République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Mohammed Seddik Ben Yahia - Jijel

Faculté des Sciences Exactes et Informatique
Département des Mathématiques

Mémoire de fin de cycle

Présenté pour l'obtention du diplôme de

Master

Spécialité : Mathématiques fondamentales

Option : Probabilités et Statistique.

Thème

Analyse de la variance et analyse de la covariance à deux facteurs

Présenté par : **Boudraa Imene**

Devant le jury :

Président	**Djeridi Zohra**	**M.C.B** Université Mohammed Seddik Ben Yahia, Jijel
Encadreur	**Laoudj Chekraoui Farida**	**M.C.A** Université Mohammed Seddik Ben Yahia, Jijel
Examinateur	**Boudjerda Khaoula**	**M.C.B** Université Mohammed Seddik Ben Yahia, Jijel

Promotion **2020/2021**

♡ Remerciements ♡

Avant toute choses, je remercie ALLAH le tout Puissant, pour m'avoir donnée la force et la patience, la santé et la volonté pour réaliser ce travail.

Je tiens à remercier sincèrement Mme Laoudj Chekraoui Farida mon encadreur, qu'elle trouve ici l'expression de ma profonde reconnaissance pour m'avoir guidées dans mon travail. Ses conseils, ses orientations, sa patience, et sa correction sérieuse de ce travail .

Mes remerciements vont également aux membres du jury pour l'intéret qu'ils ont porté à mon recherche en acceptant d'examiner ma mémoire et de l'enrichir par leurs propositions .

Je n'oublie pas de remercier vivement Le chef département et tous mes enseignants, pour leurs informations et leurs aides tout au long de rus 5 années d étude .

B.Imene

♡ *Dédicases* ♡

Je dédie ce modeste travail

A mes très chers parents qui ont bien élevé, aidé, soutenu et encouragé durant toutes ces années d'étude, qu'ALLAH les protège.

A ma grande- mère pour sa tendresse, je lui souhaite une langue vie.

Mes chers frères, et mes chères soeurs pour leur affection, compréhension et patience , et surtout mon grande frère

A tous mes amis qui m'ont toujours encouragé et surtout mon amie Nadjet

A tout la promotion 2éme Anneè Master probabilités et statistique 2020-2021

B.Imene

Table des matières

Liste des tableaux

Notations

- $\boldsymbol{E}$: Espérance .
- $\boldsymbol{Var}$: La variance .
- $\boldsymbol{Cov}$: La covariance .
- $\boldsymbol{\varepsilon}$: Le résidu .
- $\boldsymbol{ANOVA}$: L'analyse de la variance .
- $\boldsymbol{ANOVA2}$: L'analyse de la variance à deux facteurs.
- $\boldsymbol{ANCOVA}$: L'analyse de la covariance .
- $\boldsymbol{SCT}$: La somme des carrés totale .
- $\boldsymbol{SCR}$: La somme des carrés des résidus .
- $\boldsymbol{SCF}$: La somme des carrés expliqués .
- $\boldsymbol{MCF}$: Moyennes carrés associés au facteur .
- $\boldsymbol{MCR}$: Moyennes carrés résiduels .

Introduction générale

Ce travail est réalisé dans des conditions très particulières en temps de la pandémie COVID19 où l'enseignement à distance a caractérisé cette période très difficile. L'objectif initial de mon mémoire était de traiter et d'étudier l'intérêt de l'analyse de la covariance aux facteurs multiples (fixes et aléatoires) ; mais l'apprentissage scientifique n'était pas très efficient et les bases n'étaient pas toute à fait acquises, notamment l'analyse de la variance à deux facteurs. Par conséquent, l'objectif de mon mémoire a été modifié. J'ai traité et étudié en premier chapitre les fondements théoriques de l'analyse de la variance à deux facteurs fixes (section1) et j'ai tenté à démontrer l'ensemble des résultats. De plus, j'ai réalisé une application de cette méthode sous R (section 2).

Une fois, j'ai maitrisé l'analyse de la variance, j'ai pu effectuer aisément une étude sur l'analyse de la covariance. C'est l'objet du second chapitre de mon mémoire. Dans sa première section, j'ai présenté les fondements théoriques de l'analyse de la covariance à un facteur fixe puis j'ai réalisé, dans sa seconde section, une étude de cas pour montrer l'utilité de cette méthode statistique par rapport à l'analyse de la variance.

L'analyse de variance (ANOVA) est une technique statistique utilisée pour étudier le comportement d'une variable quantitative à expliquer (variable d'intérêt) en fonction d'une ou de plusieurs variables qualitatives. Autrement dit, il s'agit d'étudier l'effet d'un facteur (ou plusieurs facteurs) sur une variable d'intérêt de type quantitatif en utilisant un ensemble de modèles statistiques pour comparer les moyennes des différents échantillons indépendants. Les échantillons correspondent aux différentes modalités de la variable qualitative et les moyennes sont calculées sur la variable quantitative.

Pour la réaliser, il est nécessaire de vérifier l'indépendance des échantillons, la normalité des distributions et l'homogénéité des variances. Si nous souhaitons intégrer dans le modèle des variables explicatives quantitatives, l'emploi de l'analyse de la variance devient pas possible. C'est l'analyse de la covariance qu'il faut appliquer. C'est un modèle qui contient des variables indépendantes à la fois qualitatives (appelées facteurs) et quantitatives(appelées covariables). Il

s'agit d'un mélange de l'analyse de la variance et de la régression linéaire.

L'ajout des covariables dans le modèle permet de réduire considérablement la composante de la variabilité associée à l'erreur aléatoire ,et donc d'augmente la puissance du modèle.

Pour récapituler, ce présent travail s'organise comme suit :

- Le premier chapitre est intitulé "Analyse de la variance à deux facteurs fixes". Il contient trois sections ; à savoir : la section 1 est dédiée à la présentation des fondements théoriques de l'analyse, la section 2 aborde les différents types de l'ANOVA à deux facteurs (sans répétition, avec répétition égale et avec répétition inégale), et enfin la dernière section est consacrée à une application sous R.
- Le deuxième Chapitre est intitulé "Analyse de la covariance et son intérêt". Il est composé de deux sections : L'ensemble des fondements théoriques seront présentés dans la section 1 et la section 2 discute et démontre l'intérêt de l'analyse de la covariance à travers une étude de cas.

CHAPITRE

1 Analyse de la variance à 2 facteurs fixes

Introduction

La première application de l'analyse de la variance remonte à 1921 (On the « Probable Error » of a Coefficient of Correlation Deduced from a small Sample) publiée par Ronald Aylmer Fisher. Puis, elle est devenue très connue après la publication de livre de Fisher en 1925.

L'ANOVA est une technique statistique utilisée lorsque l'on souhaite mesurer l'effet d'un facteur ou plusieurs facteurs (variables qualitatives) sur une variable d'intérêt (variable quantitative). Elle met en évidence une influence d'un facteur (ou plusieurs facteurs) sur la variable d'intérêt en utilisant les moyennes des différents échantillons indépendants. Dans une expérimentation, les facteurs de variations sont nombreux et peuvent être connus ou inconnus (graphique1).

Lorsqu'ils sont connus, ils peuvent être non contrôlés par l'analyste. Ce sont ces facteurs qui peuvent perturber les résultats mais on ne sait pas comment les prendre en compte dans l'analyse. Dans ce cas, le rôle d'un plan d'expériences est de définir au mieux l'expérimentation de manière à minimiser au mieux ces effets non contrôlés.

Pour en ce qui concerne, les facteurs connus non étudiés dites (blocs), sont les facteurs qui ne font pas l'objet de l'étude mais peuvent perturber les résultats (comme le sexe, l'âge...). Ces facteurs, on peut les contrôler en réalisant l'étude selon tranche d'âge chez les hommes et chez les femmes séparément, etc...

Pour les facteurs connus étudiés, il peuvent être fixes ou aléatoires selon l'objectif de l'étude. Si nous utilisons la totalité des modalités du facteur ; il s'agit d'une étude ANOVA à facteurs fixes, en revanche, si nous sélectionnons uniquement une partie des modalités du facteur ; on parle d'une ANOVA à facteurs aléatoires.

Il reste les facteurs inconnus, ce sont les facteurs de variation inclus dans l'erreur aléatoire. et le meilleur modèle est celui associé à la plus faible erreur aléatoire (on cherche souvent à minimiser cette erreur) Graphe 1 : Facteurs de variations

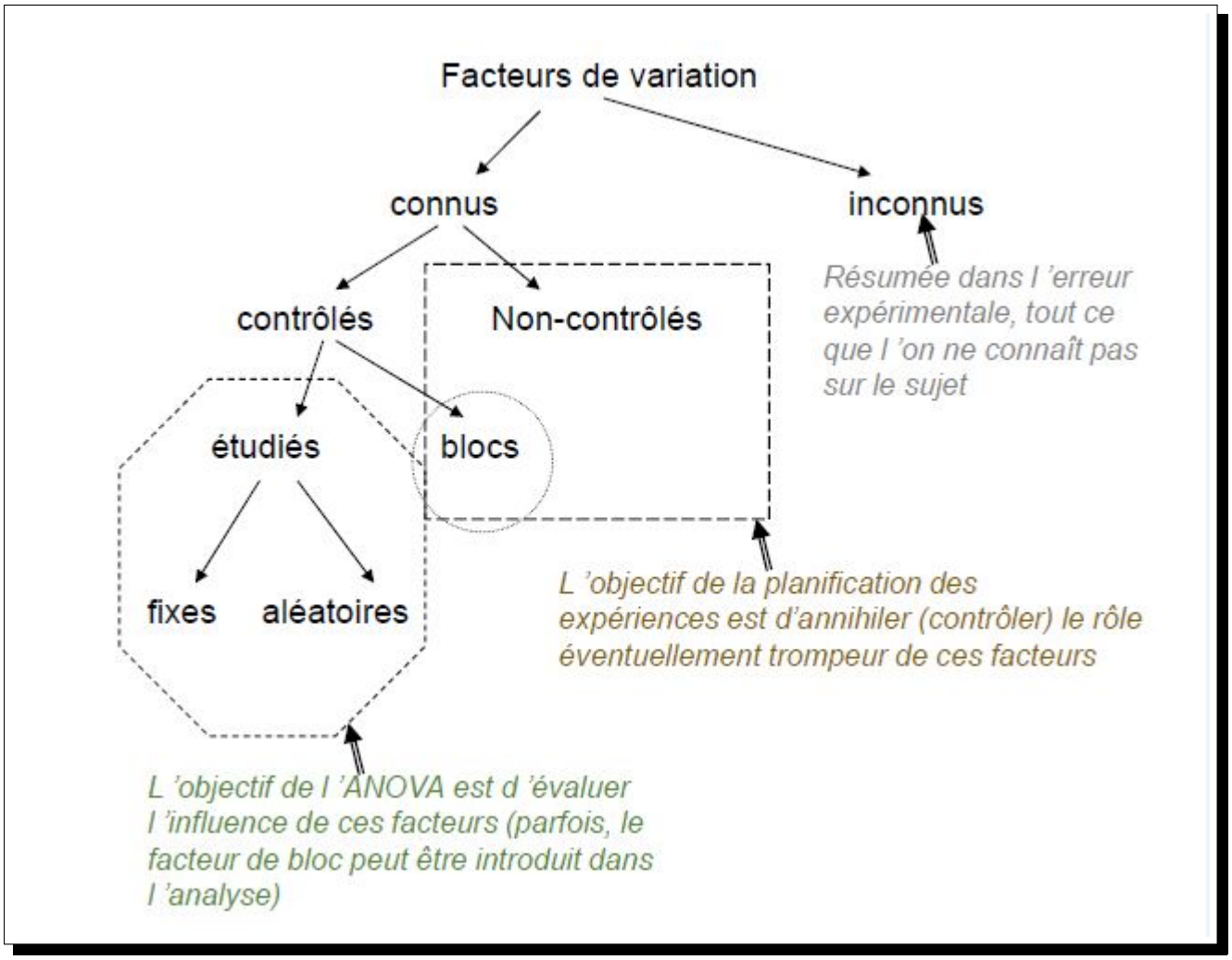

FIGURE 1.1 – Facteurs de variation

Dans ce mémoire nous traitons le cas des facteurs fixes.

1.1 Fondements théoriques de l'ANOVA à 2 facteurs fixes

L'analyse de la variance à deux facteurs est un plan d'expériences définit par deux facteurs où on cherche à expliquer une variable quantitative Y par ces facteurs notés respectivement F_1 et F_2 .Tels que F_1 à p modalités et F_2 à q modalités. A chaque couple (ij) on dispose d'un effectif n_{ij} et de mesures de Y notées y_{ijk}. On note $n_{i.} = \sum_{j=1}^{q} n_{ij}$ le nombre de mesures de Y

pour la modalité i de F_1 et on note $n_{.j} = \sum_{i=1}^{p} n_{ij}$ le nombre de mesures de Y pour la modalité j de F_2. On prend k un indice de répétition du couple (ij). Dans le cadre d'une ANOVA, il existe plusieurs plans d'expériences :

1. le plan est complet si $n_{ij} > 0$ pour tout case (ij)
2. le plan est répété si $n_{ij} > 1$ pour tout case (ij)
3. le plan est orthogonale si $n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$
4. le plan est équilibré si $n_{ij} = r$ pour tout case (ij).
   L'avantage des plans équilibrés (même effectif dans chaque échantillons) est de faire atténuer l'effet de l'hétérogénéité des variances contrairement aux autres plans.

1.1.1 Conditions d'application

L'application et la validité de l'analyse de variance repose sur le test de Fischer donc sur trois conditions qui sont :

- l'indépendance des échantillons,
- la normalité des distributions et pour vérifier l'hypothèse de la normalité, on utilise le test de Shapiro-Wilk,
- l'homogénéité des variances : Pour tester cette hypothèse on utilise le test de Baretlett.

Indépendance des échantillons

On peut aisément vérifier cette condition. Un individu statistique doit appartenir à une seule modalité de la variable qualitative et une seule seulement. On peut aussi appliquer le test de khi-deux d'indépendance en prenant les échantillons deux à deux mutuellement.

Normalité des distributions

Le test de Chapiro-Wilk permet de vérifier la normalité d'une distribution et vérifie qu'un échantillon $y_1, y_2, ..., y_p$ est issu d'une population normalement distribuée.

a. Test Shapiro-Wilk

Les hypothèse à tester sont :

$$\begin{cases} H_0 : \text{La population est normalement distribuée.} \\ H_1 : \text{La population n'est pas normalement distribuée.} \end{cases}$$

La statistique du test

$$B_{obs} = \frac{(\sum_{i=1}^{p} a_i y_{(i)})^2}{\sum_{i=1}^{p}(y_i - \bar{y})^2}$$

où

— $y_{(i)}$: désigne la i ème statistique d'ordre

— $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: est la moyenne de l échantillon.

— La constante a_i est donnée par :

$$(a_1, ..., a_n) = \frac{m^t V^{-1}}{(m^t V^{-1} V^{-1} m)^{1/2}}$$

où

$$m = (m_1, ..., m_n)^t.$$

m sont les espérances des statistiques d'ordre d' un échantillon de variables indépendantes identiquement distribuées suivants une loi normale, et V est la matrice de variances-covariances.

$$V = \begin{pmatrix} Var(y_1) & Cov(y_1, y_2) & ... & Cov(y_1, y_n) \\ Cov(y_2, y_1) & Var(y_2) & ... & Cov(y_2, y_n) \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ Cov(y_n, y_1) & Cov(y_n, y_2) & ... & Var(y_n) \end{pmatrix}$$

Décision du test

Pour un risque α

► Si la $p - valus < \alpha$ alors on rejette l'hypothèse nulle. Cela signifie que la condition de la normalité n'est pas vérifiée.

► Si la $p - valus > \alpha$ alors on ne rejette pas l'hypothèse nulle. Cela signifie que la condition de la normalité est vérifiée.

b.Présentation de la loi normale

Une loi normale est une loi de probabilité absolument continue qui dépend de deux paramètres :

son *espérance* noté μ (nombre réel) et son *écart type* noté σ (nombre réel positif) .
La densité de probabilité de la loi normale d'espérance μ et d'écart type σ est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Voici les formes de la distribution normale pour quelques valeurs de l'espérances et l'écart-type.

