Numéro d'ordre : 41984

École Doctorale Sciences Pour l'Ingénieur, Université de Lille 1

THÈSE DE DOCTORAT

Présentée par

Florence LOINGEVILLE

en vue de l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES DE LILLE 1

Discipline : MATHÉMATIQUES APPLIQUÉES
Spécialité : STATISTIQUE

Modèle linéaire généralisé hiérarchique Gamma-Poisson pour le contrôle de qualité en microbiologie

soutenance prévue pour le 22 Janvier 2016 à 14h devant le jury composé de :

Rapporteurs :	Denis ENACHESCU	Professeur, *University of Bucarest*
	Ali GANNOUN	Professeur, *Université de Montpellier 2*
Examinateurs :	Filipe MARQUES	Professeur, *University of Lisbon*
	Vincent VANDEWALLE	Professeur, *Université de Lille 2*
Directeurs de thèses :	Julien JACQUES	Professeur, *Université de Lyon 2*
	Cristian PREDA	Professeur, *Université de Lille 1*
Invités :	Philippe GUARINI	Directeur, *AGLAE*
	Olivier MOLINIER	Responsable d'exploitation, *AGLAE*

Remerciements

Je voudrais tout d'abord exprimer mes sincères remerciements à Cristian Preda, pour avoir accepté de co-diriger cette thèse, et pour s'être toujours montré disponible tout au long de ces trois années. Je le remercie pour ses conseils autant statistiques que personnels, ainsi que pour le soutien qu'il m'a apporté dans les moments difficiles de cette thèse.

Je souhaiterais également remercier Julien Jacques pour m'avoir permis de réaliser cette thèse, en m'introduisant à l'association AGLAE, me permettant ainsi d'emprunter la voie que je voulais. Je lui suis reconnaissante d'être toujours resté disponible, même depuis Lyon.

Je tiens à remercier les professeurs Denis Enaschescu et Ali Gannoun pour avoir accepté d'être les rapporteurs de ma thèse. Je leur suis reconnaissante de l'intérêt qu'ils ont porté à mon travail. Je remercie chaleureusement Vincent Vandewalle pour avoir accepté de faire partie du jury de ma soutenance.

I would like to thank Filipe Marques for his support and the advice he provides me during my PhD thesis. Our exchanges about the product of independant Gamma random variables have made significant contribution to this work. I am happy and honored that he accepted to take part in the committee of my PhD defense.

Je remercie l'association AGLAE, pour m'avoir fait confiance en me confiant ce sujet de thèse, et notamment Philippe Guarini et Olivier Molinier pour leur investissement dans ce projet.

Je remercie l'équipe Modal de l'Inria de Lille, et notamment son responsable Christophe Biernacki, pour m'avoir accueillie pendant ces trois années. J'ai eu plaisir à travailler dans cet environnement très agréable et propice à la réflexion scientifique.

Merci aux doctorants du laboratoire Paul Painlevé, notamment à Roberto, Najib, Antoine, avec qui j'ai partagé de sympathiques déjeuners et cafés tout au long de ma thèse.

Je remercie aussi l'équipe d'AGLAE pour la bonne ambiance que je retrouvais chaque fin de semaine.

Je tiens aussi à remercier toutes les personnes qui m'ont soutenue et ont rendu ces trois années plus agréables.

Merci à mes amis docteurs et futurs docteurs de la BdM : Yann, Gaetan, Thomas, Alain, et tous les autres, dont la plupart ont partagé la même aventure que moi pendant ces trois ans ! Les bons moments passés avec vous ont égayé ces trois années, et m'ont permis de retrouver le

sourire dans les moments difficiles.

Je remercie tout particulièrement Céline pour son amitié et son soutien tout au long de ces années de thèse. Nos mcdos hebdomadaires, nos séances de zumba, et nos soirées cocktails m'ont aidé à me changer les idées et à évacuer le stress !

