Analyse de variance à 2 facteurs imbriqués sur données de comptage - Application au contrôle de qualité

Florence Loingeville [1,2,3] , Julien Jacques [1,2] , Cristian Preda [1,2] , Philippe Guarini [3] & Olivier Molinier [3]

[1] *Inria Lille - Nord Europe - florence.loingeville@inria.fr*
[2] *Laboratoire Paul Painlevé / Université de Lille 1 - julien.jacques@inria.fr, cristian.preda@inria.fr*
[3] *AGLAE Hallennes-lez-Haubourdin - philippe.guarini@association-aglae.fr, olivier.molinier@association-aglae.fr*

Résumé. Le contrôle externe de qualité permet aux laboratoires chargés de l'analyse de l'environnement de vérifier, d'améliorer, et de maintenir en permanence un niveau optimal de leurs analyses. La méthode principalement utilisée pour la mise en œuvre du contrôle externe de qualité est la comparaison interlaboratoires. En microbiologie, les mesures portent sur des dénombrements bactériens. Le caractère discret des données nécessite une méthode d'analyse de la variance spécifique. Nous proposons ici une méthode d'analyse de variance à deux facteurs fixes imbriqués sur données poissoniennes.

Mots-clés. Analyse de variance, données discrètes, analyse de déviance

Abstract. External quality control enables laboratories which analyse the environment always to check, improve and maintain an optimal level of their analysis. Proficiency test is the mostly used method to carry out quality control. In microbiology, measures are bacterial counts. The discrete nature of the data requires a specific analysis of variance method. In this context, we propose a two-fixed-nested-factor analysis of variance method for Poisson data.

Keywords. Analysis of variance, discrete data, analysis of deviance

1 Introduction

Le domaine de l'analyse microbiologique de l'eau et de l'environnement est un secteur dans lequel les besoins en termes de maîtrise du résultat d'analyse sont croissants. La comparaison interlaboratoires, exercice qui consiste à soumettre un même essai à plusieurs établissements, est l'outil utilisé pour la mise en œuvre du contrôle externe de qualité. L'organisation des essais interlaboratoires peut être résumée par le plan d'expérience suivant : un matériau à analyser est envoyé à chaque laboratoire participant à l'essai, sous la forme de deux échantillons A et B. Chaque échantillon est séparé en deux réplications

(A1, A2 et B1, B2) par le laboratoire, qui réalise ensuite dans des conditions de répétabilité les mesures demandées sur les quatre réplications dont il dispose. En microbiologie, la mesure sera un dénombrement de particules (germes). À partir des mesures reçues de la part de l'ensemble des laboratoires, trois critères de qualité sont évalués:

- La capacité d'un laboratoire à répéter ses analyses (écarts entre A1 et A2, B1 et B2).
- L'hétérogénéité des préparations envoyées aux laboratoires (écarts entre A et B).
- La justesse des mesures des laboratoires : on détermine si les mesures de chaque laboratoire sont ou non cohérentes avec celles de l'ensemble des laboratoires.

L'outil statistique de base pour évaluer ces trois critères de contrôle qualité est l'analyse de variance, qui consiste à décomposer la variance totale des résultats de mesures en une variance inter-laboratoires et une variance intra-laboratoire, laquelle pouvant elle même être décomposée en une variance inter-échantillons (écart entre A et B) et intra-échantillon (écart entre A1 et A2 puis entre B1 et B2). L'objectif est alors d'évaluer la significativité des différences entre les mesures sur la base de cette décomposition.

La méthodologie actuellement utilisée en microbiologie suppose que, dans la plupart des cas, les dénombrements suivent une distribution log-normale (Norme ISO/TS 22117, 2010): les données sont transformées en valeurs logarithme décimal et une analyse de variance classique pour données gaussiennes est appliquée. Il est connu qu'appliquer une transformation aux données n'est pas toujours pertinent (R.B. O'Hara, D.J Kotze, 2010). Nous proposons une méthode alternative d'analyse de la variance à deux facteurs imbriqués, sous l'hypothèse que les données sont distribuées suivant une loi de Poisson.