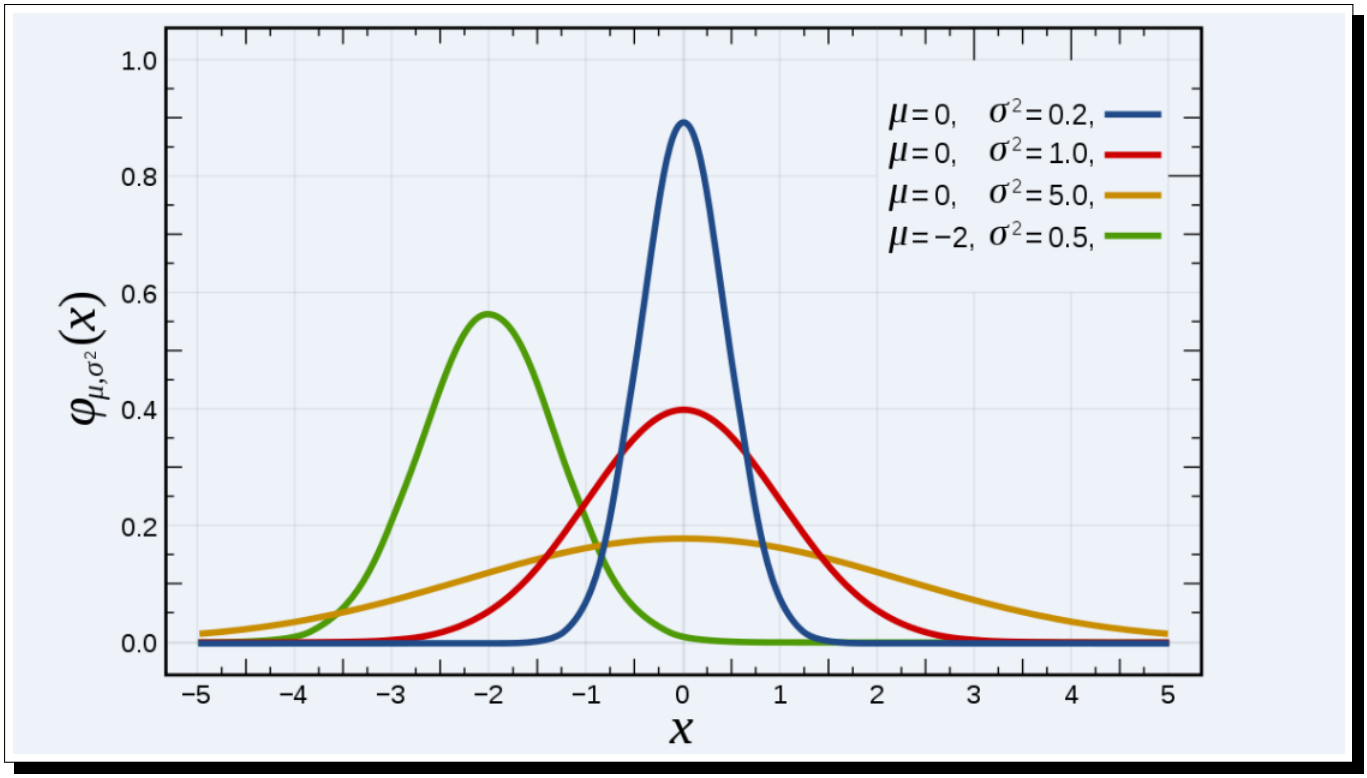

FIGURE 1.2 – Densité de probabilité de la loi normale - source WIKIpédia

Homogénéité des variances

Les hypothèses à tester sont :

$$\begin{cases} H_0 : \forall i \quad \sigma_i^2 = \sigma^2. \\ H_1 : \exists i \quad \sigma_i^2 \neq \sigma^2. \end{cases}$$

La statistique du test est :

$$B_{obs} = \frac{(n-p)lnS^2 - \sum_{i=1}^{p}(n_i-1)lnS_i^2}{1+\frac{1}{3(k-1)}(\sum_{i=1}^{p}\frac{1}{n_i-1} - \frac{1}{n-p})}$$

Où S^2 est l'estimateur non biaisé de σ^2

$$S^2 = \frac{\sum_{i=1}^{p}(n_i-1)S_i^2}{n-p}$$

avec :

$$n = \sum_{i=1}^{p} n_i \quad p \text{ échantillons à comparer}$$

S_i^2 la variance empirique du i-éme echantillon et :

$$S_i^2 = \frac{1}{n_i}\sum_{j=1}^{ni}(y_{ij} - \bar{y_i})^2$$

Sous l'hypothèse H_0 : $B_{obs} \sim \mathcal{X}^2_{p-1}$ ddl.

La région critique est :

$$B_{obs} > \mathcal{X}^2_{p-1,\alpha}$$

En cas d'absence d'homogénéité et de grande disparité entre les populations : si les résultats de l'ANOVA conduisent aux différences significatives entre les moyennes, ces différences ne peuvent pas être attribuées uniquement aux moyennes, car elles peuvent être dues aussi aux différences entre les variances. Pour contourner ce problème, nous devons respecter la condition d'homogénéité des variances.

1.1.2 Modélisation ANOVA2

Présentation du modèle complet

Le modèle complet d'analyse de variance à deux facteurs fixes s'écrire sous la forme suivante :

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{1.1}$$

avec :

- i=$\overline{1,p}$.
- j=$\overline{1,q}$.
- k=$\overline{1,r}$.

Où :

- y_{ijk} :la valeur de la variable quantitative à expliquer pour la i-ème modalité de F_1 et la j-ème modalité de F_2 case(ij).
- μ :la constante du modèle et aussi une moyenne.
- ε_{ijk} :l'erreur aléatoire.
- $\varepsilon_{ijk} \sim N(0, \sigma^2)$: hypothèse fondamentale de l'ANOVA.

Par conséquence :

$$\begin{cases} E(Y) = \mu \\ Var(Y) = Var(\varepsilon) = \sigma^2 \\ Y \sim N(\mu, \sigma^2) \end{cases}$$

$$\begin{aligned} E(Y) &= E(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}) \\ &= E(\mu) + E(\alpha_i) + E(\beta_j)E + ((\alpha\beta)_{ij}) + E(\varepsilon_{ijk}) \quad (E \text{ linéaire}) \\ E(Y) &= E(\mu) \\ &= \mu. \\ V(Y) &= V(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}) \\ &= V(\mu) + V(\alpha_i) + V(\beta_j)V + ((\alpha\beta)_{ij}) + V(\varepsilon_{ijk}) \quad (\text{l'indépendance des échantillons}) \\ V(Y) &= V(\varepsilon_{ijk}) \\ &= \sigma^2. \end{aligned}$$

Estimation des paramètres du modèle

Le modèle contient plusieurs paramètres qui nécessitent une meilleure estimation. Les méthodes d'estimation sont nombreuses et celle qui est utilisée pour l'ANOVA c'est la méthode des moindres carrés ordinaires car l'objectif est de minimiser la somme des carrés des résidus ou des erreurs aléatoires. Dans cette sous section, on présente les propriétés d'un meilleur estimateur et les différentes méthodes d'estimation.

Définition d'un estimateur

Soit $Y_1, ..., Y_n$ un échantillon aléatoire de n réalisations, on appelle estimateurs θ ; toute fonction

de observations,notée $\hat{\theta}$:

$$\hat{\theta} = f(Y_1, ..., Y_n).$$

$\hat{\theta}$ est une variable aléatoire de loi de probabilité qui dépend du paramètre inconnu θ ,et $\hat{\theta} \in \Theta$ .telle que Θ : l'ensemble des valeurs possible de θ [12]

b.Propriétés d'un estimateur

Définition d'un estimateur sans biais

On dit qu'un estimateur est sans biais si l'espérance mathématique de cet estimateur est égale au paramètre estimé : $E(\hat{\theta}) = \theta \quad \forall \quad \theta \in \Theta$ (ensemble des estimateurs).

Le biais d'un estimateur noté $B(\theta)$ est défini par :

$$\begin{aligned} B(\hat{\theta}) &= E(\hat{\theta} - \theta) \\ &= E(\hat{\theta}) - E(\theta) \\ &= E(\hat{\theta}) - \theta \end{aligned}$$

$\hat{\theta}$ est une estimateur sans biais du paramètre θ si : $E(\hat{\theta}) = \theta$ [5]

Définition d'un estimateur convergent

On dit qu'un estimateur $\hat{\theta}$ est convergent lorsque sa variance tend vers 0 quand n tend vers l'infini .

Tout estimateur sans biais dont la variance tend vers 0 est convergent :

$$\lim_{n \to +\infty} E(\hat{\theta}) = \theta$$

$$et$$

$$\lim_{n \to +\infty} Var(\hat{\theta}) = 0$$

[5]

Estimateur optimal

a- Précision d'un estimateur ou sa qualité

La précision d'un estimateur $\hat{\theta}$ se mesure par l'erreur quadratique moyenne

$$EQ(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + b_n^2(\theta)$$

avec

$$b_n^2(\theta) = E[\hat{\theta} - \theta]^2$$

[5]

b- Variance minimale

Parmi les estimateurs sans biais de θ, le plus précis est celui qui a une plus petite variance. Soient deux estimateurs sans biais $\hat{\theta}_1$ et $\hat{\theta}_2$. $\hat{\theta}_1$ est meilleur que $\hat{\theta}_2$ si $V(\hat{\theta}_2) \leq V(\hat{\theta}_2)$
Lorsque on pourrait trouver un troisième estimateur $\hat{\theta}_3$ayant une variance plus petite $V(\hat{\theta}_1)$ il faut poursuivre la recherche mais on ne peut pas améliorer indéfiniment un estimateur ! ! Nous allons voir comment peut-on régler ce problème. [5]

c- Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR)

Si Y prend ses valeurs dans un ensemble qui ne dépend pas de θ, si la densité $f(Y, \theta)$ est deux fois continûment dérivable par rapport à θ, et sous certaines conditions de régularité, tout estimateur $\hat{\theta}$ sans biais de θ dont la variance existe vérifié l'inégalité $FDCR$:

$$V(\hat{\theta}) \geq \frac{1}{I_{n(\theta)}} \quad \forall \theta \in \Theta$$

Où $I_n(\theta)$ est la quantité d'information de Fisher définie par :

$$I_{n(\theta)} = E\left(\frac{\partial lnL}{\partial \theta}\right)^2$$
$$E\left(-\frac{\partial^2 lnL}{\partial \theta^2}\right)$$

(L : est la vraisemblance).

Les conditions de régularité sont :

On suppose que l'ensemble des estimateurs Θ est un ensemble ouvert sur lequel la densité $f(Y, \theta)$ ne s'annule en aucun point y et est dérivable par rapport à θ. On suppose aussi que l'on peut intervenir dérivation par rapport à θ et intégration, et que la quantité d'information de Fisher est strictement positive. [5]

Définition d'un estimateur efficace

Un estimateur sans biais $\hat{\theta}$ est efficace si sa variance est égale à la borne inférieure de $FDCR$:

$V(\hat{\theta}) = \frac{1}{I_{n(\theta)}}$. [5]

Méthodes d'estimation courantes

a.Méthode des Moindres carrés

La méthode des moindres carrés consiste à estimer θ en minimisant la somme des carrés des résidus SCR telle que :

$$S(\hat{\theta}(y)) = min \sum_{i=1}^{n} (\hat{\varepsilon_i})^2 = min \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

C'est celle utilisée pour les modèles ANOVA . [4]
Il existe d'autres méthodes :

b.Méthode du Maximum de vraisemblance

Définition 1.1. *La statistique* $\omega \longmapsto \arg\max(\theta \longmapsto \prod_{i=1}^{n} f_\theta(X_i(\omega)))$ *s'appelle l'estimateur de maximum de vraisemblance de* θ .
$L : \theta \longmapsto \prod_{i=1}^{n} f_\theta(x_i)$ *s'appelle la fonction vraisemblance du modèle.*
$l : \theta \longmapsto \sum_{i=1}^{n} \log f_\theta(x_i)$ *s'appelle la fonction log- vraisemblance du modèle.*
pour obtenir l 'estimateur $\hat{\theta}$ *du de maximum de vraisemblance ,on maximise log- vraisemblance selon* θ *, on résolvant le système d'équation maximum de vraisemblance*

$$\frac{\partial}{\partial \theta_j} \ln(\theta_1, ..., \theta_k, y) = 0. \quad pour \quad j = \overline{1, k}$$

[12]

c. Méthode des moments

La méthode des moments consiste à estimer les paramètres inconnus en utilisant les moments d'ordre 1 et 2 : E(Y) et E(Y^2). Il s'agit de résoudre le système d'équations en égalant les moments théoriques aux moments empiriques en fonctions des paramètres inconnues. La solution des équations si elle existe et est unique, sera appelée estimateur obtenu par la méthode des moments . [12]

d.Méthode par Intervalles de confiance

Définition

Soit Y une variable aléatoire dont la loi dépend d'un paramètre réel θ inconnu et $\alpha \in [0, 1]$ un nombre donné. On appelle « intervalle de confiance » pour le paramètre θ , de niveau de

confiance $1-\alpha$, un intervalle qui a la probabilité $1-\alpha$ de contenir la vraie valeur du paramètre θ . [12]

1.1.3 Quelques éléments de la théorie des tests d'hypothèses

Hypothèses et risques d'erreurs

Les tests d'hypothèses se basent sur un certain nombre d'hypothèses concernant la nature de la population dont provient l'échantillon étudié (normalité de la variable, égalité des variances, ... etc).

Les étapes à suivre pour tester une hypothèse sont les suivantes :

- définir l'hypothèse nulle (notée H_0) à contrôler
- choisir un test statistique ou une statistique pour contrôler H_0
- définir la distribution de la statistique sous l'hypothèse « H_0 est réalisée »
- définir le niveau de signification du test ou région critique notée α
- calculer, à partir des données fournies par l'échantillon, la valeur de la statistique
- prendre une décision concernant l'hypothèse posée et faire une interprétation

Soit H_0 et H_1 deux hypothèses ; la première est appelée hypothèse nulle et la seconde H_1 est appelée hypothèse alternative.

Un test statistique est donc un mécanisme qui permet de choisir une hypothèse H_0 ou H_1 au vu des résultats trouvés.

a- Le risque de première espèce α

L'erreur de première espèce α consiste à rejeter l'hypothèse nulle H_0 alors qu'elle est vraie, soit rejeter à tort H_0 . Le risque d'erreur α est donc la probabilité que la valeur calculée de la statistique de test D appartienne à la région critique si H_0 est vraie. Dans ce cas H_0 est rejetée et H_1 est considérée comme vraie.

$$\alpha = P(rejeterH_0 \setminus H_0vraie)$$

ou bien :

$$\alpha = P(accepterH_1 \setminus H_1fausse)$$

La quantité $(1-\alpha)$ est la confiance du test.

La valeur α doit être fixée a priori par l'expérimentateur et jamais en fonction des données.

C'est un compromis entre le risque de conclure à tort et la faculté de conclure. Toutes choses étant égales par ailleurs, la région critique diminue lorsque α décroît et donc on rejette moins fréquemment H_0.

b- Le risque de seconde espèce β

L'erreur de seconde espèce β consiste à accepter l'hypothèse nulle H_0 alors qu'elle est fausse, soit accepter à tort H_0. L'erreur de seconde espèce β est la probabilité que la valeur calculée de la statistique D n'appartienne pas à la région critique si H_1 est vraie. Dans ce cas H_0 est acceptée et H_1 est considérée comme fausse. On écrit :

$$\beta = P(\text{rejeter}H_1 \setminus H_1 \text{ vraie})$$

ou bien :

$$\beta = P(\text{accepter}H_0 \setminus H_1 \text{ vraie})$$

ou bien :

$$\beta = P(\text{accepter}H_0 \setminus H_0 \text{ fausse})$$

Pour quantifier le risque β, il faut connaître la loi de probabilité de la statistique du test D sous l'hypothèse H_1, ce qui est un peu difficile. C'est pour cette raison l'erreur utilisée plus fréquemment est celle de première espèce.

c- La puissance du test

Les tests ne sont pas faits pour « démontrer » H_0 mais pour « rejeter » H_0. L'aptitude d'un test à rejeter H_0 alors qu'elle est fausse constitue la puissance du test. La puissance d'un test est notée : $(1-\beta)$; c'est la probabilité de rejeter l'hypothèse nulle alors qu'elle est fausse ou la probabilité d'accepter l'hypothèse alternative alors qu'elle est vraie.

$$(1-\beta) = P(\text{rejetter}H_0 \setminus H_0 \text{ est fausse})$$

ou

$$(1-\beta) = P(\text{accepter}H_1 \setminus H_1 \text{ est vraie})$$

La puissance d'un test augmente avec la taille de l'échantillon étudié à valeur de α constant. La puissance d'un test diminue lorsque α diminue.

