Chapitre 3

ANOVA à un facteur

Ce chapitre introduit le contexte, la notation et les propriétés probalistes d'une analyse de la variance à un facteur.

3.1 Introduction et notation

Considérons I populations indépendantes. Pour $i = 1, 2, \cdots, I$, on dispose d'un échantillon $\{Y_{i1}, Y_{i2}, \cdots, Y_{in_i}\}$ de taille n_i issu d'une loi $\mathrm{N}(\mu_i, \sigma^2)$. Notons que σ^2 ne varie pas d'un échantillon à l'autre. On fait une hypothèse d'égalité de variance . Soit $N = n_1 + n_2 + \cdots + n_I$ la taille totale de l'ensemble des échantillons. À partir de ces échantillons, on voudrait tester les hypothèses:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_I \\ H_1 &: \text{il existe } i \neq i' \text{ tels que } \mu_i \neq \mu_{i'} \end{aligned}$$

Pour $i = 1, 2, \cdots, I$, posons la moyenne échantillonale

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

On a alors $\bar{Y}_{i.} \sim \mathrm{N}(\mu_i, \sigma^2/n_i)$.

La moyenne totale est définie par

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N}\sum_{i=1}^{I} n_i \bar{Y}_{i\cdot}$$

On a alors $\bar{Y}_{\cdot\cdot} \sim \mathrm{N}(\mu, \sigma^2/N)$ où μ est la moyenne de $\{\mu_1, \mu_2, \cdots, \mu_I\}$ pondérée par $\{n_1, n_2, \cdots, n_I\}$.

$$\mu = \frac{1}{N}\sum_{i=1}^{I} n_i \mu_i$$

Sous H_0, μ est la valeur commune des μ_i.

3.2 Décomposition des carrés

Considérons la somme suivante

$$\mathrm{SST} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

C'est la somme des carrés des écarts des observations Y_{ij} par rapport à la moyenne globale $\bar{Y}_{\cdot\cdot}$. Elle mesure la variabilité totale des observations. Son nom SST fait référence à *Sum of Squares Total*.
La décomposition des carrés est une technique très importante en ANOVA. Elle consiste à décomposer SST en deux termes indépendants SSW ou *Sum of Squares Within (samples)* et SSB ou *Sum of Squares Between (samples)*.
La SSW est définie par:

$$SSW = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 = \sum_{i=1}^{I}(n_i - 1)S_i^2,$$

où

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

est la variance échantillonale du traitement i. Le terme SSW est ainsi une somme de termes qui mesurent chacun la variabilité à l'intérieur d'un échantillon. Or une hypothése cruciale

du modèle d'ANOVA est l'homogénéité des variances. Ce terme est lié alors à la mesure de cette variabilité commune. Ceci sera confirmé par les calculs plus tard.
La SSB est définie par:

$$SSB = \sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

Elle mesure l'écart entre la moyenne échantillonale et la moyenne globale. Si ces écarts sont considérable, on a tendance à favoriser l'hypothèse alternative, c'est-à-dire que les moyennes des traitements sont différents.

PROPOSITION **3.1** *La décomposition suivante est toujours vraie,*

$$SST = SSB + SSW$$

En effet, on a:

$$\begin{aligned} SST &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}((Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..}))^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) \end{aligned}$$

Or,

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\{(Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..})\} = \sum_{i=1}^{I}\{(\bar{Y}_{i.} - \bar{Y}_{..})\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})\} = 0$$

car $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.}) = 0$ pour $i = 1, 2, \cdots, I$.
donc

$$\begin{aligned} SST &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ &= SSB + SSW \end{aligned}$$

Cette décomposition est semblable à celle obtenue en régression où la somme de carrés totale est décomposé par une somme de carrés expliquée par la régression plus une somme de carrés résiduelle.

