale en plusieurs composantes : effets principaux (pour chaque facteur), interactions entre facteurs (deux à deux et trois voies) et variance résiduelle.

- **Objectif :**

  Déterminer si les différences observées entre les groupes sont statistiquement significatives et ainsi expliquer la variabilité de la variable dépendante.

- **Importance du cas non balancé :**

  Le nombre d'observations varie selon les combinaisons des modalités des facteurs, ce qui est souvent observé dans la réalité. Il est souvent difficile d'obtenir un nombre constant d'observations pour chaque combinaison.

- **Quelques comaparaisons entre le cas balancé et le cas non balancé :**

Plan balancé	**Plan non balancé**
Chaque combinaison de modalités des trois facteurs possède le même nombre d'observations.	Les effectifs varient d'une combinaison à l'autre.
Facilite la décomposition de la variance et l'interprétation des interactions.	Complexifie l'analyse et nécessite des ajustements pour corriger les disparités.
Moins de biais dans l'estimation des effets.	

Dans une ANOVA à trois facteurs, chaque observation est notée selon le facteur auquel elle appartient. Nous introduisons les notations suivantes :

- $\mathbf{Y_{ijkl}}$: valeur observée pour le i-ième individu dans la combinaison (j, k, l), où :
  - j désigne le niveau du facteur A $\quad (j = 1, \ldots, J)$
  - k désigne le niveau du facteur B $\quad (k = 1, \ldots, K)$
  - l désigne le niveau du facteur C $\quad (l = 1, \ldots, L)$
  - i est l'indice de l'observation pris dans la combinaison (j, k, l)
- $\mathbf{n_{jkl}}$: nombre total d'observations dans la combinaison (j, k, l)
- $\mathbf{N}$: nombre total d'observations dans l'expérience

- **Moyenne générale :**

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{ijk}} Y_{ijkl}$$

- **Moyennes marginales :**

$$\bar{Y}_{.j..} = \frac{1}{\sum_{l=1}^{L} \sum_{k=1}^{K} n_{jkl}} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{i=1}^{n_{jkl}} Y_{ijkl}$$

$$\bar{Y}_{..k.} = \frac{1}{\sum_{j=1}^{J} \sum_{l=1}^{L} n_{jkl}} \sum_{j=1}^{J} \sum_{l=1}^{L} \sum_{i=1}^{n_{jkl}} Y_{ijkl}$$

$$\bar{Y}_{...l} = \frac{1}{\sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk}} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jkl}} Y_{ijkl}$$

- **Moyennes pour une combinaison de deux facteurs :**

$$\bar{Y}_{.jk.} = \frac{1}{\sum_{l=1}^{L} n_{jkl}} \sum_{l=1}^{L} \sum_{i=1}^{n_{jkl}} Y_{ijkl}$$

$$\bar{Y}_{.j.l} = \frac{1}{\sum_{k=1}^{K} n_{jkl}} \sum_{k=1}^{K} \sum_{i=1}^{n_{jkl}} Y_{ijkl}$$

$$\bar{Y}_{..kl} = \frac{1}{\sum_{j=1}^{J} n_{jkl}} \sum_{j=1}^{J} \sum_{i=1}^{n_{jkl}} Y_{ijkl}$$

- **Moyenne pour une combinaison de trois facteurs :**

$$\bar{Y}_{.jkl} = \frac{1}{n_{jkl}} \sum_{i=1}^{n_{jkl}} Y_{ijkl}$$

- **Variance à l'intérieur d'un groupe :**

$$V_{jkl}(Y) = \frac{1}{n_{jkl}} \sum_{i=1}^{n_{jkl}} (Y_{ijkl} - \bar{Y}_{jkl})^2$$

- **Variances entre les groupes d'un facteur :**

$$V^1_{\text{inter}}(Y) = \frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} (\bar{Y}_{jkl} - \bar{Y}_{.j.})^2$$

où $\bar{Y}_{.j.}$ est la moyenne des observations pour le premier facteur

$$V^2_{\text{inter}}(Y) = \frac{1}{K} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} (\bar{Y}_{jkl} - \bar{Y}_{..k.})^2$$

Ici, $\bar{Y}_{..k.}$ est la moyenne des observations pour le deuxième facteur

$$V^3_{\text{inter}}(Y) = \frac{1}{L}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}(\bar{Y}_{jkl} - \bar{Y}_{\dots l})^2$$