Un grand merci à toutes les personnes qui m'ont soutenue, de près ou de loin. Je pense à Alex, Maxime, mais aussi à mes amis de la prépa Wallon, Amélie, Anne-Claire, Tatiana, Robin, Cyril, Pierre, pour les bons moments partagés au cours de ma thèse. Merci aussi à Caroline pour les sorties qui m'ont permis d'oublier un temps le monde de la thèse !

Je remercie aussi toutes les personnes avec qui j'ai eu l'occasion d'échanger, notamment dans le cadre de conférences ou de l'ASPID. Merci notamment à Matthieu, Zakaria, Carmelo, Guillaume, Clément, etc, avec qui j'ai eu plaisir à collaborer dans le cadre de l'association.

Je tiens à remercier ma famille, et notamment ma grande soeur Aude pour ses encouragements.

J'adresse mes sincères remerciements à toutes les personnes, déjà citées ici, qui ont pris la peine de relire ce manuscrit, ou du moins des parties ! Je remercie aussi les personnes qui m'ont donné des conseils pour la rédaction d'articles, la préparation de conférences, et la rédaction de la thèse.

Enfin, je réserve le plus grand des mercis à mes parents, qui m'ont toujours encouragée, et ont fait en sorte que je réalise cette thèse dans les meilleures conditions. Leur soutien indéfectible a largement contribué à l'aboutissement de ce travail.

Table des matières

Liste des tableaux

Table des figures

Notations

Chapitres 1 et 2

K :	Facteur de surdispersion,
m :	Valeur de consensus (moyenne globale),
R :	Limite de reproductibilité,
r :	Limite de répétabilité,
S_r :	Écart-type de répétabilité,
S_R :	Écart-type de reproductibilité,
S_{rk} :	Écart-type de répétabilité du laboratoire k,
S_u :	Écart-type d'hétérogénéité de lot,
S_z :	Écart-type pour l'évaluation de l'aptitude (écart-type de dispersion des résultats moyens des laboratoires),
S_L :	Écart-type de la composante Laboratoire du biais,
CV_r :	Coefficient de variation de répétabilité,
CV_R :	Coefficient de variation de reproductibilité,
CV_u :	Coefficient de variation d'hétérogénéité de lot,
a :	Nombre de laboratoires participant à un essai,
b :	Nombre de flacon par laboratoire,
n :	Nombre de réplications par flacon,
μ :	Valeur vraie du paramètre,
α_i :	Erreur due au laboratoire i,
$\beta_{j(i)}$:	Erreur due au flacon j du laboratoire i,
$\epsilon_{ij(k)}$:	Erreur de mesure,
σ_L^2 :	Variance de l'erreur interlaboratoires,
σ_u^2 :	Variance due à l'hétérogénéité de lot,

σ_r^2 : Variance due à l'erreur de mesure,

$\bar{y}_{...}$: Moyenne globale des dénombrements sur tous les laboratoires,

$\bar{y}_{i..}$: Moyenne des résultats du laboratoire i,

$\bar{y}_{ij.}$: Moyenne des résultats du flacon j du laboratoire i,

SC_T : Somme des carrés totale des écarts,

SC_L : Somme des carrés des écarts due au facteur Laboratoire,

$SC_{F(L)}$: Somme des carrés des écarts due au facteur Flacon,

SC_E : Somme des carrés des écarts de l'erreur,

CM_L : Carrés moyens du facteur Laboratoire,

$CM_{F(L)}$: Carrés moyens du facteur Flacon,

CM_E : Carrés moyens de l'erreur,

z_k : Score du laboratoire k dans une échelle normale centrée réduite,

$\mathbb{E}(X)$: Espérance de la variable aléatoire X,

$\mathbb{V}(X)$: Variance de la variable aléatoire X,

λ_{ijk} : Espérance du processus de Poisson modélisant la variabilité des résultats de dénombrements en microbiologie.

