Analyse de la Variance à 1 Facteur

Antoine Godichon-Baggioni

INSA de Rouen – Génie Mathématique - 4ème année

1 Introduction

L'analyse de variance (ANOVA) recouvre un ensemble de techniques de tests et d'estimation destinés à apprécier l'effet d'une ou plusieurs variables qualitatives sur une variable quantitative et revient dans le cas simple à comparer plusieurs moyennes d'échantillons gaussiens : on généralise le test classique d'égalité de deux moyennes au test d'égalité de p moyennes ($p \geq 2$).

Comme dans le test d'égalité de deux moyennes, on posera les hypothèses de normalité et d'indépendance du caractère quantitatif et d'égalité des variances.

La particularité de l'ANOVA, c'est que les p moyennes vont provenir de p échantillons correspondant chacun à une modalité du caractère qualitatif qui sert à stratifier la population.

On utilise dans l'ANOVA un vocabulaire particulier introduit par les agronômes qui ont été les premiers à s'intéresser à ce type de problème : la variable qualitative susceptible d'influer sur la distribution de la variable quantitative étudiée est appelée **facteur** et ses modalités **niveaux**.

Essayons d'illustrer sur un exemple la problématique de l'ANOVA à un facteur.

1.1 Le modèle d'Anova 1 à travers l'étude d'un exemple

Un forestier s'intéresse aux hauteurs moyennes de 3 forêts (extrait du livre de Azais & Bardet, 2006). Pour les estimer, il échantillonne un certain nombre d'arbres et mesure leurs hauteurs. Voici les données recueillies :

Forêt	1	2	3
	23.4	22.5	18.9
	24.4	22.9	21.1
	24.6	23.7	21.1
	24.9	24.0	22.1
	25.0	24.0	22.5
	26.2		23.5
			24.5
Nombre d'arbres	$n_1 = 6$	$n_2 = 5$	$n_3 = 7$
Moyenne	24.75	23.42	21.96

A partir de ces données, le forestier souhaite savoir si la hauteur moyenne des arbres est la même dans les 3 forêts, ou pas.

Ces données peuvent être présentées de deux manières :

1. On dispose de 3 échantillons indépendants et on désire comparer leurs moyennes : c'est l'approche "comparaison de moyennes".
2. On dispose d'un seul échantillon de longueur 18 et d'un facteur (le numéro de la forêt), et on étudie l'effet de ce facteur sur la moyenne : c'est l'approche "analyse de la variance".

Essayons de modéliser ces données. En notant Y_{ij} la hauteur du $j^{\text{ème}}$ arbre de la forêt i et μ_i la hauteur moyenne inconnue de la forêt i, on peut envisager le modèle suivant :

$$Y_{ij} \;=\; \mu_i \;+\; \varepsilon_{ij} \quad \text{avec} \quad \begin{cases} i = 1, \cdots, 3 \\ j = 1, \cdots, n_i \end{cases}$$

où ε_{ij} représente la variabilité de l'arbre j par rapport à la hauteur moyenne de la forêt i. On fait sur les variables ε_{ij} les hypothèses suivantes :

— $\mathbb{E}[\varepsilon_{ij}] = 0$, pour tout (i, j) ;
— $\mathbb{V}\text{ar}[\varepsilon_{ij}] = \sigma^2$ pour tout (i, j), *ie.* la variance de la hauteur des arbres est la même dans les 3 forêts ;
— les (ε_{ij}) sont indépendantes, ce qui est assuré par la manière dont a été fait l'échantillonnage ;
— les (Y_{ij}) (et donc les ε_{ij}) sont des variables gaussiennes.

La question auquelle on souhaite maintenant répondre est :

$$\ll\text{Les forêts sont-elles équivalentes (du point de vue la hauteur moyenne des arbres) ?}\gg$$

Ceci se traduit dans le cadre de notre modèle, par le test de l'hypothèse nulle

$$H_0 \;:\; \ll \mu_1 = \mu_2 = \mu_3 \gg$$

contre l'hypothèse alternative

$$H_1 \;:\; \ll \mu_1 \neq \mu_2 \text{ ou } \mu_2 \neq \mu_3 \text{ ou } \mu_1 \neq \mu_3 \gg$$

Un des objectif de ce cours sera de voir comment on met en œuvre un tel test.

1.2 D'autres exemples

Voici deux autres exemples, issus de domaines variés, où les techniques d'ANOVA à un facteur pourront être utilisées.