PROPOSITION **3.2** *Sous les hypothèses de normalité et d'égalité des variances, on a*

$$\frac{SSB}{\sigma^2} \sim \chi^2_{I-1}(\delta)$$

où $\delta = \sum_{i=1}^{I} n_i(\mu_i - \mu)^2$

Il est facile de voir que $SSB = \sum_{i=1}^{I} n_i \bar{Y}_{i.}^2 - N\bar{Y}_{..}^2$. Posons $Z_i = \sqrt{n_i}\bar{Y}_{i.}$ pour $i = 1, 2, \cdots, I$. On a alors

$$Z_i \sim \mathrm{N}(\sqrt{n_i}\mu_i, \sigma^2) \text{ et } \bar{Y}_{..} = \frac{1}{N}\sum_{i=1}^{I} \sqrt{n_i} Z_i$$

On en déduit que

$$\begin{aligned} SSB &= \sum_{i=1}^{I} Z_i^2 - (\frac{1}{\sqrt{N}}\sum_{i=1}^{I}\sqrt{n_i}Z_i)^2 \\ &= \sum_{i=1}^{I}(1 - \frac{n_i}{N})Z_i^2 - 2\sum_{i<j}\frac{\sqrt{n_i}\sqrt{n_j}}{N}Z_iZ_j \end{aligned}$$

SSB s'exprime alors comme une forme quadratique en $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_I)^T$.
La matrice A associée à cette forme quadratique s'exprime sous la forme $\mathbf{A} = \mathbf{I_I} - \frac{\mathbf{1}}{\mathbf{N}}\nu\nu^{\mathbf{T}}$ où $\nu = (\sqrt{\mathbf{n_1}}, \sqrt{\mathbf{n_2}}, \cdots, \sqrt{\mathbf{n_I}})^{\mathbf{T}}$.
Calculons $\mathbf{A^2}$. On a:

$$\begin{aligned} \mathbf{A^2} &= (\mathbf{I_I} - \frac{\mathbf{1}}{\mathbf{N}}\nu\nu^{\mathbf{T}}) \times (\mathbf{I_I} - \frac{\mathbf{1}}{\mathbf{N}}\nu\nu^{\mathbf{T}}) \\ &= \mathbf{I_I} - \frac{\mathbf{1}}{\mathbf{N}}\nu\nu^{\mathbf{T}} - \frac{\mathbf{1}}{\mathbf{N}}\nu\nu^{\mathbf{T}} + \frac{\mathbf{1}}{\mathbf{N^2}}\nu\nu^{\mathbf{T}}\nu\nu^{\mathbf{T}} \end{aligned}$$

On remarque que $\nu^{\mathbf{T}}\nu = \sum_{i=1}^{I}\sqrt{n_i}\sqrt{n_i} = N$. Donc les deux derniers termes de la partie droite de la dernière équation s'annulent et on a $\mathbf{A^2} = \mathbf{A}$.
D'après le théorème 1.10, SSB/σ^2 suit alors une loi du khi-deux $\chi^2_d(\delta)$ où d est l'ordre de la matrice $\mathbf{A}$, $\delta = \xi^{\mathbf{T}}\mathbf{A}\xi/\sigma^2$ et $\xi = E[\mathbf{Z}] = (\sqrt{\mathbf{n_1}}\mu_{\mathbf{1}}, \sqrt{\mathbf{n_2}}\mu_{\mathbf{2}}, \cdots, \sqrt{\mathbf{n_I}}\mu_{\mathbf{I}})^{\mathbf{T}}$.

L'ordre de la matrice $\mathbf{A}$ est égal à $I-1$. D'autre part, on a

$$\begin{aligned}\xi^{\mathbf{T}}\mathbf{A}\xi &= \xi^{\mathbf{T}}\xi - \frac{1}{N}\xi^{\mathbf{T}}\mu\mu^{\mathbf{T}}\xi \\ &= \sum_{i=1}^{I} n_i\mu_i^2 - \frac{1}{N}(\sum_{i=1}^{I} n_i\mu_i)^2 \\ &= \sum_{i=1}^{I} n_i\mu_i^2 - N\mu^2 \\ &= \sum_{i=1}^{I} n_i(\mu_i-\mu)^2\end{aligned}$$

D'où $\delta = \sum_{i=1}^{I} n_i(\mu_i-\mu)^2/\sigma^2$.