La décision aboutira à choisir H_0 ou H_1 . Il y a donc quatre cas possibles schématisés dans le tableau ci-dessous avec les probabilités correspondantes :

Décision / Réalité	Accepter H_0	Rejeter H_0
H_0vraie	$1-\alpha$	α
H_1vraie	β	$1-\beta$

TABLE 1.1 – Différentes probabilités dans un test d'hypothèses

Dans le cas de l'ANOVA2, nous avons trois tests à réaliser pour le modèle complet :

- Si l'objectif est d'étudier l'effet du facteur F_1 sur la variable Y, les hypothèses sont :

► $Test1$ $\begin{cases} H_0 : \text{Le facteur } F_1 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, i \quad \alpha_i = 0 \\ H_1 : \text{Le facteur } F_1 \text{ a un effet significatif sur Y ie :} \exists\, i \quad \alpha_i \neq 0 \end{cases}$

- Si l'objectif est d'étudier l'effet du facteur F_2 sur la variable Y , les hypothèses sont :

► $Test2$ $\begin{cases} H_0 : \text{Le facteur } F_2 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, j \quad \beta_j = 0 \\ H_1 : \text{Le facteur } F_2 \text{ a un effet significatif sur Y ie :} \exists\, j \quad \beta_j \neq 0 \end{cases}$

- Si l'objectif est d'étudier l'effet simultané des deux facteurs F_1 et F_2 sur la variable Y, les hypothèses sont :

► $Test3$ $\begin{cases} H_0 : \text{ Il n'y a pas d'effet significatif d'interaction entre } F_1 \text{ et } F_2 \text{ ie : } \forall\, (i,j)\ (\alpha\beta)_{ij} = 0 \\ H_1 : \text{ Il y a un effet significatif d'interaction entre } F_1 \text{ et } F_2 \text{ ie } \exists\, (i,j)\ (\alpha\beta)_{ij} \neq 0 \end{cases}$

Statistique du test et règles de décision

Une fois les hypothèses du test sont posées, nous devons choisir la statistique pour le réaliser (pour ANOVA, voir paragraphes suivants). C'est en comparant la valeur de cette statistique observée dans l'échantillon à la sa valeur sous l'hypothèse H_0 que nous pourrons prendre une décision c'est-à-dire donner la conclusion du test.

Règle de décision 1 : Sous l'hypothèse H_0 et pour un seuil de signification 1-α fixé (0.05 par

défaut) :

• si la valeur de la statistique D calculée ou observée est supérieure à la valeur seuil de D théorique, alors l'hypothèse H_0 est rejetée au risque d'erreur α et l'hypothèse H_1 est acceptée au risque d'erreur α.

• si la valeur de la statistique D calculée est inférieure à la valeur seuil D théorique, alors l'hypothèse H_0 ne peut être rejetée au risque d'erreur α.

Règles de décision 2 : • si αobservé est supérieure ou égale à 0,05 , l'hypothèse H_0 est acceptée

car le risque d'erreur de rejeter H_0 alors qu'elle est vraie est trop important.

• si αobservé est inférieure à 0,05 l'hypothèse H_0 est rejetée car le risque d'erreur de rejeter H_0 alors qu'elle est vraie est très faible.
Les statistiques des différents tests seront présentées dans les paragraphes suivants ainsi que les règles de décisions.

1.2 Différents types de plans d'ANOVA à 2 facteurs fixes

Il existe plusieurs types de plans d'expériences de l'ANOVA à deux facteurs fixes. Il sont les suivants :

- ANOVA2 sans répétitions, cela signifie que l'on dispose d'une seule mesure de Y_{ij} par case (i, j).
- ANOVA2 avec répétitions égales, cela signifie que l'on dispose d'un nombre égale de mesures de Y_{ij} dans chaque case (i, j), noté k et k=1, ..., r.
- ANOVA2 avec répétitions inégales, cela signifie que l'on dispose d'un nombre différent de mesures de Y_{ij} dans chaque case (i, j), noté k et k=1,...n_{ij}.

1.2.1 ANOVA à 2 facteurs fixes sans répétition

On a deux facteurs fixes F_1 et F_2 respectivement à p modalités et q modalités, Y_{ij} la mesure de la réponse Y de chaque combinaison (i, j). La taille globale de l'échantillon est n telle que :

$n = p \times q$ On s'interroge sur l'effet que peut exercer les facteur 1 et 2 sur la variable étudiée Y ?

Présentation du modèle

Le modèle de l'analyse de variance à deux facteurs sans répétitions s'écrire sous la forme :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, avec \quad i = \overline{1,p},\ j = \overline{1,q}$$

tels que :

- y_{ij} :la valeur prise par la variable à expliquer Y pour le couple (ij).
- μ : la moyenne et c'est l'effet global.
- α_i : l'effet de la modalité i du premier facteur F_1.
- β_j : l'effet de la modalité j du deuxième facteur F_2.
- ε_{ij} : l'erreur aléatoire .

Avec hypothèse fondamentale de l'indépendance des erreurs aléatoires. Les termes d'erreur ne sont donc pas corrélés entre eux. L'hypothèse la plus forte est celle qui consiste à supposer que les erreurs suivent une loi normale centrée et de variance σ^2 :

$$\begin{cases} Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \quad si \quad (i,j) \neq (i',j') \\ \varepsilon_{ij} \sim N(0, \sigma^2) \ \ \forall i,j \end{cases}$$

Par conséquence :

$$\begin{cases} E(Y) = \mu \\ Var(Y) = Var(\varepsilon) = \sigma^2 \\ Y \sim N(\mu, \sigma^2) \end{cases}$$

avec contraintes :

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

Les valeurs que peut prendre la variable Y sont présentées dans le tableau suivant :

	facteur F_2						
facteur F_1	modalité	modalité 1	modalité 2	...	modalité j	...	modalité q
	modalité 1	y_{11}	y_{12}	...	y_{1j}	...	y_{1q}
	modalité 2	y_{21}	y_{22}	...	y_{2j}	...	y_{2q}
	⋮	⋮	⋮	...	⋮	...	⋮
	modalité i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{iq}
	⋮	⋮	⋮	...	⋮	...	⋮
	modalité p	y_{p1}	y_{p2}	...	y_{pj}	...	y_{qp}

TABLE 1.2 – Les données d'ANOVA2 sans répétition

Estimation des paramètres du modèle

Pour estimer les paramètres de ce modèle μ, α_i et β_j, on utilise la méthode des moindres carrés qui consiste à minimiser la somme des carrés des résidus ou erreurs aléatoires suivante :

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p} \sum_{j=1}^{q} \varepsilon_{ij}^2$$

qui est égale aussi à :

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p} \sum_{j=1}^{q} (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

Pour trouver les estimations, on résout le système aux équations suivantes

$$\begin{cases} \frac{\partial S(\mu,\alpha_i,\beta_j)}{\partial \mu} = 0 \quad ... \quad (1). \\ \frac{\partial S(\mu,\alpha_i,\beta_j)}{\partial \alpha_i} = 0 \quad ... \quad (2). \\ \frac{\partial S(\mu,\alpha_i,\beta_j)}{\partial \beta_j} = 0 \quad ... \quad (3). \end{cases}$$

En dérivant l'équation 1 par rapport à μ, on obtient $\hat{\mu}$:

$$-2 \sum_{i=1}^{p} \sum_{j=1}^{q} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0$$

$$\sum_{i=1}^{p} \sum_{j=1}^{q} y_{ij} = pq\hat{\mu} + q \sum_{i=1}^{p} \hat{\alpha}_i + p \sum_{j=1}^{q} \hat{\beta}_j$$

avec les contraintes posées précédemment :

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

on obtient :

$$\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij} = pq\hat{\mu}$$

d'où :

$$\hat{\mu} = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}$$

En dérivant l'équation 2 par rapport à α_i, on obtient $\hat{\alpha}_i$ (i fixe)

$$-2\sum_{j=1}^{q}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0$$
$$\sum_{j=1}^{q} y_{ij} = \sum_{j=1}^{q}\hat{\mu} + \sum_{j=1}^{q}\hat{\alpha}_i + \sum_{j=1}^{q}\hat{\beta}_j.$$

Sous la contrainte :

$$\sum_{j=1}^{q}\beta_j = 0$$

on obtient :

$$\sum_{j=1}^{q} y_{ij} = q\hat{\mu} + q\hat{\alpha}_i$$
$$\hat{\alpha}_i = \frac{1}{q}\sum_{j=1}^{q} y_{ij} - \hat{\mu}.$$

On refait la même opération et en dérivant l'équation 2 par rapport à α_2, puis α_3... pour trouver les estimateurs de β_2, puis β_3...etc.

En dérivant l'équation 3 par rapport à β_j, on obtient $\hat{\beta}_j$ (j fixe) :

$$-2\sum_{i=1}^{p}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0$$
$$\sum_{i=1}^{p} y_{ij} = \sum_{i=1}^{p}\hat{\mu} + \sum_{i=1}^{p}\hat{\alpha}_i + \sum_{i=1}^{p}\hat{\beta}_j.$$

Sous la contrainte :

$$\sum_{i=1}^{p} \alpha_i = 0$$

on obtient :

$$\sum_{i=1}^{p} y_{ij} = p\hat{\mu} + p\hat{\beta}_j$$
$$\hat{\beta}_j = \frac{1}{p}\sum_{i=1}^{p} y_{ij} - \hat{\mu}$$

On refait la même opération et en dérivant l'équation 3 par rapport à β_2, puis β_3... pour trouver les estimateurs de α_2, puis α_3...etc.

En généralisant, on a :

- $\hat{\mu} = \bar{y}$ (moyenne estimée totale)
- $\hat{\alpha}_i = (\bar{y_{i.}}$ - $\bar{y})$ (Effet estimé du F1 sur Y)
- $\hat{\beta}_j = (\bar{y_{.j}}$ - $\bar{y})$ (Effet estimé du F2 sur Y)

  tels que :

$$\bullet\ \bar{y}_{i.} = \frac{1}{q}\sum_{j=1}^{q} y_{ij}$$
$$\bullet\ \bar{y}_{.j} = \frac{1}{p}\sum_{i=1}^{p} y_{ij}$$
$$\bullet\ \bar{y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij} \quad (n = p \times q)$$
$$= \frac{1}{p}\sum_{i=1}^{p} \bar{y_{i.}}$$
$$= \frac{1}{q}\sum_{j=1}^{q} \bar{y_{.j}}$$

Avec : $\bar{y}_{i.}$ est la moyenne marginale de Y liée au F2.

$\bar{y}_{.j}$ est la moyenne marginale de Y liée au F1.

$\bar{y}$ est la moyenne totale.

Équation fondamentale de l'ANOVA2 sans répétition

a.Présentation de l'équation

La somme des carrés totaux (SCT) qui est une simple addition de la somme des carrés liée au facteur F_1 (SCF_{F_1}) et la somme des carrés liée au facteur F_2 (SCF_{F_2}) et la somme des carrées des écarts ou des résidus (SCR), représente l'équation fondamentale de l'ANOVA2 sans répétitions. Car les trois sources de variations sont les deux facteurs fixes F_1 et F_2 et les facteurs inconnus (résidus). Cette équation s'écrit :

$$SCT = SCF_{F_1} + SCF_{F_2} + SCR \text{ avec :}$$

$$SCT = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y})^2$$

$$SCF_{F_1} = \sum_{i=1}^{p} q(\bar{y_{i.}} - \bar{y})^2$$

$$SCF_{F_2} = \sum_{j=1}^{q} p(\bar{y_{.j}} - \bar{y})^2$$

$$SCR = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y_{i.}} - \bar{y_{.j}} + \bar{y})^2$$

b.Démonstration de l'équation

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\bar{y})^2 &= \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\hat{y}_{ij}+\hat{y}_{ij}-\bar{y})^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\hat{y}_{ij})^2+\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{y}_{ij}-\bar{y})^2 \\
&\quad + 2\sum_{i=1}^{p}\sum_{j=1}^{q}[(y_{ij}-\hat{y}_{ij})(\hat{y}_{ij}-\bar{y})] \\
\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\hat{y}_{ij})^2+\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{y}_{ij}-\bar{y})^2 &= \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\bar{y}-\bar{y_{i.}}+\bar{y}-\bar{y_{.j}}+\bar{y})^2 \\
&\quad + \sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}+\bar{y_{i.}}-\bar{y}+\bar{y_{.j}}-\bar{y}-\bar{y})^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\bar{y_{i.}}-\bar{y_{.j}}+\bar{y})^2+\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y_{i.}}-\bar{y})^2 \\
&\quad + \sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y_{.j}}-\bar{y})^2+2\sum_{i=1}^{p}\sum_{j=1}^{q}[(\bar{y_{i.}}-\bar{y})(\bar{y_{.j}}-\bar{y})] \\
2\sum_{i=1}^{p}\sum_{j=1}^{q}[(y_{ij}-\hat{y}_{ij})(\hat{y}_{ij}-\bar{y})] &= 2\sum_{i=1}^{p}\sum_{j=1}^{q}[(\varepsilon_{ij})(\hat{y}_{ij}-\bar{y})] \\
&= 2(\sum_{i=1}^{p}\sum_{j=1}^{q}\varepsilon_{ij}\hat{y}_{ij}+\sum_{i=1}^{p}\sum_{j=1}^{q}\varepsilon_{ij}\bar{y}_{ij}) \\
&= 0 \ \text{ car } \ (\sum_{i=1}^{p}\sum_{j=1}^{q}\varepsilon_{ij}=0) \\
\sum_{i=1}^{p}\sum_{j=1}^{q}\varepsilon_{ij} &= \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\bar{y_{i.}}-\bar{y_{.j}}+\bar{y}) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}y_{ij}-\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{y_{i.}}-\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{y_{.j}}+\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{y} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}y_{ij}-\sum_{j=1}^{q}p\bar{y}-\sum_{i=1}^{p}q\bar{y}-n\bar{y} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}y_{ij}-n\bar{y}-n\bar{y}+n\bar{y} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}y_{ij}-n\frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}y_{ij} \\
&= 0
\end{aligned}$$

$$\begin{aligned} 2\sum_{i=1}^{p}\sum_{j=1}^{q}[(\bar{y_{i.}}-\bar{y})(\bar{y_{.j}}-\bar{y})] &= 2[\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{y_{i.}}\bar{y_{.j}} - \sum_{i=1}^{p}\sum_{j=1}^{q}\bar{y_{i.}}\bar{y} \\ &\quad - \sum_{i=1}^{p}\sum_{j=1}^{q}\bar{y_{.j}}\bar{y} + \sum_{i=1}^{p}\sum_{j=1}^{q}\bar{y}] \\ &= 2[n\bar{y} - n\bar{y} - n\bar{y} + n\bar{y}] \\ &= 0 \end{aligned}$$

donc :

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{p}q(\bar{y_{i.}}-\bar{y})^2 + \sum_{j=1}^{q}p(\bar{y_{.j}}-\bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\bar{y_{i.}}-\bar{y_{.j}}+\bar{y})^2$$

Alors :

$$SCT = SCF_{F_1} + SCF_{F_2} + SCR$$

Les différents tests de Fischer

Dans le cas de l'ANOVA2 sans répétition, nous avons deux tests de Fischer à réaliser :

- Si l'objectif est d'étudier l'effet du facteur F_1 sur la variable Y, les hypothèses sont :

► $Test1$ $\begin{cases} H_0 : \text{Le facteur } F_1 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, i \quad \alpha_i = 0 \\ H_1 : \text{Le facteur } F_1 \text{ a un effet significatif sur Y ie :} \exists\, i \quad \alpha_i \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_1}}{p-1}}{\frac{SCR}{(p-1)(q-1)}} \sim F_{((p-1),(p-1)(q-1))}$$

avec (p-1),(p-1)(q-1) ddl.

- Si l'objectif est d'étudier l'effet du facteur F_2 sur la variable Y , les hypothèses sont :

► $Test2$ $\begin{cases} H_0 : \text{Le facteur } F_2 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, j \quad \beta_j = 0 \\ H_1 : \text{Le facteur } F_2 \text{ a un effet significatif sur Y ie :} \exists\, j \quad \beta_j \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_2}}{q-1}}{\frac{SCR}{(p-1)(q-1)}} \sim F_{((q-1),(p-1)(q-1))}$$

Tableau de variation de l'ANOVA2 sans répétition

Les résultats des tests sont généralement présentés sous forme d'un tableau de variances :

Source de variation	SC	ddl	MC	Statistique F
Facteur F_1	SCF_{F_1}	p-1	$MCF_{F_1} = \frac{SCF_{F_1}}{P-1}$	$Fc_{F_1} = \frac{MCF_{F_1}}{MCR}$
Facteur F_2	SCF_{F_2}	q-1	$MCF_{F_2} = \frac{SCF_{F_2}}{q-1}$	$Fc_{F_2} = \frac{MCF_{F_2}}{MCR}$
Résiduelle	SCR	(p-1)(q-1)	$MCR = \frac{SCR}{(p-1)(q-1)}$	/
Totale	SCT	(pq)-1	$MCT = \frac{SCT}{(pq)-1}$	/

TABLE 1.3 – tableau d'analyse de variance à 2 facteurs sans répétition

Règles de décision

Pour le test de l'effet de F_1 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_1} > F_{\alpha,(p-1),(p-1)(q-1)}$

et

Pour le test de l'effet de F_2 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_2} > F_{\alpha,(q-1),(p-1)(q-1)}$

1.2.2 ANOVA à 2 facteurs fixes avec répétitions égales sans interaction

Supposons deux facteurs F_1 et F_2 respectivement à p modalités et q modalités , y_{ijk} la mesure de réponse y de chaque combinaison (ijk). Le nombre total des observations est $n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$ et n_{ij} est le nombre d'observations pour la modalité i de facteur F_1 et la modalité j de facteur F_2 conjointement. Le nombre des observations est identique dans toutes les cases, noté r.