Exemple 1. Les candidats à un oral ont été répartis au hasard entre trois examinateurs. Du fait des absents, le premier examinateur a fait passer l'oral à 6 étudiants, le second à 8 étudiants et le troisième à 7 étudiants. Voici les notes qu'ils ont attribués :

Examinateur	1	2	3
	10 11	8 11	10 13
	11 12	11 13	14 14
	13 15	14 15	15 16
		16 16	16
Nombres de notes	6	8	7
Moyenne	12	13	14

Tableau 1. *Notes obtenues à l'oral.*

Un candidat remarque que la moyenne des notes du premier examinateur est de 12, celle du deuxième de 13 et celle du troisième de 14. Il y a 2 points d'écarts entre la meilleure moyenne qui est de 14, et la moins bonne qui est de 12.

Avant d'entamer une procédure de recours, il se demande si une telle variation des moyennes observées peut être due au hasard seul ou si elle est révalatrice d'un réel "effet examinateur" qui influerait sur la moyenne des notes attribuées. Après tout, les tailles d'échantillons sont bien faibles !

Exemple 2. On souhaite comparer trois traitements contre l'asthme, notés A, B et C : le traitement B est un nouveau traitement que l'on souhaite mettre en compétition avec les traitements classiques A et C. On répartit par tirage au sort, les patients venant consulter dans un centre de soin, et on leur affecte l'un des trois traitements.

On mesure sur chaque patient la durée, en jours, séparant de la prochaine crise d'asthme. Voici les mesures obtenues :

Traitement	A	B	C
	26 27	29 42 44	26 26 30
	35 36	44 45 48	30 33 36
	38 38	48 52 56	38 38 39
	41 42	56 58 58	46 47 51
	45 50	60 61 63	51 56 75
	65	63 69	
Nombres de mesures	11	17	15
Moyenne	40.27	52.70	41.47

Tableau 2. *Durée séparant de la prochaine crise d'asthme.*

On se pose alors la question suivante : peut-on conclure que les traitements ont une efficacité différente pour le critère "*temps moyen séparant de la prochaine crise*" ?

Pour répondre aux questions posées dans ces deux exemples, nous mettrons en œuvre un test statistique dont l'objet sera de tester l'égalité des moyennes de chaque échantillon ou bien encore de voir si le facteur, dont on étudie l'effet, a un effet sigificatif réel. Pour cela, nous avons besoin, bien entendu, d'un modèle probabiliste et d'une statistique de test adaptée qui prendra en compte les écarts entre les moyennes observées.

2 Les données et le modèle

2.1 Les données

On cherche à étudier l'effet d'un facteur A, que l'on supposera à p niveaux, sur une variable quantitative Y. On suppose que le facteur A influe uniquement sur les moyennes des distributions de chacun des p groupes et non sur leur variance.

Pour chaque niveau i du facteur A (avec $1 \leq i \leq p$), on dispose de n_i mesures de Y, notées y_{ij} avec $j = 1, \cdots, n_i$. Dans la suite, on notera par n le nombre total d'observations, ie. $n = \sum_{i=1}^{p} n_i$.

On présente généralement les données à l'aide du tableau suivant :

Niveau du Facteur A	A_1	A_2	$\dots$	A_i	$\dots$	A_p
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{p1}
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{p2}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	y_{1n_1}	y_{2n_2}	$\cdots$	y_{in_i}	$\cdots$	y_{pn_p}
Effectifs	n_1	n_2	$\cdots$	n_i	$\cdots$	n_p
Moyennes empiriques	$\overline{y}_{1\bullet}$	$\overline{y}_{2\bullet}$	$\cdots$	$\overline{y}_{i\bullet}$	$\cdots$	$\overline{y}_{p\bullet}$

Tableau 4 *Les données en ANOVA à 1 facteur.*

2.2 Le modèle de l'ANOVA 1

On fait les hypothèses de normalité et d'indépendance suivantes :

1. Pour tout $(i,j) \in \{1,\cdots,p\}\times\{1,\cdots,n_i\}$, la donnée y_{ij} est une réalisation d'une variable aléatoire Y_{ij} de loi $\mathcal{N}(\mu_i\,,\,\sigma^2)$.
2. Les variables aléatoires (Y_{ij}) sont globalement indépendantes.

On peut résumer ces hypothèses en écrivant le modèle :

$$\boxed{Y_{ij} \;=\; \mu_i \;+\; \varepsilon_{ij} \quad \text{avec} \quad (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0,\sigma^2).} \tag{1}$$

On décrit l'effet du facteur A en supposant :

- une espérance spécifique μ_i pour chaque groupe ou chaque niveau du facteur
- et une variance intra-groupe σ^2 commune à tous les groupes.