PROPOSITION **3.3** *Sous les hypothèses habituelles de l'ANOVA, on a:*

$$\frac{SSW}{\sigma^2} \sim \chi^2_{N-I}$$

En effet, pour $i = 1, 2, \cdots, I$, d'après la proposition 1.12, on a $(n_i-1)S_i^2/\sigma^2 \sim \chi^2_{n_i-1}$. Par indépendance des échantillons, $SSW/\sigma^2 = \sum_{i=1}^{I}(n_i-1)S_i^2/\sigma^2 \sim \chi^2_{N-I}$ puisque $\sum_{i=1}^{I}(n_i-1) = N-I$.

PROPOSITION **3.4** *Sous les hypothèses habituelles de l'ANOVA, les statistiques SSB et SSW sont indépendantes et on a*

$$\frac{SST}{\sigma^2} \sim \chi^2_{N-1}(\sum_{i=1}^{I} n_i(\mu_i-\mu)^2/\sigma^2)$$

Pour $i = 1, 2, \cdots, I$, les statistiques $\bar{Y}_{i.}$ et S_i^2 sont indépendantes d'après la proposition 1.12. Les statistiques SSB et SSW sont donc indépendantes puisque la première est une fonction de $\{\bar{Y}_{1.}, \bar{Y}_{2.}, \cdots, \bar{Y}_{I.}\}$ et la deuxième est une fonction de $\{S_1^2, S_2^2, \cdots, S_I^2\}$. D'après le premier exercice des travaux pratiques # 2, $SST = SSB + SSW \sim \chi^2_{N-1}(\sum_{i=1}^{I} n_i(\mu_i-\mu)^2/\sigma^2)$.

PROPOSITION **3.5** *Sous les hypothèses habituelles de l'ANOVA, posons*

$$F = \frac{SSB/(I-1)}{SSW/(N-I)}.$$

On a alors $F \sim F_{I-1,N-I}(\sum_{i=1}^{I} n_i(\mu_i-\mu)^2/\sigma^2)$

Cette dernière proposition est une conséquence directe de la définition d'une loi de Fisher non centrée.
Sous H_0, on a $\mu_1 = \mu_2 = \cdots = \mu_I = \mu$, d'où $\sum_{i=1}^{I} n_i(\mu_i - \mu)^2/\sigma^2 = 0$. Sous cette hypothèse, les trois statistiques SST, SSB et SSW sont distribuées selon des lois du Khi-deux centrées, à $N-1$, $I-1$ et $N-I$ degrés de liberté respectivement.

3.3 Test d'égalité des moyennes avec variance connue

Dans cette section, on élabore le test d'égalité des moyennes $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ versus H_1 : il existe $i \neq i'$ tels que $\mu_i \neq \mu_{i'}$. La construction de ce test est basée sur la méthode du rapport des maximums de vraisemblance. Lorsque la variance commune σ^2 est connue, la vraisemblance totale s'écrit sous la forme

$$\begin{aligned} L(\mu_1, \mu_2, \cdots, \mu_I) &= \prod_{i=1}^{I}\prod_{j=1}^{n_i}\{\frac{1}{\sqrt{2\pi}}e^{-\frac{(Y_{ij}-\mu_i)^2}{2\sigma^2}}\} \\ &= (2\pi)^{-\frac{N}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2) \end{aligned}$$

L'estimateur du maximum de vraisemblance du vecteur $(\mu_1, \mu_2, \cdots, \mu_I)$ est $(\bar{Y}_{1.}, \bar{Y}_{2.}, \cdots, \bar{Y}_{I.})$. Sous H_0, cet estimateur devient $(\bar{Y}_{..}, \bar{Y}_{..}, \cdots, \bar{Y}_{..})$. Le rapport de vraisemblance s'écrit alors:

$$\begin{aligned} \Lambda &= \frac{L(\bar{Y}_{..}, \bar{Y}_{..}, \cdots, \bar{Y}_{..})}{L(\bar{Y}_{1.}, \bar{Y}_{2.}, \cdots, \bar{Y}_{I.})} \\ &= \frac{(2\pi)^{-\frac{N}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2)}{(2\pi)^{-\frac{N}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2)} \\ &= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2 - \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2\} \\ &= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I} n_i(\bar{Y}_{i.}-\bar{Y}_{..})^2\} \end{aligned}$$

