République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

LOUAM Manel

Titre :

Tests de comparaison et applications

Membres du Comité d'Examen :

Dr.	**Benatia Fateh**	UMKB	Président
Dr.	**Djabrane Yahia**	UMKB	Encadreur
Dr.	**Soltane Louiza**	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce modeste travail...

A celui qui a toujours garni mes chemins avec force et lumière. . . mon trés cher père.

A la plus belle perle du monde. . . ma tendre mère.

A mon frère Mohamed Je lui souhaite tout le succès. . . tout le bonheur.

A toute ma famille pour l'amour et le respect qu'ils m'ont toujours accordé.

A tous mes amis

Pour une sincérité si merveilleuse. . . jamais oubliable, en leur souhaitant tout le succès. . . tout le bonheur.

A tous les membres de ma promotion.

A tous ceux qui me sens chers et que j'ai omis de citer.

REMERCIEMENTS

Mes remerciements vont d'abord à Allah le tout puissant de m'avoir donné la force et le courage ainsi que l'audace pour dépasser toutes les difficultés. J'adresse mes plus vifs remerciements aux personnes qui m'ont aidé dans la réalisation de ce mémoire.

En premier lieu, je tiens à exprimer ma profonde gratitude à Dr.Djabrane Yahia en tant que mon encadreur, pour m'a avoir guidé, encouragé, ses précieux conseils et son orientation ficelée tout au long de mon travail. Je suis également extrêmement reconnaissant envers Monsieur Hfayed Mokhtar, le chef de département de la faculté.

Remerciements et profonde gratitude vont également aux chaque membre de Jury : Dr.Benatia Fateh et Dr.Soltane Louiza pour l'honneur qu'ils m'ont fait en portant leur attention sur ce travail. J'exprime ainsi toute ma reconnaissance à tous mes profs durant tous les années de mes études.

C'est avec chaleur et sincérité que je tiens à remercier tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce modeste travail. Je voudrais plus particulièrement exprimer ma reconnaissance envers tous mes amis qui m'ont apporté leur soutien moral pendant cet année.

Il me serait impossible, enfin, de ne pas saluer toute ma famille qui m'a gratifié de son amour et fourni les motivations. Je leur adresse toute ma gratitude du fond du cœur.

Table des matières

Table des figures

Liste des tableaux

Introduction

En 1918, Ronald Aylmer Fisher présente pour la première fois le terme variance et propose son analyse formelle dans l'article "The Correlation Between Relatives on the Supposition of Mendelian Inheritance". Sa première application de l'analyse de la variance a été publiée en 1921. Elle est devenue largement connue aprés avoir été incluse dans le livre de Fisher "Statistical Methods for Research Workers" (1^{re} édition) en 1925.

L'analyse de la variance joue un rôle tout à fait particulier en statistique. C'est depuis son origine un univers en perpétuelle expansion. La demande des praticiens a obligé les statisticiens à construire des modèles plus performants, plus souples, s'adaptant mieux à la réalité des données. Elle peut porter sur différentes domaines : économie, médecine, biométrie, biologie, psychologie,...,etc.

A titre d'exemple, ANOVA est appliquée dans la médecine, on trouve qu'afin de surveiller les coûts engendrés par les médecins en pratique privée, Santésuisse a développé l'outil statistique ANOVA, de manière à pouvoir tenir compte des différences liées à l'âge et au sexe des patients ainsi que du canton dans lequel pratique le médecin concerné. Selon Santésuisse, l'application d'ANOVA permet d'étendre et d'améliorer la comparaison entre médecins alors même que les patients et le canton de pratique sont différents. Par exemple, elle rendrait possible la comparaison entre des généralistes pratiquant à Genève et ceux pratiquant en Valais.

Cette méthode correspond à un modèle linéaire gaussien dans lequel toutes les variables explicatives sont qualitatives. Dans ce contexte elles sont appelées facteurs et leurs modalités sont appelées niveaux, des facteurs sont fixés par l'expérimentateur. On parle alors de modèle fixe. Seuls seront traités dans ce mémoire les cas de l'analyse de la variance à un facteur, et à deux facteurs. L'analyse de la variance est un test statistique qui généralise le test de Student au cadre de comparaison de plusieurs moyennes. On l'applique dès lors que l'on étudie les effets d'une ou plusieurs variables qualitatives sur une variable quantitative. Le cœur de cette méthode est la décomposition de la variabilité totale selon les différentes sources présentes les données. Il resulte de ceci le plan de notre mémoire que s'articule autour 2 chapitres.

• Le premier chapitre a pour titre : Comparaison de moyennes. Il traite en détails l'analyse de la variance tels que dans la première section nous étudions la méthode de l'analyse de la variance à un facteur, tandis que la deuxième section nous étudions la méthode de l'analyse de la variance à deux facteurs, avec des exemples dans les deux cas.

• Le deuxième chapitre, est la comparaison de variances. Où nous avons présenté les tests qui permet de vérifier l'égalité des variances entre des populations ou des niveaux de facteurs, dont on traite des exemples réelles simulés à l'aide du machine, tels que nous avons essayé d'appliquer tous ce que nous avons parlé dans ce chapitre sous le logiciel **R**.

Chapitre 1

Comparaison de moyennes

L'analyse de la variance (terme souvent abrégé par le terme anglais **ANOVA** : **AN**alysis **O**f **VA**riance) relative à la comparaison de moyennes. Elle est l'une des procédures les plus utilisées dans les applications de la statistique ainsi que dans les méthodes d'analyse de données.

C'est la comparaison de moyennes pour K échantillons ($K > 2$). Elle s'agit de comparer les espérances des variables aléatoires indépendantes gaussiennes de même variance.

Dans ce chapitre, on s'intéresse à étudier l'analyse de la variance à un facteur et à deux facteurs.

1.1 Analyse de la variance à un facteur

L'analyse de la variance à un facteur ; notée $AV(1)$ est une technique statistique qui sert à tester l'influence d'une variable qualitative (souvent appelée facteur) sur une variable quantitative (variable à expliquée Y).

Un facteur A se représente sous I niveaux $(A_1, A_2, \ldots, A_I)$. Pour chaque niveau A_i du facteur est associer J mesures d'une réponse Y qui est une variable continue.

Notons $n = I * J$ le nombre total de mesures ayant été effectuées.

1.1.1 Les données

Les données relatives à l'analyse de la variance à un facteur sont structurées comme suit :

$Facteur\ A$	Y
A_1	$y_{1,1}, \ldots, y_{1,J}$
$\vdots$	$\vdots$
A_i	$y_{i,1}, \ldots, y_{i,J}$
$\vdots$	$\vdots$
A_I	$y_{I,1}, \ldots, y_{I,J}$

TAB. 1.1 – Les données d'AV(1).

où $y_{i,j}$ désigne la $i^{ème}$ observation du $j^{ème}$ niveau ($i = \overline{1, I}$ et $j = \overline{1, J}$).

• $\overline{y}_{i,*}$ la moyenne de classe i

$$\overline{y}_{i,*} = \frac{1}{J} \sum_{j=1}^{J} y_{i,j}.$$

• $\overline{y}_{*,*}$ la moyenne totale

$$\overline{y}_{*,*} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{I} y_{i,j}.$$

• S_i^2 variance de chaque échantillon

$$S_i^2(y) = \frac{1}{J} \sum_{j=1}^{J} (y_{i,j} - \bar{y}_{i,*})^2.$$

• S^2 variance de toutes les obsevations

$$S^2(y) = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{i,j} - \bar{y}_{*,*})^2.$$

Le $*$ signifiant que l'indice n'intervient plus.

1.1.2 Modèle d'AV(1)

Soit $\mu = \frac{1}{n}\sum_{i=1}^{I} J_i.\mu_i$ et $\mu_i = \mathbb{E}[y_i]$. En terme d'observations on a le modèle :

$$\underbrace{(y_{i,j} - \bar{y}_{*,*})}_{\substack{\text{écart}\\ \text{total}}} = \underbrace{(y_{i,j} - \bar{y}_{i,*})}_{\substack{\text{écart}\\ \text{résiduel}}} + \underbrace{(\bar{y}_{i,*} - \bar{y}_{*,*})}_{\substack{\text{écart}\\ \text{factor iel}}}.$$

Le modèle théorique est :

$$(y_{i,j} - \mu) = (y_{i,j} - \mu_{i,*}) + (\mu_{i,*} - \mu).$$

Posons :

$$\begin{cases} \alpha_i &= \mu_{i,*} - \mu. \\ \epsilon_{i,j} &= y_{i,j} - \mu_{i,*}. \end{cases}$$

Nous obtenons le modèle d'AV(1) :

$$y_{i,j} = \underbrace{\mu}_{\substack{\text{moyenne}\\ \text{générale}}} + \underbrace{\alpha_i}_{\substack{\text{effet}\\ \text{principal}}} + \underbrace{\epsilon_{i,j}}_{\substack{\text{effet}\\ \text{résiduel}}}. \tag{1.1}$$

Hypothèses de modélisation

- $\sum_{i=1}^{I} \alpha_i = 0.$
- $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2), \forall i, j.$
- $cov(\epsilon_{i,j}, \epsilon_{k,l}) = 0, \forall (i,j) \neq (k,l).$

Remarque 1.1.1 *Le modèle d'AV(1) possède $I + 1$ paramètres.*

1.1.3 Estimation des paramètres

Les estimateurs $\hat{\mu}, \hat{\alpha}_i, \hat{\epsilon}_{i,j}, \hat{\sigma}^2$ des paramètres $\mu, \alpha_i, \epsilon_{i,j}, \sigma^2$ du modèle (1.1) sont donnés par les formules suivantes :

- $\hat{\mu} = \overline{y}_{*,*} = \overline{y}$.
- $\hat{\alpha}_i = \overline{y}_{i,*} - \overline{y}$.
- $\hat{\epsilon}_{i,j} = y_{i,j} - \overline{y}_{i,*}$.
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\epsilon}_{i,j}^2}{n-I}$.