En section 2, nous modélisons le problème. Les différentes configurations possibles et les tests associés sont ensuite exposés (section 3). Nous décrivons en section 4 l'un des tests, puis nous comparons sa puissance à celle du test utilisé en microbiologie (section 5). Les perspectives de ce travail sont présentées dans la section 6.

2 Modélisation du problème

Le plan d'expérience des essais interlaboratoires peut être décrit par un modèle d'analyse de la variance à deux facteurs imbriqués. Le facteur "Laboratoire" a a niveaux, et le facteur "Flacon" a b niveaux imbriqués dans chaque niveau du facteur "Laboratoire". Nous considérons n réplications par flacon.

Notons y_{ijk} le résultat de la mesure réalisée sur la $k^{ième}$ réplication du flacon j par le laboratoire i. Les résultats de mesure peuvent être décrits par le modèle linéaire suivant:

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \epsilon_{(ij)k} \quad \begin{cases} i = 1, ..., a \\ j = 1, ..., b \\ k = 1, ..., n \end{cases} \quad (1)$$

où μ est la moyenne globale, τ_i l'effet du laboratoire i, $\beta_{j(i)}$ l'effet du flacon j imbriqué dans le laboratoire i, et $\epsilon_{(ij)k}$ l'erreur de mesure entre les réplications d'un flacon pour un laboratoire. Nous considérons un plan équilibré comportant le même nombre de niveaux de "Flacon" ($b = 2$) sous chaque niveau de "Laboratoire", et un même nombre de réplications ($n = 2$). Nous considérons dans la suite de ce document que les deux facteurs sont fixes.

3 Tests des effets fixes Laboratoire et Flacon

Pour un modèle à facteurs fixes, les tests d'hypothèses s'écrivent de la façon suivante :

- Effet Laboratoire: $H_0 : \forall\, i/\ 1 \leq i \leq a\ ,\ \tau_i = 0$ contre H_1: $\exists\, i\ /\ \tau_i \neq 0$
- Effet Flacon: $H_0 : \forall (i,j)\ / 1 \leq i \leq a, 1 \leq j \leq b,\ \beta_{j(i)} = 0$ contre H_1: $\exists (i,j)\ / \beta_{j(i)} \neq 0$

Lorsque l'on travaille sur données gaussiennes, on considère les erreurs $\epsilon_{(ij)k}$ identiquement distribuées suivant une loi normale $N(0, \sigma^2)$. Cette hypothèse permet de réaliser les tests des effets Laboratoire et Flacon indépendemment l'un de l'autre, en effectuant une analyse de variance à deux facteurs imbriqués (D.C. Montgomery, 2005). Nous cherchons ici à définir des méthodes de test applicables sur données de comptage. L'hypothèse de distribution normale des erreurs n'est alors pas vérifiée. Nous considérerons dans la suite de ce document que la variable y_{ijk} suit une loi de Poisson $P(\lambda)$.

Notons L et F les propositions *"Il y a un effet Laboratoire"*, et *"Il y a un effet Flacon"*. Pour chaque configuration, $\bar{L}\,\bar{F}$, $L\,\bar{F}$, $\bar{L}\,F$, et $L\,F$, le paramètre λ s'exprime en fonction de i (λ_i), de j (λ_j), de i et de j (λ_{ij}), ou indépendemment de i et de j (λ).

	Test	H_0/H_1	Loi de y_{ijk} sous H_0	Loi de y_{ijk} sous H_1
0	Test d'existence d'un effet quelconque	$\bar{L}\ \bar{F}\ /$ $L\|\|F\|\|LF$	$y_{ijk} \sim P(\lambda)$	$y_{ijk} \sim P(\lambda_{ij})$
1	Test modèle à effet "Labo" contre modèle sans effet	$\bar{L}\bar{F}/L\bar{F}$	$y_{ijk} \sim P(\lambda)$	$y_{ijk} \sim P(\lambda_i)$
2	Test modèle à effet "Flacon" contre modèle sans effet	$\bar{L}\bar{F}/\bar{L}F$	$y_{ijk} \sim P(\lambda)$	$y_{ijk} \sim \frac{1}{2}P(\lambda_1) + \frac{1}{2}P(\lambda_2)$
3	Test modèle à deux effets contre modèle effet "Labo"	$L\bar{F}/LF$	$y_{ijk} \sim P(\lambda_i)$	$y_{ijk} \sim P(\lambda_{ij})$
4	Test modèle à deux effets contre modèle effet"Flacon"	$\bar{L}F\ /LF$	$y_{ijk} \sim \frac{1}{2}P(\lambda_1) + \frac{1}{2}P(\lambda_2)$	$\mathrm{y}_{ijk} \sim P(\lambda_{ij})$