On s'interroge sur l'effet de F1 et F2 sur la variable étudiée Y, sans interaction .

Présentation du modèle

Le modèle de l'analyse da variance à deux facteurs avec répétitions égales et sans interaction s'écrit comme suit :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad i = \overline{1,p},\ j = \overline{1,q},\ k = \overline{1,r}$$

Tels que :

- y_{ijk} :la valeur prise par la variable à expliquer Y pour le couple (ij) pour K ème individu statistique.
- μ : la moyenne totale et c'est l'effet global.
- α_i : l'effet de la modalité i du premier facteur F_1.
- β_j : l'effet de la modalité j du deuxième facteur F_2.
- ε_{ijk} : l'erreur aléatoire .

Avec hypothèse fondamentale de l'indépendance des erreurs aléatoires. Les termes d'erreur ne sont donc pas corrélés entre eux. L'hypothèse la plus forte est celle qui consiste à supposer que les erreurs suivent une loi normale centrée et de variance σ^2 :

$$\begin{cases} Cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \quad si \quad (i,j,k) \neq (i',j',k') \\ \varepsilon_{ijk} \sim N(0,\sigma^2) \ \ \forall i,j,k \end{cases}$$

Par conséquence :

$$\begin{cases} E(Y) = \mu \\ Var(Y) = Var(\varepsilon) = \sigma^2 \\ Y \sim N(\mu, \sigma^2) \end{cases}$$

avec contraintes :

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

Les réalisations de la variable aléatoire Y sont présentées dans un tableau à double

entrées comme suit :

	facteur F_2						
facteur F_1	modalité	modalité 1	modalité 2	...	modalité j	...	modalité q
	modalité 1	$y_{111}, y_{112}, ..., y_{11r}$	$y_{121}, y_{122}...y_{12r}$	...	$y_{1j1}, y_{1j2}, ..., y_{1jr}$	...	$y_{1q1}, y_{1q2}..., y_{1qr}$
	modalité 2	$y_{211}, y_{212}..., y_{21r}$	$y_{221}, y_{222}, ..., y_{22r}$	...	$y_{2j1}, y_{2j2}, ...y_{2jr}$	...	$y_{2q1}, y_{2q2}, ..., y_{2qr}$
	⋮	⋮	⋮	...	⋮	...	⋮
	modalité i	$y_{i11}, y_{i12}, ..., y_{i1r}$	$y_{i21}, y_{i22}, ..., y_{i2r}$	...	$y_{ij1}, y_{ij2}, ..., y_{ijr}$	...	$y_{iq1}, y_{iq2}, ..., y_{iqr}$
	⋮	⋮	⋮	...	⋮	...	⋮
	modalité p	$y_{p11}, y_{p12}, ..., y_{p1r}$	$y_{p21}, y_{p22}, ..., y_{p2r}$	...	$y_{pj1}, y_{pj2}, ..., y_{pjr}$	...	$y_{qp1}, y_{pq2}, ..., y_{pqr}$

TABLE 1.4 – les données d'ANOVA2 avec répétitions égales

Estimation des paramètres du modèle

Pour estimer les paramètres du modèle, on utilise la méthode des moindres carrés. On suit les mêmes étapes que la sous section précédente (de la page 15 à la page17) : On cherche à minimiser $S(\mu, \alpha_i, \beta_j)$:

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}\varepsilon_{ijk}^2$$

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \mu - \alpha_i - \beta_j)^2$$

En dérivant par rapport à μ, on obtient $\hat{\mu}$:

$$-2\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0$$

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}y_{ijk} = pqr\hat{\mu} + qr\sum_{i=1}^{p}\hat{\alpha}_i + pr\sum_{j=1}^{q}\hat{\beta}_j$$

avec les contraintes suivantes

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

on obtient :

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} = pqr\hat{\mu}$$
$$\hat{\mu} = \frac{1}{pqr}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk}$$

En dérivant par rapport à α_i, on obtient $\hat{\alpha}_i$ (i fixe ; le ième paramètre)

$$-2\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0$$
$$\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} = qr\hat{\mu} + rq\hat{\alpha}_i + r\sum_{j=1}^{q}\hat{\beta}_j.$$

sous la contrainte :

$$\sum_{j=1}^{q}\beta_j = 0$$

on obtient :

$$\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} = qr\hat{\mu} + qr\hat{\alpha}_i$$

D'où :

$$\hat{\alpha}_i = \frac{1}{qr}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} - \hat{\mu}.$$

En dérivant par rapport à β_j, on obtient $\hat{\beta}_j$ (j fixe , le jème paramètre)

$$-2\sum_{i=1}^{p}\sum_{k=1}^{r}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0$$
$$\sum_{i=1}^{p}\sum_{k=1}^{r} y_{ij} = pr\hat{\mu} + r\sum_{i=1}^{p}\hat{\alpha}_i + pr\hat{\beta}_j.$$

Sous la contrainte :

$$\sum_{i=1}^{p}\alpha_i = 0$$

on obtient :

$$\sum_{i=1}^{p}\sum_{k=1}^{r} y_{ijk} = pr\hat{\mu} + pr\hat{\beta}_1$$

$$\hat{\beta}_j = \frac{1}{pr}\sum_{i=1}^{p}\sum_{k=1}^{r} y_{ijk} - \hat{\mu}$$

Nous trouvons :

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = (\bar{y_{i..}} - \bar{y})$
- $\hat{\beta}_j = (\bar{y_{.j.}} - \bar{y})$

Tels que

$$\bullet\ \bar{y}_{i..} = \frac{1}{qr}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk}.$$

$$\bullet\ \bar{y}_{.j.} = \frac{1}{pr}\sum_{i=1}^{p}\sum_{k=i}^{r} y_{ijk}.$$

$$\bullet\ \bar{y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} \qquad (n = p \times q \times r)$$

Avec : $\bar{y}_{i..}$ est la moyenne marginale de Y liée au F2.

$\bar{y}_{.j.}$ est la moyenne marginale de Y liée au F1.

$\bar{y}$ est la moyenne totale estimée.

l'équation fondamentale de l'analyse de variance

L'équation fondamentale de l'ANOVA à 2 facteurs fixes à répétitions égales sans interaction est : $SCT = SCF_{F_1} + SCF_{F_2} + SCR$

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y})^2 = \sum_{i=1}^{p} qr(\bar{y_{i..}} - \bar{y})^2 + \sum_{j=1}^{q} pr(\bar{y_{.j.}} - \bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y_{i.}} - \bar{y_{.j}} + \bar{y})^2$$

$$SCT = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y})^2$$
$$SCF_{F_1} = \sum_{i=1}^{p} qr(\bar{y_{i..}} - \bar{y})^2.$$
$$SCF_{F_2} = \sum_{j=1}^{q} pr(\bar{y_{.j.}} - \bar{y})^2.$$
$$SCR = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y_{i.}} - \bar{y_{.j}} + \bar{y})^2.$$

Sa démonstration est analogue à celle de la sous section précédente.

Les différents tests de Fischer

Dans le cas de l'ANOVA2 avec répétitions égales sans interaction, nous avons aussi deux tests de Fischer à réaliser :

- Si l'objectif est d'étudier l'effet du facteur F_1 sur la variable Y, les hypothèses sont :

► $Test1$ $\begin{cases} H_0 : \text{Le facteur } F_1 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, i \quad \alpha_i = 0 \\ H_1 : \text{Le facteur } F_1 \text{ a un effet significatif sur Y ie :} \exists\, i \quad \alpha_i \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_1}}{p-1}}{\frac{SCR}{pq(r-1)}} \sim F_{((p-1),pq(r-1))}$$

- Si l'objectif est d'étudier l'effet du facteur F_2 sur la variable Y, les hypothèses sont :

► $Test2$ $\begin{cases} H_0 : \text{Le facteur } F_2 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, j \quad \beta_j = 0 \\ H_1 : \text{Le facteur } F_2 \text{ a un effet significatif sur Y ie :} \exists\, j \quad \beta_j \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_2}}{q-1}}{\frac{SCR}{pq(r-1)}} \sim F_{((q-1),pq(r-1))}$$

Tableau d'ANOVA 2 avec répétitions égales sans interaction

Les résultats des tests sont généralement présentés sous forme d'un tableau de variances :

Source de variation	SC	ddl	MC	Statistique F
Facteur F_1	SCF_{F_1}	p-1	$MCF_{F_1} = \frac{SCF_{F_1}}{P-1}$	$Fc_{F_1} = \frac{MCF_{F_1}}{MCR}$
Facteur F_2	SCF_{F_2}	q-1	$MCF_{F_2} = \frac{SCF_{F_2}}{q-1}$	$Fc_{F_2} = \frac{MCF_{F_2}}{MCR}$
Résiduelle	SCR	pq(r-1)	$MCR = \frac{SCR}{pq(r-1)}$	/
Totale	SCT	pqr-1	$MCT = \frac{SCT}{(pqr)-1}$	/

Règles de décision

Pour le test de l'effet de F_1 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_1} > F_{\alpha,(p-1),pq(r-1)}$

et

Pour le test de l'effet de F_2 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_2} > F_{\alpha,(q-1),pq(r-1)}$

1.2.3 ANOVA à 2 facteurs fixes avec répétitions égales avec interaction

Supposons deux facteurs F_1 et F_2 respectivement à p modalités et q modalités , y_{ijk} la mesure de réponse y de chaque combinaison (ijk). Le nombre total des observations est $n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$ et n_{ij} est le nombre d'observations pour la modalité i de facteur F_1 et la modalité j de facteur F_2 conjointement. Le nombre de'observations est identique dans toutes les cases, noté r.
On s'interroge sur l'effet de F1 et F2 sur la variable étudiée Y, mais aussi sur l'effet de ces deux facteurs simultanément sur Y (interaction)

Présentation du modèle

Le modèle de l'analyse da variance à deux facteurs avec répétitions égales et avec interaction s'écrit comme suit :

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad i = \overline{1,p},\ j = \overline{1,q},\ k = \overline{1,r}$$

Tels que :

- y_{ijk} :la valeur prise par la variable à expliquer Y pour le couple (ij) pour K ème individus statistique.
- μ : la moyenne totale et c'est l'effet global.
- α_i : l'effet de la modalité i du premier facteur F_1.
- β_j : l'effet de la modalité j du deuxième facteur F_2.
- $(\alpha\beta)_{ij}$:l'effet de l'interaction du facteur F_1 et du facteur F_2 sur la variable Y.
- ε_{ijk} : l'erreur aléatoire .

Avec hypothèse fondamentale de l'indépendance des erreurs aléatoires. Les termes d'erreur ne sont donc pas corrélés entre eux. L'hypothèse la plus forte est celle qui consiste à supposer que les erreurs suivent une loi normale centrée et de variance σ^2 :

$$\begin{cases} Cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \quad si \quad (i,j,k) \neq (i',j',k') \\ \varepsilon_{ijk} \sim N(0,\sigma^2) \ \ \forall i,j,k \end{cases}$$

Par conséquence :

$$\begin{cases} E(Y) = \mu \\ Var(Y) = Var(\varepsilon) = \sigma^2 \\ Y \sim N(\mu,\sigma^2) \end{cases}$$

avec les contraintes :

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

$$\sum_{i=1}^{p} (\alpha\beta)_{ij} = 0 \quad j = \overline{1,q}$$

$$\sum_{j=1}^{q} (\alpha\beta)_{ij} = 0 \quad i = \overline{1,p}$$

Les réalisations de la variable aléatoire Y sont présentées dans un tableau à double entrées comme le tableau ci-dessus (voir ANOVA2 à répétitions égales sans interaction).

Estimation des paramètres du modèle

Pour estimer les paramètres du modèle avec interaction, on utilise également la méthode des moindres carrés : on minimise la somme des carrés des résidus :

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}\varepsilon_{ijk}^2$$

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2$$

L'estimateurs $\hat{\mu}, \hat{\alpha}$ et $\hat{\beta}$ on les retrouve de la même manière que dans le modèle à répétitions égales et sans interaction. Et pour estimer l'effet conjoint des deux facteurs sur Y :

On dérive par rapport à $(\alpha\beta)_{ij}$ pour obtenir $\widehat{(\alpha\beta)}_{ij}$: (i fixe et j fixe)

$$-2\sum_{k=1}^{r}(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \widehat{(\alpha\beta)}_{ij}) = 0$$

$$\sum_{k=1}^{r} y_{ijk} = \sum_{k=1}^{r}\hat{\mu} + \sum_{k=1}^{r}\hat{\alpha}_i + \sum_{k=1}^{r}\hat{\beta}_j + \sum_{k=1}^{r}\widehat{(\alpha\beta)}_{ij}$$

$$\sum_{k=1}^{r} y_{ijk} = r\hat{\mu} + r\hat{\alpha}_i + r\hat{\beta}_j + r\widehat{(\alpha\beta)}_{ij}$$

on obtient :

$$\widehat{(\alpha\beta)}_{ij} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$$

$$\widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}$$

nous trouvons :

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = (\bar{y}_{i..}$ - $\bar{y})$
- $\hat{\beta}_j = (\bar{y}_{.j.}$ - $\bar{y})$
- $\widehat{(\alpha\beta)}_{ij}$=$(\bar{y}_{ij.}$- $\bar{y}_{i..}$- $\bar{y}_{.j.}$+$\bar{y})$

Tels que :

$$\bullet\ \bar{y}_{ij.} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk},$$
$$\bullet\ \bar{y}_{i..} = \frac{1}{qr}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk},$$
$$\bullet\ \bar{y}_{.j.} = \frac{1}{pr}\sum_{i=1}^{p}\sum_{k=i}^{r} y_{ijk},$$
$$\bullet\ \bar{y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk} \qquad (n = p \times q \times r)$$

Avec : $\bar{y}_{ij.}$ est la moyenne estimée de Y dans la case ij.

$\bar{y}_{i..}$ est la moyenne marginale de Y liée au F2.

$\bar{y}_{.j.}$ est la moyenne marginale de Y liée au F1.

$\bar{y}$ est la moyenne totale estimée.

l'équation fondamentale de l'ANOVA2 à répétitions égales et avec interaction

L'équation fondamentale de l'ANOVA à 2 facteurs fixes à répétitions égales avec interaction est :

donc :

$$SCT = SCF_{F_1} + SCF_{F_2} + SCF_{F_2} + SCF_{F_1 \times F_2} + SCR.$$

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y})^2 = \sum_{i=1}^{p} qr(\bar{y_{i..}} - \bar{y})^2 + \sum_{j=1}^{q} pr(\bar{y_{.j.}} - \bar{y})^2 + r\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y_{ij.}} - \bar{y_{i..}} - \bar{y_{.j.}} + \bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y_{ij.}})^2$$

Avec :

$$SCT = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y})^2$$

$$SCF_{F_1} = \sum_{i=1}^{p} qr(\bar{y_{i..}} - \bar{y})^2$$

$$SCF_{F_2} = \sum_{j=1}^{q} pr(\bar{y_{.j.}} - \bar{y})^2$$

$$SCF_{F_1 \times F_2} = r\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y_{ij.}} - \bar{y_{i..}} - \bar{y_{.j.}} + \bar{y})^2$$

$$SCR = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y_{ij.}})^2$$

Les différents tests de Fischer

Dans le cas de l'ANOVA2 avec répétitions égales avec interaction, nous avons trois tests de Fischer à réaliser :

- Si l'objectif est d'étudier l'effet du facteur F_1 sur la variable Y, les hypothèses sont :

► $Test1$ $\begin{cases} H_0 : \text{Le facteur } F_1 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, i \quad \alpha_i = 0 \\ H_1 : \text{Le facteur } F_1 \text{ a un effet significatif sur Y ie :} \exists\, i \quad \alpha_i \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_1}}{p-1}}{\frac{SCR}{pq(r-1)}} \sim F_{((p-1),pq(r-1))}$$

- Si l'objectif est d'étudier l'effet du facteur F_2 sur la variable Y , les hypothèses sont :

► $Test2$ $\begin{cases} H_0 : \text{Le facteur } F_2 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, j \quad \beta_j = 0 \\ H_1 : \text{Le facteur } F_2 \text{ a un effet significatif sur Y ie :} \exists\, j \quad \beta_j \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_2}}{q-1}}{\frac{SCR}{pq(r-1)}} \sim F_{((q-1),pq(r-1))}$$

- Si l'objectif est d'étudier l'effet simultané des deux facteurs sur la variable Y, les hypothèses sont :

► $Test3$ $\begin{cases} H_0 : \text{Il n'y a pas d'effet intéraction significatif sur Y ie :} \forall\, i \quad \alpha_i = 0 \\ H_1 : \text{Il y a un effet intéraction significatif sur Y ie : } \exists\, (i,j) \quad (\alpha\beta)_{ij} \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_1 \times F_2}}{(p-1)(q-1)}}{\frac{SCR}{pq(r-1)}} \sim F_{((p-1)(q-1),pq(r-1))}$$

Tableau d'analyse de variance

Les résultats des tests sont généralement présentés sous forme d'un tableau de variances :

Source de variation	SC	ddl	MC	Statistique F
Facteur F1	SCF_{F1}	p-1	$MCF_{F1} = \frac{SCF_{F1}}{P-1}$	$Fc_{F1} = \frac{MCF_{F1}}{MCR}$
Facteur F2	SCF_{F2}	q-1	$MCF_{F2} = \frac{SCF_{F2}}{q-1}$	$Fc_{F2} = \frac{MCF_{F2}}{MCR}$
Interaction F1F2	SCF_{F1F2}	(p-1)(q-1)	$MCF_{F1F2} = \frac{SCF_{F1F2}}{(p-1)(q-1)}$	$Fc_{F1F2} = \frac{MCF_{F1F2}}{MCR}$
Résiduelle	SCR	pq(r-1)	$MCR = \frac{SCR}{pq-(r-1)}$	/
Totale	SCT	n-1	$MCT = \frac{SCT}{n-1}$	/

Règles de décision

Pour le test de l'effet de F_1 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_1} > F_{\alpha,(p-1),pq(r-1)}$

et

Pour le test de l'effet de F_2 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_2} > F_{\alpha,(q-1),pq(r-1)}$

et

Pour le test de l'effet de l'interaction sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_{12}} > F_{\alpha,(q-1)(q-1),pq(r-1)}$

1.2.4 ANOVA à 2 facteurs fixes avec répétitions inégales sans interaction

Supposons deux facteurs F_1 et F_2 respectivement à p modalités et q modalités, y_{ijk} la mesure de réponse y de chaque combinaison (ijk). Le nombre total des observations est $n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$ et n_{ij} est le nombre d'observations pour la modalité i de facteur F_1 et la modalité j de facteur F_2 conjointement. Le nombre des observations est différent d'une case à l'autre.
On s'interroge sur l'effet de F1 et F2 sur la variable étudiée Y, sans interaction ?
Cette sous section concorde avec la sous section ci-dessus de l'ANOVA à 2 facteurs fixes avec répétitions égales sans interaction. Seul élément qui diffère est le nombre d'observations par case et ceci influence évidemment l'équation fondamentale.