Chapitres 3 et 4

EIL : Essai Interlaboratoires,

λ_{ijk} : Espérance du processus de Poisson modélisant la variabilité des résultats de dénombrements en microbiologie,

$\mathbb{E}(X)$: Espérance de la variable aléatoire X,

$\mathbb{V}(X)$: Variance de la variable aléatoire X,

y_{ijk} : Résultat de la mesure effectuée par le laboratoire i sur la réplication k du flacon j,

$\exp(\alpha_i)$: Effet du laboratoire i,

$\exp(\beta_{ij})$: Effet du flacon j du laboratoire i,

$\exp(\gamma_{ijk})$: Effet de la $k^{\text{ième}}$ réplication du flacon j par le laboratoire k,

μ : Moyenne globale des dénombrements bactériens,

X : Matrice de design des effets fixes,

Z : Matrice de design de l'effet aléatoire Laboratoire,

F :	Matrice de design de l'effet aléatoire Flacon,
R :	Matrice de design de l'effet aléatoire Réplication,
$\tilde{Z}$:	Matrice de design de l'ensemble des effets aléatoires (Laboratoire, Flacon, réplication),
u_1^2 :	Coefficient de dispersion associé à l'effet Laboratoire,
u_2^2 :	Coefficient de dispersion associé à l'effet Flacon,
u_3^3 :	Coefficient de dispersion associé à l'effet Réplication,
ϕ :	Coefficient de surdispersion du modèle,
η :	Prédicteur linéaire,
Σ :	Matrice diagonale de la variance du vecteur de travail,
D :	Matrice d'information de Fisher des effets aléatoires (Laboratoire, Flacon, Réplication),
D_1^{-1} :	Matrice d'information de Fisher de l'effet Laboratoire,
D_2^{-1} :	Matrice d'information de Fisher de l'effet Flacon,
D_3^{-1} :	Matrice d'information de Fisher de l'effet Réplication,
z_k :	Score du laboratoire k dans une échelle normale centrée réduite,
$\tilde{\lambda}$:	Paramètre d'intensité d'une distribution Gamma généralisée,
r :	Paramètre de forme d'une distribution Gamma généralisée,
p :	Paramètre de profondeur d'une distribution Gamma généralisée entière,
$\tilde{\theta}$:	Paramètre de décalage d'une distribution Gamma généralisée décalée.

Introduction générale

Le travail présenté dans cette thèse est issu d'une problématique posée par l'Association AGLAE. AGLAE, *Association Générale des Laboratoires d'Analyses et d'Essais*, a pour but de contribuer à l'amélioration des analyses, notamment chimiques, microbiologiques et biologiques, dans les domaines de l'environnement et de la biologie médicale. AGLAE est accréditée par le *Comité Français d'Accréditation* (*Cofrac*) pour des activités d'organisateurs de comparaisons interlaboratoires. Cette thèse s'inscrit dans le cadre d'une collaboration avec Inria (centre Inria Lille - Nord Europe) et plus spécifiquement l'équipe-projet MODAL, axée sur la conception de modèles probabilistes génératifs pour les données complexes. Le thème principal de la thèse est la mise en place d'outils statistiques spécifiques au contrôle de qualité en analyse microbiologique.

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. Après avoir mis les pratiques analytiques sous contrôle par l'utilisation de méthodes normalisées et la validation des rendements de leurs milieux de culture, les laboratoires doivent, dans le cadre d'une accréditation, procéder à des contrôles de qualité interne et externe quantitatifs, afin d'assurer le suivi de leur procédure analytique. 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. Dans le domaine qui concerne AGLAE, les essais sont principalement des mesures physiques, chimiques (dosages) ou biologiques (dénombrements et mesures de doses toxiques). L'objectif principal assigné à ces essais est l'évaluation et l'amélioration des performances analytiques des laboratoires [ISO, 2010a]. En ce sens, ils constituent un moyen essentiel du système d'assurance qualité des laboratoires d'analyses, répondant aux exigences de [AFNOR, 2000].