Table 1: Tests des effets Laboratoire et Flacon

Il est donc nécessaire de distinguer les cinq tests de la Table 1.
Pour effectuer ces tests, il nous faut estimer les paramètres λ, λ_i, λ_{ij}, λ_1, et λ_2 .
Le paramètre λ correspond au cas $\bar{L}\,\bar{F}$ (H_0 des tests 0, 1 et 2). Nous considérons alors les mesures y_{ijk} effectuées par l'ensemble des laboratoires indépendantes et identiquement distribuées suivant une loi de Poisson $P(\lambda)$. Nous estimons λ par:

$$\hat{\lambda} = \bar{y}_{...} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{2}\sum_{k=1}^{2} y_{ijk}}{4a} \tag{2}$$

Les estimateurs $\hat{\lambda}_i$ de λ_i (configuration $L\,\bar{F}$, hypothèses H_1 du test 1 et H_0 du test 3) et $\hat{\lambda}_{ij}$ de λ_{ij} (configuration $L\,F$, hypothèses H_1 des tests 0 et 3, H_1 du test 4) sont:

$$\hat{\lambda}_i = \bar{y}_{i..} = \frac{\sum_{j=1}^{2}\sum_{k=1}^{2} y_{ijk}}{4} \tag{3}$$

$$\hat{\lambda}_{ij} = \bar{y}_{ij.} = \frac{\sum_{k=1}^{2} y_{ijk}}{2} \tag{4}$$

Dans le cas $\bar{L}\,F$ (H_1 du test 2 et H_0 du test 4), les flacons 1 et 2 n'étant pas identifiés, la variable y_{ijk} suit une loi mélange de deux lois de poisson de paramètres λ_1 et λ_2, dont nous connaissons les proportions. En effet, 50% des variables y_{ijk} suivent une loi de Poisson $P(\lambda_1)$ et les 50 % restants une loi de Poisson $P(\lambda_2)$. Les paramètres λ_1 et λ_2 sont alors estimés par un algorithme EM (A.P. Dempster, N.M. Laird, D.B. Rubin, 1977).

Pour effectuer l'un des tests de la Table 1, nous formons son rapport de vraisemblance, $\Lambda = \frac{L_{H0}}{L_{H_1}}$, où L_{H0} et L_{H1} correspondent aux vraisemblances du modèle (1) sous les hypothèses H_0 et H_1. Nous calculons ainsi la déviance associée à un test, D=-2 log(Λ).

En fonction du problème, nous pourrons utiliser un ou plusieurs des tests de la Table 1. Dans le cadre de la problématique introduite en section 1, le test 0 permet d'identifier un éventuel effet. Si l'on détecte un effet (rejet de H_0), le test 3 permet alors d'identifier si cet effet est uniquement dû au laboratoire. Lorsque l'on rejette l'hypothèse H_0 du test 3, nous ne pouvons cependant pas identifier si nous sommes en présence d'un effet Flacon seul (modèle $\bar{L}F$) ou d'un effet Flacon et d'un effet Laboratoire (modèle LF). Il nous faudra travailler sur un modèle à facteurs aléatoires pour le déterminer.
Les tests 2 et 4 peuvent quant à eux être utilisés lorsque les laboratoires qui participent à l'essai reçoivent deux flacons présentant des niveaux de concentration différents, ou issus de deux cuves différentes.