L'objet de cette étude sera de savoir si, au vu des données du Tableau 4, les moyennes des p échantillons sont égales ou différentes. Autrement dit, on souhaite savoir si les moyennes empiriques observées $(\overline{y}_{i\bullet})$ diffèrent à cause de différences réelles entre les moyennes (μ_i), ou bien si les différences entre les $(\overline{y}_{i\bullet})$ peuvent raisonnablement être attribuées aux seules fluctuations d'échantillonnage.

Remarque. On décompose parfois μ_i en

$$\boxed{\mu_i \;=\; \mu \;+\; \alpha_i \quad \text{avec} \quad \sum_{i=1}^{p} n_i\alpha_i = 0} \tag{2}$$

où

- μ représente un effet global inconnu du facteur ;
- α_i représente l'effet principal (spécifique) inconnu du niveau i du facteur A.

Le modèle s'écrit alors

$$\boxed{Y_{ij} \;=\; \underbrace{\mu \;+\; \alpha_i}_{\mu_i} \;+ \varepsilon_{ij} \quad \text{avec} \quad (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0,\sigma^2).} \tag{3}$$

Lorsque l'égalité des moyennes ne sera pas retenue, le problème se posera d'estimer, selon le contexte, soit les (μ_i), soit μ et les (α_i).

2.3 La dimension du modèle de l'ANOVA 1

Définition 2.1 On appellera **dimension du modèle** dans le contexte de l'ANOVA, la dimension de l'espace dans lequel vit l'espérance des variables aléatoires (Y_{ij}). Cette dimension est égale

— au nombre de paramètres d'espérance envisagés dans la modélisation

moins

— le nombre de contraintes d'identifiabilité nécessaires (indépendantes) à l'estimation des dits paramètres.

Remarque. Le modèle de l'ANOVA 1 est de dimension p. On le notera donc (M_p). On a en effet
- soit p paramètres (les (μ_i)) et aucune contrainte ;
- soit $(p+1)$ paramètres (μ et les (α_i)) et une contrainte : $\sum_{i=1}^{p} n_i\alpha_i = 0$.

3 Test de l'effet du facteur

3.1 Introduction - Comparaison de modèles

On veut savoir si le facteur A influe réellement sur la variable d'intérêt Y. On fait sur les données du Tableau 4, les hypothèses de normalité et d'indépendance des p échantillons, c'est à dire qu'on suppose que pour tout couple (i,j), la donnée y_{ij} est une réalisation d'une variable aléatoire Y_{ij} de loi $\mathcal{N}(\mu_i\,,\,\sigma)$, les variables aléatoires Y_{ij} étant de plus globalement indépendantes.

Pour tester l'absence d'effet du facteur, on va tester l'hypothèse nulle

$$H_0 : \ll \mu_1 = \ldots = \mu_p \gg$$

contre l'alternative

$$H_1 : \ll \exists\,(i,j) \text{ tq } \mu_i \neq \mu_j \gg$$

3.2 L'approche comparaison de modèles

L'égalité $\ll \mu_1 = \ldots = \mu_p \gg$ permet de définir un sous-modèle du modèle complet de l'ANOVA 1. En notant μ cette moyenne commune, ce sous-modèle s'écrit :

$$Y_{ij} \;=\; \mu + \varepsilon_{ij} \quad \text{avec} \quad (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$$

Ce sous-modèle étant de dimension 1 (un seul paramètre et aucune contrainte), on le notera (M_1). Autrement dit, tester l'absence d'effet du facteur A sur Y, c'est tester

$$H_0 : \ll \text{ Modèle } (M_1) : Y_{ij} = \mu + \varepsilon_{ij} \;\text{ avec }\; (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0,\sigma^2) \gg$$

contre

$$H_1 : \ll \text{ Modèle } (M_p) : Y_{ij} \;=\; \mu_i + \varepsilon_{ij} \;\text{ avec }\; (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0,\sigma^2) \gg$$

3.3 Estimation des paramètres des deux modèles

3.3.1 Dans le modèle complet (M_p).

Dans ce modèle, il nous faut estimer les (μ_i) et σ^2 :

1. On estime μ_i (pour tout $i = 1, \cdots, p$) par $\widehat{\mu}_i = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} Y_{ij} \;=\; \overline{Y}_{i\bullet}$
2. On prédit, pour tout (i,j), Y_{ij} par $\widehat{Y}_{ij} = \widehat{\mu}_i = \overline{Y}_{i\bullet}$
3. Les résidus (estimations des ε_{ij}) sont définis par les $\widehat{\varepsilon}_{ij} = Y_{ij} - \overline{Y}_{i\bullet}$
4. La somme des carrés résiduelle vaut $SCR(M_p) = \displaystyle\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$
5. Enfin, on estime σ^2 par $S^2 \;=\; \dfrac{SCR(M_p)}{n-p}$

