2017 ■ Vol. 13 ■ no. 1

L'analyse de variance pour des groupes inégaux et la solution du n harmonique : une question d'équité
The unweighted "harmonic mean" solution for unbalanced anova designs : A detailed argument

Louis Laurencelle[a], ✉

[a]Université du Québec à Trois-Rivières

Abstract ■ The treatment of unbalanced designs in analysis of variance (anova) has a long and still controversial history, an issue being the choice between the so-called "harmonic mean" or unweighted solution and the classical weighted solution. We here argue in favour of the unweighted, i.e. equally weighted solution, based on the following reasons. The classical solution gives more weight to the means obtained from a more numerous group of data, thus inducing a positive bias in the computation of the between-group mean square, irrespective of the groups' effect sizes. Indeed, this differential weighing is at variance with the determination and handling of effect sizes, whose values are kept free of the various group sizes implied, so that the final 'weighted' F statistic cannot stand for a truthful reflection of those. Besides, the oft-quoted argument around the demographic representativeness of the various groups compared is specious in the context of most anova applications, the purpose of anova being to compare groups/conditions one to the other, whatever their sample sizes. Finally, in the cases of two- or multi-way designs, the weighted solution precludes the calculation of truly orthogonal and additive variance components, the 'linear regression' alternatives for this problem being complex and essentially arbitrary. The "harmonic mean" solution preserves orthogonality and additivity in the variance decomposition for multi-dimensional designs, is congruent with effect sizes and entails no differential bias in the calculation of the F test whatever the sample sizes. On the other hand, it suffers from a positive bias in the F's significance, a bias negligible for mildly unbalanced group sizes and aptly corrected by RANKIN's (1974) modified degrees of freedom.

Keywords ■ Anova, Unbalanced design, Harmonic mean solution.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

✉ louis.laurencelle@gmail.com

LL: 0000-0003-3448-2872

10.20982/tqmp.13.1.p095

Introduction

L'analyse de variance pour des groupes à tailles inégales peut être effectuée de différentes façons, notamment en appliquant la solution "classique" qui consiste à pondérer chaque moyenne de groupe par la taille correspondante, ou bien en utilisant un facteur de pondération commun pour tous les groupes, le "n harmonique", c'est-à-dire la moyenne harmonique des tailles de groupes.

Nous tenterons ici de démontrer que :

— l'argument de représentativité, celle associée à la taille du groupe par rapport à la taille de la sous-population correspondante, est spécieux et ne doit pas être considéré ;

— la solution pondérée classique biaise de manière différentielle l'évaluation du carré moyen des groupes en faveur des groupes plus nombreux et ne traite pas avec équité les groupes plus petits ;

— le test F concluant l'analyse de variance doit refléter directement et sans biais les grandeurs d'effets estimées dans les groupes, lesquelles sont totalement franches des tailles de groupes ;
— la solution non pondérée, dite par n harmonique, n'introduit pas de biais attribuable aux tailles de groupes inégales mais elle produit un test F positivement biaisé, le biais n'apparaissant sensible que pour des ratios de tailles élevés. Ce biais peut être contenu par une estimation réduite des degrés de liberté du carré moyen des groupes ou être complètement contourné grâce au recours à un test permutationnel ou par rééchantillonnage ;
— la solution pondérée crée, dans le cas d'un devis factoriel à groupes inégaux, des estimations d'effets corrélées, la décomposition orthogonale des effets observés - l'apanage de l'analyse de variance - étant ainsi compromise ; semblablement à l'orthogonalité, l'additivité des effets dans les devis factoriels d'analyse n'est pas respectée par la solution pondérée.

Il convient d'abord de faire un bref rappel du modèle de base de l'analyse de variance.

Le modèle algébrique de l'analyse de variance, en rappel

Reprenant une notation paradigmatique en analyse de variance (Kirk, 1994 ; Winer, Brown & Michels, 1991 ; Sokal & Rohlf, 1981), le modèle d'analyse à une dimension et k groupes s'écrit :

$$X_{i,j} = \mu + \tau_j + \varepsilon_{i,j} (j = 1 \text{ à } k); \tag{1}$$

μ est une constante générale (la moyenne de la "population"), τ_j l'effet (ou décalage) apporté par la condition "j", $\varepsilon_{i,j}$ l'"erreur" associée à la mesure[1], la valeur observée $X_{i,j}$ reflétant la somme de ces trois composants. Or, sous l'hypothèse d'une variation aléatoire du composant ε, indépendante de τ_j, la moyenne d'un groupe de données associées à la condition j représente :

$$\bar{X}_j = \mu + \tau_j + \bar{\varepsilon}_{i,j}, \tag{2}$$

alors que leur variance est :

$$s_j^2 = s_{\varepsilon_j}^2. \tag{3}$$

Quant à la moyenne globale G des données de l'étude, elle est calculée par :

$$G = \sum_{j=1}^{n} n_j \bar{X}_j / N, \tag{4}$$

où $N = \sum_{j=1}^{n} n_j$, et elle représente :

$$G = \mu + \bar{\tau} + \bar{\bar{\varepsilon}} \tag{5}$$

le terme $\bar{\tau}$ s'annulant dans le modèle paramétrique dit "à effets déterminés". Pour plus de commodité et sans perdre de généralité, nous invoquerons ici le modèle dit "à effets aléatoires", pour lequel $\bar{\tau} \neq 0$.

Pour faire court, le test de significativité des différences apportées par les k conditions comparées est réalisé par un quotient F de "carrés moyens", soit :

$$F = CM_{groupes} / CM_{intragroupe}, \tag{6}$$

le carré moyen au dénominateur, obtenu par la moyenne pondérée des variances intragroupe (3), représentant justement le paramètre σ_ε^2.