Où $\bar{Y}_{\dots l}$ est la moyenne des observations pour le troisième facteur

- **Variance intra en considérant l'ensemble des groupes :**

$$V_{\text{intra}}(Y) = \frac{1}{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_{jkl}}(Y_{ijkl} - \bar{Y}_{jkl})^2$$

- **Variance totale :**

$$V(Y) = \frac{1}{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_{jkl}}(\bar{Y}_{ijkl} - \bar{Y})^2$$

- **Modélisation :**

  Dans une ANOVA à trois facteurs sans interactions, le modèle statistique s'écrit sous la forme :

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + \varepsilon_{ijkl}$$

  - μ est l'effet commun
  - α_i est l'effet du facteur *A*
  - β_j est l'effet du facteur *B*
  - γ_k est l'effet du facteur *C*
  - $\varepsilon_{i,j,k}$ est l'erreur aléatoire associée à chaque observation.

- **Contraintes du modèle :**

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \sum_{k=1}^{K} \gamma_k = 0$$

- **Hypothèses :**

  • Les erreurs ε_{jk} sont supposées indépendantes et identiquement distribuées suivant une loi normale centrée réduite :

$$\varepsilon_{i,j,k} \sim \mathcal{N}\left(0, \sigma^2\right)$$

  • L'homoscédasticité est supposée :

$$\text{Var}\left(\varepsilon_{i,j,k}\right) = \sigma^2 \quad \forall i, j, k$$

  • L'indépendance des erreurs est respectée entre toutes les observations.

- **Estimateurs du modèle :**

$$\hat{\mu} = \bar{Y}$$
$$\hat{\alpha}_j = \bar{Y}_{.j..} - \bar{Y}$$
$$\hat{\beta}_k = \bar{Y}_{..k.} - \bar{Y}$$
$$\hat{\gamma}_l = \bar{Y}_{...l} - \bar{Y}$$
$$\hat{\epsilon}_{ijkl} = Y_{ijkl} - \hat{Y}_{ijkl}$$

- **Interprétation :**
  - Si $\hat{\alpha}_i > 0$, cela signifie que le niveau *i* du facteur *A* a une influence positive sur *Y*.
  - Si $\hat{\beta}_j < 0$, cela signifie que le niveau *j* du facteur *B* réduit en moyenne *Y*.
  - Si $\hat{\gamma}_k \approx 0$, cela indique que le facteur *C* n'a pas d'impact significatif.

- **Sommes des carrés :**

  On calcule la variation totale :

$$SC_{\text{totale}} = \sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jkl}} (Y_{ijkl} - \bar{Y})^2$$

- **Variation due à chaque facteur :**

$$SC_A = n_{.j..} \sum_{j=1}^{J} (\bar{Y}_{.j..} - \bar{Y})^2$$

$$SC_B = n_{..k.} \sum_{k=1}^{K} (\bar{Y}_{..k.} - \bar{Y})^2$$

$$SC_C = n_{...l} \sum_{l=1}^{L} (\bar{Y}_{...l} - \bar{Y})^2$$

- **Variation résiduelle :**

$$SC_{résiduelle} = SC_{totale} - SC_A - SC_B - SC_C$$

- **Degrés de Liberté :**
  - Pour A : $n_A = J - 1$
  - Pour B : $n_B = K - 1$
  - Pour C : $n_C = L - 1$
  - Pour l'erreur : $n_R = JKL - J - K - L + 2$
  - Total : $n_{TOT} = JKL - 1$

- **Calcul des Carrés Moyens :**

$$CM_A = \frac{SC_A}{J-1}, \quad CM_B = \frac{SC_B}{K-1}, \quad CM_C = \frac{SC_C}{L-1}, \quad CM_R = \frac{SC_R}{n_R}$$

- **Calcul des Statistiques de Fisher :**

$$F_A = \frac{CM_A}{CM_R}, \quad F_B = \frac{CM_B}{CM_R}, \quad F_C = \frac{CM_C}{CM_R}$$

Ces valeurs sont comparées aux valeurs critiques de la loi de Fisher à un seuil α donné (souvent 5%).