A partir des résultats récoltés lors d'un essai, AGLAE a pour mission d'évaluer trois critères de contrôle de qualité : la capacité d'un laboratoire à répéter ses analyses, l'hétérogénéité des préparations envoyées aux laboratoires, et la justesse des mesures 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-laboratoires. Cette variance intra-laboratoire peut ensuite elle-même être décomposée en une variance inter-échantillons et une variance intra-échantillons. L'objectif est alors d'évaluer la significativité des différences entre les mesures sur la base de cette décomposition de la variance totale. Dans le cas de mesures quantitatives continues, l'hypothèse de distribution gaussienne permet de réaliser des tests statistiques évaluant la significativité des différences de mesures entre laboratoires, entre échantillons, et entre réplications de l'échantillon. L'un des objectifs de la thèse est de définir une méthodologie d'analyse de variance similaire lorsque les mesures ne sont plus des variables aléatoires quantitatives continues mais des variables aléatoires quantitatives discrètes, comme c'est le cas lors des analyses microbiologiques, où les mesures sont des dénombrements bactériens.

La maîtrise statistique de l'erreur commise lors d'un dénombrement effectué sur un échantillon introduit la notion de surdispersion. Théoriquement, si l'on répète un nombre de fois suffisant le même dénombrement bactérien, on devrait observer une répartition des résultats conforme à une variable aléatoire de Poisson. En pratique, il n'en est pourtant pas toujours ainsi. Il arrive souvent que la dispersion observée soit plus élevée que celle attendue d'après la loi de Poisson, à cause de la non-vérification de l'hypothèse d'homogénéité du milieu échantillonné [Tillett and Lightfoot, 1995], ainsi que de la variabilité des conditions de réalisation du dénombrement indirect des volumes élémentaires contenant l'analyte recherché. Ce dénombrement indirect engendre en effet une certaine incertitude sur le plan métrologique (volumétrie, température). Mais surtout, le rendement de récupération des germes cibles n'est jamais de 100% et l'incertitude générée par la procédure analytique est souvent loin d'être négligeable [Niemelä, 2002]. La question de la représentation mathématique de cette réalité a fait l'objet de mûres réflexions quant aux variables aléatoires à utiliser [Jarvis, 1989]. Le modèle le plus adapté demeure la loi binomiale négative [Johnson and Nier, 1953], [BCR, 1993], [McCullagh and Nelder, 1989]. Cette distribution est idéalement utilisée pour modéliser la variation excédentaire (surdispersion) du caractère aléatoire de la loi de Poisson [ISO, 2000] et [El-Shaarawi et al., 1981]. L'un des objectifs de cette thèse est de prendre en compte la notion de surdispersion dans la méthode développée.

Ce manuscrit est divisé en 4 chapitres :

Le **Chapitre 1** présente le contexte et la problématique de la thèse. Pour cela, nous présentons dans un premier temps l'activité de l'association AGLAE, ainsi que ses objectifs. Nous détaillons, dans une deuxième partie, la méthode utilisée par AGLAE pour mener à bien son activité dans le domaine de la microbiologie. Nous soulignons finalement la nécessité du développement d'une nouvelle méthode, en exposant les inconvénients et les limites de la méthode utilisée par AGLAE.

Après avoir exposé les problèmes rencontrés lors de l'application d'une transformation aux données, nous proposons, dans le **Chapitre 2**, une méthode semblable à l'analyse de la variance, applicable directement sur des données de comptage, telles que celles issues de la microbiologie. Nous présentons la méthode d'analyse de déviance, et la formulons pour un modèle linéaire à deux facteurs fixes imbriqués. Nous exposons ensuite les tests de significativité des effets basés sur la déviance, ainsi que la puissance de ces tests. Nous concluons en présentant les apports de cette méthode par rapport à celle qui est actuellement utilisée, ainsi que ses limites.