Le problème d'équité, comme nous souhaitons le nommer, réside entièrement dans le numérateur du test, $CM_{groupes}$. Le calcul habituel de ce terme, inspiré des protocoles d'étude à groupes égaux, est :

$$CM_{groupes} = \frac{\sum_{j=1}^{n} n_j (\bar{X}_j - G)}{k-1}, \qquad \text{(Pondéré ;7)}$$

l'expansion en espérance étant :

$$E\{CM_{groupes}\} = \frac{\sum_{j=1}^{k} n_j [(\mu + \tau_j + \bar{\varepsilon}_j) - (\mu + \bar{\tau} + \bar{\bar{\varepsilon}})]^2}{k-1} \tag{8}$$

et donnant finalement :

$$E\{CM_{groupes}\} = \frac{\sum_{j=1}^{k} n_j (\tau_j - \bar{\tau})^2}{k-1} + \frac{\sum_{j=1}^{k} n_j (\bar{\varepsilon}_j - \bar{\bar{\varepsilon}})^2}{k-1} \tag{9}$$

en admettant la non-corrélation entre l'effet τ_j et l'erreur ε.

L'ingrédient $(\bar{\varepsilon}_j - \bar{\bar{\varepsilon}})^2$ dans le terme de droite de (9) dénote un estimateur de la variance d'une moyenne de n_j éléments ε, de sorte que, par un théorème connu, $n_j \times (\bar{\varepsilon}_j - \bar{\bar{\varepsilon}})^2$ estime le paramètre σ_ε^2, d'où nous obtenons l'estimateur pondéré de la variance des moyennes de groupes :

$$E\{CM_{groupes}\} = \frac{\sum_{j=1}^{k} n_j (\tau_j - \bar{\tau})^2}{k-1} + \sigma_\varepsilon^2 \tag{10}$$

et, si les tailles n_j des groupes sont toutes égales à n, nous avons :

$$E\{CM_{groupes}\} = \frac{n \sum_{j=1}^{k} (\tau_j - \bar{\tau})^2}{k-1} + \sigma_\varepsilon^2 \tag{11}$$

qu'on peut exprimer équivalemment par :

$$E\{CM_{groupes}\} = n\sigma_\tau^2 + \sigma_\varepsilon^2 \tag{12}$$

1. Cette « erreur » reflète collectivement la variabilité interindividuelle des sources échantillonnales de même que « l'erreur de mesure ».

Le calcul par le truchement de la moyenne harmonique des tailles, notée n_h et familièrement désignée "n harmonique", consiste à substituer n_h à n dans la formule (11) afin d'obtenir ce qu'il convient d'appeler l'estimateur non pondéré [2] de l'effet des groupes, lequel est alors :

$$CM_{groupes} = \frac{n_{\mathrm{h}} \sum_{j=1}^{k} \left(\bar{X}_j - G\right)^2}{k-1} \quad \text{(Non pondéré,13)}$$

et il correspond alors à :

$$E\{CM_{groupes}\} = \frac{n_{\mathrm{h}} \sum_{j=1}^{k} \left(\tau_j - \bar{\tau}\right)^2}{k-1} + \tilde{\sigma}^2_{\bar{\varepsilon}}, \quad (14)$$

où :

$$n_h = k/(1/n_1 + 1/n_2 + \cdots + 1/n_k) \quad (15)$$

et :

$$G = \frac{\sum_{j=1}^{k} \bar{X}_j}{k}, \quad (16)$$

l'estimateur $\tilde{\sigma}^2_{\bar{\varepsilon}}$ de $\sigma^2_{\bar{\varepsilon}}$ devenant alors approximatif.

La représentativité, un faux argument

Dans certaines études, notamment les sondages rapportant des données à caractère démographique, l'échantillon observé comportera souvent des sous-groupes correspondant à autant de strates de la population sondée, ces sous-groupes pouvant être de tailles inégales. Bien sûr, l'estimation d'une moyenne (ou d'une proportion) représentative de la population exigera que les moyennes des sous-groupes soient assemblées en pondérant chacune par la taille de la sous-population correspondante : les ouvrages sur l'échantillonnage documentent très bien cette question (Cochran, 1977 ; Kish, 1965, etc.). Ici, toutefois, il est question d'analyse de variance et de comparaison des moyennes dans le but de confronter les unes aux autres les caractéristiques des groupes : que l'étude soit à caractère démographique ou non, il importe avant tout de savoir si telle condition, représentée par un sous-groupe, présente des données pareilles ou significativement différentes par rapport à telle ou telle autre condition : la taille des sous-groupes n'intervient pas et ne doit pas intervenir dans la logique de cette comparaison ni dans sa conclusion.

Le problème d'équité et le calcul du $CM_{groupes}$

Le but et la fonction de l'analyse de variance consistent essentiellement à comparer des moyennes, le plan d'analyse le plus simple touchant la comparaison des moyennes de deux ou quelques groupes. Dans un cadre de recherche de type expérimental, il est habituel de former et comparer des groupes de tailles égales ; c'est aussi le cas usuel des études dites "essais cliniques randomisés". On peut dire de ces comparaisons qu'elles sont équitables en ce que les conditions de traitement ou de mesure que représentent les divers groupes se traduisent par des statistiques - les moyennes - ayant un même poids d'évidence. D'un autre côté, dans les études quasi expérimentales ou celles dites appliquées, le cas est plus fréquent dans lequel la taille des groupes varie, et varie parfois considérablement. Même les groupes égaux des études expérimentales sont susceptibles de devenir inégaux, soit par attrition des personnes, perte ou mauvais enregistrement de données ou par d'autres causes. La comparaison de groupes de tailles égales est équitable de façon évidente, chaque moyenne ayant même représentativité numérique, même poids échantillonnal et une variance comparable. Qu'en est-il du cas de groupes à tailles diverses, et comment doit-on le traiter ?

Nous illustrons notre propos en prenant l'exemple d'une étude à $k = 4$ groupes totalisant $N = 40$ participants, l'un des groupes présentant une moyenne de 15 et les trois autres, de 10 (en unités arbitraires). Le tableau 1 fournit les données et calculs.