Présentation du modèle

Le modèle de l'analyse da variance à deux facteurs avec répétitions inégales et sans interaction s'écrit comme suit :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad i = \overline{1,p},\ j = \overline{1,q},\ k = \overline{1,n_{ij}}$$

Tels que :

- y_{ijk} :la valeur prise par la variable à expliquer Y pour le couple (ij) pour le K ème individu statistique.
- μ : la moyenne totale et c'est l'effet global.
- α_i : l'effet de la modalité i du premier facteur F_1.
- β_j : l'effet de la modalité j du deuxième facteur F_2.
- ε_{ijk} : l'erreur aléatoire .
- k : le nombre d'observations par case (ij).

Avec hypothèse fondamentale de l'indépendance des erreurs aléatoires. Les termes d'erreur ne sont donc pas corrélés entre eux. L'hypothèse la plus forte est celle qui consiste à supposer

que les erreurs suivent une loi normale centrée et de variance σ^2 :

$$\begin{cases} Cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \quad si \quad (i,j,k) \neq (i',j',k') \\ \varepsilon_{ijk} \sim N(0, \sigma^2) \ \ \forall i,j,k \end{cases}$$

Par conséquence :

$$\begin{cases} E(Y) = \mu \\ Var(Y) = Var(\varepsilon) = \sigma^2 \\ Y \sim N(\mu, \sigma^2) \end{cases}$$

avec contraintes :

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

Les réalisations de la variable aléatoire Y sont présentées dans un tableau à double entrées comme suit :

		facteur F_2					
facteur F_1	modalité	modalité 1	modalité 2	...	modalité j	...	modalité q
	modalité 1	$y_{111}, y_{112}, ..., y_{11k}$	$y_{121}, y_{122} ... y_{12k}$	...	$y_{1j1}, y_{1j2}, ..., y_{1jk}$	...	$y_{1q1}, y_{1q2} ..., y_{1qk}$
	modalité 2	$y_{211}, y_{212} ..., y_{21k}$	$y_{221}, y_{222}, ..., y_{22k}$	...	$y_{2j1}, y_{2j2}, ... y_{2jk}$	...	$y_{2q1}, y_{2q2}, ..., y_{2qk}$
	⋮	⋮	⋮	...	⋮	...	⋮
	modalité i	$y_{i11}, y_{i12}, ..., y_{i1k}$	$y_{i21}, y_{i22}, ..., y_{i2k}$	...	$y_{ij1}, y_{ij2}, ..., y_{ijk}$	...	$y_{iq1}, y_{iq2}, ..., y_{iqk}$
	⋮	⋮	⋮	...	⋮	...	⋮
	modalité p	$y_{p11}, y_{p12}, ..., y_{p1k}$	$y_{p21}, y_{p22}, ..., y_{p2k}$	...	$y_{pj1}, y_{pj2}, ..., y_{pjk}$	...	$y_{qp1}, y_{pq2}, ..., y_{pqk}$

TABLE 1.5 – Les données d'ANOVA2 avec répétitions inégales

Estimation des paramètres du modèle

Pour estimer les paramètres du modèle, on utilise la méthode des moindres carrés. On suit les mêmes étapes que la sous section précédente (de la page 15 à la page17) : On cherche à

minimiser $S(\mu, \alpha_i, \beta_j)$:

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} \varepsilon_{ijk}^2$$

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} (y_{ijk} - \mu - \alpha_i - \beta_j)^2$$

On obtient les résultats suivants :

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = (\bar{y_{i..}} - \bar{y})$
- $\hat{\beta}_j = (\bar{y_{.j.}} - \bar{y})$

Tels que :

$$\bullet\ \bar{y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk},$$

$$\bullet\ \bar{y}_{i..} = \frac{1}{q} \sum_{j=1}^{q} \bar{y_{ij.}},$$

$$\bullet\ \bar{y}_{.j.} = \frac{1}{p} \sum_{i=1}^{p} \bar{y_{ij.}},$$

$$\bullet\ \bar{y} = \frac{1}{p} \sum_{i=1}^{p} \bar{y_{i..}}$$

$$= \frac{1}{q} \sum_{j=1}^{q} \bar{y_{.j.}}$$

Avec : $\bar{y}_{ij.}$ est la moyenne de Y dans la case ij

$\bar{y}_{i..}$ est la moyenne marginale de Y liée au F2.

$\bar{y}_{.j.}$ est la moyenne marginale de Y liée au F1.

$\bar{y}$ est la moyenne totale estimée.

l'équation d'analyse de variance

la variabilité se décompose de la façon suivante :

$$(y_{ijk} - \bar{y}) = (\bar{y_{i..}} - \bar{y}) + (\bar{y_{.j.}} - \bar{y}) + (y_{ijk} - \bar{y_{i..}} - \bar{y_{.j.}} + \bar{y})$$

Et l'équation fondamentale de l'ANOVA à 2 facteurs à répétitions inégales sans interaction est presque identique à celle présentée pour le modèle de l'ANOVA à 2 facteurs à répétitions égales sans interaction. Comme les observations ou le nombre de mesures de Y peut différer entre les cases, on conserve les triple sommes comme suit :

$$SCT = SCF_{F_1} + SCF_{F_2} + SCF_{F_2} + SCR.$$

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y})^2 = \sum_{i=1}^{p}\sum_{k=1}^{n_{ij}} q(\bar{y_{i..}}-\bar{y})^2 + \sum_{j=1}^{q}\sum_{k=1}^{n_{ij}} p(\bar{y_{.j.}}-\bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y_{i.}}-\bar{y_{.j}}+\bar{y})^2$$

Avec :

$$SCT = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y})^2$$

$$SCF_{F_1} = \sum_{i=1}^{p}\sum_{k=1}^{n_{ij}} q(\bar{y_{i..}}-\bar{y})^2.$$

$$SCF_{F_2} = \sum_{j=1}^{q}\sum_{k=1}^{n_{ij}} p(\bar{y_{.j.}}-\bar{y})^2.$$

$$SCR = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y_{i.}}-\bar{y_{.j}}+\bar{y})^2.$$

Les différents tests de Fischer

Dans le cas de l'ANOVA2 avec répétitions inégales sans interaction, nous avons aussi deux tests de Fischer à réaliser :

- Si l'objectif est d'étudier l'effet du facteur F_1 sur la variable Y, les hypothèses sont :

► $Test1$ $\begin{cases} H_0 : \text{Le facteur } F_1 \text{ n'a pas d'effet significatif sur Y ie :} \forall\ i \quad \alpha_i = 0 \\ H_1 : \text{Le facteur } F_1 \text{ a un effet significatif sur Y ie :} \exists\ i \quad \alpha_i \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_1}}{p-1}}{\frac{SCR}{n-pq}} \sim F_{((p-1),(n-pq))}$$

- Si l'objectif est d'étudier l'effet du facteur F_2 sur la variable Y , les hypothèses sont :

► $Test2$ $\begin{cases} H_0 : \text{Le facteur } F_2 \text{ n'a pas d'effet significatif sur Y ie :} \forall\ j \quad \beta_j = 0 \\ H_1 : \text{Le facteur } F_2 \text{ a un effet significatif sur Y ie :} \exists\ j \quad \beta_j \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_2}}{q-1}}{\frac{SCR}{n-pq}} \sim F_{((q-1),n-pq)}$$

Tableau d'analyse de variance

Les résultats se représentent dans un tableau de variation :

Source de variation	SC	ddl	MC	Statistique F
Facteur F_1	SCF_{F_1}	p-1	$MCF_{F_1} = \frac{SCF_{F_1}}{p-1}$	$Fc_{F_1} = \frac{MCF_{F_1}}{MCR}$
Facteur F_2	SCF_{F_2}	q-1	$MCF_{F_2} = \frac{SCF_{F_2}}{q-1}$	$Fc_{F_2} = \frac{MCF_{F_2}}{MCR}$
Résiduelle	SCR	n-pq	$MCR = \frac{SCR}{n-pq}$	/
Totale	SCT	n-1	$MCT = \frac{SCT}{n-1}$	/

Règles de décision

Pour le test de l'effet de F_1 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_1} > F_{\alpha,(p-1),n-pq}$

et

Pour le test de l'effet de F_2 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_2} > F_{\alpha,(q-1),n-pq}$

1.2.5 ANOVA à 2 facteurs fixes avec répétitions inégales avec interaction

Dans cette sous section, on s'interesse aussi à l'effet d'interaction sur la variable à expliquer Y. Le nombre des observations est différent d'une case à l'autre.

Cette sous section concorde avec la sous section ci-dessus de l'ANOVA à 2 facteurs fixes avec répétitions égales avec interaction. Seul élément qui diffère est le nombre d'observations par case et ceci influence évidemment l'équation fondamentale.

Présentation du modèle

Le modèle de l'analyse da variance à deux facteurs avec répétitions inégales et avec interaction s'écrit comme suit :

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad i = \overline{1,p},\ j = \overline{1,q},\ k = \overline{1,n_{ij}}$$

Tels que :

- y_{ijk} :la valeur prise par la variable à expliquer Y pour le couple (ij) pour K ème individus statistique.
- μ : la moyenne totale et c'est l'effet global.
- α_i : l'effet de la modalité i du premier facteur F_1.
- β_j : l'effet de la modalité j du deuxième facteur F_2.
- $(\alpha\beta)_{ij}$:l'effet de l'interaction du facteur F_1 et du facteur F_2 sur la variable Y.
- ε_{ijk} : l'erreur aléatoire .

Avec hypothèse fondamentale de l'indépendance des erreurs aléatoires. Les termes d'erreur ne sont donc pas corrélés entre eux. L'hypothèse la plus forte est celle qui consiste à supposer que les erreurs suivent une loi normale centrée et de variance σ^2 :

$$\begin{cases} Cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \quad si \quad (i,j,k) \neq (i',j',k') \\ \varepsilon_{ijk} \sim N(0,\sigma^2) \ \ \forall i,j,k \end{cases}$$

avec les contraintes :

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

$$\sum_{i=1}^{p} (\alpha\beta)_{ij} = 0 \quad j = \overline{1,q}$$

$$\sum_{j=1}^{q} (\alpha\beta)_{ij} = 0 \quad i = \overline{1,p}$$

Les réalisations de la variable aléatoire Y sont présentées dans un tableau à double entrées comme le tableau ci-dessus (ANOVA2 à répétitions inégales sans interaction).

Estimation des paramètres du modèle

Pour estimer les paramètres du modèle ANOVA à 2 facteurs à répétitions inégales et avec interaction, on utilise également la méthode des moindres carrés : on minimise la somme des carrés des résidus :

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} \varepsilon_{ijk}^2$$

$$S(\mu, \alpha_i, \beta_j) = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} (y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2$$

L'estimateurs $\hat{\mu}, \hat{\alpha}$ et $\hat{\beta}$ on les retrouve de la même manière que dans le modèle à répétitions égales et sans interaction.
nous trouvons :

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = (\bar{y}_{i..}$ - $\bar{y})$
- $\hat{\beta}_j = (\bar{y}_{.j.}$ - $\bar{y})$
- $\widehat{(\alpha\beta)}_{ij}$=$(\bar{y}_{ij.}$- $\bar{y}_{i..}$- $\bar{y}_{.j.}$+$\bar{y})$

Tels que :

$$\bullet\ \bar{y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk},$$

$$\bullet\ \bar{y}_{i..} = \frac{1}{q} \sum_{j=1}^{q} \bar{y_{ij}}.,$$

$$\bullet\ \bar{y}_{.j.} = \frac{1}{p} \sum_{i=1}^{p} \bar{y_{ijk}},$$

$$\bullet\ \bar{y} = \frac{1}{p} \sum_{i=1}^{p} \bar{y_{i}}..$$

$$= \frac{1}{q} \sum_{j=1}^{q} \bar{y_{.j}}.$$

Avec : $\bar{y}_{ij.}$ est la moyenne estimée de Y dans la case ij.
$\bar{y}_{i..}$ est la moyenne marginale de Y liée au F2.
$\bar{y}_{.j.}$ est la moyenne marginale de Y liée au F1.

$\bar{y}$ est la moyenne totale estimée.

l'équation fondamentale de l'ANOVA2 à répétitions inégales et avec interaction

L'équation fondamentale de l'ANOVA à 2 facteurs fixes à répétitions inégales avec interaction est : donc :

$$SCT = SCF_{F_1} + SCF_{F_2} + SCF_{F_1 \times F_2} + SCR.$$

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y})^2 = \sum_{i=1}^{p}\sum_{k=1}^{n_{ij}} q(\bar{y_{i..}}-\bar{y})^2 + \sum_{j=1}^{q}\sum_{k=1}^{n_{ij}} p(\bar{y_{.j.}}-\bar{y})^2 + r\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(\bar{y_{ij.}}-\bar{y_{i..}}-\bar{y_{.j.}}+\bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y_{ij.}})^2$$

Avec :

$$SCT = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y})^2$$

$$SCF_{F_1} = \sum_{i=1}^{p}\sum_{k=1}^{n_{ij}} q(\bar{y_{i..}}-\bar{y})^2$$

$$SCF_{F_2} = \sum_{j=1}^{q}\sum_{k=1}^{n_{ij}} p(\bar{y_{.j.}}-\bar{y})^2$$

$$SCF_{F_1 \times F_2} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(\bar{y_{ij.}}-\bar{y_{i..}}-\bar{y_{.j.}}+\bar{y})^2$$

$$SCR = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y_{ij.}})^2$$

Les différents tests de Fischer

Dans le cas de l'ANOVA2 avec répétitions inégales avec interaction, nous avons trois tests de Fischer à réaliser :

- Si l'objectif est d'étudier l'effet du facteur F_1 sur la variable Y, les hypothèses sont :

► $Test1$ $\begin{cases} H_0 : \text{Le facteur } F_1 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, i \quad \alpha_i = 0 \\ H_1 : \text{Le facteur } F_1 \text{ a un effet significatif sur Y ie :} \exists\, i \quad \alpha_i \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_1}}{p-1}}{\frac{SCR}{n-pq)}} \sim F_{((p-1),n-pq)}$$

- Si l'objectif est d'étudier l'effet du facteur F_2 sur la variable Y , les hypothèses sont :

► $Test2$ $\begin{cases} H_0 : \text{Le facteur } F_2 \text{ n'a pas d'effet significatif sur Y ie :} \forall\, j \quad \beta_j = 0 \\ H_1 : \text{Le facteur } F_2 \text{ a un effet significatif sur Y ie :} \exists\, j \quad \beta_j \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_2}}{q-1}}{\frac{SCR}{n-pq}} \sim F_{((q-1),n-pq)}$$

- Si l'objectif est d'étudier l'effet simultané des deux facteurs sur la variable Y, les hypothèses sont :