3.3.2 Dans le sous-modèle (M_1).

Dans ce modèle, il nous faut estimer μ et σ^2 :

1. On estime μ par $\widehat{\mu} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij} \;=\; \overline{Y}_{\bullet\bullet}$
2. On prédit, pour tout (i,j), Y_{ij} par $\widehat{Y}_{ij} = \widehat{\mu} = \overline{Y}_{\bullet\bullet}$;
3. Les résidus sont les $\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{Y}_{ij} = Y_{ij} - \overline{Y}_{\bullet\bullet}$
4. La somme des carrés résiduelle vaut $SCR(M_1) = \displaystyle\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2$
5. On estime σ^2 par $S^2 \;=\; \dfrac{SCR(M_1)}{n-1}$

Cet estimateur S^2 ne sera un bon estimateur de la variance σ^2 du modèle (M_p) que lorqu'on aura égalité des (μ_i).

3.4 Construction du test

On veut tester, au risque δ, l'hypothèse nulle H_0 : $\ll \mu_1 = \ldots = \mu_p \gg$ contre l'hypothèse alternative H_1 : $\ll \exists\,(i,j)$ tq $\mu_i \neq \mu_j \gg$, ce qui revient à comparer le sous-modèle

$$(M_1) : Y_{ij} = \mu + \varepsilon_{ij} \;\text{ avec }\; (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

au modèle complet

$$(M_p) : Y_{ij} = \mu_i + \varepsilon_{ij} \;\text{ avec }\; (\varepsilon_{ij}) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

La construction du test va reposer sur le théorème suivant :

Théorème 3.1 *Dans le cadre du modèle complet d'ANOVA 1, on a*

$$SCR(M_1) \;=\; \sum_{i=1}^{p}\sum_{j=1}^{n_i}\underbrace{\big(\widehat{Y}_{ij}(M_p) - \widehat{Y}_{ij}(M_1)\big)}_{\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}}^2 + \; SCR(M_p)$$

et $SCR(M_p) \;\sim\; \sigma^2 \bullet \chi^2_{n-p}$. *De plus sous* H_0, $\quad SCR(M_1) \;\sim\; \sigma^2 \bullet \chi^2_{n-1}$ *et*

$$Z \;=\; \frac{\big(SCR(M_1) - SCR(M_p)\big)/(p-1)}{SCR(M_p)/\big(n-p\big)}$$

suit une loi de Fisher $F\big(p-1\,;\,n-p\big)$.

Remarque 3.2 Le résultat $Z \underset{H_0}{\sim} F\big(p-1\,;\,n-p\big)$ est bien évidemment faux sous H_1, ce qui fait de Z une statistique de test pour tester H_0 contre H_1, ie. (M_1) contre (M_p).
Le résultat $SCR(M_p) \sim \sigma^2 \bullet \chi^2_{n-p}$ entraîne que $S^2 = \dfrac{SCR(M_p)}{n-p}$ est un estimateur sans biais de σ^2.

Construction du test, au risque δ, de $H_0 : \ll \mu_1 = \ldots = \mu_p \gg$ contre $H_1 : \ll \exists\,(i,j)$ tq $\mu_i \neq \mu_j \gg$

Hypothèses. Celles du modèle complet d'ANOVA 1.
Statistique de test utilisée et loi sous H_0. On utilise la statistique

$$Z \;=\; \frac{\big(SCR(M_1) - SCR(M_p)\big)/(p-1)}{SCR(M_p)/\big(n-p\big)}$$

qui, sous H_0, suit une loi de Fisher $F(p-1\,;\,n-p)$.

Construction de la zone de rejet. On fixe un risque δ et on calcule $f_{p-1,n-p,\delta}$ tel que

$$\mathbb{P}\Big[F(p-1\,;\,n-p) \;\leq\; f_{p-1,n-p,\delta}\Big] \;=\; 1-\delta$$

La zone de rejet de H_0 au risque δ est de la forme : $\big\{Z > f_{p-1,n-p,\delta}\big\}$.

Stratégie de décision. On calcule la valeur z de Z sur les données $(y_{ij})_{1\leq i\leq p, 1\leq j\leq n_i}$. On a

$$z \;=\; \frac{\big(scr(M_1) - scr(M_p)\big)/(p-1)}{scr(M_p)/\big(n-p\big)}$$

et on adopte la stratégie suivante :

— si $z \;\leq\; f_{p-1,n-p,\delta}$ alors on accepte H_0 au risque δ et on considère qu'il n'y a pas d'effet significatif du facteur ;
— si $z \;>\; f_{p-1,n-p,\delta}$ alors on rejette H_0 au risque δ et on considère que l'effet du facteur sur Y est significatif.