Le passage de l'avant dernière ligne à la dernière ligne se fait du fait que

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

Cette égalité a été établie lors de la décomposition de SST en somme $SST = SSB + SSW$. Donc on rejette H_0 si le rapport $\Lambda = \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2\}$ est petit. C'est à dire si $\sum_{i=1}^{I} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2/\sigma^2$ est grand. On reconnaît l'expression de SSB/σ^2.

PROPOSITION **3.6** *Lorsque la variance est connue, on rejette H_0 au seuil α si et seulement si*

$$\frac{SSB}{\sigma^2} > \chi^2_{I-1,\alpha}$$

En effet, d'après la section précédente, on a vu que $SSB/\sigma^2 \sim \chi^2_{I-1}$ sous H_0.

3.4 Test d'égalité des moyenne avec variance inconnue

Dans cette section, on élabore le test d'égalité des moyennes $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ versus H_1 : il existe $i \neq j$ tels que $\mu_i \neq \mu_j$. La variance est supposée commune à tous les échantillons mais inconnue. Dans le paragraphe précédent, on a vu qu'on rejette H_0 lorsque SSB/σ^2 est grande si la variance est connue. Il est donc naturel de rejetter H_0 lorsque $SSB/\hat{\sigma}^2$ est grande si la variance est inconnue, où $\hat{\sigma}^2$ est un estimateur convenable de σ^2.
Cherchons alors le "meilleur" estimateur possible pour σ^2. On sait que pour $i = 1, 2, \cdots, I$, $(n_i - 1)S_i^2 \sim \chi^2_{n_i-1}$ et par conséquent $E[S_i^2] = \sigma^2$. Chacun de $S_1^2, S_2^2, \cdots, S_I^2$ est un estimateur sans biais de σ^2. Il est donc naturel de chercher le meilleur estimateur sans biais de σ^2 parmi les combinaisons linéaires de $S_1^2, S_2^2, \cdots, S_I^2$. Un tel estimateur s'écrit sous la forme $\tilde{\sigma}^2 = \sum_{i=1}^{I} a_i S_i^2$. On a $\sigma^2 = E[\tilde{\sigma}^2] = \sum_{i=1}^{I} a_i\sigma^2$, on en déduit que $\sum_{i=1}^{I} a_i = 1$ qu'on peut écrire encore

$$a_I = 1 - (a_1 + a_2 + \cdots + a_{I-1}) = 1 - \sum_{i=1}^{I-1} a_i.$$

D'autre part $\mathrm{Var}[\tilde{\sigma}^2] = \sum_{i=1}^{I} a_i^2 \mathrm{Var}[S_i^2]$. Or, pour $i = 1, 2, \cdots, I$, on a

$$\mathrm{Var}[(n_i - 1)\frac{S_i^2}{\sigma^2}] = \mathrm{Var}[\chi^2_{n_i-1}] = 2(n_i - 1).$$

On en déduit que $\mathrm{Var}[S_i^2] = 2\sigma^4/(n_i - 1)$ et que

$$\mathrm{Var}[\tilde{\sigma}^2] \;=\; \sum_{i=1}^{I} a_i^2 \frac{\sigma^4}{n_i - 1}$$

$$= \sigma^4\{\sum_{i=1}^{I-1}\frac{a_i^2}{n_i-1}+\frac{(1-(a_1+a_2+\cdots+a_{I-1}))^2}{n_I-1}\}$$