► $Test3$ $\begin{cases} H_0 : \text{Il n'y a pas d'effet intéraction significatif sur Y ie :} \forall\, i \quad \alpha_i = 0 \\ H_1 : \text{Il y a un effet intéraction significatif sur Y ie : } \exists\, (i,j) \quad (\alpha\beta)_{ij} \neq 0 \end{cases}$

Sous l'hypothèse nulle, la statistique du test de Fischer est :

$$F = \frac{\frac{SCF_{F_1 \times F_2}}{(p-1)(q-1)}}{\frac{SCR}{n-pq}} \sim F_{((p-1)(q-1),n-pq)}$$

Tableau d'analyse de variance

Les résultats se représentent dans un tableau de variation :

Source de variation	SC	ddl	MC	Statistique F
Facteur F1	SCF_{F1}	p-1	$MCF_{F1} = \frac{SCF_{F1}}{P-1}$	$Fc_{F1} = \frac{MCF_{F1}}{MCR}$
Facteur F2	SCF_{F2}	q-1	$MCF_{F2} = \frac{SCF_{F2}}{q-1}$	$Fc_{F2} = \frac{MCF_{F2}}{MCR}$
Interaction F1F2	SCF_{F1F2}	(p-1)(q-1)	$MCF_{F1F2} = \frac{SCF_{F1F2}}{(p-1)(q-1)}$	$Fc_{F1F2} = \frac{MCF_{F1F2}}{MCR}$
Résiduelle	SCR	n-pq	$MCR = \frac{SCR}{n-pq}$	/
Totale	SCT	n-1	$MCT = \frac{SCT}{n-1}$	/

Règles de décision

Pour le test de l'effet de F_1 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_1} > F_{\alpha,(p-1),n-pq}$

et

Pour le test de l'effet de F_2 sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_2} > F_{\alpha,(q-1),n-pq}$

et

Pour le test de l'effet de l'interaction sur Y, on rejette H_0 au seuil α si la quantité de Fischer observée est supérieure à la valeur théorique (lue dans la table de la loi de Fischer-Snédécor) :
$Fobs_{F_{12}} > F_{\alpha,(q-1)(q-1),n-pq}$

1.3 Application

Pour notre application de l'analyse de la variance à deux facteurs fixes, on a choisi d'utiliser les données du célèbre exemple (vitamine et calories), de B. Falissard, publiées dans Comprendre et utiliser, les statistiques dans les sciences de la vie. Edition Masson 2005. [6]
L'objectif de notre étude est de chercher si les régimes alimentaires spécifiques ont une influence sur le poids. B. Falissard propose des données sur des rats de laboratoire. Le gain de poids des rats est désigné par la variable, Poids, c'est la variable à expliquer Y, exprimée en grammes. Les deux facteurs sont les variables Calorie ($F1$) et Vitamine($F2$). La variable Calorie est codée en 1 et 2. Elle vaut 1 si les rats n'ont pas suivi un régime hypercalorique (modalité $C1$) et 2 s'ils

ont suivi ce régime (modalité $C2$). La variable Vitamine vaut 1 si les rats n'ont pas reçu de compléments vitaminés (modalité $V1$)et 2 s'ils ont reçu ces compléments (modalité $V2$). et on mesure le poids des rats après une certaine période.Les valeurs observées sont données dans le tableau suivant :

vitamine / Calorie	V_1	V_2
C_1	84,66,66,56,82,79,62,89	62,59,84,74,73,74,75,74
C_2	87,89,92,101,77,95,88,91	103,90,107,116,95,112,96,92

TABLE 1.6 – Tableau des valeurs observées des poids des rats

On constate que plusieurs mesures de poids sont prises par case (modalité croisée ij) , modalités de $F1$ et $F2$ simultanément, et leur nombre est égal par case ($r = 8$). Pour répondre à la problématique, la méthode adéquate est l'analyse de la variance à deux facteurs fixes à répétitions égales.

1.3.1 Vérification des conditions d'application

1- Les échantillons sont indépendants : un rat (individu statistique) appartient à une seule modalité et une seule seulement, il ne peut pas appartenir à deux modalités différentes. Un rat est sous un régime hypercalorique ou pas et aussi un rat prend des compléments vitaminés ou pas

2- Normalité des résidus :

```
> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.97069, p-value = 0.5187
```

$p - value = 0.52 > 0.05$, on ne rejette pas l'hypothèse nulle au seuil 0.05 cela signifie que l'hypothèse de la normalité des résidus est acceptée.

Le graphique Q-Q plot vient confirmer les résultats de ce test, les observations forment une droite.

On note trois observations atypiques 21,24 et 29 (voir graphique distance de cook).

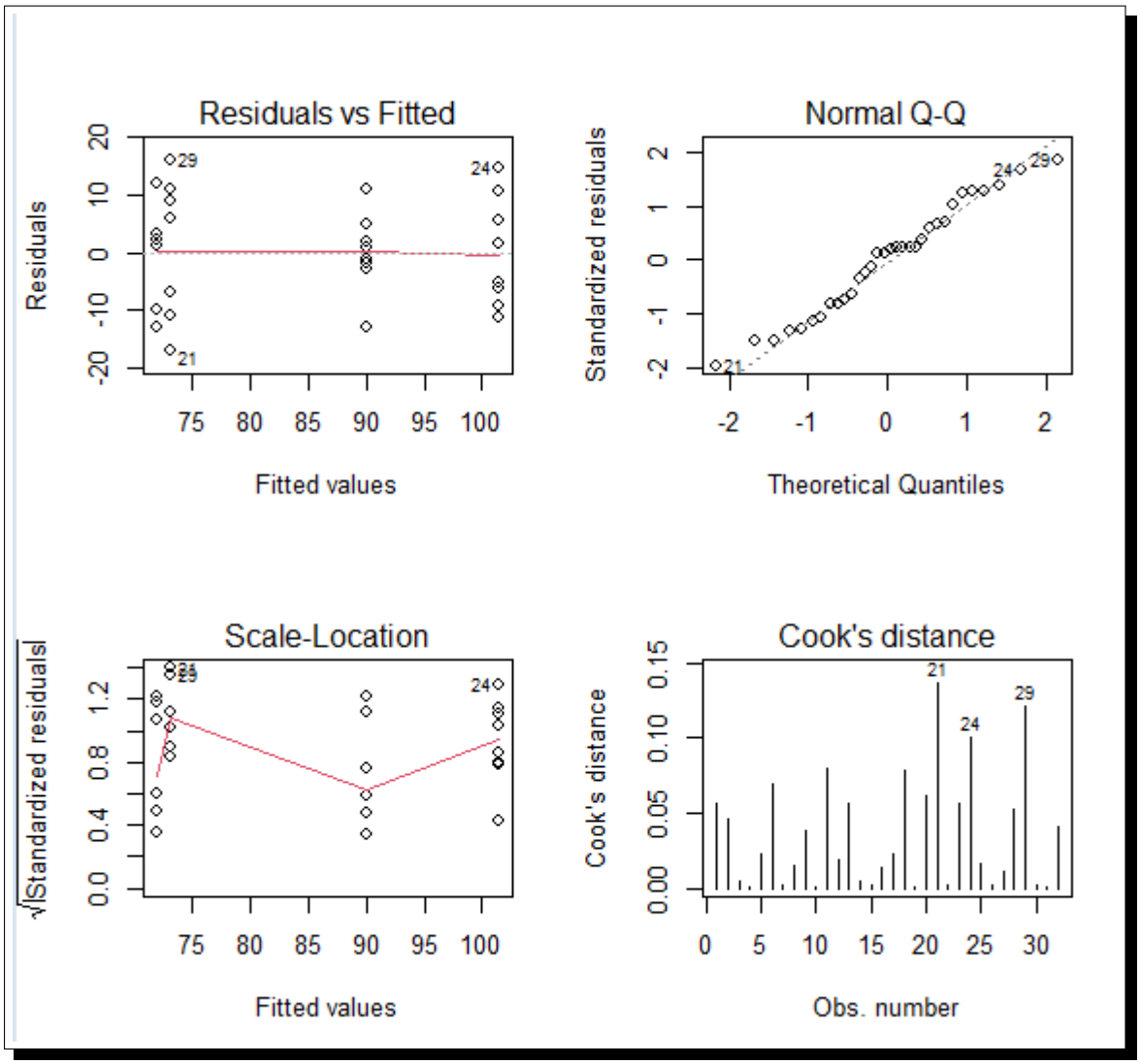

FIGURE 1.3 – Le graphique Q-Q plot

3- Homogénéité des variances : Les résultats du test de Bartlett sont les suivants :

```
> bartlett.test(residus~ cal ,data=dat)

        Bartlett test of homogeneity of variances

data:  residus by cal
Bartlett's K-squared = 0.52299, df = 1, p-value = 0.4696

> bartlett.test(residus~vit , data=dat)

        Bartlett test of homogeneity of variances

data:  residus by vit
Bartlett's K-squared = 0.16001, df = 1, p-value = 0.6891
```

Pour le facteur Calorie, la p-value du test Bartelett est supérieure à 0.05 (0.47) et que la quantité observée de Bartlett (0.52) est inférieure à la valeur théorique tabulée de khi-deux à 1 ddl (3.84). Cela signifie que les différences observées entre les variances des différents échantillons n'est statistiquement pas significatives et que l'hypothèse de l'homoscadasticité (égalité des variances) est vérifiée.
Pour le facteur Vitamine, la p-value du test Bartelett est supérieure à 0.05 (0.69) et que la quantité observée de Bartlett (0.16) est inférieure à la valeur théorique khi-deux à 1 ddl (3.84) tabulée. Cela signifie que les différences observées entre les variances des différents échantillons n'est statistiquement pas significatives et que l'hypothèse de l'homoscadasticité (égalité des variances) est vérifiée.
On peut appliquer l'ANOVA2 et ses résultats sont certainement correctes.

1.3.2 Résultats de l'analyse de la variance

Les résultats sont présentés dans le tableau suivant :

Source de variation	ddl	SC	MC	F	Pr($> F$)
calorie	1	4325	4325	50.061	1.07e-07
vitamine	1	210	210	2.432	0.1301
calorie ×vitamine	1	312	312	3.169	0.0675
Résiduelle	28	2419	86	/	/

TABLE 1.7 – Tableau de variation

le degré de liberté des sommes des carrés des résidus est : pq(r-1) = 2*2*7 = 28 ddl
le degré de liberté des sommes des carrés liée au F1 est : p-1 = 1 ddl
le degré de liberté des sommes des carrés liée au F2 est : q-1= 1 ddl
SC : Somme des carrés
MC : carrés moyens = SC/ddl

1.3.3 Interprétation des résultats

Le tableau ci-dessus (tableau de variation) illustre clairement que le facteur Vitamine n'exerce pas un effet statistiquement significatif sur la prise du poids des rats au seuil 5%.En effet, pour le facteur Vitamine, la p-value du test de Fischer est largement supérieure à 0.05.
La différence de poids médian des rats entre les deux échantillons $V1$ et $V2$ n'est pas très prononcée (graphique 1.3)

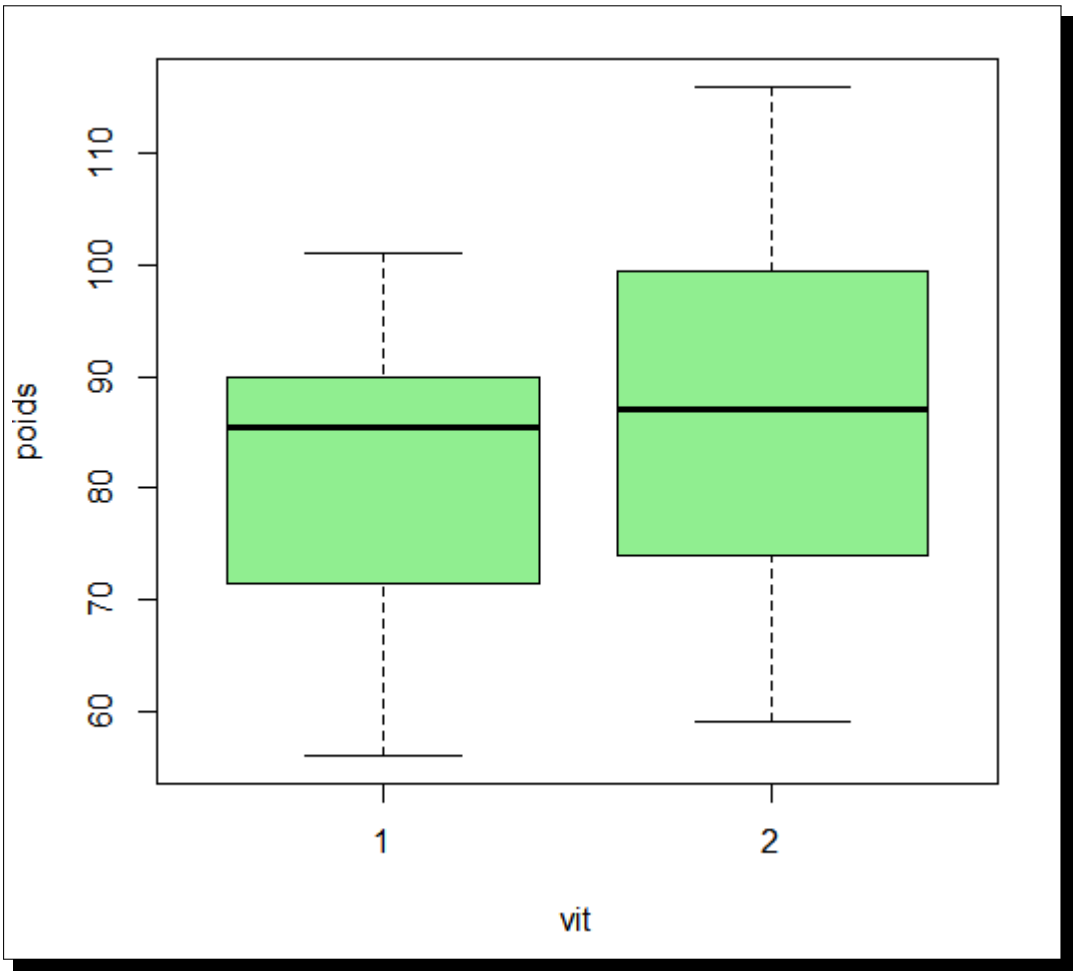

FIGURE 1.4 – Les boites à moustaches de la variable piods en fonction de la variable vitamine

Pour en ce qui concerne le facteur Calorie, il exerce un effet statistiquement significatif sur la prise du poids des rats au seuil 5%.En effet, pour le facteur Calorie, la p-value du test de Fischer est largement inférieure à 0.05.
La différence de poids médian des rats entre les deux échantillons $C1$ et $C2$ est remarquable (graphique 1.4)

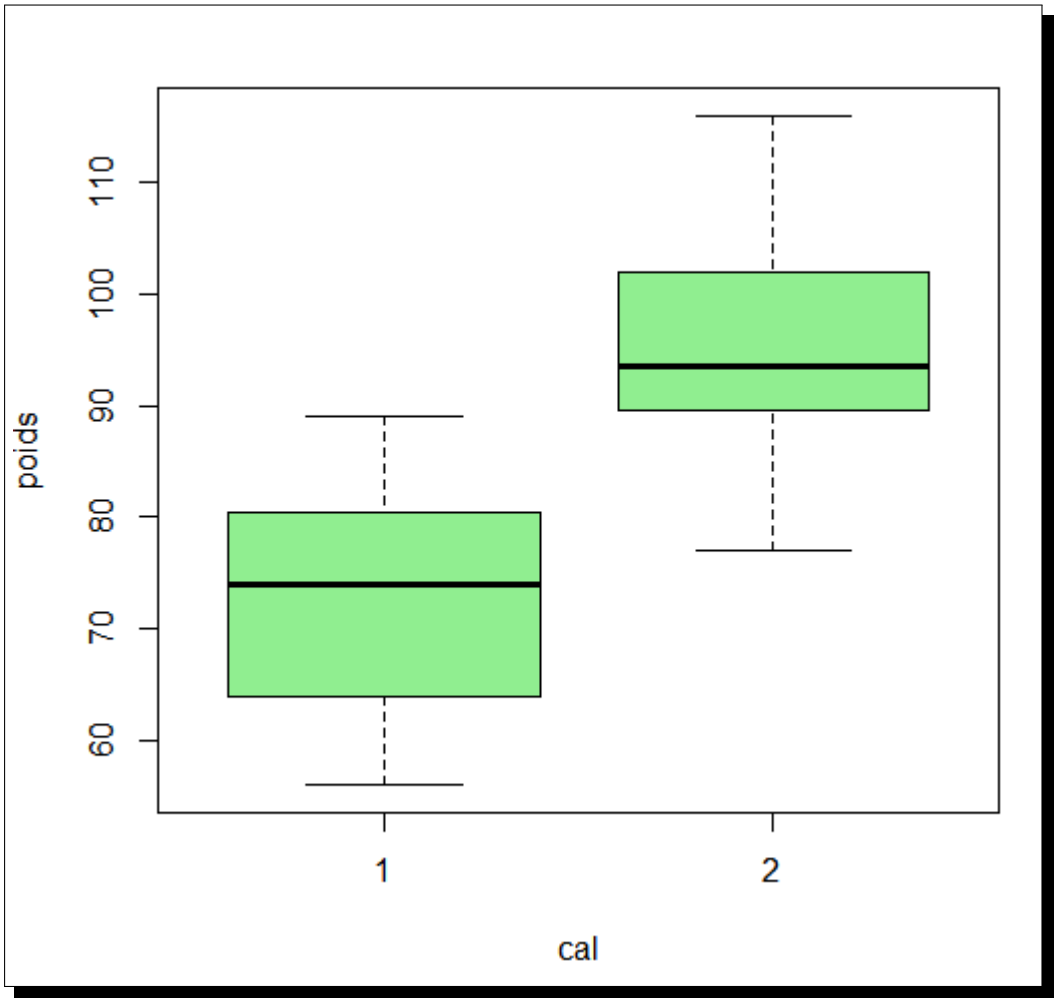

FIGURE 1.5 – Les boites à moustaches de la variable piods en fonction de la variable calorie

L'effet simultané des deux facteurs est statistiquement significatif au seuil10%.et c'est due certainement à l'effet du régime hypercalorique.