Après avoir souligné les limites d'un modèle à facteurs fixes au chapitre 2, nous proposons dans le **Chapitre 3** un modèle linéaire généralisé hiérarchique pour pour expliquer la variabilité des résultats de dénombrement d'un essai interlaboratoires en microbiologie. Dans un premier temps, nous présentons les différents modèles développés au fil du temps, ayant permis d'aboutir à l'apparition des modèles linéaires généralisés hiérarchiques. Nous formulons ensuite le problème de la surdispersion en microbiologie en des termes statistiques. Nous pouvons alors, dans une troisième partie, aboutir à la formulation d'un modèle linéaire généralisé hiérarchique Gamma-Poisson à trois facteurs aléatoires, pour modéliser la variabilité des résultats de dénombrement en microbiologie. Nous présentons la méthode utilisée pour l'estimation des paramètres. Nous étudions l'adéquation du modèle proposé aux données réelles d'AGLAE. Nous proposons des tests de significativité des effets aléatoires. Nous proposons finalement une nouvelle méthode de scoring pour évaluer les laboratoires qui participent à un essai.

Dans le **Chapitre 4**, nous nous intéressons à la distribution de l'intensité du modèle de Poisson proposé au chapitre 3. Pour ce faire, nous étudions le produit de variables aléatoires indépendantes de lois Gamma. Comme la caractérisation d'une telle variable aléatoire est complexe, nous proposons ici une distribution presque exacte du produit de variables aléatoires indépendantes de lois gamma généralisées. Nous présentons des applications pratiques de l'utilisation de cette distribution presque exacte dans le cadre de notre problématique.

Une conclusion générale ainsi que des perspectives viennent terminer ce manuscrit.

Chapitre 1

Contexte et Objectifs

Sommaire

1.1 Introduction

Dans le cadre d'une accréditation, les laboratoires doivent vérifier, améliorer et maintenir la qualité de leurs analyses physico-chimiques [AFNOR, 2000] et biologiques [ISO, 2012a]. AGLAE est une association dont l'objectif est de contribuer à l'amélioration des analyses, notamment chimiques, microbiologiques et biologiques, dans les domaines de l'environnement et de la biologie médicale. Pour cela, AGLAE propose une large gamme d'essais interlaboratoires. Dans le domaine de l'environnement, ces essais portent sur des eaux destinées à la consommation humaine, des eaux de surface continentales et marines, des eaux résiduaires, des boues de station d'épuration, des sédiments, ou encore des sols pollués. Dans le domaine médical, les essais portent sur la cytobactériologie des urines, la bactériologie des selles et du sang, la quantification des endotoxines bactériennes sur eaux à usage médical, ou encore la microbiologie des eaux à usage médical.

Dans ce chapitre, nous introduisons le contexte de cette thèse, ainsi que ses enjeux. Nous présentons ensuite la méthode actuellement utilisée à AGLAE, puis mettons en évidence ses limites.

Dans un premier temps, nous introduisons l'activité d'AGLAE, basée sur les essais interlaboratoires (section 1.2), ainsi que les notions de dénombrement en microbiologie, et de surdispersion (section 1.3). Dans les sections 1.4 à 1.6, nous présentons la méthode actuellement mise en place à AGLAE pour le contrôle de qualité en microbiologie. Nous présentons ensuite les méthodes de calcul des incertitudes de mesures (section 1.7) et d'évaluation de la performance analytique des laboratoires (section 1.8). Enfin, en section 1.9, nous pointons les limites de la méthode utilisée par AGLAE.

1.2 Les essais interlaboratoires

1.2.1 Définition et objectifs

Pour mettre en place un contrôle externe de qualité, l'outil utilisé par AGLAE est la comparaison interlaboratoires. Toute comparaison interlaboratoires se planifie, s'exécute et s'exploite comme un plan d'expérience. Sa finalité détermine donc la mise en œuvre de chacune des étapes à réaliser. Les essais interlaboratoires constituent un moyen essentiel du système d'assurance qualité des laboratoires d'analyses [AFNOR, 2000], [Lawrence et al., 2013].