Pour minimiser $\mathrm{Var}[\tilde{\sigma}^2]$, on la dérive par rapport à $(a_1, a_2, \cdots, a_{I-1})$. Pour $i = 1, 2, \cdots, I$, on a

$$\frac{\partial}{\partial a_i}\mathrm{Var}[\tilde{\sigma}^2] = 2\sigma^4\{\frac{a_i}{n_i-1}-\frac{1-(a_1+a_2+\cdots+a_{I-1})}{n_I-1}\}$$

En mettant $\partial\mathrm{Var}[\tilde{\sigma}^2]/\partial a_i = 0$, on obtient:

$$\frac{a_1}{n_1-1} = \frac{a_2}{n_2-1} = \cdots = \frac{a_I}{n_I-1}.$$

Et comme on doit avoir $a_1 + a_2 + \cdots + a_I = 1$, on en déduit que $a_i = (n_i - 1)/(N - I)$. Le meilleur estimateur sans biais s'écrit alors

$$\tilde{\sigma}^2 = \sum_{i=1}^{n}\frac{n_i-1}{N-I}S_i^2 = \frac{SSW}{N-I} = MSW.$$

On rejette alors H_0 si SSB/SSW est grand.

PROPOSITION **3.7** *Lorsque la variance est inconnue, on rejette H_0 au seuil α si et seulement si*

$$F = \frac{SSB/(I-1)}{SSW/(N-I)} = \frac{MSB}{MSW} > F_{I-1,N-I,\alpha}$$

Ce dernier résultat est une conséquence directe de la proposition 3.5.
Les statistiques MSB et MSW définies respectivement par $MSB = SSB/(I-1)$ et $MSW = SSW/(N-I)$ sont appelées *mean squares between (samples)* et *mean squares within (samples)*. Cette dernière est aussi notée MSE où *mean square error* car $E[MSE] = \sigma^2$.
La méthode du rapport de vraisemblances avec variance inconnue donne le même test que la proposition 3.7.

3.5 Table d'ANOVA

La table d'ANOVA existe bien avant les ordinateurs. Elle permet de résumer tous les calculs nécéssaires pour effectuer le test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ versus H_1 : il existe $i \neq i'$

tels que $\mu_i \neq \mu_{i'}$ lorsque la variance est inconnue. Cette situation étant très fréquente en pratique. Aujourd'hui, la majorité de logiciels d'analyses statistiques fournissent directement les éléments de cette table. Dans le cas d'ANOVA à un facteur, cette table se présente comme suit:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Statistics	p-value
Between Treatments	SSB	$I-1$	$MSB = \frac{SSB}{I-1}$	$F_{obs} = \frac{MSB}{MSW}$	$P_{H_0}[F > F_{obs}]$
Within Treatments	SSW	$N-I$	$MSW = \frac{MSW}{N-I}$	***	***
Total	SST	$N-1$	***	***	***

EXEMPLE **3.1** *Un chercheur a mené une expérience dans le but de comparer trois types d'engrais pour tomates. Le tableau suivant résume les données:*

Type d'engrais	Taille de l'échantillon	Moyenne échantillonnale	Écart-type échantillonnal
A	21	6.50	1.25
B	21	4.75	1.05
C	21	5.10	1.40

On a alors $n_1 = n_2 = n_3 = 21$. *Donc* $N = 63$. *La moyenne globale est égale à:*

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I} n_i \bar{Y}_{i\cdot} = \frac{(21 \times 6.50) + (21 \times 4.75) + (21 \times 5.10)}{63} = 5.45.$$

Calculons ensuite SSB et SSW:

$$\begin{aligned} SSB &= \sum_{i=1}^{I} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= 21(6.50 - 5.45)^2 + 21(4.75 - 5.45)^2 + 21(5.10 - 5.45)^2 = 36.015 \\ SSW &= \sum_{i=1}^{I} (n_i - 1) S_i^2 \\ &= 20(1.25)^2 + 20(1.05)^2 + 20(1.40)^2 = 92.500 \end{aligned}$$

On obtient le tableau d'ANOVA suivant

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Statistics	p-value
Between Treatments	36.015	2	18.0075	11.6805	0.00005
Within Treatments	92.500	60	1.5417	***	***
Total	128.515	62	***	***	***