Conclusion de chapitre 1

L'analyse de la variance à deux facteurs présente plusieurs avantages ; à savoir :

- elle est très facile à appliquer ,
- les conditions sont faciles à vérifier,
- la décomposition de la variance est toujours valable, quelle que soit la distributions des variables d'intérêt Y,
- les résultats sont faciles à interpréter,
- lorsqu'on utilise le test de Fischer pour la prise de décision, on fait l'hypothèse de la normalité de ces distributions. Si les distributions s'écartent légèrement de la normalité, l'ANOVA est assez robuste.

CHAPITRE

2

Analyse de la covariance

Introduction

L'analyse de covariance est une technique statistique entant dans le carde du modèle linéaire générale, où la variable à expliquer de type quantitatives et les variables explicatives sont de type à la fois quantitatives et qualitatives . Le modèle d'analyse de la covariance base consiste à ajouter à un modèle d'analyse de variance une ou plusieurs variables quantitatives comme variables explicatives . Elle peut être vue comme une mélange être l'analyse de variances et le régression linéaire.

l'objectif d'analyse de la covariance est d 'expliquer la variable réponse (variable quantitative) par des variables qualitatives (les facteurs) et des variables quantitatives (covariables).

Ce chapitre a moins de pages car l'analyse de la covariance a presque les mêmes fondement théorique que l'analyse de la variance .

2.1 Analyse de la covariance à deux facteurs fixes

2.1.1 Présentation du modèle

Le variable quantitative Y est expliquée par deux variables qualitatives à P modalités et q modalités (deux facteurs) et une variable quantitative Z (covariable) le modèle linéaire s'écrit :

$$Y_i = \mu + \alpha_i + \beta_i + (\alpha\beta)_i + \gamma z_i + \varepsilon_i \quad i = \overline{1,n} \tag{2.1}$$

Telle que

- Y_i :La réponse
- μ : la moyenne générale .
- α_i : l'effet du facteur X_1 .
- β_i : l'effet différentiel du facteur X_2.
- $(\alpha\beta)_i$: l'effet de l'interaction entre les deux facteurs.
- γ : coefficient de régression (pente) entre Y_i et Z_i.
- Z_i : valeurs connus de la covariable Z .

2.1.2 Les conditions d'applications

Pour appliquer l'analyse de covariance il faut vérifier les conditions suivantes :

1. Les condition d'analyse de variances ANOVA.
2. La relation linéaire entre la variable dépendante Y et la covariable Z.
3. Il n'ya pas d'interaction significative entre le facteur F et la covariable.
4. La covariable est fixe et mesurée sans erreur.

2.2 Application de l'ANCOVA à 2 facteurs fixes

2.2.1 Présentation des données

Supposons que nous désirions évaluer une nouvelle méthode d'enseignement dans un collège. Les élèves ont été attribués aléatoirement soit à l'ancienne méthode,soit à la nouvelle. Et chaque méthode d'enseignement a été enseignée par des enseignants différents et les élèves étaient assignés à l'un ou l'autre des enseignants et ceci de manière aléatoire .
Les élèves passent deux tests : un avant la nouvelle méthode d'enseignement à vouloir étudier ($pré-test$), et un second test après avoir suivi la nouvelle méthode ($post-test$). La variable Z_i représente le score au pré-test de chaque élève.
La variable X_{1i} est une variable représente la méthode d'enseignement codée (-1 pour l'ancienne et 1 pour la nouvelle),
X_{i2} est une variable codant cette fois l'enseignant (-1 pour l'enseignant A et 1 pour l'enseignant B).
Enfin, Y_i représente le score d'aptitude au post-test de chaque élève. n= 40 élèves.

Les données sont issues d'une étude réalisée par JUDD C., MCCLELLAND G. et al. (2018) sont présentées dans le tableau suivant. [9]

L'objectif de cette section est de comparer l'analyse de la variance à l'analyse de la covariance afin de montrer les avantages de l'analyse de la covariance.

Les programmes réalisés par moi même sont présentés en annexe.

Y_i	X_{1i}	X_{2i}	Z_i	Y_i	X_{1i}	X_{2i}	Z_i
58	1	-1	50	57	1	-1	49
63	1	-1	49	61	1	-1	52
65	1	-1	53	57	1	-1	50
56	1	-1	47	67	1	-1	51
60	1	-1	53	56	1	-1	46
50	-1	- 1	49	62	-1	-1	54
58	-1	-1	51	55	-1	-1	48
52	-1	-1	50	63	-1	-1	52
55	-1	-1	47	50	-1	-1	46
57	-1	-1	52	58	-1	-1	51
61	1	1	47	59	1	1	49
71	1	1	53	65	1	1	54
68	1	1	52	60	1	1	46
58	1	1	48	65	1	1	51
68	1	1	51	65	1	1	49
47	-1	1	46	62	-1	1	51
56	-1	1	51	51	-1	1	49
63	-1	1	53	54	-1	1	51
53	-1	1	48	58	-1	1	50
52	-1	1	47	54	-1	1	54

TABLE 2.1 – Données expérimentales hypothétiques n=40

2.2.2 Présentation du modèle et ses hypothèses

Le modèle

$$Y_i = \mu + \alpha_i + \beta_i + (\alpha\beta)_i + \gamma Z_i + \varepsilon_i \quad i = \overline{1,n}$$

Tels que :

- Y_i :Le score d'aptitude post-test de l'élève i.
- μ : la moyenne totale des scores post-test.
- α_i : paramètre qui mesure l'effet de la méthode d'enseignement (facteur X_1) sur le score post-test.
- β_j : paramètre qui mesure l'effet de l'enseignant (facteur X_2) sur le score post-test.
- $(\alpha\beta)_i$: paramètre qui mesure l'effet de l'interaction des deux facteurs sur le score post-test.
- γ : coefficient de régression qui mesure l'effet du niveau initial de l'élève sur les résultats post-test Y_i.
- Z_i : le score pré-test obtenu par l'élève i.
- ε_i : l'erreur aléatoire.

Les hypothèses du modèle de la covariance

Les hypothèses sont semblables au modèle de la régression linéaires multiples.

1. le vecteur Z_i est déterminée sans erreurs.
2. $\varepsilon_i \sim N(0.\sigma^2)$.
3. $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$:l'erreurs est indépendante .
4. $Cov(Z_i, \varepsilon_j) = 0 \quad \forall i \neq j$:l'erreurs sont indépendante de covariable .
5. $\varepsilon \sim N(0.\sigma^2 I_n)$: les erreurs suivent une loi normale multidimensionnelle.

2.2.3 Vérification des conditions d'application

Condition 1 : Hypothèse de linéarité entre Y et Z

La relation linéaire ente Y (score post-test) et Z (score pré-test) est démontrée par le modèle de régression simple ci-dessous.

La p-valeur $(0.002746 < 0.05)$ nous indique que le coefficient de Zi est statistiquement différent de 0 alors la relation linéaire ente ces deux variables est statistiquement significative au seuil 5%.

```
> mod=lm(postt~prett ,data=dat)
> mod$coef
(Intercept)       prett
  -4.524336    1.265487
> anova(mod)
Analysis of Variance Table

Response: postt
          Df Sum Sq Mean Sq F value    Pr(>F)
prett      1 361.93  361.93  16.075 0.0002746 ***
Residuals 38 855.57   22.52
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le nuage de points confirme ce constat.

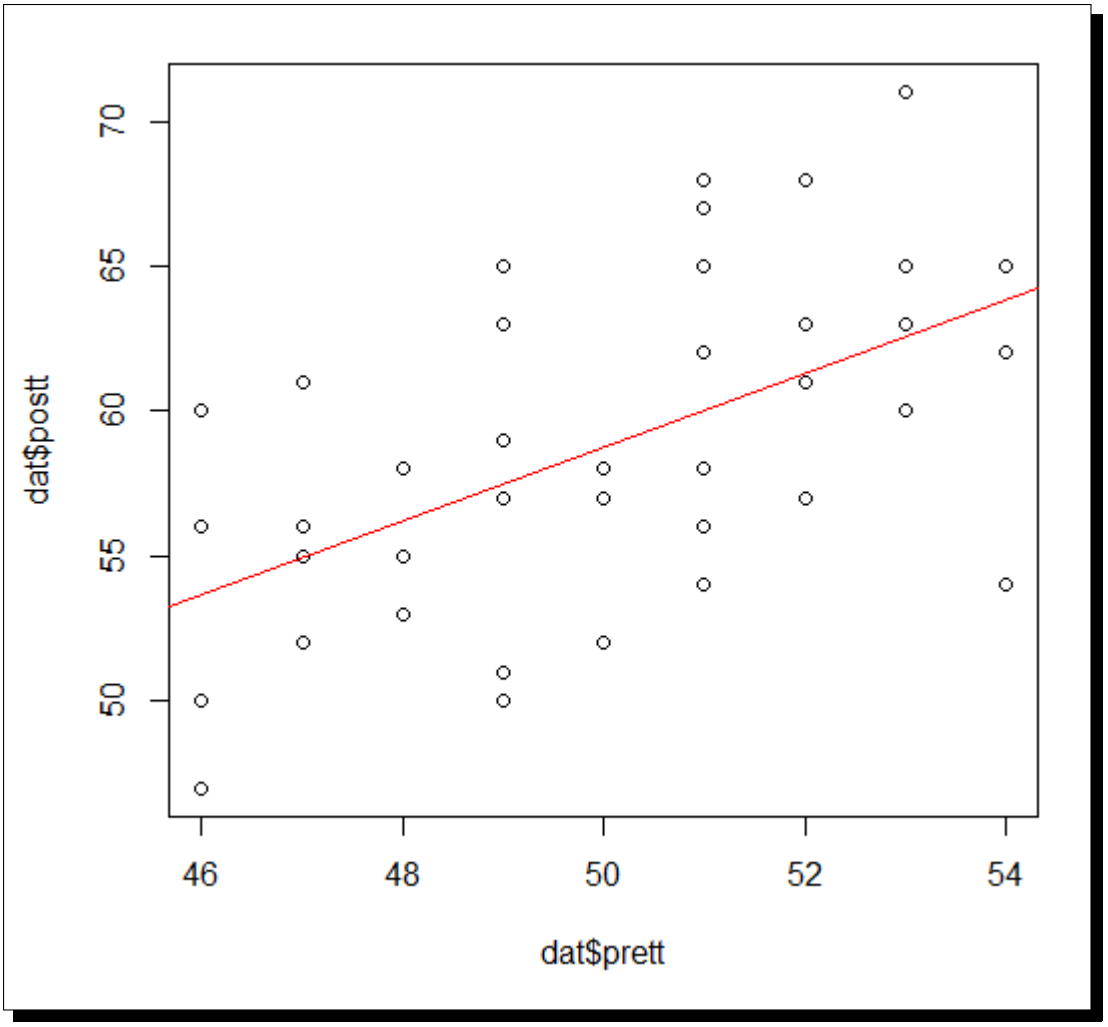

FIGURE 2.1 – Représentation de la droits de régression des moindres carrés sur le nuage de points

Condition2 : Absence d'interaction entre Z et X Pour éviter le problème de la colinéarité, l'interaction entre la variable Z et les variables X ne doit pas être significative.

```
> ano=aov(postt~ prett+meth+ens+ meth*ens + prett*meth+ prett*ens + prett*meth*ens ,data=dat)
> summary(ano)
               Df Sum Sq Mean Sq F value   Pr(>F)
prett           1  361.9   361.9  33.646 1.94e-06 ***
meth            1  422.5   422.5  39.277 5.02e-07 ***
ens             1   22.5    22.5   2.092   0.1578
meth:ens        1   62.5    62.5   5.810   0.0219 *
prett:meth      1    2.0     2.0   0.190   0.6662
prett:ens       1    0.2     0.2   0.022   0.8838
prett:meth:ens  1    1.6     1.6   0.146   0.7048
Residuals      32  344.2    10.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D'après les résultats du test statistique, l'interaction de la variable Z et X1 n'est statistiquement pas significative au seuil 5% ;toutes choses égales par ailleurs. (p-valeurs $0.6662 > 0.05$). Aussi, l'interaction de la variable Z et X2 n'est statistiquement pas significative au seuil 5% toutes choses égales par ailleurs. (p-valeurs $0.8838 > 0.05$).

Condition3 : Normalité des résidus Selon le test statistique de SHAPIRO-WILK, $p - value = 0.5852 > 0.05$, on ne rejette pas l'hypothèse nulle au seuil 0.05 et cela signifie que l'hypothèse de la normalité des résidus est acceptée.

```
> mod.1=lm(postt~prett+meth*ens ,data=dat)
> residus=residuals(mod.1)
> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.97716, p-value = 0.5852
```

Le graphique Q-Q plot vient confirmer les résultats de ce test, les observations forment une droite.

On note trois observations atypiques 40, 36 et 24.

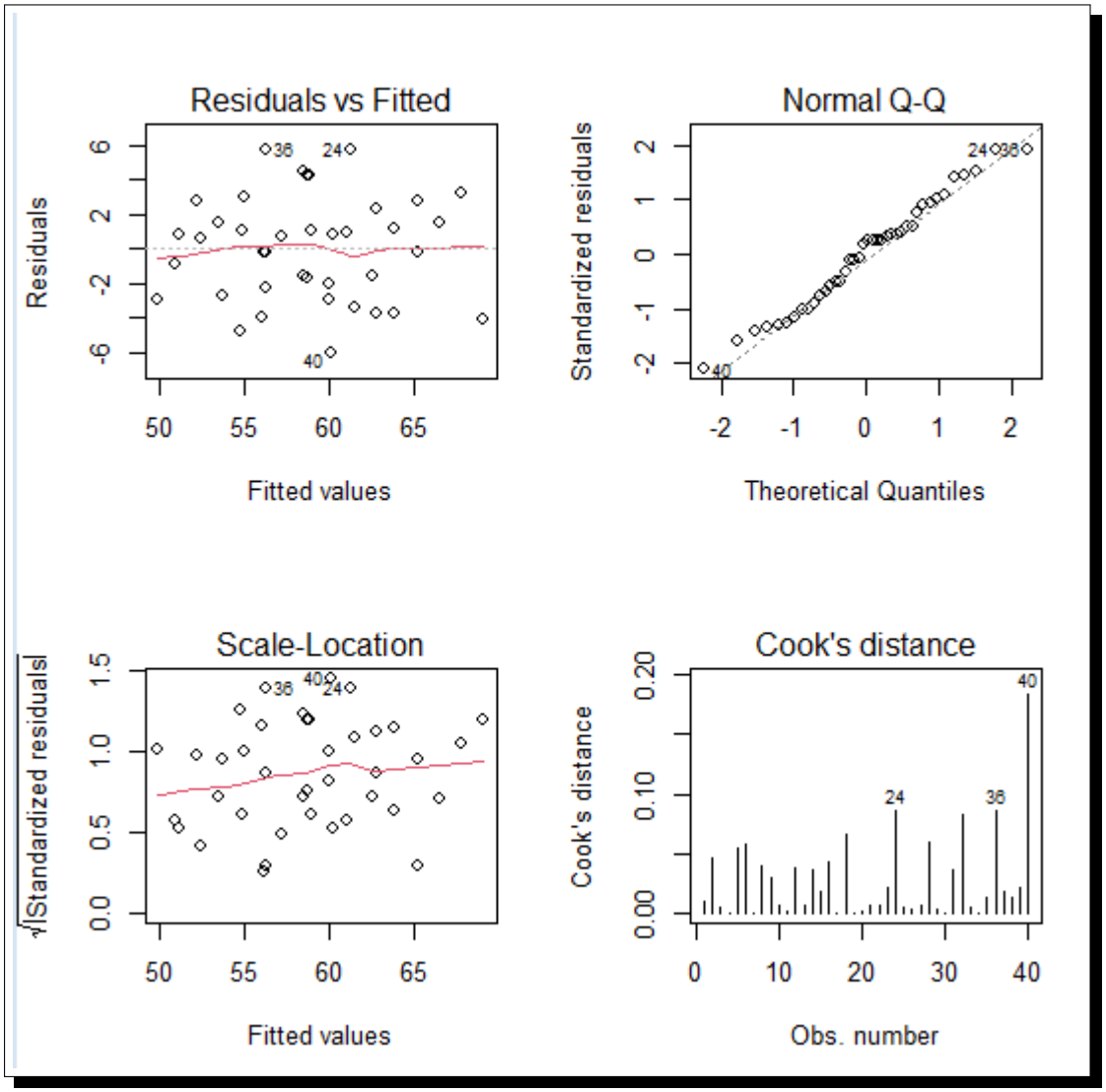

FIGURE 2.2 – Le graphique Q-Q plot

Condition4 : Homogénéité des variances

Pour le facteur Méthode ($X1$), la p-value du test Bartelett est supérieure à 0.05 (0.0.6513) et que la quantité observée de Bartlett (0.20) est inférieure à la valeur théorique tabulée de khi-deux à 1 ddl (3.84). Cela signifie que les différences observées entre les variances des différents échantillons n'est statistiquement pas significatives et que l'hypothèse de l'homoscadasticité (égalité des variances) est vérifiée.