L'organisation optimale des essais interlaboratoires peut être résumée par le plan d'expérience suivant (Figure 1.1) : 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, A_1, A_2 et B_1, B_2, par le laboratoire, qui réalise ensuite dans des conditions de répétabilité les mesures demandées sur les quatre réplications dont il dispose.

Suivant la nature de l'analyte, la mesure pourra être un dosage (analyse chimique), un dénombrement bactérien (analyse microbiologique), ou encore l'évaluation d'une dose inhibitrice pour différentes dilutions du matériau (analyse écotoxicologique).

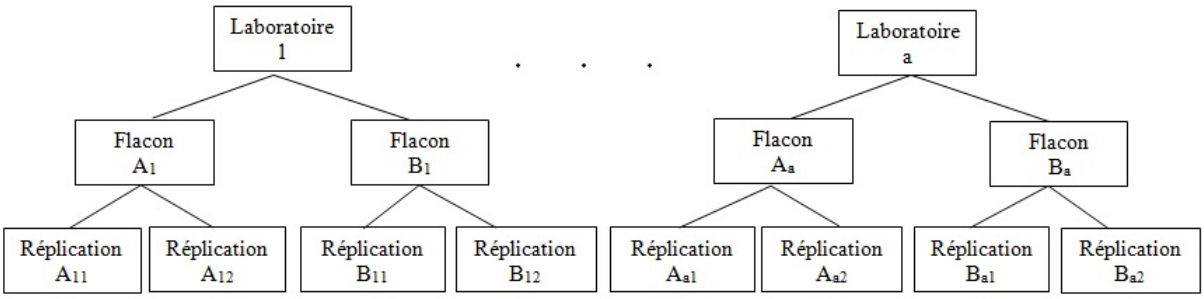

FIGURE 1.1 – Schéma de l'organisation optimale des essais interlaboratoires

A partir des résultats de mesures qu'elle reçoit, AGLAE a pour mission d'évaluer trois critères de qualité :

- la capacité d'un laboratoire à répéter ses analyses (répétabilité) : les écarts de mesure entre les réplications A_1 et A_2 puis B_1 et B_2 sont-ils acceptables ?
- l'hétérogénéité des préparations envoyées aux laboratoires : les écarts de mesure entre les échantillons A et B sont-ils acceptables ?
- la justesse des mesures des laboratoires : les mesures de chaque laboratoire sont-elles cohérentes avec la valeur vraie du matériau ?

Précisons que, pour ce dernier point, étant donné la nature environnementale ou biologique des analyses et la nature des matériaux analysés, la quantité à mesurer ne peut pas toujours être définie précisément à l'avance. La justesse est alors évaluée en comparant les résultats du laboratoire considéré avec les résultats de l'ensemble des autres laboratoires ; ceci en partant du principe qu'une mesure sera satisfaisante si elle n'est pas atypique vis-à-vis de la distribution des résultats de l'ensemble des laboratoires.

1.2.2 Conception et organisation

Les essais interlaboratoires mis en œuvre par AGLAE sont conçus et organisés, sur le plan technique, selon les deux règles suivantes :

Travail sur des matériaux proches de ceux analysés en routine. Dans la plupart des cas, les matériaux sont des échantillons prêts à l'emploi préparés à partir de matrices naturelles. Les participants reçoivent alors des échantillons qu'ils introduisent dès réception dans leur système analytique. Des échantillons à reconstituer sont également employés afin d'améliorer la stabilité de certains paramètres. Il s'agit d'échantillons qui doivent subir une phase de préparation avant introduction dans le système analytique.

Maîtrise optimale de la qualité des échantillons envoyés. L'efficacité du système d'essais d'aptitude est directement liée à la qualité des échantillons mis à disposition pour analyse. Il est donc absolument nécessaire de maîtriser cette qualité, aussi bien en termes de stabilité qu'en termes d'homogénéité. Les points suivants permettent une gestion optimale de la qualité des matériaux envoyés :

- La mise en place d'un processus de fabrication approprié.
  Les principes de fabrication des échantillons utilisés par AGLAE permettent de conduire à la préparation de matériaux stables et homogènes. Une fois ces matériaux préparés, ils sont envoyés aux participants le même jour en livraison express.