3.5 Interprétations du test d'ANOVA.

Remarque 3.3 La statistique de test Z peut se voir comme le rapport de deux estimateurs de σ^2 : un toujours bon et un qui ne l'est que sous H_0. En effet, notons $SCM = SCR(M_1) - SCR(M_p)$. On peut montrer que (cf. feuille TD 6) :

$$\mathbb{E}\Big[\frac{SCM}{p-1}\Big] \;\;=\;\; \sigma^2 \;+\; \frac{1}{p-1}\sum_{i=1}^{p} n_i(\mu_i - \mu)^2 \;\;\text{où}\;\; \mu = \frac{1}{n}\sum_{i=1}^{p} n_i\mu_i$$

La quantité $\sum_{i=1}^{p} n_i(\mu_i-\mu)^2$ est nulle ssi $\forall\, i = 1, \cdots, p,\;\; \mu_i = \mu$, c'est à dire lorsqu'on a égalité des p moyennes, ie. lorsque H_0 est vraie. On peut donc déduire que sous H_0, $SCM\,/\,(p-1)$ est un estimateur sans biais de σ^2.

Ainsi tester l'absence d'effet du facteur A, c'est comparer deux estimateurs de σ^2 :

- un qui n'est bon que sous H_0, celui donné par $SCM/(p-1)$
- un qui est toujours bon, celui obtenu dans le modèle (M_p) et donné par $SCR(M_p)/(n-p)$.

Lorsque H_0 est vraie, la valeur de Z doit être comparable à 1, lorsque H_0 est fausse, la variable Z doit prendre de grandes valeurs, ce qui explique la forme de la zone de rejet.

Remarque 3.4 On peut aussi voir la somme des carrés $SCM = SCR(M_1) - SCR(M_p)$ comme une mesure de la réduction d'erreur, quand on passe du sous-modèle (M_1) au modèle (M_p), ie. quand on ajoute dans le modèle, les effets spécifiques du facteur A à la constante (pas d'effet du facteur).

Remarque 3.5 La statistique Z peut s'interpréter comme le rapport de la variabilité **inter-groupe** sur la variabilité **intra-groupe**. En effet, la quantité

$$SCM \;=\; \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \;=\; \sum_{i=1}^{p} n_i(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$$

mesure l'écart des moyennes des groupes à la moyenne générale : c'est une mesure de variabilité entre les groupes. La quantité $SCR(M_p)$, quant à elle, mesure l'écart de chaque individu à la moyenne du groupe auquel il appartient : c'est une mesure de variabilité à l'intérieur de chaque groupe. Sous H_0, la variabilité inter-groupe est comparable à la variabilité intra-groupe, sous H_1, elle est d'autant plus grande que les (μ_i) sont différentes.

3.6 Table de l'Anova 1.

On présente généralement les résultats sous la forme d'une table d'analyse de la variance :

Source de variation	Somme de carrés	Degrés de liberté (ddl)	Statistique de Test
Facteur A (inter-groupe)	$SCM = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$	$p-1$	$Z = \dfrac{\dfrac{SCM}{(p-1)}}{\dfrac{SCR(M_p)}{(n-p)}}$
Résiduelle (intra-groupe)	$SCR(M_p) = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$	$n-p$	
Totale	$SCR(M_1) = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2$	$n-1$	

Exercice E.1 *Reprendre les différents exemples et dire au risque* 5% *si il y a un effet significatif dû au facteur.*

4 Comparaisons multiples de moyennes

Le rejet de l'hypothèse nulle H_0 : $\ll \mu_1 = \ldots = \mu_p \gg$ ne signifie pas que toutes les moyennes (μ_i) sont différentes. Lorsque l'effet du facteur est significatif, on désire parfois pousser plus loin l'analyse en classant les différentes moyennes ou bien en les comparant à un témoin.
On peut aussi chercher à comparer deux moyennes choisies a priori, par exemple μ_1 et μ_2 : on testera alors $H_0 : \ll \mu_1 = \mu_2 \gg$ contre $H_1 : \ll \mu_1 \neq \mu_2 \gg$, ce qui est équivalent à tester $H_0 : \ll \mu_1 - \mu_2 = 0 \gg$ contre $H_1 : \ll \mu_1 - \mu_2 \neq 0 \gg$. Plus généralement, on peut chercher à tester l'égalité à 0 de contrastes entre les paramètres (μ_i).