Remarque 1.1.2 *$\hat{\mu}, \hat{\alpha}_i, \hat{\sigma}^2$ sont des estimateurs sans biais de μ, α_i, σ^2 (resp).*

1.1.4 Décomposition de la variabilité

Remarquons tout d'abord l'égalité suivante :

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

En élevant au carré, il vient :

$$\begin{aligned}(y_{i,j} - \overline{y})^2 &= [(y_{i,j} - \overline{y}_{i,*}) + (\overline{y}_{i,*} - \overline{y})]^2 \\ &= (y_{i,j} - \overline{y}_{i,*})^2 + (\overline{y}_{i,*} - \overline{y})^2 + 2(y_{i,j} - \overline{y}_{i,*})(\overline{y}_{i,*} - \overline{y}).\end{aligned}$$

Par double sommation, on obtient :

$$\begin{aligned}\sum_{j=1}^{J}(y_{i,j} - \overline{y})^2 &= \sum_{j=1}^{J}(y_{i,j} - \overline{y}_{i,*})^2 + \sum_{j=1}^{J}(\overline{y}_{i,*} - \overline{y})^2 + 2\sum_{j=1}^{J}(y_{i,j} - \overline{y}_{i,*})(\overline{y}_{i,*} - \overline{y}) \\ &= \sum_{j=1}^{J}(y_{i,j} - \overline{y}_{i,*})^2 + J(\overline{y}_{i,*} - \overline{y})^2 + \underbrace{2(\overline{y}_{i,*} - \overline{y})\underbrace{\sum_{j=1}^{J}(y_{i,j} - \overline{y}_{i,*})}_{\overline{=}\,0}}_{(1,1)}\end{aligned}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j}-\overline{y})^2=\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j}-\overline{y}_{i,*})^2+\sum_{i=1}^{I}J(\overline{y}_{i,*}-\overline{y})^2.$$

(1,1) est due au fait que $(\overline{y}_{i,*}-\overline{y}_{*,*})$ ne dépend pas de j on met donc en facteur. D'autre part, cette quantité est nulle en raison de la nullité de la somme des écarts par rapport à la moyenne d'un échantillon (quantité toujours nulle par définition de la moyenne).
Donc on obtient l'équation fondamentale de l'analyse de la variance suivante :

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j}-\overline{y})^2}_{}\quad=\quad\underbrace{\sum_{i=1}^{I}J(\overline{y}_{i,*}-\overline{y})^2}_{}\quad+\quad\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j}-\overline{y}_{i,*})^2}_{}.$$

Variabilité	:	totale	inter-groupe = due au facteur	intra-groupe = résiduelle
Notations	:	SCE_T	SCE_A	SCE_R

On a alors la relation fondamentale de l'$AV(1)$:

$$SCE_T=SCE_A+SCE_R. \tag{1.2}$$

1.1.5 Les degrés de liberté

Les degrés de liberté (**ddl**) des différentes quantités de variation sont définis par la relation :

ddl = nombre de variables aléatoires − nombre de relations entre ces variables.

On y associe des degrés de liberté de (1.2) par :

$$n-1=(I-1)+(n-I).$$

★ Les sommes des carrés des écarts peuvent être divisées par leur nombres de degré de liberté respectifs, on obtient alors, les sommes des carrés moyennes :

$$\begin{cases} MC_T = SCE_T / (n-1). \\ MC_A = SCE_A / (I-1). \\ MC_R = SCE_R / (n-I). \end{cases}$$

1.1.6 Tableau d'analyse de la variance et test d'AV(1)

Nous construisons le tabeau d'analyse de la variance à partir des informations précédentes :

Variation	ddl	SCE	MC	F
Factorielle	$I-1$	SCE_A	$MC_A = \frac{SCE_A}{I-1}$	$\frac{MC_A}{MC_R}$
Résiduelle	$n-I$	SCE_R	$MC_R = \frac{SCE_R}{n-I}$	
Totale	$n-1$	SCE_T		

TAB. 1.2 – Tableau d'AV(1).

On souhaite de tester l'hypothèse d'égalité des moyennes :

$$\begin{cases} H_0 \ : \ \mu_1 = \mu_2 = \cdots = \mu_I, \\ H_1 \ : \ \exists (i \neq j) \in \{1, 2, \ldots, I\} \, | \mu_i \neq \mu_j. \end{cases} \Longleftrightarrow \begin{cases} H_0 \ : \ \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0, \\ H_1 \ : \ \exists i \in \{1, 2, \ldots, I\} \, | \alpha_i \neq 0. \end{cases}$$

Sous l'hypothèse H_0, on a :

$MC_A = \frac{SCE_A}{I-1} \sim \mathcal{X}^2_{I-1}$,

$MC_R = \frac{SCE_R}{n-I} \sim \mathcal{X}^2_{n-I}$,

où $\mathcal{X}^2_h$ est la loi de Khi-deux à h ddl.

Alors, la statistique F du test est définie par :

$$F := \frac{(n-I)SCE_A}{(I-1)SCE_R} \sim \mathcal{F}(I-1, n-I).$$

Donc, la région critique du test au niveau $\alpha \in (0,1)$ fixé s'écrit :

$$R.C. : F \geq f_{1-\alpha}(I-1, n-I).$$

où $f_{1-\alpha}$ est le fractile d'ordre $1-\alpha$ de la loi Fisher-Snédécor à $(I-1)$ et $(n-I)$ ddl.

Remarque 1.1.3 *Si on rejette H_0 avec le test ANOVA à un facteur précédent, on conclut qu'au moins deux moyennes sont différentes $(\exists \mu_i, \mu_{i'} | \mu_i \neq \mu_{i'})$. Dans ce cas, nous utilisons le test de comparaison deux à deux des couples $(\mu_i, \mu_{i'})$.*

1.1.7 Test de Bonferroni

On souhaite tester :

$$\begin{cases} H_0 : & \mu_i = \mu_{i'} \\ H_1 : & \mu_i \neq \mu_{i'} \end{cases}, i \neq i'.$$

Nombre de comparaisons

On a m tests de Student à faire :

$$\begin{aligned} m &= \complement_I^2 = \frac{I!}{(I-2)!2!} \\ &= \frac{(I \times (I-1))}{2}. \end{aligned}$$

Statistique de test

Le test de Bonferroni est un test de Student basé sur la statistique suivante :

$$T := \frac{\bar{y}_{i,*} - \bar{y}_{i',*}}{\left[(\frac{1}{n_i} + \frac{1}{n_{i'}})MCR\right]^{\frac{1}{2}}} \sim \mathcal{T}(n-I). \tag{1.3}$$

avec : $MCR = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{i,j} - \bar{y}_{i,*})^2$, $\mathcal{T}(h)$ est la loi de Student à h ddl.

Règle de décision

On accepte $H_0(\mu_i = \mu_{i'})$ si :

$$|T| < t_{(1-\frac{\alpha}{2m})}(n-I).$$

où $t_{(1-\frac{\alpha}{2m})}$ est le fractile d'ordre $1-\frac{\alpha}{2m}$ de la loi de Student à $n-I$ ddl.

Exemple 1.1.1 *Des chercheurs veulent tester le temps de réaction entre certains groupes du gens à propos de leurs avis sur la boisson la plus favorable (Eau, Jus, Café) [11], chaque groupe se compose de 5 personnes, le premier groupe prend de l'eau, le deuxième groupe prend du jus et le dernier groupe prend du café.*

Y-a-t-il une différence significative entre les boissons ?

C-à-d, le type de boisson a-t-il un effet sur le temps de réaction ?

(Temps de réaction mesuré en centièmes de seconde).

		
31	19	12
31	20	13
32	21	14
33	21	14
33	22	15

TAB. 1.3 – Données des temps de réaction pour chaque type de boisson.

On a : $I = 3$, $J = 5$, $\bar{y}_1 = 32$, $\bar{y}_2 = 20.6$, $\bar{y}_3 = 13.6$, $\bar{y} = 22.07$.