Comme on peut voir, tout en conservant les mêmes moyennes, les tailles égales de la ligne A donnent des carrés moyens égaux. Dans le cas de tailles inégales, aux lignes B et C, les deux estimateurs de carrés moyens diffèrent. Alors que la solution en n harmonique produit des carrés moyens de même grandeur, la solution pondérée pondère différentiellement les moyennes comparées selon leurs tailles : l'impact de la moyenne divergente, au groupe 1, va être réduit lorsque la taille sera plus faible (en B) ou être accru quand elle sera plus forte (en C). Les graphiques de la figure 1 donnent une image plus complète de ces effets différentiels : les mêmes moyennes qu'au tableau 1 y sont pondérées par les tailles de groupe de groupe $n_1 = 1$ à $19, n_2 = n_3 = 10, n_4 = 20 - n_1$, pour un N total de 40.

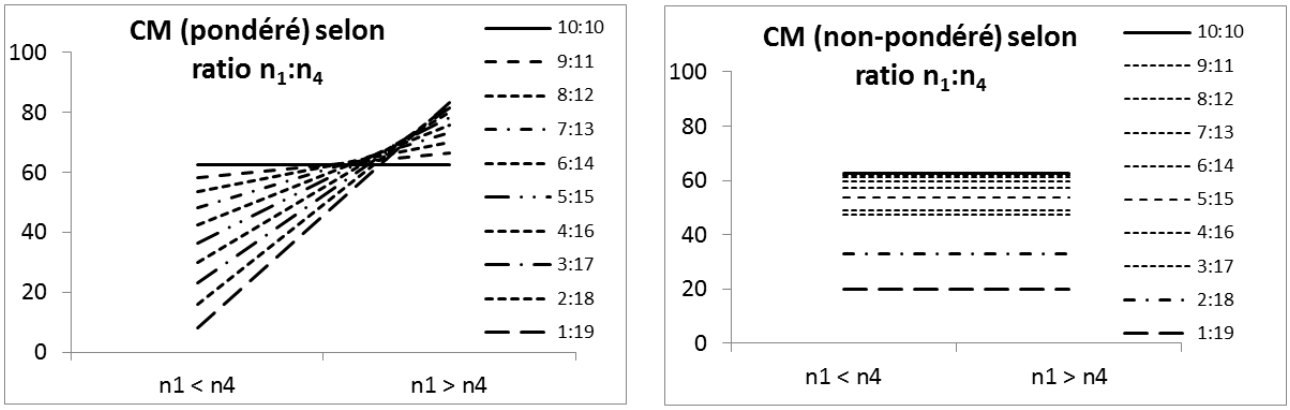
Dans les cas d'inégalités de tailles, la solution non pondérée propose une solution égale, disons équitable, et qui reflète directement les grandeurs d'effet obtenues, ce quelle que soit la répartition des tailles n_j entre les groupes : c'est ce qu'indique la figure 1, à droite. Le seul effet, important, de l'inégalité de tailles est d'amoindrir ou gruger la valeur du carré moyen ($CM_{groupes}$), via la diminution du n_h, ce pour tenir compte de la plus grande variabilité globale des moyennes. Dans le cas du calcul pondéré, illustré à gauche, le CM sera plus fort si la taille du groupe divergent (le groupe 1 dans notre exemple) est plus grande, et plus faible si la taille est plus petite. Le CM pondéré, et par conséquent le test F qui en découle, sont donc biaisés par la répartition des tailles de groupes, un argument que nous reprenons plus loin relativement à l'impact des gran-

2. Plus précisément, il s'agit ici d'une pondération égale.

Tableau 1 ■ Carrés moyens des groupes en formules pondérée (formule 7) et non pondérée (formule 13) selon trois configurations des données de 4 groupes.

Groupe	1	2	3	4	n_h	CM pond.	CM non-pond.
Moyenne	15	10	10	10			
n (A)	10	10	10	10	10,00	62,50	62,50
n (B)	5	10	10	15	8,57	36,46	53,57
n (C)	15	10	10	5	8,57	78,13	53,57

Figure 1 ■ Valeurs des CM pondérés et non-pondérés



deurs d'effets sur le calcul et sur la puissance.

Test de l'effet versus grandeur d'effet, ou test t versus analyse de variance

Que ce soit par le moyen d'une étude expérimentale destinée à vérifier l'effet de quelques conditions de traitement sur des cobayes ou dans le cadre d'une recherche descriptive comparant divers regroupements de participants, l'anova a pour but d'établir la crédibilité des différences observées entre une série de moyennes, ces différences devant refléter les "effets" de ces conditions ou regroupements. Et qu'est-ce qu'un effet ? Dans le cas d'une ou de plusieurs moyennes, on peut quantifier un effet ou des effets de diverses manières, par exemple : $\bar{X} - \mu$ (par rapport à une valeur de référence), $\bar{X}_1 - \bar{X}_2$ (par rapport à une autre moyenne), $var(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ (pour un ensemble de moyennes) , le fameux d de Cohen (1988), $\left(\bar{X}_1 - \bar{X}_2\right)/s$, étant une variante de celles-ci. Donc, pour autant que le résultat statistique de l'anova et la conclusion qui doit s'ensuivre sont supposés refléter les "effets" mesurés, ces effets étant identifiés ci-dessus sans référence à la taille des groupes qui les ont produits, comment justifier le calcul pondéré (7) ? Voyons maintenant où et comment cette pondération intervient.