3.6 Inférence statistique pour le paramètre σ^2

On a vu que

$$(N-I)\frac{MSE}{\sigma^2} \sim \chi^2_{N-I}$$

où MSE est l'estimateur de σ^2 défini par:

$$MSE = \frac{1}{N-I}\sum_{i=1}^{I}(n_i-1)S_i^2$$

On en déduit alors que l'intervalle de confiance au niveau $1-\alpha$ pour σ^2 s'écrit:

$$[(N-I)\frac{MSE}{\chi_{N-I,\alpha/2}}, (N-I)\frac{MSE}{\chi^2_{N-I,1-\alpha/2}}]$$

Soit le test d'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ vs $H_0 : \sigma^2 \neq \sigma_0^2$.

On rejette H_0 si $(N-I)MSE/\sigma_0^2 \geq \chi^2_{N-I,\alpha/2}$ ou $(N-I)MSE/\sigma_0^2 \leq \chi^2_{N-I,1-\alpha/2}$

Soit le test d'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 > \sigma_0^2$.

On rejette H_0 si $(N-I)MSE/\sigma_0^2 \geq \chi^2_{N-I,\alpha}$

Soit le test d'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 < \sigma_0^2$.

On rejette H_0 si $(N-I)MSE/\sigma_0^2 \leq \chi^2_{N-I,1-\alpha}$

3.7 Inférence statistique pour les paramètres $\{\mu_1, \mu_2, \cdots, \mu_I\}$

On a vu que Pour $i = 1, 2, \cdots, I$, on a

$$\frac{\bar{Y}_{i.} - \mu_i}{\sqrt{\sigma^2/n_i}} \sim \mathrm{N}(0, 1).$$

On en déduit que

$$\frac{\bar{Y}_{i.} - \mu_i}{\sqrt{MSE/n_i}} \sim t_{N-I},$$

puisque que les statistiques MSE et $\bar{Y}_{i.}$ sont indépendantes. À partir de cette dernière relation, on déduit l'intervalle de confiance et les régions de rejet suivantes: L'intervalle de confiance au niveau de confiance α pour μ_i s'écrit:

$$[\bar{Y}_{i.} - t_{N-I,\alpha/2}\sqrt{\frac{MSE}{n_i}}, \bar{Y}_{i.} + t_{N-I,\alpha/2}\sqrt{\frac{MSE}{n_i}}]$$

Soit le test d'hypothèse $H_0 : \mu_i = \mu^*$ vs $H_1 = \mu_i \neq \mu^*$.

$$\text{On rejette } H_0 \text{ si } |\frac{\bar{Y}_{i.} - \mu^*}{\sqrt{MSE/n_i}}| > t_{N-I,\alpha/2}$$

Soit le test d'hypothèse $H_0 : \mu_i = \mu^*$ vs $H_1 = \mu_i > \mu^*$.

$$\text{On rejette } H_0 \text{ si } \frac{\bar{Y}_{i.} - \mu^*}{\sqrt{MSE/n_i}} > t_{N-I,\alpha}$$

Soit le test d'hypothèse $H_0 : \mu_i = \mu^*$ vs $H_1 = \mu_i < \mu^*$.

$$\text{On rejette } H_0 \text{ si } |\frac{\bar{Y}_{i.} - \mu^*}{\sqrt{MSE/n_i}}| < t_{N-I,1-\alpha}$$

3.8 Inférence statistique pour des combinaisons linéaires $\sum_{i=1}^{I} c_i\mu_i$:

En pratique, il arrive très souvent de s'intéresser à une combinaison linéaire des moyennes théoriques $\mu_1, \mu_2, \cdots, \mu_I$, soit $\sum_{i=1}^{I} c_i\mu_i$. Il est naturel d'estimer cette combinaison linéaire par $\sum_{i=1}^{I} c_i\bar{Y}_{i.}$ On a alors

$$E[\sum_{i=1}^{I} c_i\bar{Y}_{i.}] = \sum_{i=1}^{I} c_i\mu_i \,\text{et}\, Var[\sum_{i=1}^{I} c_i\bar{Y}_{i.}] = \sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{n_i}$$