On veux tester au niveau de signification $\alpha = 0.05$,

$$\begin{cases} H_0 : & \mu_1 = \mu_2 = \mu_3. \\ H_1 : & \textit{Au moins deux moyennes sont différentes.} \end{cases}$$

Avec

Variation	ddl	SCE	MC	F
Factorielle	2	862.53	431.27	359.39
Résiduelle	12	14.4	1.2	
Totale	14	876.93		

TAB. 1.4 – Tableau d'AV(1) de l'exemple (1.1.1) .

On a : $F = 359.39 > f_{1-\alpha} = 3.89$. *Alors, on rejette* H_0; *il y a donc une différence significative entre les boissons.*

D'aprés le test de Bonferroni, on a $(m = 3)$ *tests de comparaison :*

1. $\begin{cases} H_0 : \ \mu_1 = \mu_2 \\ H_1 : \ \mu_1 \neq \mu_2 \end{cases} \Longrightarrow \ T_1 = 16.45 > t_{0.991}(12) = 2.68. \quad \textit{Alors :} \ \ \mu_1 \neq \mu_2.$

2. $\begin{cases} H_0 : \ \mu_1 = \mu_3 \\ H_1 : \ \mu_1 \neq \mu_3 \end{cases} \Longrightarrow \ T_2 = 26.56 > t_{0.991}(12) = 2.68. \quad \textit{Alors :} \ \ \mu_1 \neq \mu_{3.}$

3. $\begin{cases} H_0 : \ \mu_2 = \mu_3 \\ H_1 : \ \mu_2 \neq \mu_3 \end{cases} \Longrightarrow \ T_3 = 10.10 > t_{0.991}(12) = 2.68. \quad \textit{Alors :} \ \ \mu_2 \neq \mu_3.$

En conclusion, c'est la boisson qui fait la différence pas les gens.

1.2 Analyse de la variance à deux facteurs

Cette section est consacré à l'étude des situations expérimentales dans lesquelles l'effet de deux facteurs (variables qualitatives) est étudiée simultanément, c'est-à-dire dans le même protocole expérimental. En cela l'analyse de la variance à deux facteurs notée $AV(2)$ est une extension à la situation précédente dans laquelle on n'étudiait qu'un seul facteur à la fois $AV(1)$. Selon le regroupement des données, il résulte une analyse de la variance à deux facteurs avec répétitions et sans répétitions, alors dans la suite nous étudions ces deux types d'$AV(2)$.

1.3 Analyse de la variance à deux facteurs avec répétitions

L'étude de test $AV(2)$ avec répétitions simultanée d'un facteur A à I niveaux $(A_1, A_2, \ldots, A_I)$ et d'un facteur B à J niveaux $(B_1, B_2, \ldots, B_J)$ sur une variable quantitative Y.
Chaque couple (A_i, B_j) effectue K mesures d'une réponse Y qui est une variable continue.
Notons $n = I * J * K$ le nombre total de mesures ayant été effectuées.

Les données d'$AV(2)$ sont formulées comme suit :

	$Facteur\ B$				
$Facteur\ A$	B_1	$\ldots$	B_j	$\ldots$	B_J
A_1	$y_{1,1,1} \ldots y_{1,1,K}$	$\ldots$	$y_{1,j,1} \ldots y_{1,j,K}$	$\ldots$	$y_{1,J,1} \ldots y_{1,J,K}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	$y_{i,1,1} \ldots y_{i,1,K}$	$\ldots$	$y_{i,j,1} \ldots y_{i,j,K}$	$\ldots$	$y_{i,J,1} \ldots y_{i,J,K}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	$y_{I,1,1} \ldots y_{I,1,K}$	$\ldots$	$y_{I,j,1} \ldots y_{I,j,K}$	$\ldots$	$y_{I,J,1} \ldots y_{I,J,K}$

TAB. 1.5 – Les données d'AV(2) avec répétitions.

Où $y_{i,j,k}$ est la $k^{ème}$ réalisation de la variable quantitative Y, lorsque on fixe le premier facteur à la $i^{ème}$ niveau et le deuxième facteur à la $j^{ème}$ niveau ($i = \overline{1, I}$,$j = \overline{1, J}$,$k = \overline{1, K}$).

1.3.1 Modèle d'AV(2)

Nous introduisons le modèle :

$$\begin{array}{lccccccccc} y_{i,j,k} & = & \underbrace{\mu} & + & \underbrace{\alpha_i} & + & \underbrace{\beta_j} & + & \underbrace{(\alpha\beta)_{i,j}} & + & \underbrace{\epsilon_{i,j,k}}. \\ & & \text{moyenne} & & \text{effet} & & \text{effet} & & \text{effet} & & \text{effet} \\ & & \text{générale} & & \text{de A} & & \text{de B} & & \text{d'interaction} & & \text{résiduel} \end{array} \tag{1.4}$$

Avec :

- $\alpha_i = \mu_{i,*} - \mu$.
- $\beta_j = \mu_{*,j} - \mu$.
- $(\alpha\beta)_{i,j} = \mu_{i,j} - \mu_{i,*} - \mu_{*,j} + \mu$.
- $\epsilon_{i,j,k} = y_{i,j,k} - \mu_{i,j}$.

Hypothèses de modélisation :

- $\sum_{i=1}^{I} \alpha_i = 0$ et $\sum_{j=1}^{J} \beta_j = 0, \sum_{i=1}^{I} (\alpha\beta)_{i,j} = 0, \forall j = \overline{1, J}$ et $\sum_{j=1}^{J} (\alpha\beta)_{i,j} = 0, \forall i = \overline{1, I}$.
- $\epsilon_{i,j,k} \sim \mathcal{N}(0, \sigma^2), \forall i, j, k$.
- $cov(\epsilon_{i,j,k}; \epsilon_{r,s,t}) = 0, \forall (i, j, k) \neq (r, s, t)$.

1.3.2 Estimation des paramètres

Les estimateurs $\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j, \widehat{(\alpha\beta)}_{i,j}, \hat{\sigma}^2$ des paramètres $\mu, \alpha_i, \beta_j, (\alpha\beta)_{i,j}, \sigma^2$ du modèle (1.4) sont donnés par les formules suivantes :

- $\hat{\mu} = \overline{y}_{*,*,*}$.
- $\widehat{\alpha}_i = \overline{y}_{i,*,*} - \overline{y}_{*,*,*}$.
- $\widehat{\beta}_j = \overline{y}_{*,j,*} - \overline{y}_{*,*,*}$.
- $\widehat{(\alpha\beta)}_{i,j} = \overline{y}_{i,j,*} - \overline{y}_{i,*,*} - \overline{y}_{*,j,*} + \overline{y}_{*,*,*}$.
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{\epsilon}^2_{i,j,k}}{IJ(K-1)}$.

Remarque 1.3.1 *$\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j, \widehat{(\alpha\beta)}_{i,j}, \hat{\sigma}^2$ sont des estimateurs sans biais de $\mu, \alpha_i, \beta_j, (\alpha\beta)_{i,j}, \sigma^2$ (resp).*

1.3.3 Décomposition de la variabilité

Notons par :

$$\begin{array}{rcll}
\overline{y}_{i,*,*} & := & \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{i,j,k} & : \text{ moyenne de la ligne } i. \\
\overline{y}_{*,j,*} & := & \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{i,j,k} & : \text{ moyenne de la colonne } j. \\
\overline{y}_{i,j,*} & := & \frac{1}{K}\sum_{k=1}^{K} y_{i,j,k} & : \text{ moyenne de la case } (i,j). \\
\overline{y}_{*,*,*} = \overline{y} & := & \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{i,j,k} & : \text{ moyenne totale.}
\end{array}$$

On a alors :

$$y_{i,j,k} - \overline{y} = (\overline{y}_{i,*,*} - \overline{y}) + (\overline{y}_{*,j,*} - \overline{y}) + (\overline{y}_{i,j,*} - \overline{y}_{*,j,*} - \overline{y}_{i,*,*} - \overline{y}) + (y_{i,j,k} - \overline{y}_{i,j,*}),$$

L'équation d'$AV(2)$ s'écrit donc :

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{i,j,k}-\overline{y})^2}_{SCE_T} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{y}_{i,*,*}-\overline{y})^2}_{SCE_A} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{y}_{*,j,*}-\overline{y})^2}_{SCE_B}$$

$$+ \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{y}_{i,j,*}-\overline{y}_{*,j,*}-\overline{y}_{i,*,*}+\overline{y})^2}_{SCE_{AB}} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{i,j,k}-\overline{y}_{i,j,*})^2}_{SCE_R}.$$

Avec :

Variabilité	SCE	ddl
totale	$SCE_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{i,j,k}-\overline{y})^2$	$d_T := IJK-1$
due à A	$SCE_A = JK\sum_{i=1}^{I}(\overline{y}_{i,*,*}-\overline{y})^2$	$d_A := I-1$
due à B	$SCE_B = IK\sum_{j=1}^{J}(\overline{y}_{*,j,*}-\overline{y})^2$	$d_B := J-1$
due à l'interaction $A\times B$	$SCE_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{y}_{i,j,*}-\overline{y}_{*,j,*}-\overline{y}_{i,*,*}+\overline{y})^2$	$d_{AB} := (I-1)(J-1)$
résiduelle	$SCE_R = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{i,j,k}-\overline{y}_{i,j,*})^2$	$d_R := IJ(K-1)$

TAB. 1.6 – Somme des carrés et ddl pour les variabilités d'AV(2) avec répétitions.