C'est un argument pédagogique traditionnel que d'affirmer et de montrer que l'analyse de variance (ou anova) est une généralisation du test t de Student pour la comparaison des moyennes de deux ou plusieurs groupes. Quant au volet savant de l'argument, il concerne la parenté directe qu'il y a entre le t correspondant, avec $\nu = n_1 + n_2 - 2$ degrés de liberté, et le F de l'anova pour deux groupes, doté de 1 degré de liberté au numérateur et ν au dénominateur, soit, pour le seuil de significativité α,

$$F_{1,\nu[1-\alpha]} = t^2_{\nu[1-\alpha/2]}. \tag{17}$$

Le *test t pour deux moyennes*. Or, le test t pour deux groupes, soit :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)\times s_1^2 + (n_2-1)\times s_2^2}{n_1+n_2-2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}, \tag{18}$$

est une fonction strictement proportionnelle de l'expression $\bar{X}_1 - \bar{X}_2$, dans laquelle les deux moyennes exercent une influence égale. Une expression équivalente à (18) serait :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_\varepsilon \sqrt{1/n_1 + 1/n_2}}, \tag{19}$$

où :

$$\hat{\sigma}_{\varepsilon}^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2} = CM_{intragroupe}. \tag{20}$$

Escamotant l'estimateur d'erreur $\hat{\sigma}_{\varepsilon}$, nous retenons la seule différence "standardisée" des moyennes, que nous noterons t', soit :

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{1/n_1 + 1/n_2}}. \tag{21}$$

Rappelant la définition du n harmonique plus haut (éq. 15), on montre que :

$$t' = \frac{\sqrt{n_h}\,(\bar{X}_1 - \bar{X}_2)}{\sqrt{2}}. \tag{22}$$

Or, puisque $s^2(A, B) = (A - B)^2/2$, nous obtenons enfin :

$$(t')^2 = n_h s^2(\bar{X}_1,\ \bar{X}_2), \tag{23}$$

expression qui indique que le 'numérateur' du test t sur deux moyennes reflète la différence simple entre celles-ci, différence modulée par la moyenne harmonique des tailles correspondantes.

Nous allons montrer maintenant que, si d'une part, le $CM_{groupes}$ de l'anova pour deux groupes selon la formule 7 est algébriquement équivalent la formule 23 ci-dessus, se comportant ainsi comme le test t, ce n'est plus le cas de l'anova pour trois groupes ou davantage, où un jeu de pondérations complexes en fonction des tailles de groupes vient s'immiscer dans le calcul.

L'anova pour deux moyennes. Réécrite pour le cas de 2 moyennes, la formule 7 devient :

$$CM_{groupes} = \frac{n_1\left(\bar{X}_1 - \bar{X}_{\mathrm{G}}\right)^2 + n_2\left(\bar{X}_2 - \bar{X}_{\mathrm{G}}\right)^2}{2\ -\ 1}, \tag{24}$$

où :

$$\bar{X}_{\mathrm{G}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}. \tag{25}$$

Après quelques manipulations et réductions, l'équation 24 devient :

$$CM_{groupes} = \left(\bar{X}_1 - \bar{X}_2\right)^2 \left(\frac{n_1 n_2}{n_1 + n_2}\right) \tag{26}$$

et :

$$CM_{groupes} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)^2}{2} \left(\frac{2n_1 n_2}{n_1 + n_2}\right) \tag{27}$$

$$= \frac{\left(\bar{X}_1 - \bar{X}_2\right)^2}{2} \cdot \frac{2}{\left(1/n_1 + 1/n_2\right)} \tag{28}$$

$$= s^2(\bar{X}_1, \bar{X}_2) \cdot n_h = (t')^2, \tag{29}$$

un résultat qui, rappelons-le, démontre que l'anova pour deux moyennes procède comme le test t pour cette situation et *ne pondère pas les moyennes* en fonction de la taille des groupes.

L'anova pour trois moyennes ou plus. Prenant maintenant le cas de trois moyennes, cas généralisable à quatre moyennes ou plus, nous réécrivons la formule 7 comme suit :

$$CM_{groupes} = \frac{\sum_{j=1}^{3} n_j \left(\bar{X}_j\ -\ \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{N}\right)^2}{3 - 1}. \tag{30}$$

Retenons le premier terme de la somme, soit celui associé au groupe $j = 1$, ce qui donne :

$$n_1 \left[\frac{\left(N \cdot \bar{X}_1 - n_1\bar{X}_1 - n_2\bar{X}_2 - n_3\bar{X}_3\right)}{N}\right]^2 \tag{31}$$

$$= n_1 \left[\frac{(n_2 + n_3) \cdot \bar{X}_1 - n_2\bar{X}_2 - n_3\bar{X}_3}{N}\right]^2, \tag{32}$$

une expression qui montre à l'évidence que chaque moyenne, ici $\bar{X}_1$, est pondérée, c'est-à-dire traitée de façon singulière et différentielle, selon un système complexe de poids basé sur les tailles de groupes[3]. On est loin du calcul parfaitement égalitaire, et équitable, fourni par la formule 13, basée sur la solution dite du n harmonique.

Biais de significativité et correction de Rankin

Sous l'hypothèse nulle, c.-à-d. en l'absence de différences systématiques d'une condition à l'autre, la solution pondérée est toujours juste et respecte le seuil α prescrit (Box, 1954). La solution non pondérée utilisant le n harmonique, quant à elle, a comme inconvénient peu connu un biais de significativité positif, soit, sous l'hypothèse nulle, $P(F|\alpha) \geq \alpha$: ce biais est d'autant plus important que l'inégalité des tailles de groupes augmente. Rankin (1974) se penche sur ce problème, et il propose une correction destinée à rétablir le niveau de significativité prescrit. Cette correction s'applique aux degrés de liberté ($dl_{groupes}$) du carré moyen des groupes, comme suit :

$$dl_{groupes} = (k - 1) \cdot e, \tag{33}$$

3. Supposant par exemple que la taille du groupe 1 est minime, p. ex. $n_1 = 1$, non seulement sa composante et sa contribution au CM seront-elles minimisées par le facteur n_1 précédant le crochet mais, en outre, sa participation au calcul des composantes 2 et 3, à l'intérieur des crochets, sera presque occultée, rendant ainsi proportionnellement inopérante la valeur de la moyenne $\bar{X}_1$ correspondante.

où :

$$e = \left(1 + \frac{C^2}{k-1}\right)^{-1} \text{ et } C^2 = \frac{k-2}{k}\sum_{j=1}^{k}\frac{(n_j - n_{\text{h}})^2}{n_j^2}; \tag{34}$$

l'élément-clé de la correction ressort évidemment comme la quasi variance des n_j calculée dans la quantité C^2. Les degrés de liberté obtenus par (33) comportant souvent une partie fractionnaire, l'utilisateur doit alors recourir à l'interpolation pour estimer la valeur à partir d'une table de la distribution F (ou d'un logiciel la fournissant) : l'interpolation harmonique est recommandée.