Pour le facteur Enseignant ($X2$), la p-value du test Bartelett est supérieure à 0.05 (0.7318)

et que la quantité observée de Bartlett (0.11) est inférieure à la valeur théorique khi-deux à 1 ddl (3.84) tabulée. Cela signifie que les différences observées entre les variances des différents échantillons n'est statistiquement pas significatives et que l'hypothèse de l'homoscadasticité (égalité des variances) est vérifiée.

```
> bartlett.test(residus~ meth ,data=dat)

        Bartlett test of homogeneity of variances

data:  residus by meth
Bartlett's K-squared = 0.20425, df = 1, p-value = 0.6513

> bartlett.test(residus~ens , data=dat)

        Bartlett test of homogeneity of variances

data:  residus by ens
Bartlett's K-squared = 0.11742, df = 1, p-value = 0.7318
```

2.2.4 Résultats : Comparaison de l'ANOVA et l'analyse de la covariance

Résultats du modèle ANOVA2

L'application du modèle ANOVA classique sur nos conduit aux résultats suivants :

$$postt = 58.75 + 3.5meth + 0.75ens + 1.25(meth * ens) \quad ... \ (M_1)$$

Source de variation	ddl	SC	MC	F	Pr($> F$)
Méthode	1	422.5	422.5	21.42	4.65e-05
Enseignant	1	22.5	22.5	1.141	0.2926
Méthode ×Enseignant	1	62.5	62.5	3.169	0.0835
Résiduelle	36	710.00	19.7	/	/
Total	39	1217.50	/	/	/

TABLE 2.2 – Modèle M1 : ANOVA à deux facteurs

Il résulte que le facteur $X1$ (Méthode) influence les scores post-test des élevés. Son

effet est statistiquement significatif au seuil 5% ($p_value 4.65e-05 < 0.05$). Alors que l'effet du facteur (Enseignant) n'est statistiquement pas significatif seuil 5% ($p_value 0.2926 > 0.05$).
Aussi l'effet de l'interaction des deux facteurs n'est statistiquement pas significatif seuil 5% ($p_value 0.08 > 0.05$).
Ce qui est intéressant à souligner est la valeur des sommes des carrés des résidus (710), elle est relativement très élevée. On tente d'améliorer le modèle en retirant l'interaction du modèle M1 :

$$postt = 58.75 + 3.5meth + 0.75ens \quad ... \ (M_2)$$

Source de variation	ddl	SC	MC	F	Pr(> F)
Méthode	1	422.5	422.5	20.236	6.56e-05
Enseignant	1	22.5	22.5	1.078	0.306
Résiduelle	37	772.5	20.9	/	/
Total	39	1217.50	/	/	/

TABLE 2.3 – Modèle M2 : ANOVA à deux facteurs

Lorsque nous supprimant l'interaction du M1, le modèle ne s'améliore pas (SCR=772.5). La valeur de la somme des carrés des résidus augmente (M2). la différence entre la SCR du M1 et la SCR du M2 est statistiquement significative au seuil10%

```
> anova(anova.1,anova.2)
Analysis of Variance Table

Model 1: postt ~ meth * ens
Model 2: postt ~ meth + ens
  Res.Df   RSS Df Sum of Sq     F  Pr(>F)
1     36 710.0
2     37 772.5 -1     -62.5 3.169 0.08349 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En revanche, lorsqu'on introduit la variable explicative (score pré-test), le modèle est nettement meilleur. La somme des carrés des résidus s'est nettement baissée (348.07).
L'interaction devient statistiquement significative ($p_value 0.017 < 0.05$).

$$postt = -4.524 + 1.265prett + 3.250meth + 0.750ens + 1.250(meth * ens) \quad ... \ (M_3)$$

Source de variation	ddl	SC	MC	F	Pr($> F$)
Pré-test	1	361.93	361.93	36.39	6.99e-07
Méthode	1	422.5	422.5	42.48	1.61e-07
Enseignant	1	22.5	22.5	2.26	0.142
Méthode ×Enseignant	1	62.5	62.5	6.28	0.017
Résiduelle	35	348.07	9.94	/	/
Total	39	1217.50	/	/	/

TABLE 2.4 – Modèle M3 : L'analyse de la covariance

le modèle s'améliore nettement (SCR=348.07). La valeur de la somme des carrés des résidus a nettement baissée dans M3. la différence entre la SCR du M1 et la SCR du M3 est statistiquement significative au seuil 5% p-value $(0.000000699) < (0.05)$.

```
> anova(anova.1 ,ancov)
Analysis of Variance Table

Model 1: postt ~ meth * ens
Model 2: postt ~ prett + meth * ens
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     36 710.00
2     35 348.07  1    361.93 36.394 6.992e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pour approfondir notre recherche, on applique une ANOVA en prenant comme variable à expliquer la différence entre le score post-test et le score pré-test (modèle M4).

$$diff = 8.75 + 3.5meth + 0.75ens + 1.25(meth * ens) \quad ... \ (M_4)$$

Les résultats sont présentés dans le tableau suivant :

Source de variation	ddl	SC	MC	F	Pr($> F$)
Méthode	1	422.5	422.5	41.79	1.67e-07
Enseignant	1	22.5	22.5	2.22	0.1445
Méthode ×Enseignant	1	62.5	62.5	6.18	0.0177
Résiduelle	36	364.00	9.94	/	/
Total	39	1217.50	/	/	/

TABLE 2.5 – Modèle M4 : ANOVA avec la différence des deux scores

En prenant en compte le niveau initial des élevés, l'interaction entre les deux facteurs (Méthode*Enseignant) devient statistiquement significatif ; p-value (0.0177) $<$ (0.05), après avoir n'était pas significative dans M1. ceci témoigne l'intérêt de prendre en compte le niveau initial des élevés dans le modèle (variable quantitative).
La somme des carrés résiduelle SCR du modèle M4 est faible mais reste plus elevée que celle du modèle M3 (modèle de l'analyse de la covariance). Statistiquement, le meilleure modèle est celui de l'analyse de la covariance M3.
Comme les modèles M3 et M4 ne sont pas emboités, on renforce la comparaison par les critères AIC.

Le critère AIC s'applique aux modèles estimés par une méthode du maximum de vraisemblance.

Définition

Le critère d'information d'Akaike est défini par :

$$AIC = 2k - 2\log(l).$$

où l est la vraisemblance maximisée et k le nombre de paramètres dans le modèle. [8]

Critère d'information bayésien BIC

Définition

Le critère d'information bayésien BIC est défini par :

$$BIC = k\log(n) - 2\log(l)$$

On retient également le modèle ayant le plus petit BIC. [8]

Le meilleur modèle est celui qui correspond à la plus faible valeur AIC et BIC.

$$\begin{cases} M_3 : postt = -4.524 + 1.265prett + 3.250meth + 0.750ens + 1.250(meth * ens) \\ M_4 : diff = 8.75 + 3.250meth + 0.750ens + 1.250(meth * ens) \end{cases}$$

```
> ancd=glm(diff~meth*ens , data=dt)
> anc=glm(postt~prett+meth*ens   , data=dat)
> c(AIC(anc),AIC(ancd))
[1] 212.0561 211.8461
> c(BIC(anc),BIC(ancd))
[1] 222.1894 220.2905
```

On remarque que les valeurs de l'AIC et du BIC du modèle M_4 sont inférieures aux valeurs AIC et BIC du modèle M_3 mais leur différences ne sont pas statistiquement significatives au seuil 5% c'est à dire $(222.19 - 220.29 = 1.89 < 3.84khi - deux1ddl)$.

On conclut que le meilleur modèle est celui de l'analyse de la covariance M3.

Conclusion générale

L'analyse de variance et l'analyse de covariance sont deux modèles statistiques entrant dans le cadre des modèles linéaires généralisés mais avec des caractéristiques différentes :

L'analyse de la variance à deux facteurs présente plusieurs avantages ; à savoir :

- elle est très facile à appliquer ,
- les conditions sont faciles à vérifier,
- la décomposition de la variance est toujours valable, quelle que soit la distributions des variables d'intérêt Y,
- les résultats sont faciles à interpréter,
- lorsqu'on utilise le test de Fischer pour la prise de décision, on fait l'hypothèse de la normalité de ces distributions. Si les distributions s'écartent légèrement de la normalité, l'ANOVA est assez robuste.

  En revanche, si les distributions s'écartent fortement de la normalité, l'ANOVA devient très sensible. Pour contourner ce problème, on peut toujours effectuer un changement de variable en prenant comme variable à expliquer log Y au lieu de Y par exemple. ou bien, on utilise un équivalent non paramétrique de l'ANOVA. De plus, si on souhaite intégrer une variable explicative ou plusieurs variables explicatives quantitatives, elle devient non valable. et plusieurs autres raisons nous laissent donc conclure que le modèle M3 est le meilleur :

1. La différence entre les valeurs de l'AIC et du BIC du modèle M_3 sont inférieurs à l'AIC et l'BIC du modèle M_4 est très faible,
2. Le modèle de l'analyse de la covariance M3 est meilleur car il est associé au minimum des erreurs aléatoires,
3. Le modèle de l'analyse de la covariance M3 est meilleur car il renseigne clairement sur l'effet de la covariable exercé sur la variable à expliquer (effet du niveau initial des élevés sur leur réussite au nouveau enseignement).

Résumé

L'analyse de variance (ANOVA) est une technique statistique utilisée pour étudier le comportement d'une variable quantitative à expliquer (variable d'intérêt) en fonction d'une ou de plusieurs variables qualitatives. Autrement dit, il s'agit d'étudier l'effet d'un facteur (ou plusieurs facteurs) sur une variable d'intérêt de type quantitatif en utilisant un ensemble de modèles statistiques pour comparer les moyennes des différents échantillons indépendants.

Si nous souhaitons intégrer dans le modèle des variables explicatives quantitatives, l'emploi de l'analyse de la variance devient pas possible. C'est l'analyse de la covariance qu'il faut appliquer. C'est un modèle qui contient des variables indépendantes à la fois qualitatives (appelées facteurs) et quantitatives(appelées covariables). Il s'agit d'un mélange de l'analyse de la variance et de la régression linéaire.

L'ajout des covariables dans le modèle permet de réduire considérablement la composante de la variabilité associée à l'erreur aléatoire ,et donc d'augmente la puissance du modèle.

Abstract

L'analysis of variance (ANOVA) is a statistical technique used to study the behavior of a quantitative variable to be explained (variable of interest) depending on one or more qualitative variables. In other words, it is about studying the effect of a factor (or several factors) on a quantitative variable of interest using a set of statistical models to compare the means of different samples independent.

If we want to integrate quantitative explanatory variables into the model, the use of analysis of variance becomes not possible. This is the analysis of covariance that should be applied. It is a model that contains independent variables that are both qualitative (called factors) and quantitative (called covariates). It is a mixture of analysis of variance and analysis of linear regression.

Adding covariates in the model significantly reduces the component of the variability associated with the random error, and therefore increases the power of the model.

ANNEXE1

Le programme sous R :

Simulation de modèle d'analyse de la variance à deux facteur fixe avec répétition égale et interaction

```
dat=read.table("tab.csv" ,dec="," , sep=";" , header=TRUE)
dat
mod=lm(poids∼cal*vit , data=dat)
residus=residuals(mod)
shapiro.test(residus)
bartlett.test(residus∼ cal ,data=dat)
bartlett.test(residus∼vit , data=dat)
anova=aov(poids∼cal*vit , data=dat)
par(mfrow=c(2,2))
plot(anova,which=1)
plot(anova ,which=2)
plot(anova,which=3)
plot(anova ,which=4)
boxplot(poids∼cal ,pch=16 , cex=0.5 ,col="light green" ,data=dat )
boxplot(poids∼vit ,pch=16 ,cex=0.5,col=" light green" ,data=dat)
summary(anova)
```

ANNEXE2

Le programme sous R :

Simulation de modèle d'analyse de covariance

```
dat=read.table("TAL.csv",dec="," , sep=" ;" , header=TRUE)
dat
mod=lm(postt ~ prett , data=dat)
anova=(mod)
ano= (postt ~ prett+meth+ens+ meth*ens + prett*meth+ prett*ens + prett*meth*ens ,
data=dat)
summary(ano)
mod.1=lm(postt ~ prett+meth+ens , data=dat)
residus=residuals(mod.1)
shapiro.test(residus)
bartlett.test(residus ~ meth , data=dat)
bartlett.test(residus ~ ens , dat=dat)
anova.1 =aov(postt ~ meth*ens , data=dat)
summary(anova.1)
anova.2 =aov(postt ~ meth+ens , data=dat)
summary(anova.2)
anova(anova.1,anova.2)
# comparaison des deux modèles d'analyse de la variance #
ancov=(postt ~ prett+meth*ens , data=dat)
summary(ancov)
anova(anova.1 , ancov)
# comparaison entre le modèle d'analyse de variance et le modèle d'analyse de covariance #
plot(ancov ,which=1)
plot(ancov ,which=2)
```

```
plot(ancov ,which=3)
plot(ancov ,which=4)
dt=read.table("T.csv",dec="," , sep=";" , header=TRUE)
dt
ancdif=aov( diff ~ meth*ens , data=dt)
ancd= glm( diff ~ meth*ens , data=dt)
ancd
anc = glm( postt ~ prett+meth*ens , data=dat)
anc
c(AIC(anc) ,AIC(ancd))
c(BIC(anc) ,BIC(ancd))
```

Bibliographie

[1] **Antoine Godichon Baggioni** : *Analyse de la variance à 2 facteurs* , INSA de Rouen .

[2] **Antoine Godichon Baggioni** : *Analyse de covariance ,cours statistique* ,INSA de Rouen ,2017 - 2018 .

[3] **Broc Guillaume** : *Stats faciles avec R* ,2016 .Édition de boeck .

[4] **C.Chouquet** : *Modèles linéaire* . Laboratoire de statistique et probabilités -Université Paul Sabatier - Toulouse ,M1 IMAT -2009 -2010 .

[5] **Farida Laoudj Chekraui** :*Cours de statistique mathématique*, Université Mohamed Sedik ben Yahia- Jijel

[6] **Frédéric Bertrand** : *Exemples classiques de dispositifs expérimentaux* ,T.D.n^0 8, 1^{er}-Année - 2012 - 2013 ,http ://irma.math.unistra.fr .

[7] **Frédéric Bertrand** : *l'analyse de la covariance* ,2016-2017 .

[8] **Frédéric Bertrand** et **Myrian Mauny - Bertrand** : *Choix du modèle*,Université de Strasbourg France -Master1 - 2017 .

[9] **Judd Ch-McClelland G.ET al** :*ANALYSE DES DONNÉES.une approche par comparaison des modéles* .2018 édition deboeck.

[10] **Nathaniel E-Helwig** : *Analysis of covariance* , University of Minnesota ,04 Jan 2017 .

[11] **Lajmi Lakhal-Chaieb** : *Planification des expacriences* ,SCTT - 4100 /SST-7230 ,2015 .

[12] **Gherda Mebrouk** : *Statistique inferentielle* ,Troisième Année Spécialité Mathématiques,Université Mohammed Sedik Benyahia -Jijel ,2019-2020 .

[13] **Rafik Abdesselam** : *Statistique et informatique pour la science des données* ,Support2 : ANOVA- ANCOVA .Université Lumiére Lyon2 , Décembre 2020 .

[14] **Ricco-Rakotomalala** : *Partique de la régression linéaire multiple diagnostic et sélection de variables* , Université Lumiére Lyon2 , 22-May-2015