1.3.4 Tableau d'analyse de la variance

Nous résumons ces informations dans le tableau ci-dessous :

Variation	ddl	SCE	MC	F
Facteur A	d_A	SCE_A	MC_A=SCE_A/d_A	F_A=MC_A/MC_R
Facteur B	d_B	SCE_B	MC_B=SCE_B/d_B	F_B=MC_B/MC_R
Facteur $A \times B$	d_{AB}	SCE_{AB}	MC_{AB}=SCE_{AB}/d_{AB}	F_{AB}=MC_{AB}/MC_R
Résiduelle	d_R	SCE_R	MC_R=SCE_R/d_R	
Totale	d_T	SCE_T		

TAB. 1.7 – Tableau d'AV(2) avec répétitions.

1.3.5 Test AV(2) avec répétitions

1. $$\begin{cases} (H_0)_A & : \quad \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0. \\ (H_1)_A & : \quad \exists i_0 \in [1, I] | \alpha_{i_0} \neq 0. \end{cases}$$

• Si $F_A > f_{1-\alpha}(I-1, IJ(K-1))$, on a $(H_0)_A$ est rejetée et on conclut que le facteur A a une influence significative sur le caractère étudié.

2. $$\begin{cases} (H_0)_B & : \quad \beta_1 = \beta_2 = \cdots = \beta_J = 0. \\ (H_1)_B & : \quad \exists j_0 \in [1, J] | \beta_{j_0} \neq 0. \end{cases}$$

• Si $F_B > f_{1-\alpha}(J-1, IJ(K-1))$, on a $(H_0)_B$ est rejetée et on conclut qu'il existe une influence significative sur le caractère étudié.

Remarque 1.3.2 *Lorsque l'hypothèse nulle est rejetée, nous pouvons procéder à des comparaisons multiples des différents effets des niveaux du deux facteurs A et B (voir la statistique (1.3)).*

3. $$\begin{cases} (H_0)_{AB} & : \quad (\alpha\beta)_{1,1} = (\alpha\beta)_{1,2} = \cdots = (\alpha\beta)_{1,J} = (\alpha\beta)_{2,1} = \cdots = (\alpha\beta)_{I,J} = 0. \\ (H_1)_{AB} & : \quad \exists (i_0, j_0) \in [1, I] \times [1, J] | (\alpha\beta)_{i_0, j_0} \neq 0. \end{cases}$$

• Si $F_{AB} > f_{1-\alpha}((I-1)(J-1), IJ(K-1))$, on a $(H_0)_{AB}$ est rejetée on conclut que l'interaction des deux facteurs a une influence significative sur le caractère étudié.

Exemple 1.3.1 *On garde la même idée de l'exemple précédent (1.1.1) et on ajoute une autre variable indépendante : Effet du moment de la journée "La matinée" - "Le soir".*

			
	32 33 33 34 34	30 32 29 31 34	27 28 27 30 31
	33 33 35 37 32	31 32 30 31 33	30 32 29 28 29

TAB. 1.8 – Données des temps de réaction pour chaque type de boisson dans la matinée et le soir.

Trois effets à tester :

1- Effet du moment de la journée sur le temps de réaction ?

2- Effet du type de boisson sur le temps de réaction ?

3- Effet d'interaction entre type de boisson et moment de la journée sur le temps de réaction ?

On a : $I = 2,\ J = 3,\ K = 5,\ n = 30.$

Le tableau d'ANOVA est comme suit :

Variation	ddl	SCE	MC	F
Facteur A	1	3.33	3.33	1.31
Facteur B	2	101.27	50.63	19.86
Facteur A×B	2	0.87	0.43	0.17
Résiduelle	24	61.2	2.55	
Totale	29	166.71		

TAB. 1.9 – Tableau d'AV(2) avec répétitions de l'exemple (1.2.1).

- $F_A = 1.31 < f_{1-0.05}(1, 24) = 4.26.$

  $\Longrightarrow$ *acceptation de* H_0.

  $\Longrightarrow$ *pas de différence significative entre les groupes quant à leur moyenne.*

  $\Longrightarrow$ *pas d'effet "Moment de la journée" sur "Le temps de réaction".*

- $F_B = 19.86 > f_{1-0.05}(2, 24) = 3.40.$

  $\Longrightarrow$ *rejet de* H_0.

  $\Longrightarrow$ *différence significative entre les groupes quant à leur moyenne.*

  $\Longrightarrow$ *effet "Type de boisson" sur "Le temps de réaction".*

- $F_{AB} = 0.17 < f_{1-0.05}(2, 24) = 3.40.$

  $\Longrightarrow$ *acceptation de* H_0.

  $\Longrightarrow$ *pas de différence significative entre les groupes quant à leur moyenne.*

  $\Longrightarrow$ *pas d'effet d'interaction de "Type de boisson" et "Moment de la journée" sur "Le temps de réaction".*

1.4 Analyse de la variance à deux facteurs sans répétitions

L'étude de test $AV(2)$ sans répétitions simultanée d'un facteur A à I niveaux $(A_1, A_2, \ldots, A_I)$ et d'un facteur B à J niveaux $(B_1, B_2, \ldots, B_J)$ sur une variable quantitative Y.
Chaque couple (A_i, B_j) effectue une seule mesure d'une réponse Y qui est une variable continue. Notons $n = I * J$ le nombre total de mesures ayant été effectuées.
Les données sont alors ;

	$Facteur\ B$				
$Facteur\ A$	B_1	$\ldots$	B_j	$\ldots$	B_J
A_1	$y_{1,1}$	$\ldots$	$y_{1,j}$	$\ldots$	$y_{1,J}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	$y_{i,1}$	$\ldots$	$y_{i,j}$	$\ldots$	$y_{i,J}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	$y_{I,1}$	$\ldots$	$y_{1,j}$	$\ldots$	$y_{I,J}$

TAB. 1.10 – Les données d'AV(2) sans répétitions.

où $y_{i,j}$ désigne la $i^{ème}$ observation du $j^{ème}$ niveau ($i = \overline{1, I}$ et $j = \overline{1, J}$).

1.4.1 Modèle d'AV(2)

Nous introduisons le modèle le plus simple qu'est d'additionner les effets du facteur A avec les effets du facteur B.

$$\begin{array}{ccccccccc} y_{i,j,k} & = & \underbrace{\mu} & + & \underbrace{\alpha_i} & + & \underbrace{\beta_j} & + & \underbrace{\epsilon_{i,j,k}}. \\ & & \text{moyenne} & & \text{effet} & & \text{effet} & & \text{effet} \\ & & \text{générale} & & \text{de A} & & \text{de B} & & \text{résiduel} \end{array} \tag{1.5}$$

Hypothèses de modélisation

- $\sum_{i=1}^{I} \alpha_i = 0.$
- $\sum_{j=1}^{J} \beta_j = 0.$
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \forall i, j.$
- $cov(\epsilon_{i,j}; \epsilon_{k,l}) = 0, \forall (i,j) \neq (k,l).$

Remarque 1.4.1 *Le modèle possède $I + J + 1$ paramètres.*

1.4.2 Estimation des paramètres

Les estimateurs $\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j, \hat{\sigma}^2$ des paramètres $\mu, \alpha_i, \beta_j, \sigma^2$ du modèle (1.5) sont donnés par les formules suivantes :

- $\hat{\mu} = \overline{y}_{*,*,*} = \overline{y}.$
- $\hat{\alpha}_i = \overline{y}_{i,*,*} - \overline{y}.$
- $\hat{\beta}_j = \overline{y}_{*,j,*} - \overline{y}.$
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{\epsilon}^2_{i,j,k}}{(I-1)(J-1)}.$

Remarque 1.4.2 *$\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j, \hat{\sigma}^2$ sont des estimateurs sans biais de $\mu, \alpha_i, \beta_j, \sigma^2$ (resp).*

1.4.3 Décomposition de la variabilité

Soit le modèle (1.5), on a :

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

On obtient :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j}-\overline{y})^2 \;=\; J\sum_{i=1}^{I}(\overline{y}_{i,*}-\overline{y})^2 \;+\; I\sum_{j=1}^{J}(\overline{y}_{*,j}-\overline{y})^2 \;+\; \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j}-\overline{y}_{i,*}-\overline{y}_{*,j}+\overline{y})^2.$$

totale	due à A	due à B	résiduelle
SCE_T	SCE_A	SCE_B	SCE_R
$d_T := IJ-1$	$d_A := I-1$	$d_B := J-1$	$d_R := (I-1)(J-1)$

1.4.4 Tableau d'analyse de la variance

Le tableau d'AV(2) sans répétitions est comme suit :

Variation	ddl	SCE	MC	F
Facteur A	d_A	SCE_A	$MC_A = SCE_A/d_A$	$F_A = MC_A/MC_R$
Facteur B	d_B	SCE_B	$MC_B = SCE_B/d_B$	$F_B = MC_B/MC_R$
Résiduelle	d_R	SCE_R	$MC_R = SCE_R/d_R$	
Totale	d_T	SCE_T		

TAB. 1.11 – Tableau d'AV(2) sans répétitions.