Définition 4.1 *Un contraste entre les paramètres* $(\mu_i)_{i=1,\cdots,p}$ *est une combinaison linéaire des* (μ_i) *de la forme* $\sum_{i=1}^{p} c_i\,\mu_i$ *où les* (c_i) *sont des coefficients réels constants vérifiant la condition* $\sum_{i=1}^{p} c_i = 0$.

Exemple. Les quantités $(\mu_1 - \mu_2)$, $(\mu_1 - \mu_3)$ et $\mu_2 - \frac{1}{2}(\mu_1 + \mu_3)$ sont des contrastes.

Pour un contraste donné ψ, nous allons tester l'hypothèse nulle

$$H_0 \ : \ \ll \psi \ = \ \sum_{i=1}^{p} c_i \, \mu_i \ = \ 0 \gg$$

contre l'hypothèse alternative

$$H_1 \ : \ \ll \psi \ = \ \sum_{i=1}^{p} c_i \, \mu_i \ \neq \ 0 \gg$$

Soit $\widehat{\psi} = \sum_{i=1}^{p} c_i \, \overline{Y}_{i\bullet}$ l'estimateur sans biais du contraste $\sum_{i=1}^{p} c_i \, \mu_i$. On a alors le théorème suivant.

Théorème 4.2 *Dans la cadre du modèle complet d'ANOVA 1 et sous H_0,*

$$\boxed{Z \ = \ \frac{\displaystyle\sum_{i=1}^{p} c_i \, \overline{Y}_{i\bullet}}{\sqrt{\dfrac{SCR(M_p)}{n-p} \left(\displaystyle\sum_{i=1}^{p} \dfrac{c_i^2}{n_i} \right)}} \ \sim \ T_{n-p}}$$

Remarque. Sous H_1, la variable aléatoire Z ne suit plus une loi de Student T_{n-p}, ce qui fait de Z une statistique de test. Notons de plus que ce résultat est vrai quelque soit le contraste considéré.

Exercice E.2 *Démontrer que* $\displaystyle\sum_{i=1}^{p} c_i \, \overline{Y}_{i\bullet} \ \sim \ \mathcal{N}\left(\sum_{i=1}^{p} c_i \, \mu_i \, , \, \sigma \sqrt{\sum_{i=1}^{p} \frac{c_i^2}{n_i}} \right)$.

Preuve du théorème. Du résultat de l'exercice précédent et en utilisant le fait que $SCR(M_p)/\sigma^2$ suit une loi du khi-deux à $(n-p)$ degrés de liberté et que $SCR(M_p)$ et $\sum_{i=1}^{p} c_i \, \overline{Y}_{i\bullet}$ sont des variables aléatoires indépendantes (voir la remarque ci-dessous), on déduit "en studentisant" que

$$\frac{\displaystyle\sum_{i=1}^{p} c_i \, \overline{Y}_{i\bullet} - \sum_{i=1}^{p} c_i \, \mu_i}{\sqrt{\dfrac{SCR(M_p)}{n-p} \left(\displaystyle\sum_{i=1}^{p} \dfrac{c_i^2}{n_i} \right)}} \ \sim \ T_{n-p}$$

On conclut à la preuve du théorème en remarquant que $\sum_{i=1}^{p} c_i \, \mu_i = 0$ sous l'hypothèse H_0.

Remarque 4.3 Puisque les variables aléatoires (Y_{ij}) sont globalement indépendantes, on en déduit que les paquets de variables aléatoires $(Y_{i1}, \ldots, Y_{in_i})_{1 \leq i \leq p}$ sont aussi indépendants. Par suite, les variables aléatoires $(\overline{Y}_{i\bullet})$ sont indépendantes, puisque construites à partir de paquets disjoints de variables aléatoires globalement indépendantes. Il en va de même pour les variables aléatoires $\left(\sum_{i=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2 \right)$. Enfin, puisque pour tout i, les variables $\overline{Y}_{i\bullet}$ et $\sum_{i=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2$ sont indépendantes, on en déduit que les variables $SCR(M_p)$ et $\sum_{i=1}^{p} c_i \, \overline{Y}_{i\bullet}$ le sont aussi.

4.1 Test au risque δ de $H_0 : \ll\psi = 0\gg$ contre $H_1 : \ll\psi \neq 0\gg$

Hypothèses. Cadre du modèle complet d'ANOVA 1.

Statistique de test utilisée et loi sous H_0**.** On utilise la statistique de test

$$Z \;=\; \frac{\displaystyle\sum_{i=1}^{p} c_i\,\overline{Y}_{i\bullet}}{\sqrt{\dfrac{SCR(M_p)}{n-p}\left(\displaystyle\sum_{i=1}^{p}\dfrac{c_i^2}{n_i}\right)}}$$

qui suit sous H_0 une loi de Student T_{n-p}.