Prenons pour exemple une situation présentant $k = 4$ groupes de tailles $n_j = 10$, 10, 8 et 20, pour $N = 48$. Nous obtenons $n_h \approx 10,667, C^2 \approx 0,169, e \approx 0,947$ et $dl_{groupes} \approx (4-1) \times 0,947 \approx 2,841$. Les degrés de liberté de la variance intragroupe étant $dl_{intragroupe} = N - k = 48-4 = 44$, nous avons à estimer la valeur de $F_{2,841;44[0,95]}$ pour le seuil α de 5%, ce entre $F_{2;44}$ et $F_{3;44}$. Pour ce seuil, une table du F donne $F_{2;44} = 3,209$ et $F_{3;44} = 2,816$, d'où, par interpolation harmonique, nous calculons :

$$\begin{aligned} F_{2,841;44} &\approx F_{2;44} + (F_{3;44} - F_{2;44}) \\ &\quad \times (1/2,841 - 1/2)/(1/3 - 1/2) \\ &\approx 2,860. \end{aligned}$$

Nous avons exploré cette proposition de Rankin par d'extensives expérimentations Monte Carlo. Le tableau 2 illustre les résultats.

Tel qu'attendu, les inégalités de tailles n'affectent en rien la solution pondérée dans le cas où l'hypothèse nulle est vraie et il n'existe aucune différence systématique entre les groupes, cas illustré ici. Quant à la solution par n harmonique calculée selon (13), son application simple, en utilisant les degrés de liberté habituels, soit $dl_{groupes} = k - 1$, tend à devenir sur-significative à mesure que les tailles s'éloignent les unes des autres. L'effet est petit, voire négligeable, pour de petites inégalités, disons $n_{max}/n_{min} \leq 1,5$, et il devient gênant au-delà, créant un surplus artificiel de "puissance". La correction (33) de Rankin (1974), on le voit, attaque correctement le problème. En résumé, la solution de l'anova à tailles inégales par la méthode dite du n harmonique peut, si les différences de tailles sont importantes, nécessiter une correction des degrés de liberté associés au carré moyen inter-groupes, grâce à quoi le biais de sur-significativité qu'elle entraîne est effacé[4]. Cette correction n'est pas requise lorsque la significativité de l'effet (reflété par un quotient F) est estimée par ré-échantillonnage (Edgington, 1980 ; Robert & Casella, 1999 ; Laurencelle, 2001) plutôt qu'à partir de la loi F, une approche "validante" déjà endossée par R. A. R. A. (Fisher (1971).

Puissance (ou sensibilité) du F selon les solutions pondérée et non pondérée (Rankin)

Qu'en est-il maintenant de la puissance du test F global, dans le contexte d'une anova à tailles inégales ? Pour documenter cette facette du problème, nous n'avons trouvé rien de paru. L'exploration de configurations paramétriques variées ayant produit des résultats semblables, nous en donnons un exemple illustré à la figure 2. L'échantillon global comporte 40 participants répartis en $k = 4$ groupes de tailles variables, allant de l'égalité complète ($n_j = 10, 10, 10, 10$) à une inégalité extrême ($n_j = 1, 1, 1, 37$). Quant à la grandeur d'effet, elle est de $\theta = 0,75$[5] , générée à partir du vecteur de moyennes $\mu_j = 0, 0, 0, 1$ pour une variance d'erreur paramétrique (intragroupe) de 1. Les tailles allouées aux quatre groupes ont été tour à tour affectées en ordre direct (p. ex. 8, 8, 8, 16) et inversé (p. ex. 16, 8, 8, 8) par rapport à l'ordre des moyennes, ce afin d'estimer le biais éventuel de la pondération donnée à la quatrième moyenne, $\mu_4 = 1$.

À partir d'un maximum observé pour le cas de tailles égales[6], la puissance de l'anova non pondérée tend à décroître à mesure qu'augmentent les inégalités, une tendance que reflète aussi la valeur du n_h ; le même scénario se produit aussi pour la solution pondérée pour des tailles en ordre inversé. Dans le cas de tailles en ordre direct, la solution pondérée avantage la moyenne μ_4 plus forte, ce qui ressort par un sursaut de puissance maintenu jusqu'à la répartition '4, 4, 4, 28', l'avantage restant tout de même sensible aux inégalités plus fortes. Un aspect remarquable de ces résultats découle du contraste entre les deux ordres, direct versus inversé, pour chaque mode de calcul : pour le calcul pondéré, les ratios de 'puissance' s'échelonnent de 1,20 à 2,41 pour une médiane de 1,92, tandis qu'avec la solution Rankin, nous observons 1,01 à 1,23 pour une médiane de 1,05, cette seconde série apparaissant plus équilibrée : rappelons qu'ici, pour toutes ces configurations, la grandeur d'effet est la même.