- **Décision :**
  - Si $F > F_{\text{lue}}$, alors le facteur a une influence significative sur Y
  - Si $F < F_{\text{lue}}$, alors le facteur n'a pas d'influence significative sur Y

- **Modélisation :**

  Dans le cadre de l'ANOVA à 3 facteurs, il est possible que les facteurs interagissent simultanément pour influencer la variable à expliquer.
  Tester ces interactions permet de :

  • D'éffectuer une analyse plus réaliste des effets des facteurs

  • De révéler des éventuels effets combinés

  Le modèle s'écrit sous la forme :

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} + \epsilon_{ijkl}$$

- **Détails des Paramètres du Modèle :**

μ: Effet commun à toutes les observations

$\alpha_j,\ \beta_k,\ \gamma_l$: Effets principaux des facteurs A, B, C

$(\alpha\beta)_{jk}$: Interaction entre les niveaux *j* de A et *k* de B

$(\alpha\gamma)_{jl}$: Interaction entre les niveaux *j* de A et *l* de C

$(\beta\gamma)_{kl}$: Interaction entre les niveaux *k* de B et *l* de C

$(\alpha\beta\gamma)_{jkl}$: Interaction triple entre les trois facteurs

ϵ_{ijkl}: Terme d'erreur aléatoire $\sim \mathcal{N}(0, \sigma^2)$

Contraintes sur les effets spécifiques

$$\sum_{j=1}^{J} \alpha_j = 0, \quad \sum_{k=1}^{K} \beta_k = 0, \quad \sum_{l=1}^{L} \gamma_l = 0$$

Contraintes sur les interactions doubles

$$\sum_{j=1}^{J} (\alpha\beta)_{jk} = 0, \quad \forall k \in \{1, \ldots, K\}, \quad \sum_{k=1}^{K} (\alpha\beta)_{jk} = 0, \quad \forall j \in \{1, \ldots, J\}$$

$$\sum_{j=1}^{J} (\alpha\gamma)_{jk} = 0, \quad \forall l \in \{1, \ldots, L\}, \quad \sum_{l=1}^{L} (\alpha\gamma)_{jl} = 0, \quad \forall j \in \{1, \ldots, J\}$$

$$\sum_{k=1}^{K} (\beta\gamma)_{kl} = 0, \quad \forall l \in \{1, \ldots, L\}, \quad \sum_{l=1}^{L} (\beta\gamma)_{kl} = 0, \quad \forall k \in \{1, \ldots, K\}$$

Contraintes sur l'interaction triple

$$\sum_{j=1}^{J}(\alpha\beta\gamma)_{jkl} = 0, \quad \forall (k,l) \in \{1,\ldots,K\} \times \{1,\ldots,L\}$$

$$\sum_{k=1}^{K}(\alpha\beta\gamma)_{jkl} = 0, \quad \forall (j,l) \in \{1,\ldots,J\} \times \{1,\ldots,L\}$$

$$\sum_{l=1}^{L}(\alpha\beta\gamma)_{jkl} = 0, \quad \forall (j,k) \in \{1,\ldots,J\} \times \{1,\ldots,K\}$$

Estimation des paramètres du modèle

Les estimations des paramètres dans une ANOVA 3 avec interactions sont obtenues en utilisant la MCO. Les estimations des paramètres sont données par :

$$
\begin{aligned}
\hat{\mu} &= Y_{\bullet,\bullet,\bullet,\bullet} = \overline{Y} \\
\hat{\alpha}_j &= Y_{\bullet,j,\bullet,\bullet} - \overline{Y}, \quad 1 \leq j \leq J \\
\hat{\beta}_k &= Y_{\bullet,\bullet,k,\bullet} - \overline{Y}, \quad 1 \leq k \leq K \\
\hat{\gamma}_l &= Y_{\bullet,\bullet,\bullet,l} - \overline{Y}, \quad 1 \leq l \leq L \\
(\alpha\beta)_{jk} &= Y_{\bullet,j,k,\bullet} - Y_{\bullet,j,\bullet,\bullet} - Y_{\bullet,\bullet,k,\bullet} + \overline{Y}, \\
(\alpha\gamma)_{jl} &= Y_{\bullet,j,\bullet,l} - Y_{\bullet,j,\bullet,\bullet} - Y_{\bullet,\bullet,\bullet,l} + \overline{Y}, \\
(\beta\gamma)_{kl} &= Y_{\bullet,\bullet,k,l} - Y_{\bullet,\bullet,k,\bullet} - Y_{\bullet,\bullet,\bullet,l} + \overline{Y}, \\
(\alpha\beta\gamma)_{jkl} &= Y_{\bullet,j,k,l} - Y_{\bullet,j,k,\bullet} - Y_{\bullet,j,\bullet,l} - Y_{\bullet,\bullet,k,l} \\
&\quad + Y_{\bullet,j,\bullet,\bullet} + Y_{\bullet,\bullet,k,\bullet} + Y_{\bullet,\bullet,\bullet,l} - \overline{Y}, \\
\hat{\varepsilon_{ijkl}} &= Y_{ijkl} - \hat{Y_{ijkl}}.
\end{aligned}
$$