- La définition d'une période raisonnable de début de traitement des échantillons.
  Même si les procédures de fabrication des matériaux permettent d'assurer *a priori* que les lots d'échantillons envoyés aux participants sont suffisamment stables et homogènes, c'est finalement la qualité des échantillons au moment où les laboratoires réalisent leurs analyses qui est importante. C'est pourquoi des contrôles des lots d'échantillons envoyés sont régulièrement effectués par des laboratoires tiers mandatés pour cette tâche. Plus précisément, AGLAE a défini, pour chaque paramètre, une période durant laquelle la qualité des échantillons est garantie optimale et sous contrôle. C'est pendant cette période que ces contrôles sont effectués, période qui est indiquée aux participants comme la *période raisonnable de début de traitement* des échantillons.

- La mesure systématique de la qualité des lots préparés.
  La mesure de la qualité des lots d'échantillons est essentiellement effectuée par l'examen statistique des données produites par les participants. Les essais d'AGLAE sont organisés en conséquence, avec l'envoi de plusieurs échantillons appartenant à un même lot et avec la collecte des dates de début de traitement des échantillons auprès des participants. La réalisation de mesures répétées sur plusieurs échantillons du lot permet une évaluation directe du degré d'homogénéité de celui-ci. Lorsque la maîtrise du matériau est suffisamment bonne (risque d'hétérogénéité réduit), deux échantillons appartenant à des lots différents peuvent être envoyés. Le degré d'homogénéité de lot n'est alors plus évalué directement. Seule une image globale de son interférence sur l'objectif de l'essai reste accessible (cette image suffit généralement à détecter les éventuels défauts). Des impossibilités techniques peuvent parfois conduire à ne pas pouvoir envoyer deux échantillons appartenant à un même lot, alors que le risque d'hétérogénéité existe. Les deux lots envoyés parallèlement devront alors obligatoirement être préparés de façon identique.
  L'examen statistique de la répartition des résultats produits par les participants en fonction de la date (et éventuellement de l'heure) de début de traitement des échantillons permet une évaluation directe de la stabilité globale du lot d'échantillons.

1.2.3 Évaluation de l'exactitude

Le rôle premier des essais interlaboratoires est de quantifier les performances analytiques des laboratoires. Mais leur rôle est aussi de quantifier la précision de l'analyse réalisée par la profession. La précision d'une analyse s'exprime en terme d'*exactitude* ([ISO, 1994a]), terme qui recouvre en réalité deux notions bien distinctes :

- la notion de *fidélité*, qui correspond à l'étroitesse de l'accord de mesures répétées,

- la notion de *justesse*, qui correspond à l'écart à la valeur vraie (le biais analytique).

Connaître la *valeur vraie* d'un paramètre sur un matériau environnemental n'est pas toujours possible. La difficulté est encore plus grande lorsqu'il s'agit d'un matériau d'origine biologique qui fait l'objet d'un essai quantitatif reposant sur la mise en culture. Le niveau du paramètre considéré comme référence est généralement la *valeur de consensus* issue des données des participants ayant réalisé l'analyse durant la période raisonnable de début de traitement des échantillons. En pratique, il s'agit de la moyenne des données produites dans la période raisonnable de début de

traitement des échantillons constituant une population homogène. L'écart-type pour l'évaluation de l'aptitude définit l'échelle de variation acceptable parmi les laboratoires, pour chaque essai particulier. Cet écart-type est également calculé à partir des résultats des laboratoires ayant débuté leur analyse dans la période recommandée.