1.4.5 Test AV(2) sans répétitions

$$1 \quad \begin{cases} (H_0)_A & : \quad \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0. \\ (H_1)_A & : \quad \exists i_0 \in [1, I] | \alpha_{i_0} \neq 0. \end{cases}$$

• Si $F_A > f(I-1, (I-1)(J-1))$, alors le facteur A a une influence significative sur le caractère étudié.

$$2 \quad \begin{cases} (H_0)_B & : \quad \beta_1 = \beta_2 = \cdots = \beta_J = 0. \\ (H_1)_B & : \quad \exists j_0 \in [1, J] | \beta_{j_0} \neq 0. \end{cases}$$

• Si $F_B > f(J-1, (I-1)(J-1))$, alors le facteur B a une influence significative sur le caractère étudié.

Remarque 1.4.3 *Lorsque l'hypothèse $(H_0)_A$ (resp $(H_0)_B$) est rejetée, nous pouvons procéder à des comparaisons multiples des différents effets des niveaux du facteur (voir la statistique (1.3)).*

Remarque 1.4.4 *L'analyse de la variance à plusieurs facteurs $AV(p \geq 3)$ répond aux mêmes objectifs que l'analyse de la variance à un facteur et l'étude théorique reste la même.*

Chapitre 2

Comparaison de variances

L'une des conditions à satisfaire pour effectuer un test paramétrique est l'égalité des variances des populations dont sont extraits les échantillons pour réaliser son plan expérimental.

Les tests de comparaison de variances peuvent être utilisés aussi pour comparer les variabilités sur différentes populations.

Dans ce chapitre, nous présentons des tests traitent les K échantillons ensembles.

Les hypothèses à tester sont :

$$\begin{cases} H_0 : & \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2, \\ H_1 : & \exists i \neq j | \sigma_i^2 \neq \sigma_j^2. \end{cases} \tag{2.1}$$

2.1 Test de Bartlett

Le test de Bartlett sert à éprouver la comparaison de K variances. C'est une généralisation du test de Fisher et les hypothèses à tester sont les mêmes de test (2.1). On suppose que les K échantillons sont gaussiens de taille n_k $(1 \leq k \leq K)$, alors la statistique du test [6] définie par :

$$B = \frac{1}{C}\left[\hat{V}\ln(S_R^2) - \sum_{k=1}^{K}(V_k.\ln S_k^2)\right].$$

Avec :

- $S_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_{i,k} - \bar{X}_k)^2$: variance empirique.
- $V_k = n_k - 1$: nombre de ddl associé à S_k^2.
- $\hat{V} = \sum_{k=1}^{K} V_k$: chaque V_k est supérieur à 5.
- $S_R^2 = \frac{1}{\hat{V}} \sum_{k=1}^{K} V_k . S_k^2$: variance résiduelle.
- $C := 1 + \frac{1}{3(K-1)} \left[\sum_{k=1}^{K} \frac{1}{V_k} - \frac{1}{\hat{V}} \right]$

Sous l'hypothèse H_0, la variable B suit une loi du $\mathcal{X}^2$ à $(K-1)$ ddl. Alors, pour un risque α fixé on a la région critique du test s'écrit :

$$R.C. : B \geq b_{1-\alpha}(K-1).$$

où $b_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{X}^2$ à $(K-1)$ ddl.

Remarque 2.1.1 *(**Instruction** R) : La fonction à utiliser est **bartlett.test()**.*

Définition 2.1.1 *La p-value est la probabilité, sous H_0, d'obtenir une statistique aussi extrême que la valeur observée sur l'échantillon (c'est le risque de rejetter H_0).*

Exemple 2.1.1 *En reprenant l'exemple (1.1.1) du chapitre précédent, et on teste l'égalité des K variances sous* **R** *de la manière suivante :*

>tab = read.table("boisson.txt",header=TRUE,sep="\t") #Lecture de données tabulaires.

>names(tab) #Définir les noms du tableau.

[1] "Temps" "Type"

>bartlett.test(Temps~Type,tab) #Effectuer le test.

Bartlett test of homogeneity of variances

data : Temps by Type

Bartlett's K-squared = 0.08005, df = 2, p-value = 0.9608.

On remarque que la p-value est supérieur à 0.05, donc l'hypothèse d'égalité des variances est acceptée.

Nous allons utiliser des diagrammes en boîte à moustaches[1] pour comparer les variances intra-groupes, par la commande suivante :

> boxplot(tab$Temps~tab$Type,ylab="Temps (en jours)",xlab="Type",
col=c("blue","yellow","brown")) #Tracer un diagramme en boîte à moustaches.

Voici le graphe :

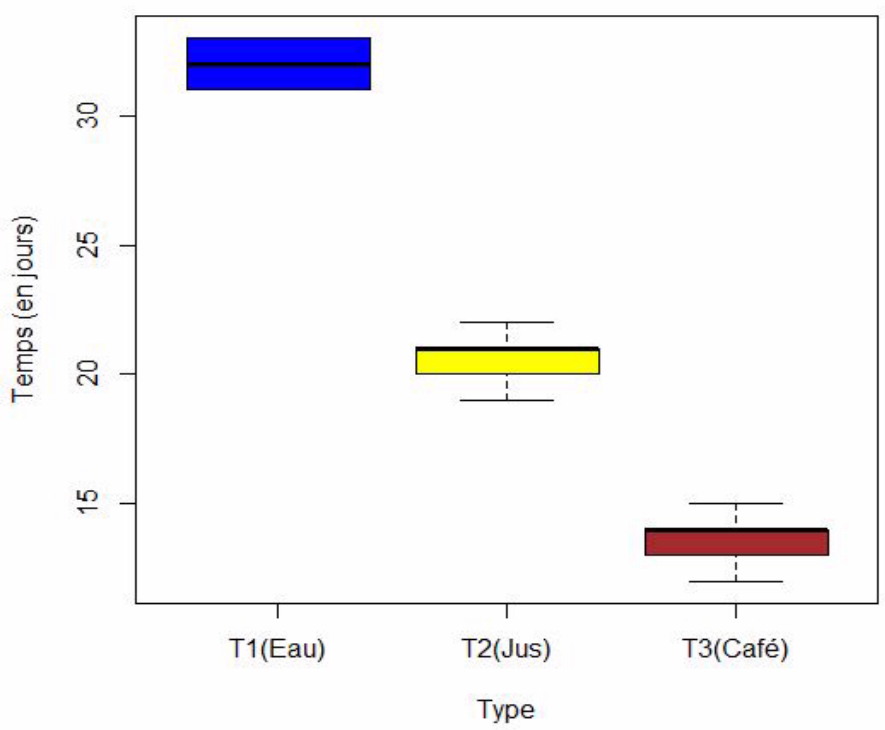


FIG. 2.1 – Boîtes à moustaches des temps de réaction pour chaque type de boisson.

[1]La boîte à moustaches (en anglais box-plot) est un outil graphique trés pratique représentant une distribution empirique à l'aide de quelques paramètres de localisation tels que la médiane (M), le 1^{er} et la $3^{ème}$ quartiles ; Q_1 et Q_3(resp).

2.2 Test de Levene

Soient K échantillons de taille $n_1, \ldots, n_k$. On définie une nouvelle variable aléatoire comme suit :

$$Z_{i,k} = \left| X_{i,k} - \bar{X}_k \right|.$$

où $\bar{X}_k$ est la moyenne empirique de l'échantillon k et $\bar{X}$ est la moyenne empirique totale. Sous l'hypothèse H_0 la statistique définie [7] par :

$$L = \frac{(n-K)\sum_{k=1}^{K} n_k(\bar{Z}_k - \bar{Z})^2}{(K-1)\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Z_{i,k} - \bar{Z}_k)^2} \sim F(K-1, n-K).$$

Alors pour α fixé, On a la région critique du test définie comme suit :

$$R.C. : L \geq l_{1-\alpha}(K-1, n-K).$$

où $l_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi Fisher-Snédécor à $K-1$ et $n-K$ ddl.

Remarque 2.2.1 *Le test de Levene consiste à effectuer une analyse de la variance sur une variable transformée $Z_{i,k}$ et la statistique de Levene est alors le rapport entre les variances inter-groupes et intra-groupes.*

Remarque 2.2.2 *(**Instruction** **R**) : Nous allons utiliser une fonction de la librairie **car** (Il faut installer ce package sur votre système). Nous allons utiliser la fonction **leveneTest** pour vérifier la condition d'application de l'égalité des variances intra-groupes.*

Exemple 2.2.1 *Considérons cinq traitements* $(T_1, \ldots, T_5)$ *contre les boutons de fièvre, dont un placebo, ont été administrés par tirage au sort à trente patients (six patients par groupe de traitement) [2]. Le délai (en jours) entre l'apparition des boutons et la cicatrisation complète a été recueilli chez chaque patient.*

Traitement				
$T_1(placebo)$	T_2	T_3	T_4	T_5
5	4	6	7	9
8	6	4	4	3
7	6	4	6	5
7	3	5	6	7
10	5	4	3	7
8	6	3	5	6

TAB. 2.1 – Données des délais pour chaque traitement.