Degrés de Liberté

Nous introduisons les degrés de liberté (ddl) associés à chaque ligne du tableau de l'ANOVA :

Table: Tableau des degrés de liberté (ddl)

Source	**Degrés de liberté**
Facteur A	$n_A = J - 1$
Facteur B	$n_B = K - 1$
Facteur C	$n_C = L - 1$
Interaction AB	$n_{AB} = (J-1)(K-1)$
Interaction AC	$n_{AC} = (J-1)(L-1)$
Interaction BC	$n_{BC} = (K-1)(L-1)$
Interaction ABC	$n_{ABC} = (J-1)(K-1)(L-1)$
Résiduelle	$n_R = JKL(N-1)$
Totale	$n_{TOT} = JKLN - 1$

- J, K, L = nombre de niveaux pour A, B, C
- N = répétitions par combinaison

ddl = (Nombre de paramètres estimés) - (Nombre de contraintes)

Critères d'influence des facteurs sur Y

- Test des facteurs pris séparément

  Dans ce cas, on procède comme dans le cas du modèle sans interaction en faisant attention à prendre en compte les nouveaux degrés de liberté du modèle comme spécifié dans le tableau.

- Tests des interactions entre les facteurs

  Nous supposons tester le critère d'influence entre les facteurs A et B, étant donné que le nombre de cas à tester est assez grand.

  On veut tester :

$$H_0 : (\alpha\beta)_{jk} = 0 \quad \forall (j,k)$$
$$H_1 : \exists (j',k') \mid (\alpha\beta)_{j'k'} \neq 0$$

Statistique de Test

La statistique de test est donnée par :

$$F_{AB} = \frac{\text{SCE}_{AB}/n_{AB}}{\text{SCR}/n_R}$$

Avec :

- SCE_{AB} : Somme des carrés expliquée par l'interaction AB
- $n_{AB} = (J-1)(K-1)$: ddl de l'interaction AB
- SCR : Somme des carrés résiduelle
- $n_R = JKL(N-1)$: ddl résiduels

$$\text{Sous } H_0, \quad F_{AB} \sim \mathcal{F}(n_{AB},\, n_R)$$

Règle de Décision

Pour l'interaction AB :

$$\begin{cases} F_{AB} > F_{\alpha}(n_{AB}, n_R) & \Rightarrow \text{Rejet de } H_0 \\ F_{AB} \leq F_{\alpha}(n_{AB}, n_R) & \Rightarrow \text{Non rejet de } H_0 \end{cases}$$

Procédure identique pour :

- Toutes interactions (AC, BC, ABC)

Test du facteur d'interaction triple

Dans le cas de l'interaction entre trois facteurs (A, B et C), nous voulons tester l'influence de l'interaction triple $(\alpha\beta\gamma)_{jkl}$.

Formulation des Hypothèses :

$$H_0 : (\alpha\beta\gamma)_{ijk} = 0 \quad \forall (i, j, k) \quad \text{(pas d'interaction triple)}$$
$$H_1 : \exists (i', j', k') \quad \text{tel que} \quad (\alpha\beta\gamma)_{i'j'k'} \neq 0$$

Cela signifie qu'il existe au moins une combinaison de niveaux des trois facteurs où l'interaction triple est significative.

Statistique de Test

La statistique de test pour l'interaction triple est :

$$F_{ABC} = \frac{\frac{\text{SCE}_{ABC}}{(J-1)(K-1)(L-1)}}{\frac{\text{SCR}}{N-JKL}}$$

Avec :

- SCE_{ABC} : Somme des Carrés expliquée par l'interaction triple (A × B × C).
- $N - JKL$: Degrés de liberté associés à l'interaction triple.
- SCR : Somme des Carrés des Résidus (erreur non expliquée par le modèle).

