

Plans d'expériences

Séance 1: Rappels d'ANOVA

On dispose de p échantillons de tailles respectives n_1, \dots, n_p correspondant chacun à un niveau différent d'un facteur A. L'effectif total est noté $n = n_1 + \dots + n_p$.

Hypothèse

On suppose dans la suite que l'on a la même variance dans le groupe d'observations.

Le facteur peut être une dose d'engrais, d'insecticide, de médicament, des temps d'exposition à la lumière...etc...

Chaque échantillon serait le résultat de l'application d'un niveau du facteur A.

⇒ On souhaite savoir si il **existe** un niveau du facteur A qui aurait un effet sur la population.

Hypothèse testée.

Ainsi on teste l'égalité des p moyennes

$$\begin{cases} H_0 : m_1 = m_2 = \dots = m_p. \\ H_1 : \exists i, j / m_i \neq m_j. \end{cases}$$

Modèle sous-jacent.

On note X_{ik} la k ème observation du i ème échantillon, $i \in \{1, \dots, p\}$ et $k \in \{1, \dots, n_p\}$. On suppose que

$$X_{ik} = m_i + \epsilon_{ik},$$

où m_i est la moyenne (théorique non observée) de l'échantillon i et ϵ_{ik} , le terme d'erreur.

Hypothèse Gaussienne

$$\epsilon_{ik} \sim N(0, \sigma^2) \quad \text{avec} \quad \sigma > 0.$$

Modèle sous-jacent.

$$\begin{pmatrix} X_{11} \\ \vdots \\ X_{1n_1} \\ X_{21} \\ \vdots \\ X_{2n_2} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots \\ \vdots & \vdots & \cdots \\ 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \cdots \\ \vdots & 1 & \cdots \\ 0 & 0 & \cdots \end{pmatrix} \begin{pmatrix} m_1 \\ \vdots \\ m_p \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{pn_p} \end{pmatrix}.$$

Décomposition de la variance : Version théorique.

Notons par m la moyenne (théorique non observée) de l'ensemble des observations. Pour une observation on écrit :

$$X_{ik} - m = (m_i - m) + (X_{ik} - m_i).$$

En élevant au carré et en faisant la somme sur toutes les observations et en divisant par l'effectif total n , on obtient :

$$\sum_{i=1}^p \sum_{k=1}^{n_i} (X_{ik} - m)^2 = \sum_{i=1}^p n_i (m_i - m)^2 + \sum_{i=1}^p \sum_{k=1}^{n_i} (X_{ik} - m_i)^2.$$

ou bien

$$\sum_{i=1}^p \sum_{k=1}^{n_i} (X_{ik} - m)^2 = \sum_{i=1}^p n_i (m_i - m)^2 + \sum_{i=1}^p \sum_{k=1}^{n_i} \epsilon_{ik}^2.$$

SC totale

SC factorielle

SC des résidus

C'est la formule habituelle de la décomposition de la variance simplement appliquée au modèle considéré.

Décomposition de la variance : Version empirique.

Comme on ne connaît pas les m_i et m , on doit les estimer par $\bar{X}_i = \sum_{k=1}^{n_i} X_{ik}$ et $\bar{X} = \sum_{i=1}^p \sum_{k=1}^{n_i} X_{ik}$. On obtient de la même façon que précédemment

$$\sum_{i=1}^p \sum_{k=1}^{n_i} (X_{ik} - \bar{X})^2 = \sum_{i=1}^p n_i (m_i - \bar{X})^2 + \sum_{i=1}^p \sum_{k=1}^{n_i} (X_{ik} - \bar{X}_i)^2.$$

 S_T^2 S_F^2 S_R^2

Test.

Sous H_0 et les hypothèses gaussianité et d'homoscédasticité, on a

$$\frac{S_T}{\sigma^2} \sim \chi_{n-1}^2, \quad \frac{S_F}{\sigma^2} \sim \chi_{p-1}^2, \quad \frac{S_R}{\sigma^2} \sim \chi_{n-p}^2.$$

★ Ainsi on a :

$$T := \frac{S_F^2/p - 1}{S_R^2/n - p} \sim F_{p-1, n-p}.$$

★ Règle de décision :

Si $T > F_{p-1, n-p, 1-\alpha}$, on rejette H_0 avec un risque de première espèce α .

Exemple pratique (voir fiche 1).

On veut savoir si la moyenne en algèbre deuxième année des étudiants ayant suivi l'option math en Terminale S est égale à celle des étudiants n'ayant pas suivi cette option en Terminale.

- ⇒ Programmer votre test d'ANOVA en utilisant le langage SAS IML (voir poly résumant les principales commandes SAS IML)
- ⇒ Pour cela répondre aux différentes questions de la fiche 1.

Comparaisons multiples de moyennes (contrastes).

Si le test d'ANOVA révèle une différence entre les moyennes, on peut essayer de comparer des sous-groupes.

Définition

Un contraste est une combinaison de plusieurs moyennes m_i dont la somme des coefficients est égale à zéro :

$$\begin{cases} C = c_1 m_1 + \dots + c_p m_p = \sum_{i=1}^p c_i m_i \\ \sum_{i=1}^p c_i = 0. \end{cases}$$

On estime le contraste par

$$\hat{C} = \sum_{i=1}^p c_i \bar{X}_i.$$

Estimation de la variance du contraste.