L'objectif ici est de tester l'égalité des variances.

```
> tab= read.table("Boutons de fièvres.txt",header=TRUE,sep="\t")
> names(tab)
[1] "Délai" "Traitement"
> library(car)      #Chargé du package.
> leveneTest(Délai~factor(Traitement),data=tab)      #Effectuer le test.
Levene's Test for Homogeneity of Variance (center = median)
       ddl  F_emp    p_value
group  4    0.5851   0.6763
       25
```

La probabilité critique est de 0.6763 et dépasse donc le seuil habituel de 5%. On ne rejette pas significativement l'hypothèse nulle. Alors il est raisonnable de supposer l'égalité des variances intra-groupe.

À cette étape, il serait bon de tracer des diagrammes en boîte à moustaches des délais de cicatrisation pour chaque traitement. Il est clair que le minimum, les quartiles, la médiane et le maximum, partagent chaque série dans la figure en quatre groupes constitués chacun d'environ 25% de ces valeurs (Voir annexe [2.3]).

Pour cela, nous tapons la ligne de commande suivante :

> boxplot(tab$Délai~tab$Traitement,ylab="Délai (en jours)",xlab="Traitement", col=c("blue","yellow","brown","pink","green")) #Tracer un diagramme en boîte à moustaches.

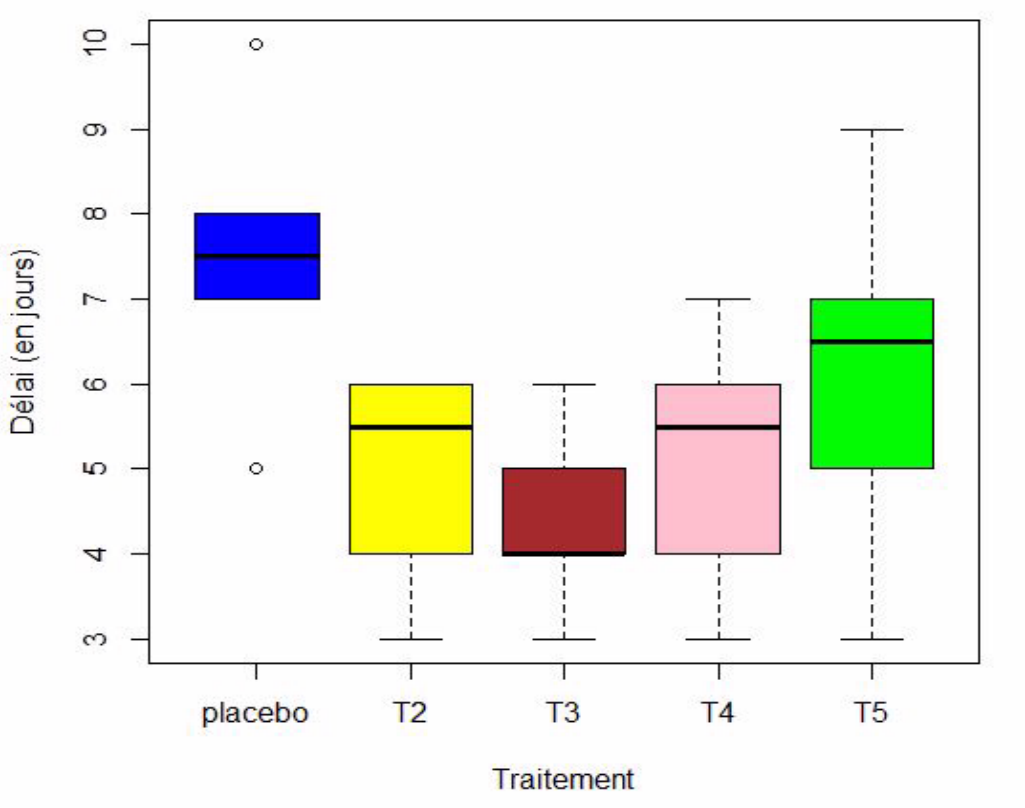


FIG. 2.2 – Boîtes à moustaches des délais de cicatrisation pour chaque traitement.

2.3 Test de Hartley

En statistique, le test de Hartley également connu sous le nom de test $F_{\max}$ ou $F_{\max}$ de Hartley, est utilisé dans l'analyse de la variance pour vérifier que les différents groupes ont même variance. Le test pour lequel nous intéressons à réaliser sera formulé d'aprés le test (2.1).

Deux conditions doivent être réunies pour utiliser ce test sont :

1. Les distributions de $(X_i)_{i=1,\dots,n}$ doivent être normales.
2. Les effectifs doivent être parfaitement équilibrés c-à-d, $n_1 = n_2 = \cdots = n_K = N$.

La statistique du test consiste à prendre le rapport entre la plus grande variance empirique et la plus petite [6] :

$$\begin{aligned} H = F_{\max} &= \frac{S^2_{\max}}{S^2_{\min}} \\ &:= \frac{\max(\hat{S}^2_1,\dots,\hat{S}^2_k)}{\min(\hat{S}^2_1,\dots,\hat{S}^2_k)}. \end{aligned}$$

où $F_{\max}$ suit une loi de Hartley de $(K, N-1)$ ddl.

Alors pour α fixé, on a la région critique du test définie comme suit :

$$R.C. : H \geq h_{1-\alpha}(K, N-1).$$

où le quantile $h_{1-\alpha}(K, N-1)$ à K et $N-1$ ddl est lue dans une table spécifique (Annexe [2.3]).

Exemple 2.3.1 *Soient les données d'AV*(1)*, avec une variable d'intérêt* Y *et un facteur* A *à* 3 *niveaux :* "1","2" *et* "3"*.[10].*

A	1	1	1	1	1	1	1	2	2	2
Y	8.14	9.05	1.26	9.13	6.32	0.97	2.78	5.46	9.57	9.64
A	2	2	2	3	3	3	3	3	3	3
Y	1.57	9.70	9.57	4.85	8.01	1.41	4.21	9.15	7.92	9.59

Nous pouvons désormais calculer le $F_{\max}$ de Hartley en fonction **R** *comme suit :*

```
> DATA=data.frame(A=factor(c("1","1","1","1","1","1","1","2","2","2","2","2","2",
+ "3","3","3","3","3","3","3")),Y=c(8.14,9.05,1.26,9.13,6.32,0.97,2.78,5.46,9.57,
9.64,1.57,9.70,9.57,4.85,8.01,1.41,4.21,9.15,7.92,9.59))
> # Adaptation de la base de données.
> DATA = na.omit(DATA)
> DATA = as.data.frame(DATA)
> DATA[,1] = as.factor(DATA[,1])
> # Récupération des informations descriptives.
> n = dim(DATA)[1]
> P = dim(DATA)[2]-1
> biblioY = summary(DATA[,1])
> K = length(biblioY)
> # Création de la matrice de résultats finaux.
> Resultats = matrix(0,P,K+10)
> # Application du Fmax de Hartley pour toutes les variables de la base de données.
> variance = matrix(0,1,K)
> # Calcul des différentes variances.
> for (k in 1 :K) {variance[k] = var(DATA[which(DATA[,1]==names(biblioY[k])),2])}
> # Recherche des valeurs minimales et maximales des variances.
> MAXv = max(variance)
> MINv = min(variance)
> # Calcul du Fmax de Hartley
> Fmax = MAXv/MINv
# Affichage des résultats.
> MINv
[1] 9.120548
```

> MAXv

[1] 13.20251

#Fmax_Hartley.

> Fmax

[1] 1.447557

Si nous comparons la statistique de test $F_{\max} = 1.447557$ *à la table de la loi de Hartley pour les degrés de liberté suivantes :* $(3, 20-1) = (3, 19)$,

- *Au niveau de signification* $\alpha = 5\%$, *on trouve* $h(3, 19) = 2.95$.
- *Au niveau de signification* $\alpha = 1\%$, *on trouve* $h(3, 19) = 3.8$.

Nous en concluons que nous ne pouvons rejeter H_0 *et donc que les variances sont égales.*

Remarque 2.3.1 *Il existe autres tests possibles d'égalité des variances : test de Cochran, Test de O'Brien, Test de Brown-Forsythe. Pour plus de détails voir [7].*

Conclusion

En conclusion, nous pouvons dire que l'ANOVA vérifie l'hypothèse selon laquelle les moyennes de deux populations ou plus sont égales. ANOVA évalue l'importance d'un ou plusieurs facteurs en comparant les moyennes des variables de réponse pour les différents niveaux de facteurs. L'hypothèse nulle stipule que toutes les moyennes de la population sont égales, tandis que l'hypothèse alternative stipule qu'au moins l'une d'elles diffère des autres.