Construction de la zone de rejet. On fixe un risque δ et on calcule $t_{n-p,\delta/2}$ tel que

$$\mathbb{P}\Big[|T_{n-p}| \;<\; t_{n-p,\delta/2}\Big] \;=\; 1-\delta$$

La zone de rejet de H_0 au risque δ est alors de la forme : $\big\{|Z| > t_{n-p,\delta/2}\big\}$.

Stratégie de décision. On calcule la valeur z de Z sur les données $(y_{ij})_{1\leq i\leq p, 1\leq j\leq n_i}$. On a

$$z \;=\; \sum_{i=1}^{p} c_i\,\overline{y}_{i\bullet} \Big/ \sqrt{\frac{scr(M_p)}{n-p}\sum_{i=1}^{p}\frac{c_i^2}{n_i}}$$

et on adopte la stratégie suivante : si $|z| \leq t_{n-p,\delta/2}$ alors on accepte H_0 au risque δ et on considère que le contraste ψ est nul, sinon (ie. $|z| > t_{n-p,\delta/2}$) alors on rejette H_0 au risque δ et on considère que le contraste ψ est différent de 0.

5 Robustesse aux hypothèses

La méthodologie de l'Anova 1 est plus ou moins robuste au non respect des hypothèses de modélisation, à savoir la normalité, l'homoscédasticité et l'indépendance des erreurs.
On peut dire que :

1. la méthodologie est robuste à la non normalité des échantillons.
2. la non homogénéité des variances peut être contournée.
3. le plus grave est bien sûr le non respect de l'indépendance des erreurs, il faut alors avoir recours à d'autres modèles que celui de l'Anova 1.

Une analyse des résidus devrait être effectuée avant toute utilisation du modèle pour essayer de le valider, en vérifiant de manière descriptive ou par des tests adéquats les hypothèses du modèle.
Par exemple, on peut tester l'homogénéité des variances, c'est à dire tester $H_0 : \ll\sigma_1^2 = \ldots = \sigma_p^2\gg$ contre $H_1 : \ll\exists(i,j), \sigma_i^2 \neq \sigma_j^2\gg$ où σ_i^2 désigne la variance de l'échantillon i, à l'aide du **test de Bartlett** mais qui est sensible à la non normalité ou bien à l'aide du **test de Cochran**, qui est robuste à la non normalité mais qui ne s'applique que lorsque les p échantillons ont la même taille

Lorsque l'égalité des variances n'est pas vérifiée, on peut utiliser le test non paramétrique de Kruskal-Wallis pour savoir si les distributions des p échantillons sont identiques ou pas.

6 Test de comparaison de p variances

L'homogénéité des variances entre groupes est cruciale en analyse de la variance, mais n'est que rarement testée. Elle peut cependant être testée de différentes manières. La solution la plus simple serait d'effectuer les $p(p-1)/2$ comparaisons 2 à 2 des variances de tous les groupes grâce au test classique d'égalité des variances de deux échantillons gaussiens en testant pour tout couple (i,j) l'hypothèse H_0 : $\ll \sigma_i^2 = \sigma_j^2 \gg$ contre l'alternative H_1 : $\ll \sigma_i^2 \neq \sigma_j^2 \gg$. Mais on est alors confronté au problème des tests multiples : si l'on choisit d'effectuer chaque test à un niveau de 5%, on ne peut rien garantir sur le niveau global après avoir effectué les $p(p-1)/2$ tests. Il existe d'autres procédures de tests (plus ou moins robustes aux hypothèses de modélisation sous-jacentes) permettant de tester globalement l'égalité des variances telles que le test de Bartlett (sensible à la non normalité) ou le test de Levene ou bien encore le test de Cochran. On présente dans ce qui suit le test de Bartlett.

On fait les hypothèses de normalité et d'indépendance des p échantillons, c'est à dire qu'on suppose que les données (y_{ij}) sont les réalisations de variables aléatoires (Y_{ij}) de loi $\mathcal{N}(\mu_i, \sigma_i)$, les variables (Y_{ij}) étant globalement indépendantes. On pose l'hypothèse nulle

$$H_0 \ : \ \ll \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_p^2 \gg$$

que l'on veut tester contre l'hypothèse alternative

$$H_1 \ : \ \ll \exists \ i \, , \, j \in \{1, \ldots, p\} \, , \ \sigma_i^2 \neq \sigma_j^2 \gg$$

Notons (S_i^2) les estimateurs sans biais des variances (σ_i^2) des p échantillons. On rappelle que

$$\forall \ i \in \{1, \ldots, p\} \, , \quad S_i^2 \ = \ \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2$$

et par conséquent, on peut réécrire $SCR(M_p)$ sous la forme $SCR(M_p) \ = \ \sum_{i=1}^{p} (n_i - 1) \, S_i^2$.