On remarque que :

$$\text{Var}(\hat{C}) = \sum_{i=1}^p c_i^2 \text{Var}(\bar{X}_i) = \sigma^2 \sum_{i=1}^p \frac{c_i^2}{n_i},$$

⇒ On estime la variance de \hat{C} par

$$\hat{\sigma}_C^2 = \hat{\sigma}^2 \sum_{i=1}^p \frac{c_i^2}{n_i}.$$

On a montré que (Scheffé) :

$$\mathbb{P} \left(C \notin \left[\hat{C} \pm \sqrt{(p-1)F_{p-1, n-p, 1-\alpha} \cdot \hat{\sigma}_C} \right] \right) = \alpha$$

Exemple d'application : Concentration de sulfure dans des veines de charbon (coal seams). Voir feuilles distribuées.

- ⇒ Comparaisons des moyennes deux à deux : On voit si zéro est inclus dans les différents intervalles.
- ⇒ Vérifier certaines quantifications entre les moyennes : Tester des hypothèses du type $m_i = 2m_j$.

Deuxième façon de procéder : On vérifie si pour tout i, j

$$\frac{|\bar{X}_i - \bar{X}_j|}{\hat{\sigma} \sqrt{n_i^{-1} + n_j^{-1}}} > \sqrt{(p-1)F_{p-1, n-p, 1-\alpha}}$$

Si oui, alors on décide $m_i \neq m_j$.

ANOVA à deux facteurs.

Soit X_{ijk} la k ème observation ayant la modalité i pour le facteur A et j pour le facteur B. ($i \in \{1, \dots, p\}$ et $j \in \{1, \dots, q\}$)

Soit $p \times q$ échantillons de tailles n correspondant aux différentes combinaisons des modalités de deux facteurs A et B.

- ⇒ \bar{X}_i représente la moyenne des observations de modalité i .
- ⇒ \bar{X}_{ij} représente la moyenne des observations ayant pour modalité i et j pour les facteurs A et B.

ANOVA à deux facteurs.

De la même manière que pour de la décomposition de la variance où l'on considère 1 facteur, on obtient pour 2 facteurs :

$$\begin{aligned}
 \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X})^2 &= nq \sum_{i=1}^p (\bar{X}_{i.} - \bar{X})^2 + np \sum_{j=1}^q (\bar{X}_{.j} - \bar{X})^2 \\
 &+ n \sum_{i=1}^p \sum_{k=1}^{n_i} (\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2 \\
 &+ \sum_{i=1}^p \sum_{k=1}^{n_i} (X_{ijk} - \bar{X}_{ij})^2.
 \end{aligned}$$

$$SC_T^2 = SC_A^2 + SC_B^2 + SC_{AB}^2 + SC_R^2.$$

ANOVA à deux facteurs.

Le terme SC_{AB}^2 permet de mesurer l'interaction entre les deux facteurs :

Ce terme est nul quand les variations relatives au premier facteur sont indépendantes des modalités du deuxième facteur :

$$\bar{X}_{ij} - \bar{X}_{.j} = \bar{X}_{i.} - \bar{X}$$

et inversement...

- Si on suppose que l'interaction est nulle, on dit alors que le modèle d'analyse de la variance est **additif**.

ANOVA à deux facteurs.

On montre que :

$$\frac{SC_A^2/p-1}{SC_R/pq(n-1)} \sim F_{p-1,pq(n-1)} \Rightarrow \text{Teste l'effet du facteur A.}$$

$$\frac{SC_B^2/q-1}{SC_R/pq(n-1)} \sim F_{q-1,pq(n-1)} \Rightarrow \text{Teste l'effet du facteur B.}$$

$$\frac{SC_{AB}^2/(p-1)(q-1)}{SC_R/pq(n-1)} \sim F_{(p-1)(q-1),pq(n-1)} \Rightarrow \text{Teste l'interaction entre les deux facteurs.}$$

ANOVA à trois facteurs.

De la même façon que dans les cas d'ANOVA à un ou deux facteurs on peut écrire dans le cas où l'on a trois facteurs A, B et C :

$$SC_T^2 = SC_A^2 + SC_B^2 + SC_C^2 + SC_{AB}^2 + SC_{AC}^2 + SC_{BC}^2 + SC_{ABC}^2 + SC_R^2.$$

⇒ On peut aussi tester les effets des différents facteurs et leur interaction.

Il faut définir de façon précise

- ★ Les buts de l'expérience. (étude de l'efficacité d'un médicament, d'un engrais...)
- ★ Les unités expérimentales. (malade, parcelle de terre...)
- ★ Les objets : Les combinaisons des facteurs inclus dans l'étude. (médicament A ou B, type et dose d'engrais...).
- ⇒ Ne pas oublier le niveau témoin : Malade auquel on administre un placebo, parcelle sans engrais...
- ★ Les observations : recueillies à l'issue de l'expérimentation.
- ★ Analyse des résultats : Par exemple en effectuant une ANOVA.

Bibliographie de cours

Droesbeke J.J., Fine J., Saporta G. *Plans d'expériences-Applications à l'entreprise*. Editions Technip 1997.

Bailey R.A. *Design of comparative experiments*. Cambridge University Press 2008.

Lorenzen T.J. *Design of experiments- A no-name approach*. Marcel Dekker INC 1993.