Cette analyse est appelée "Analyse de la variance" parce que sa procédure s'appuie sur les variances pour déterminer si les moyennes sont différentes. La procédure compare la variance entre les moyennes des groupes et la variance à l'intérieur des groupes afin de déterminer si les groupes font tous partie d'une population plus élargie ou de populations distinctes ayant des caractéristiques propres.

Généralement, ANOVA est utile en sciences sociales dans l'analyse de certaines données, organisées en blocs de même taille. Il s'agit dans ce cas, d'ANOVA à un seul facteur. ANOVA à deux facteurs est en revanche fréquente dans l'exploitation d'enquêtes d'usage psychologique. C'est dans l'exploitation des résultats d'enquêtes d'usage que l'analyse de la variance est indispensable et qu'elle prouve toute son utilité pour le psychologue.

L'analyse de la variance est la meilleure du point de vue statistique parce qu'elle crée des groupes de même nature, et les limites de classes sont des valeurs réelles de la distribution. L'une des inconvénients d'ANOVA est qu'il est peu de stabilité en augmentant le nombre de classes.

Bibliographie

[1] Bertrand, F., & Maumy, M. (2011). Analyses de la variance. Manuel de cours, université de Strasbourg. 140p.

[2] De Micheaux, P. L., Drouilhet, R., & Liquet, B. (2011). Le logiciel R : Maitriser le langage-Effectuer des analyses statistiques. Springer Science & Business Media.

[3] Faraway, J. J. (2002). Practical regression and ANOVA using R.

[4] Larsen, R. J., & Marx, M. L. (2017). An introduction to mathematical statistics and its applications (Vol. 5). Pearson.

[5] Maumy-Bertrand, M., & Bertrand, F. (2010). Initiation à la statistique avec R : Cours, exemples, exercices et problèmes corrigés. Dunod.

[6] NICOLAS, Savy, Cours de Statistiques. Université de Bretagne Occidentale, Formation Ingénieur des Techniques de l'Agro-Alimentaire, Première Année.

[7] Rakotomalala, R. (2013). Comparaison de populations : Tests paramétriques.

[8] Sahai, H., & Ojeda, M. M. (2004). Analysis of Variance for Random Models, Volume 2 : Unbalanced Data : Theory, Methods, Applications, and Data Analysis (Vol. 2). Springer Science & Business Media.

[9] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

[10] https ://lemakistatheux.wordpress.com/category/tests-statistique-indices-de-liaison-et-coefficients-de-correlation/le-fmax-de-hartley.

[11] https ://youtu.be/ITf4vHhyGpc.

Annexe A : Logiciel *R*

Qu'est-ce-que le langage R ?

• Le langage **R** est un langage de programmation et un environnement mathématique utilisé pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques trés utiles et de qualité professionnelle.

• **R** a été crée par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Référence des fonctions de R utilisées

read.table(file) : lit un fichier au format tabulaire et en fait un data frame ; le séparateur de colonne par défaut **sep=""** désigne n'importe quel espacement ; utilisez **header=TRUE** pour prendre la première ligne comme titre (header) de colonne.

data(x) : charge de données spécifiées.

xlab=, **ylab=** : titre des axes (caractères).

col : couleur(s) des symboles et lignes ; exemple, col="red","blue".

main : titre du graphique (caractères).

library(x) : charge des packages additionnels.

data.frame(...) : crée un data frame avec les arguments (nommés ou non).

factor(x,levels=) : transforme un vecteur **x** en factor (les niveaux sont indiqués par **levels=**).

na.omit(x) : supprime les observations avec des valeurs manquantes (**NA** : not available) ; supprime les lignes correspondantes si **x** est une matrice ou un data.frame.

as.data.frame(x), as.factor(x) : conversion de type.

dim(x) : récupère ou définit (**dim(x) <- c(3,2)**) les dimensions d'un objet.

summary(a) : donne un « résumé » de **a**, généralement un résumé statistique, mais c'est une fonction générique (fonctionne différemment selon la classe de **a**).

length(x) : nombre d'éléments dans **x.**

matrix(x,nrow=,ncol=) : crée une matrice (tous les éléments sont de même type) ; les éléments se répètent s'ils sont trop courts.

boxplot(x) : diagramme en boîte [boîte à moustaches] ; la boîte et son milieu montrent les 3 quartiles ; les moustaches (whisker) un intervalle de confiance de 95% pour la médiane (s'il y a des valeurs en dehors, elles sont affichées).

max(x) : maximum des éléments de **x.**

min(x) : minimum des éléments de **x.**

var(x) : variance des éléments de **x** (calculé avec n-1 au dénominateur) ; si **x** est une matrice ou un data.frame, la matrice de variance est calculée.

bartlett.test, **levene.test** : tests d'égalité des variances.

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$ANOVA$	:	Analyse de la variance.
$AV(1)$	:	Analyse de la variance à un facteur.
$AV(2)$	:	Analyse de la variance à deux facteurs.
ddl	:	Degré de liberté.
SCE	:	Somme des carrés des écarts.
MC_R	:	Moyenne des carrés résiduelle.
$R.C$	:	Région critique.
H_0	:	Hypothèse nulle.
H_1	:	Hypothèse alternative.
F_{emp}	:	Fonction de répartition empirique.
$resp$	:	Respectivement.
indép	:	Indépendant.
c-à-d	:	C'est à dire.
min	:	Minimum.
max	:	Maximum.

$E(.)$	:	Espérence.
$cov(.)$	:	Covariance.
$\sum_{i=1}^{I}$	:	La somme de i jusqu'à I.
χ^2	:	Loi du Khi-deux.
$\|.\|$	:	La valeur absolue.
$\exists$	:	Il existe.
$\forall$	:	Pour tout, quel que soit.
C_n^k	:	Combinaison de k parmi n, tels que $\mathsf{C}_n^k = \frac{n!}{k!(n-k)!}$.
$N(\mu, \sigma^2)$	:	Loi normale d'esperence μ et de variance σ^2.
ln	:	La fonction logarithme.
α	:	Risque de premier espèce.
S^2	:	Variance empirique.
$\neq$	:	Différent (de).
$\hat{}$	:	Puissance.
$\sim$	:	Environ.
$:=$	:	Signifie "est égal, par définition, à".

Annexe C : Boîte à moustaches

La boîte à moustaches une traduction de Box & Whiskers Plot, est une invention de JOHN TUKEY (1977) pour représenter schématiquement une distribution.

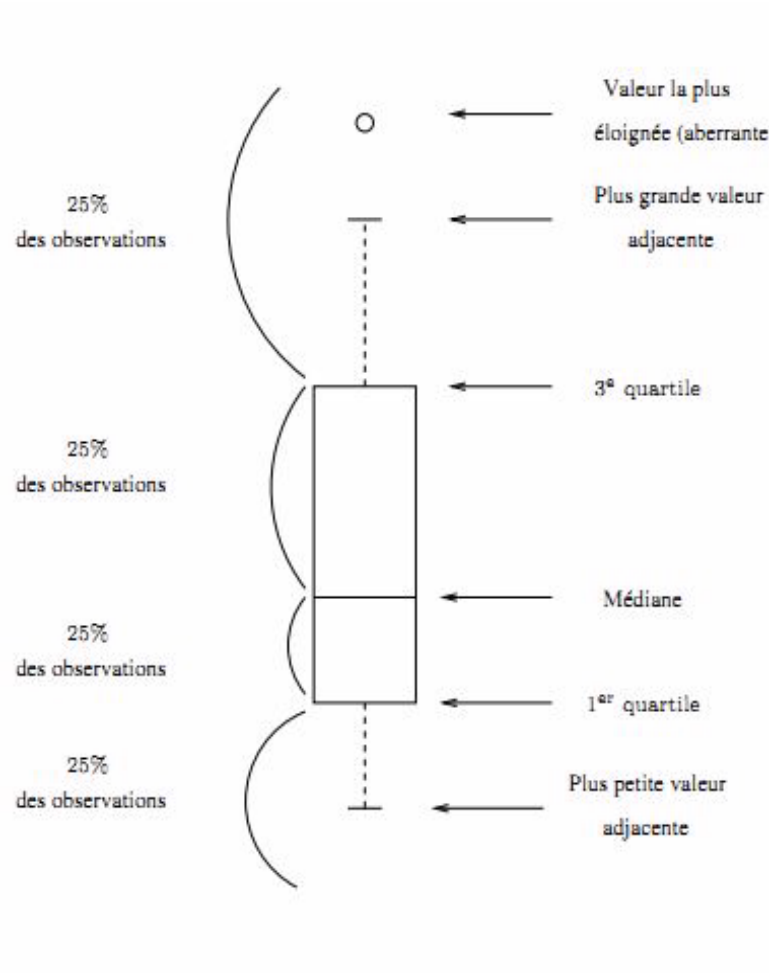


FIG. 2.3 – Boîte à moustaches et explications associées.