Sous l'hypothèse H_0, on montre que la statistique

$$\frac{2.3026}{C} \ \left[(n-p) \ \ln \left(\frac{SCR(M_p)}{n-p} \right) \ - \ \sum_{i=1}^{p} (n_i - 1) \ln \left(S_i^2 \right) \right] \text{ avec } C = 1 + \frac{1}{3(p-1)} \left[\sum_{i=1}^{p} \frac{1}{n_i - 1} \ - \frac{1}{p-1} \right]$$

suit approximativement un χ^2 à $(p-1)$ degrés de liberté. Bien entendu, sous H_1, cette variable aléatoire ne suit plus un $\chi^2(p-1)$. Ce résultat est donc suffisant pour construire un test de H_0 contre H_1.

Exercice E.3 *Reprendre les exemples 1, 2 et 3 et tester au risque* 5% *l'égalité des variances.*

7 Estimation des effets

On veut construire des estimateurs sans biais des effets, c'est à dire construire des estimateurs des paramètres inconnus $\mu_1, \mu_2, \ldots, \mu_p$, ou bien μ et $\alpha_1, \alpha_2, \ldots, \alpha_p$ selon le contexte.

7.1 Estimation des paramètres $\mu_1, \mu_2, \ldots, \mu_p$

On utilise les résultats d'estimation bien connus dans le cadre d'un échantillon gaussien.

Théorème 7.1 *Sous les hypothèses de normalité et d'indépendance des p échantillons, pour tout* $i \in \{1, \dots, p\}$, $\overline{Y}_{i\bullet}$ *est un estimateur sans biais du paramètre* μ_i *et*

$$\overline{Y}_{i\bullet} \ \sim \ \mathcal{N}\left(\mu_i \, , \, \frac{\sigma_i}{\sqrt{n_i}}\right)$$

De plus, $S_i^2 \ = \ \dfrac{1}{n_i - 1} \displaystyle\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$ *est un estimateur sans biais de* σ_i^2*, indépendant de* $\overline{Y}_{i\bullet}$*, et on a*

$$\frac{(n_i - 1)}{\sigma_i^2} \, S_i^2 \ \sim \ \chi^2_{n_i - 1}$$

Conséquence. Il est possible de bâtir des intervalles de confiance pour les paramètres (μ_i). En effet, on déduit du théorème précédent que

$$\frac{\sqrt{n_i} \left(\overline{Y}_{i\bullet} - \mu_i\right)}{S_i} \ \sim \ T_{n_i - 1}$$

et par suite l'intervalle de confiance au niveau de confiance $(1 - \delta)$ de μ_i est :

$$\mathrm{IC}_{(1-\delta)}(\mu_i) = \left[\overline{y}_{i\bullet} \ \pm \ \frac{s_i \, t_{n_i - 1, \delta/2}}{\sqrt{n_i}}\right],$$

où s_i est la réalisation de S_i sur les données et $t_{n_i - 1, \delta/2}$ est tel que $\mathbb{P}\Big[|T_{n_i - 1}| \ < \ t_{n_i - 1, \delta/2}\Big] \ = \ 1 - \delta$.

7.2 Estimation des paramètres μ et $\alpha_1, \alpha_2, \dots, \alpha_p$

Comme on l'a vu dans le paragraphe précédent, les paramètres (μ_i) sont facilement estimables par les $(\overline{Y}_{i\bullet})$. En revanche, les (α_i) et μ ne le sont pas directement car il existe une ambigüité pour les définir. En effet, les $(\mu + \alpha_i)$ peuvent s'obtenir d'une infinité de manière.

Pour remédier à cela, on introduit une contrainte qui est généralement la suivante : on suppose que l'effet moyen est nul, c'est à dire que $\displaystyle\sum_{i=1}^{p} n_i \, \alpha_i \ = \ 0$.
On a alors les résultats suivants.

Théorème 7.2 *Sous les hypothèses de normalité et d'indépendance des p échantillons et sous la contrainte* $\sum_{i=1}^{p} n_i \, \alpha_i \ = \ 0$*,*

- $\overline{Y}_{\bullet\bullet}$ *est un estimateur sans biais du paramètre* μ *;*
- *pour tout* $i \in \{1, \dots, p\}$, $\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)$ *est un estimateur sans biais du paramètre* α_i*.*