La puissance statistique est généralement diminuée, on le voit, quand les groupes sont de tailles inégales et ce, davantage dans la solution non pondérée que dans la pondérée. Toutefois, la solution pondérée est affectée de façon différentielle par cette inégalité : si les conditions le permettent, la puissance peut passer du simple au double ou davantage en favorisant le groupe plus nombreux,

4. Dans le même contexte d'une anova à tailles inégales, Thomas et Hultquist (1978) proposent une méthode de calcul pour l'intervalle de confiance de la variance expérimentale, σ_τ^2.

5. Rappelons que, en anova, une mesure de la grandeur d'effet est $\theta = \sum(\mu_j - \mu)^2/\sigma^2$.

6. Mentionnons que, pour la configuration de moyennes présentée et des tailles égales, la puissance *calculée* est de 0,571, telle qu'indiquée à la figure 2.

Tableau 2 ■ Significativité du test F à $\alpha = 0,05$ pour l'analyse de variance de k groupes inégaux sous l'hypothèse nulle, selon trois solutions†. Les valeurs imprimées sont basées sur 2×10^6 itérations ($IC_{95} = \{0,0497..0,0503\}$)

		Calcul non pondéré (n harmonique)	
Répartition des tailles n_j	Calcul pondéré	$dl_{gr} = k - 1$	$dl_{gr} = (k-1) \cdot e$
$k = 3 (N = 30)$			
8, 10, 12	0,0501	0,0505	0,0501
5, 10, 15	0,0498	0,0531	0,0495
2, 10, 18	0,0501	0,0626	0,0502
1, 1, 28	0,0500	0,0567	0,0497
$k = 5 (N = 50)$			
8, 9, 10, 11, 12	0,0500	0,0508	0,0501
6, 8, 10, 12, 14	0,0501	0,0532	0,0498
2, 5, 10, 15, 18	0,0498	0,0711	0,0490
2, 2, 2, 22, 22	0,0497	0,0655	0,0495
1, 1, 1, 1, 46	0,0499	0,0584	0,0502
$k = 10 (N = 100)$			
2,4,6,6,7,9,12,16,18,20	0,0500	0,0758	0,0496
5,5,5,5,5,15,15,15,15,15	0,0501	0,0625	0,0504
3,3,3,3,3,17,17,17,17,17	0,0502	0,0728	0,0506
1,1,1,1,1,19,19,19,19,19	0,0501	0,0852	0,0505
1,1,1,1,1,1,1,1,1,19	0,0498	0,0555	0,0500

Note. Les deux solutions dites non pondérées diffèrent uniquement par les degrés de liberté du carré moyen des groupes ($dl_{groupes}$), auxquels on applique ou non la correction basée sur l'indice de variation e de RANKIN (1974).

créant ainsi un biais de conclusion, biais complètement indépendant de la grandeur d'effet. Cela dit, les données affichées à la figure 2 en témoignent et comme le recommandent tous les manuels de méthodologie et de statistique appliquée, mieux vaut un contexte à groupes égaux ou, sinon, des groupes à inégalités légères.

Contrastes et comparaisons versus grandeurs d'effet

Comparaisons. La sanction fournie par le quotient F d'une analyse de variance pour trois groupes ou plus ne termine pas le travail du chercheur, du moins elle ne le fait pas souvent. Voire, elle n'en est même pas une condition. Le chercheur, par exemple, peut devoir déterminer si chacun des deux groupes dits 'expérimentaux' diffère du groupe témoin (par sa moyenne), ou repérer lesquelles parmi les six comparaisons deux à deux de quatre moyennes paraissent statistiquement sérieuses, etc. Nous sommes placés dans le contexte de comparaisons multiples, contexte pour lequel plusieurs techniques de décision sont proposées dans la littérature, presque toutes incorporant un contrôle de risque tout à fait indépendant du F global, la plupart utilisant une formule de type t de Student, c'est-à-dire sans pondération différentielle des moyennes. À titre d'exemple, pour la comparaison de $k-1$ groupes expérimentaux à un $k$ième groupe témoin, la technique de DUNNETT (1955 ; voir aussi SOKAL et ROHLF, 1981, ou WINER et al., 1991) s'impose, la formule pour des groupes inégaux prenant la forme suivante :

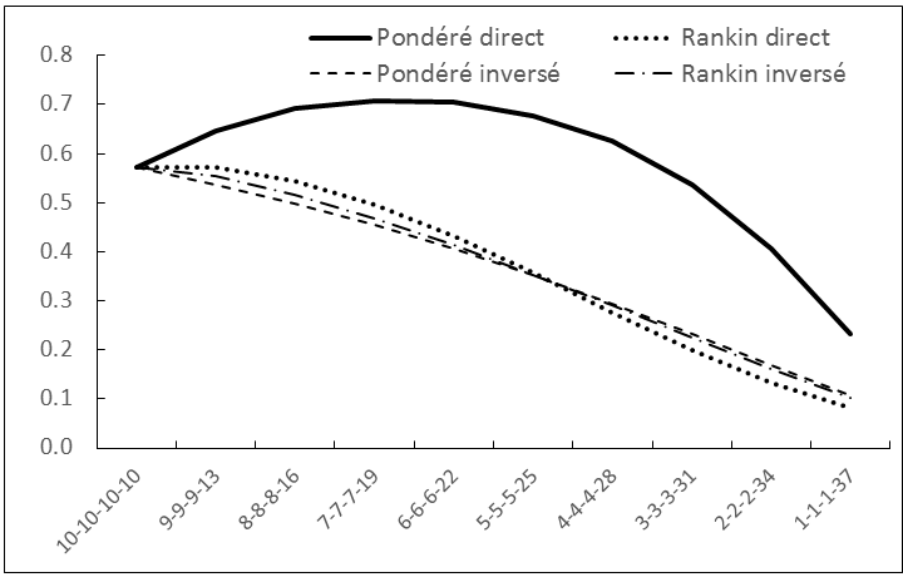
$$t = \frac{\bar{X}_{\mathrm{E1}} - \bar{X}_{\mathrm{T}}}{\hat{\sigma}_e \sqrt{\frac{1}{n_{\mathrm{E1}}} + \frac{1}{n_{\mathrm{T}}}}}, \tag{35}$$

une statistique approximativement distribuée comme la loi t_D de Dunnet, forme utilisant l'équivalent de la moyenne harmonique des n_j concernés, tout comme dans le classique test t de Student. Cette statistique (35), comme toutes celles qui lui sont semblables, est strictement proportionnelle à la grandeur d'effet brute, ici $\bar{X}_{\mathrm{E1}} - \bar{X}_{\mathrm{T}}$, sans que les tailles n_j y apportent une pondération différentielle.