4 Test d'existence d'un effet

Nous allons expliciter ici le test 0 de la Table 1. Les hypothèses de ce test sont:

$H_0 : \bar{L}\bar{F}$: $y_{ijk} \sim P(\lambda)$ $\qquad$ $H_1 : L\bar{F}\,||\bar{L}F\,||LF$: $y_{ijk} \sim P(\lambda_{ij})$

Le rapport de vraisemblance nous permet de comparer le modèle LF (H_1), et le modèle $\bar{L}\bar{F}$ (H_0). Les estimateurs de λ sont $\hat{\lambda}$ (expression (2)) sous H_0, et $\hat{\lambda}_{ij}$ (expression (4))

sous H_1. Nous pouvons alors former le rapport de vraisemblance du test 0, puis obtenir l'expression de sa déviance D_0:

$$D_0 = -4 \sum_{i=1}^{a} \sum_{j=1}^{2} \bar{y}_{ij.} \log \left(\frac{\bar{y}_{...}}{\bar{y}_{ij.}} \right) \tag{5}$$

Dans le cadre de notre étude, il peut arriver que les mesures rapportées par un laboratoire sur l'un des flacons pour un essai soient nulles (typiquement lorsque λ est faible). Dans ce cas, $\bar{y}_{ij.} = 0$, et la déviance D_0 du jeu de données est indéfinie. En observant que $\lim_{\bar{y}_{ij.} \to 0} \left(\bar{y}_{ij.} \log(\frac{\bar{y}_{...}}{\bar{y}_{ij.}}) \right) = 0$, l'expression (5) devient:

$$D_0 = -2 \log(\Lambda_0) = -4 \sum_{i=1}^{a} \sum_{\substack{j=1 \\ \forall\, j \,/\, \bar{y}_{ij.} \neq 0}}^{2} \bar{y}_{ij.} \log \left(\frac{\bar{y}_{...}}{\bar{y}_{ij.}} \right) \tag{6}$$

D'après un théorème asymptotique (Lehmann, 1986), sous H_0, la distribution de la déviance tend vers une loi du χ^2 à $2a-1$ degrés de liberté lorsque le nombre de réplications n tend vers l'infini. Etant donné le petit nombre de réplications considérées ici ($n = 2$), le théorème de Lehmann ne peut être appliqué. Cependant, pour de grandes valeurs de λ (typiquement $\lambda > 10$), la loi de Poisson peut être approchée par une loi normale. Or, puisque dans le cas de données gaussiennes la déviance suit une loi du χ^2 quelque que soit n (J.J Droesbeke, M. Lejeune, G. Saporta, 2005), la loi de notre déviance D_0 pourra être approchée par une loi du χ^2 à $2a-1$ degrés de liberté. Pour des valeurs de λ plus faibles, nous proposons d'approcher la loi de la déviance (sous H_0) par simulation.

5 Puissance des tests

Pour évaluer les tests proposés, nous comparons leurs puissances à celle de la méthode utilisée en microbiologie. Nous présentons ici les courbes de puissance du test 0 pour $\lambda = 1$ et $\lambda = 15$. Pour nous placer sous H_1, nous introduisons un effet Laboratoire. Pour cela, nous simulons les résultats des mesures de 14 laboratoires suivant $\mathrm{P}(\lambda)$, et celles d'un laboratoire suivant $\mathrm{P}(\lambda + \delta)$. Nous effectuons 1000 tests sur 1000 jeux de données ainsi simulés, pour différentes valeurs de δ. Pour $\lambda = 1$, nous comparons la puissance du test basé sur des simulations de la loi de la déviance sous H_0 (courbe bleue, 10^4 simulations par p-valeur), avec celle du test d'ANOVA (courbe rouge). Pour $\lambda = 15$, nous étudions les puissances des trois tests suivants : le test où la loi de la déviance est assimilée à une distribution du χ^2 à $2a-1$ degrés de liberté (courbe verte), le test où la loi de la déviance est approchée par simulation (courbe bleue), et le test d'ANOVA (courbe rouge). Sur la *Figure 1*, nous constatons que, pour les deux valeurs de λ considérées, la méthode de test proposée ici est beaucoup plus puissante que la méthode utilisée en microbiologie (ANOVA sur données ayant subi une transformation log-normale). Pour $\lambda = 15$, le test basé sur la loi du χ^2 donne des résultats très proches de ceux obtenus avec le test par simulation. Cela s'explique par le fait que, à $\lambda = 15$, la loi de Poisson peut être approchée par une loi normale. On note qu'ici la méthode d'ANOVA détecte mal les effets.