Annexe D : Table statistique

Ci-dessous la table de la loi de Hartley :

		Nombre de groupes										
ddl	α	2	3	4	5	6	7	8	9	10	11	12
2	0,05	39	87,5	142	202	266	333	403	475	550	626	704
	0,01	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605
3	0,05	15,4	27,8	39,2	50,7	62	72,9	83,5	93,9	104	114	124
	0,01	47,5	85	120	151	184	216	249	281	310	337	361
4	0,05	9,6	15,5	20,6	25,2	29,5	33,6	37,5	41,4	44,6	48	51,4
	0,01	23,2	37	49	59	69	79	89	97	106	113	120
5	0,05	7,15	10,8	13,7	16,3	18,7	20,8	22,9	24,7	26,5	28,2	29,9
	0,01	14,9	22	28	33	38	42	46	50	54	57	60
6	0,05	5,82	8,38	10,4	12,1	13,7	15	16,3	17,5	18,6	19,7	20,7
	0,01	11,1	15,5	19,1	22	25	27	30	32	34	36	37
7	0,05	4,99	6,94	8,44	9,7	10,8	11,8	12,7	13,5	14,3	15,1	15,8
	0,01	8,89	12,1	14,5	16,5	18,4	20	22	23	24	26	27
8	0,05	4,43	6	7,18	8,12	9,03	9,8	10,5	11,1	11,7	12,2	12,7
	0,01	7,5	9,9	11,7	13,2	14,5	15,8	16,9	17,9	18,9	19,8	21
9	0,05	4,03	5,34	6,31	7,11	7,8	8,41	8,95	9,45	9,91	10,3	10,7
	0,01	6,54	8,5	9,9	11,1	12,1	13,1	13,9	14,7	15,3	16	16,6
10	0,05	3,72	4,85	5,67	6,34	6,92	7,42	7,87	8,28	8,66	9,01	9,34
	0,01	5,85	7,4	8,6	9,6	10,4	11,1	11,8	12,4	12,9	13,4	13,9
12	0,05	3,28	4,16	4,79	5,3	5,72	6,09	6,42	6,72	7	7,25	7,48
	0,01	4,91	6,1	6,9	7,6	8,2	8,7	9,1	9,5	9,96	10,2	10,6
15	0,05	2,86	3,54	4,01	4,37	4,68	4,95	5,19	5,4	5,59	5,77	5,93
	0,01	4,07	4,9	5,5	6	6,4	6,7	7,1	7,3	7,5	7,8	8
20	0,05	2,46	2,95	3,29	3,54	3,76	3,94	4,1	4,24	4,37	4,49	4,59
	0,01	3,32	3,8	4,3	4,6	4,9	5,1	5,3	5,5	5,6	5,8	5,9
30	0,05	2,07	2,4	2,61	2,78	2,91	3,02	3,12	3,21	3,29	3,36	3,39
	0,01	2,63	3	3,3	3,4	3,6	3,7	3,8	3,9	4	4,1	4,2
60	0,05	1,67	1,85	1,96	2,04	2,11	2,17	2,22	2,26	2,3	2,33	2,36
	0,01	1,96	2,2	2,3	2,4	2,4	2,5	2,5	2,6	2,6	2,7	2,7

Annexe E : Biographie

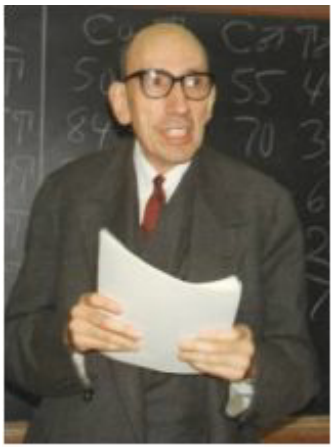
Ronald Aylmer Fisher

Sir Ronald Aylmer Fisher, est un biologiste et statisticien britannique. Dans le domaine des statistiques, il a introduit de nombreux concepts clés tels que le maximum de vraisemblance, l'information de Fisher et l'analyse de la variance, les plans d'expériences ou encore la notion de statistique exhaustive.

Date et lieu de naissance : 17 février 1890, East Finchley (Angleterre).

Date et lieu de décès : 29 juillet 1962 (à 72 ans), Adélaïde (Australie).

Howard Levene



Howard Levene, est un professeur américain de statistiques mathématiques et de génétique à l'Université Columbia. Le créateur du test d'homogénéité de la variance de Levene (Publié en 1960).

Date de naissance : né en 1914.

Date de décès : est décédé le 2 juillet 2003.

Carlo Emilio Bonferroni

Carlo E.Bonferroni, était un mathématicien italien, spécialiste en théorie des probabilités. Il est surtout connu pour les inégalités de Bonferroni (une généralisation de l'union), et pour la correction de Bonferroni dans les statistiques (qu'il n'a pas inventées mais qui utilise ses inégalités).

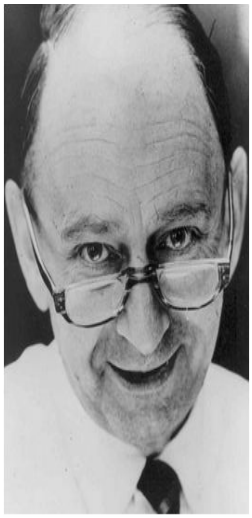
Date et lieu de naissance : 28 janvier 1892 Bergame.

Date et lieu de décès : 18 août 1960 Florence.

Activités : Mathématicien, statisticien.

H.O.Hirschfeld

Hermann Otto Hirschfeld connu aussi sous le nom de H.O.Hartley et plus familièrement HOH, fut un statisticien germano-américain. Il inventa le test d'Hartley d'égalité des variances en 1950. En 1967, avec J.N.K. Rao il publia une méthode de Maximum de vraisemblance, pour trouver les composants de la variance dans les modèles mixes.



Il contribua significativement à la programmation, à l'optimisation et à l'échantillonnage des sondages. Il fonda le département de Statistiques de l'université A&M du Texas. Il fut le $74^{ème}$ président de l'American Statistical Association.

Date et lieu de naissance : 13 avril 1912, Berlin, Allemagne.

Date et lieu de décès : 1980, Durham, Caroline du Nord, Ètats-Unis

Livres : Contributions to Survey Sampling and Applied Statistics : Papersin Honor of H. O. Hartley.

Maurice Bartlett

Maurice Stevenson Bartlett était un statisticien anglais qui a contribué à l'analyse des modèles temporels et spatiaux. Il est à l'origine de la Méthode de Bartlett. Il est aussi connu pour ses travaux sur l'inférence statistique.

Date et lieu de naissance : 18 juin 1910, Londres, Royaume-Uni.

Date et lieu de décès : 8 janvier 2002, Exmouth, Royaume-Uni.

Livres : An introduction to stochastic processes, with special reference to methods and applications,...,etc.

George Snedecor

George Waddel Snedecor est un statisticien américain. Il a été un pionnier des méthodes d'analyse de la variance des plans d'expériences et a élaboré des tests statistiques. Le test F de Fisher-Snedecor crée en 1920 grâce aux travaux de Ronald Aylmer Fisher et basé sur la table de distribution de George Waddell Snedecor.

Date et lieu de naissance : 20 octobre 1881, Memphis, Tennessee, États-Unis.

Date et lieu de décès : 15 février 1974, Amherst, Massachusetts, États-Unis.

Renommé pour : loi de Fisher-Snedecor.

Distinctions : prix Samuel Wilks (1970).

Livres : Statistical methods,. . . ,etc.

William Sealy Gosset

William Gosset connu sous le pseudonyme Student est un statisticien anglais.

Il a ainsi inventé le test de Student.

Date et lieu de naissance : 13 juin 1876, Canterbury, Royaume-Uni.

Date et lieu de décès : 16 octobre 1937, Beaconsfield, Royaume-Uni.

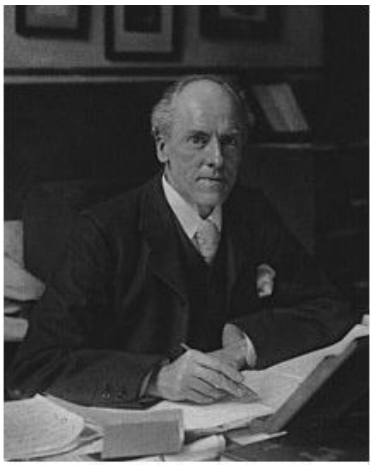
Livres : Student : A Statistical Biography of William Sealy Gosset.

Enseignement : Université d'Oxford, New College, Winchester College.

Karl Pearson

Karl Pearson, mathématicien britannique, est un des fondateurs de la statistique moderne.

Il est aujourd'hui principalement connu pour avoir développé le coefficient de corrélation et le Test du $\mathcal{X}^2$.



Date et lieu de naissance : 27 mars 1857, Islington Londres (Angleterre).

Date et lieu de décès : 27 avril 1936 (à 79 ans) Coldharbour, Surrey (Angleterre).

Renommé pour : Fonction de Pearson.