Contrastes. Par contraste[7], nous désignons ici un composé linéaire de moyennes de forme $C = c_1 \times \bar{X}_1 + c_2 \times \bar{X}_2 + \cdots + c_k \times \bar{X}_k$ contraint par $c_1 + c_2 + \cdots + c_k = 0$ tel que, sous l'hypothèse nulle, $E\{C\} = 0$; les comparaisons de moyennes deux à deux sont un cas particulier de contraste. La comparaison conjointe de deux groupes expérimentaux

7. Certains manuels plus anciens faisaient, à propos d'une collection de contrastes formés à partir du même vecteur de moyennes, la distinction entre des contrastes orthogonaux ou non orthogonaux, ceux du premier ensemble étant soi-disant épargnés de la nécessité d'un contrôle global du taux d'erreur de type I. Cette distinction spécieuse a été largement abandonnée aujourd'hui et, pour ceux qui accréditent l'argument d'un contrôle global du taux d'erreur, tous les contrastes d'un ensemble, orthogonal ou pas, doivent s'y soumettre. Voir cependant LAURENCELLE (2007).

FIGURE 2 ■ Puissance du test F d'analyse de variance correspondant à des effets $\mu = 0, 0, 0, 1$ pour les solutions Pondérées et Rankin (non pondérées) sous des répartitions à inégalités croissantes (indiquées en abscisses), selon un ordre Direct (la taille plus forte coïncide avec l'effet dominant, $\mu_4 = 1$) ou un ordre Inversé (données de simulation sur 2×10^6 échantillons)



à un témoin, selon les coefficients $c_j = 1, 1, -2$ donne un exemple simple, le carré moyen approprié étant ici :

$$(C_{1,1,-2})^2 = \frac{\left(\bar{X}_1 + \bar{X}_2 - 2\bar{X}_3\right)^2}{\frac{1^2}{n_1} + \frac{1^2}{n_2} + \frac{(-2)^2}{n_3}}, \tag{36}$$

lequel se réduirait à $\left(\bar{X}_1 + \bar{X}_2 - 2\bar{X}_3\right)^2 \times n/6$ si $n_j = n$ pour tous j.

Prenons une autre exemple, celui d'un contraste basé sur le modèle polynomial linéaire, au premier degré. Avec $k = 5$ groupes (ou niveaux de condition ordonnés et 'équidistants'), les coefficients bruts du contraste sont $c_j = -2, -1, 0, 1, 2$, le numérateur du contraste s'écrivant :

$$C_{lin} = -2 \times \bar{X}_1 - 1 \times \bar{X}_2 + 0 \times \bar{X}_3 + 1 \times \bar{X}_4 + 2 \times \bar{X}_5. \tag{37}$$

Le carré moyen correspondant, doté de 1 degré de liberté, s'obtient par :

$$C_{lin}^2 / [(-2)^2/n_1 + (-1)^2/n_2 + (0)^2/n_3 + 1^2/n_4 + 2^2/n_5], \tag{38}$$

expression qui se réduit à $C_{lin}^2/10 \times n$ lorsque $n_j = n$ pour tous j. Ici comme ci-dessus pour les comparaisons deux à deux, la gestion des tailles (inégales ou non) n'encourt aucune pondération directe des moyennes individuelles mais utilise plutôt une moyenne harmonique particulière, utilisée globalement au dénominateur et modulée par les coefficients du contraste. Il est en outre impensable, sinon impossible, de pondérer chaque moyenne par sa taille sans fausser complètement la portée interprétative du contraste.

Additivité et orthogonalité des effets

Telle que proposée initialement par R. A. FISHER (1925/1970), l'analyse de variance consiste en une décomposition additive de la variance totale d'un tableau de données et elle sous-tend en principe la propriété d'orthogonalité. Prenons le cas d'un plan d'analyse simple, à k groupes. Le tableau est constitué de N données réparties en k colonnes (ou 'groupes') de n_j lignes (ou 'participants') chacune ; la variance totale, avec $N - 1$ degrés de liberté, se divisera exactement en une variance inter-colonnes comportant $k - 1$ degrés de liberté et une autre intra-colonnes, amalgame des variances inter-participants, avec $N - k$ degrés de liberté. Cette décomposition est exactement respectée par la solution pondérée, que les tailles n_j soient égales ou inégales, tandis qu'elle ne l'est pas dans la solution dite par n harmonique. C'est là un inconvénient reconnu de la solution non pondérée, lequel est, comme nous avons vu, compensé par sa propriété que, si l'hypothèse nulle n'est pas avérée, les effets (ou bris d'égalité) des moyennes paramétriques des 'groupes' sont équitablement respectés.