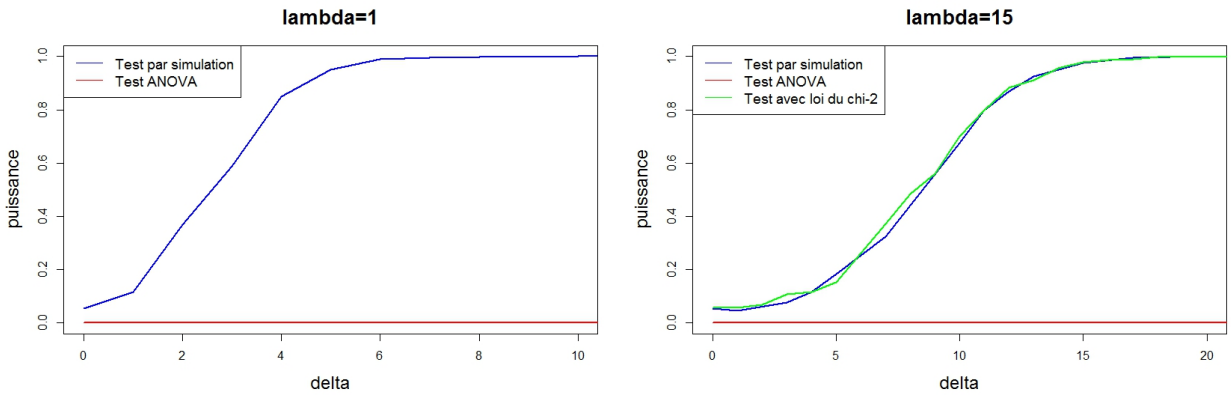

Figure 1: Comparaison des puissances des trois méthodes

6 Conclusion

Nous proposons ici une méthode d'analyse de variance à deux facteurs fixes imbriqués sur données de poisson. Les tests de puissance montrent que cette méthode est plus pertinente que celle qui est utilisée en microbiologie. En pratique, il s'avère parfois que la dispersion des résultats de dénombrements de germes soit supérieure à celle attendue d'après le modèle de poisson (BCR Information, 1993). Il pourrait alors être nécessaire de définir des méthodes de test similaires à partir d'une loi de probabilité adaptée à la surdispersion, telle que la loi binomiale négative.

Bibliographie

[1] BCR Information (1993), *Statistical Analysis of Certification Trials for Microbiological Reference Materials*, Report EUR 15008 EN.
[2] P. McCullagh, J.A. Nelder (1989), *Generalized Linear Models*, pp. 21-43.
[3] C.E. McCulloch, S.R. Searle (2001), *Generalized, Linear, and Mixed Models (Wiley Series in Probability and Statistics)*, pp. 28-68.
[4] A.P. Dempster, N.M. Laird, D B. Rubin (1977), *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Vol. 39, n.1, pp. 1-38.
[5] J.J. Droesbeke, M. Lejeune, G. Saporta (2005), *Modèles statistiques pour données qualitatives*, pp. 83-90.
[6] R.B. O'Hara, D.J Kotze (2010), *Do not log-transform count data*, British Ecological Society, Methods in Ecology and Evolution, vol. 1, pp. 118-122.
[7] ISO/TS 22117 (2010), *Microbiologie des aliments - Exigences spécifiques et lignes directrices pour les essais d'aptitude par comparaison interlaboratoires.*
[8] E.L. Lehmann, J.P. Romano (2005), *Testing Statistical Hypotheses*, pp. 504-517.
[9] D.C. Montgomery (2005), *Design and Analysis of Experiments*, pp. 60-118, pp. 525-536.