Distribution de la Statistique sous H_0

Sous l'hypothèse nulle H_0, la statistique F_{ABC} suit une loi de Fisher :

$$F_{ABC} \sim F\left((J-1)(K-1)(L-1), N-JKL\right)$$

Règle de Décision :

- Si $F_{ABC} > F_\alpha$, on rejette H_0 : l'interaction triple est significative.
- Sinon, on ne peut pas conclure à une interaction significative.

Tests de Significativité des Interactions

Hypothèses :

$$H_0 : (\alpha\beta)_{jk} = 0 \quad \forall (j,k) \quad \text{(pas d'interaction)}$$
$$H_1 : \exists (j',k') \mid (\alpha\beta)_{j'k'} \neq 0 \quad \text{(interaction significative)}$$

Statistique de Test :

$$F = \frac{CM_{\text{Interaction}}}{CM_{\text{Résiduelle}}} = \frac{\dfrac{SCE_{\text{Interaction}}}{ddl_{\text{Interaction}}}}{\dfrac{SCR}{ddl_{\text{Résiduelle}}}}$$

Sous H_0 :

$$F \sim F(ddl_{\text{Interaction}}, ddl_{\text{Résiduelle}})$$

Formules des Statistiques F

Interactions à Deux Facteurs :

$$F_{AB} = \frac{\frac{SCE_{AB}}{(J-1)(K-1)}}{\frac{SCR}{N-JKL}} \quad F_{AC} = \frac{\frac{SCE_{AC}}{(J-1)(L-1)}}{\frac{SCR}{N-JKL}} \quad F_{BC} = \frac{\frac{SCE_{BC}}{(K-1)(L-1)}}{\frac{SCR}{N-JKL}}$$

Interaction Triple (A × B × C) :

$$F_{ABC} = \frac{\frac{SCE_{ABC}}{(J-1)(K-1)(L-1)}}{\frac{SCR}{N-JKL}}$$

Décomposition de la Variance

$$SCT = SCE_A + SCE_B + SCE_C + SCE_{AB} + SCE_{AC} + SCE_{BC} + SCE_{ABC} + SCR$$

Sommes des Carrés :

$$SCT = \sum (Y_{ijkl} - \bar{Y})^2 \quad SCR = \sum (Y_{ijkl} - \hat{Y}_{ijkl})^2$$

Tableau Récapitulatif de l'ANOVA

Source	SC	ddl	CM	F
A	SCE_A	$J-1$	CM_A	$F_A = \frac{CM_A}{CM_R}$
B	SCE_B	$K-1$	CM_B	$F_B = \frac{CM_B}{CM_R}$
C	SCE_C	$L-1$	CM_C	$F_C = \frac{CM_C}{CM_R}$
$A \times B$	SCE_{AB}	$(J-1)(K-1)$	CM_{AB}	$F_{AB} = \frac{CM_{AB}}{CM_R}$
$A \times C$	SCE_{AC}	$(J-1)(L-1)$	CM_{AC}	$F_{AC} = \frac{CM_{AC}}{CM_R}$
$B \times C$	SCE_{BC}	$(K-1)(L-1)$	CM_{BC}	$F_{BC} = \frac{CM_{BC}}{CM_R}$
$A \times B \times C$	SCE_{ABC}	$(J-1)(K-1)(L-1)$	CM_{ABC}	$F_{ABC} = \frac{CM_{ABC}}{CM_R}$
Erreur (Résidus)	SCR	$N-JKL$	CM_R	-
Total	SCT	$N-1$	-	-

En conclusion, l'ANOVA à trois facteurs constitue un outil puissant pour analyser des données complexes, même en présence de groupes non équilibrés. Cette étude a mis en évidence l'importance de choisir des méthodes adaptées pour le calcul des sommes des carrés et des degrés de liberté afin de garantir la validité des tests statistiques. L'analyse des interactions permet d'obtenir une vision plus réaliste des effets des facteurs, offrant ainsi une meilleure compréhension des phénomènes étudiés. Grâce à ses applications pratiques dans divers domaines, l'ANOVA à trois facteurs s'avère être une méthode essentielle pour explorer et interpréter des relations complexes entre variables.

Merci de votre aimable attention