Puisque $\{\bar{Y}_{1.}, \bar{Y}_{2.}, \cdots, \bar{Y}_{I.}\}$ sont des variables aléatoires indépendantes distribuées selon la loi normale, on en déduit que:

$$\frac{\sum_{i=1}^{I} c_i \bar{Y}_{i.} - \sum_{i=1}^{I} c_i \mu_i}{\sqrt{\sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{n_i}}} \sim \mathrm{N}(0,1).$$

et par la suite, que

$$\frac{\sum_{i=1}^{I} c_i \bar{Y}_{i.} - \sum_{i=1}^{I} c_i \mu_i}{\sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}} \sim t_{N-I}$$

par indépendance de MSE et $\{\bar{Y}_{1.}, \bar{Y}_{2.}, \cdots, \bar{Y}_{I.}\}$. Cette dernière relation nous permet de construire l'intervalle de confiance et les régions de rejets suivantes:
L'intervalle de confiance au niveau de confiance α pour $\sum_{i=1}^{I} c_i \mu_i$ s'écrit:

$$[\sum_{i=1}^{I} c_i \bar{Y}_{i.} - t_{N-I,\alpha/2} \sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}, \sum_{i=1}^{I} c_i \bar{Y}_{i.} + t_{N-I,\alpha/2} \sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}]$$

Soit le test d'hypothèse $H_0 : \sum_{i=1}^{I} c_i \mu_i = b$ vs $H_1 = \sum_{i=1}^{I} c_i \mu_i \neq b$.

$$\text{On rejette } H_0 \text{ si } |\frac{\sum_{i=1}^{I} c_i \bar{Y}_{i.} - b}{\sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}}| > t_{N-I,\alpha/2}$$

Soit le test d'hypothèse $H_0 : \sum_{i=1}^{I} c_i \mu_i = b$ vs $H_1 = \sum_{i=1}^{I} c_i \mu_i > b$.

$$\text{On rejette } H_0 \text{ si } \frac{\sum_{i=1}^{I} c_i \bar{Y}_{i.} - b}{\sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}} > t_{N-I,\alpha}$$

Soit le test d'hypothèse $H_0 : \sum_{i=1}^{I} c_i \mu_i = b$ vs $H_1 = \sum_{i=1}^{I} c_i \mu_i < b$.

$$\text{On rejette } H_0 \text{ si } \frac{\sum_{i=1}^{I} c_i \bar{Y}_{i.} - b}{\sqrt{MSE \sum_{i=1}^{I} \frac{c_i^2}{n_i}}} < t_{N-I,1-\alpha}$$

3.9 Aspect informatique

Les données brutes pour ajuster un modèle d'analyse de variance à un facteur se présente sour la forme d'une matrice de dimension $N \times 2$. Une ligne correspond à une unité expérimentale et une colonne à une variable. Une colonne identifie l'échantillon auquel l'unité appartient,

c'est-à-dire la modalité du facteur à l'étude. L'autre colonne donne la valeur de la variable à l'étude, y.
Par exemple un fichier de données sur l'usure (wear) de trois sortes de revêtement pour le bois (brand) a la forme suivante.

```
Wear    Brand
0.87    Ajax
1.62    Ajax
...     ...
2.18    Champion
2.2     Champion
2.22    Champion
2.67    Champion
```

Plusieurs logiciels sont disponibles pour calculer la table ANOVA d'une analyse de variance à un facteur. Sur SAS on retrouve ANOVA, GLM et MIXED et sur R la fonction aov peut faire ce travail. Par exemple la table ANOVA produite avec aov pour l'exemple de l'usure de trois revêtement (on a $n_i = 10$ pour les 3 échantillons) est

```
           Df  Sum Sq Mean Sq F value  Pr(>F)
Brand       2 0.84201 0.42100  5.0562 0.01364 * \\
Residuals  27 2.24813 0.08326
```