Pour les plans d'analyse comportant deux ou plusieurs facteurs, l'enjeu est tout autre, en ce sens que, pour ces plans et dans le cas de tailles inégales, ni l'additivité ni surtout l'orthogonalité des effets ne sont respectées par la solution pondérée[8]. Ce problème peut être et a été approché de deux manières, et ce depuis longtemps. L'une consiste à invoquer le "modèle linéaire général" et procéder par régression linéaire, en convertissant par la méthode habituelle les paramètres de structure du plan d'analyse, p. ex. A, B et A×B dans un plan à deux dimensions, en prédicteurs multiples, *puis en évaluant séquentiellement* les parts du coefficient de détermination (R^2) qui sont attribuables à ces paramètres. La littérature propose trois solutions, dénotées I, II et III (Overall & Spiegel, 1969 ; Howell, 2008), qui produiront les carrés moyens principaux, p. ex. CM_A et CM_B, de valeurs toutes différentes selon la solution choisie : reste alors à l'utilisateur de choisir la solution qui lui convient[9]. Ces solutions ne sont pas orthogonales, la séquence de calcul déterminant la valeur des résultats ; elles ne sont pas additives non plus, notamment par le fait que la variance totale du tableau n'est pas épuisée. C'est aussi le cas de l'autre manière, celle du n harmonique, qui n'explique pas entièrement la variance totale. Cependant, l'utilisation d'un facteur de pondération unique, le n_{harm}, fait en sorte que la variance des effets (p. ex. A, B et A×B) est décomposée exactement et les carrés moyens résultants sont orthogonaux. Cette manière, moins équivoque que l'autre, est endossée par Keppel (1992), Howell (2008) et d'autres, en plus d'être de calcul plus facile. Howell note que la solution par régression désignée III et celle du n_{harm} fournissent des résultats voisins. Gosslee et Lucas (1965), citant Box (1954), mentionnent enfin que les inégalités de tailles n'affectent que modérément le seuil de signification, la correction de Rankin (1974), discutée plus haut, réglant définitivement ce problème.

Un essai de conclusion

Qu'il ait planifié ou non son étude mettant en jeu des groupes égaux, le chercheur se voit souvent confronté à des inégalités petites ou importantes d'un groupe à l'autre : l'analyse doit-elle prendre en compte ou non ces inégalités ou, en termes plus clairs, un groupe de taille plus importante doit-il peser plus fort sur le calcul des effets par rapport à un groupe moins nombreux ?

L'examen des méthodes de calcul en lice révèle que toutes celles qui reflètent expressément les inégalités de tailles entre les groupes trahissent ou biaisent la valeur des grandeurs d'effets, lesquelles sont indépendantes des tailles, et produisent, dans le cas de plans d'analyse à deux ou plusieurs dimensions, des composantes de variance (ou carrés moyens) non-orthogonales et non-additives. Ces conséquences, qui n'ont qu'une importance mineure *dans les cas où l'hypothèse nulle est vraie*, deviennent critiques dans des études où le chercheur s'ingénie à la contredire et veut faire apparaître des effets. C'est ainsi, en prenant le point de vue et le parti du chercheur, que nous posons que chaque groupe, chaque condition comparée, doit être traité(e) pour ce qu'il ou elle représente, donc également, et non pas sous l'influence différentielle du nombre d'éléments (ou de participants) qui le ou la représentent. Le calcul de l'analyse de variance par la solution non pondérée, dite aussi du n harmonique, respecte ce point de vue, est non biaisé par rapport aux grandeurs d'effets comme vis-à-vis du seuil de significativité convenu (grâce à la correction de Rankin (1974), ou à un calcul par ré-échantillonnage), et permet une décomposition orthogonale et additive des effets.

Références

Box, G. E. P. (1954). Some theorems on quadratic forms applied inthe study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*, 290–302.

Cochran, W. G. (1977). *Sampling techniques (3e édition)*. New York : Wiley.

Cohen, J. (1988). *Statistical power analysis for the social sciences*. Hillsdale NJ : Lawrence Erlbaum.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, *50*, 1096–1121.

Edgington, E. S. (1980). *Randomization tests*. New York : Marcel Dekker.

Fisher, R. A. (1925/1970). *Statistical methods for the research workers (14e édition)*. New York : Hafner Press.

Fisher, R. A. ((1971). *The design of experiments (7e édition)*. New York : Hafner Press.

Gosslee, D. G. & Lucas, H. L. (1965). Analysis of variance of disproportionate data when interaction is present. *Biometics*, *21*, 115–133.

Howell, D. C. (2008). *Méthodes statistiques en sciences humaines*. Bruxelles : de Boeck.

Keppel, G. (1992). *Introduction to design and analysis : A student's handbook*. New York : Freeman.

Kirk, R. E. (1994). *Experimental design : procedures for the behavioural sciences*. PacificGrove (CA) : Brooks / Cole.

Kish, L. (1965). *Survey sampling*. New York : Wiley.

8. Font exception les tableaux dans lesquels les tailles inégales sont proportionnelles à travers les niveaux d'une dimension, la décomposition orthogonale y étant alors observée.

9. Le logiciel IBM SPSS (version 24) offre les trois solutions et propose par défaut la solution III.

Laurencelle, L. (2001). *Hasard, nombres aléatoires et méthode Monte Carlo*. Québec : Presses de l'Université du Québec.

Laurencelle, L. (2007). Inventer ou estimer la puissance statistique ? *Quelques considérations utiles pour le chercheur*. *The Quantitative Methods for Psychology*, *3*, 35–42.

Overall, J. E. & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin*, *72*, 311–322.

Rankin, N. O. (1974). The harmonic mean method for one-way and two-way analyses of variance. *Biometrika*, *61*, 117–122.

Robert, C. P. & Casella, G. (1999). *Statistical Monte Carlo methods*. New York : Springer.

Sokal, R. R. & Rohlf, F. J. (1981). *Biometry (2e édition)*. New York : Freeman.

Thomas, J. D. & Hultquist, R. A. (1978). Interval estimation for the unbalanced case of the one-way random effects model. *Annals of Statistics*, *6*, 582–587.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991). *Statistical principles in experimental design (3e édition)*. New York : McGraw-Hill.

Citation

Laurencelle, L. (2017). The unweighted "harmonic mean" solution for unbalanced anova designs : A detailed argument. *The Quantitative Methods for Psychology*, *13*(1), 95–104. doi :10.20982/tqmp.13.1.p095

Received: 04/11/2016 ~ Accepted: 29/11/2016