Accéder à la fidélité de l'analyse est, en revanche, toujours possible. Aussi, étant donné que la reproductibilité et l'impact de la qualité des matériaux sur l'objectif de l'essai ne peuvent être jugés objectivement que par rapport à la répétabilité, il appartient à tout organisme d'essais interlaboratoires de mettre en œuvre une mesure de cette borne inférieure de la fidélité, dès que cela est possible. Sauf exception justifiée par des contraintes particulières, il est demandé aux laboratoires de procéder à des mesures répétées.

1.3 Le dénombrement en microbiologie

La nature des modèles mathématiques employés pour exploiter les résultats d'une étude interlaboratoires est un élément déterminant qui se doit d'être renseigné et justifié [ISO, 2005]. Pour travailler à bon escient, il faut donc comprendre et maitriser la variabilité du résultat de mesure, en l'occurrence le comptage de germes bactériens pour culture sur milieu solide. Le présent paragraphe détaille la conceptualisation que se fait le biostatisticien de la mesure faite par le microbiologiste.

1.3.1 Notion d'analyse quantitative

Pour le microbiologiste, procéder à une analyse quantitative correspond à dénombrer les germes présents ou potentiellement présents dans un matériau. Pour cela, il procède à une succession d'opérations que l'on peut schématiquement résumer de la façon suivante :

1. réalisation d'un prélèvement sur le matériau,
2. ensemencement sur un milieu de culture solide (gélose),
3. réalisation de la mesure proprement dite (Figure 1.2) : comptage des colonies qui se sont développées sur la gélose (une colonie correspondant à une bactérie présente dans le prélèvement),
4. calcul du résultat,
5. interprétation des résultats,
6. formulation d'un avis, voire prise de décision.

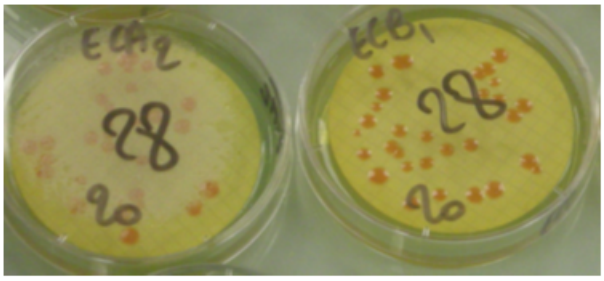

FIGURE 1.2 – Dénombrement bactérien sur boîte de Petri en microbiologie

En revanche, d'un point de vue biostatistique, la démarche d'analyse quantitative n'a pas la même signification. En effet, une conceptualisation microscopique de l'analyse consiste à "discrétiser" le matériau d'essai en le décomposant en une somme de volumes élémentaires de la taille de l'analyte. Le processus devient alors binaire : le volume élémentaire est occupé par l'analyte, ou bien ne l'est pas. Cette autre représentation de la démarche analytique quantitative amène la personne chargée de son exploitation à différencier sa démarche : réaliser un échantillonnage (c'est-à-dire prélever un certain nombre de volumes élémentaires), puis comptabiliser les volumes élémentaires contenant l'analyte.

Maîtrise statistique de la dispersion liée à l'échantillonnage. Lorsque l'on sélectionne au hasard n individus parmi une population en contenant N, dont T possèdent un caractère particulier, que l'on appellera *individus particuliers*, la proportion d'individus particuliers sélectionnés n'est pas forcément représentative de celle de la population. Si k est le nombre d'individus particuliers sélectionnés, le rapport k/n est susceptible de varier d'une sélection à l'autre. Un simple calcul de probabilité nous permet de déterminer dans quelle proportion. La probabilité d'observer k individus particuliers parmi les n éléments prélevés est :

$$P(X = k) = \frac{\binom{N\pi}{k}\binom{N-N\pi}{n-k}}{\binom{N}{n}},$$

où

- $\pi = k/N$ est la proportion d'individus particuliers dans la population,
- $\binom{N}{n}$ est le nombre d'échantillons qu'il est possible de créer,
- $\binom{N\pi}{k}$ est le nombre de groupes de t individus particuliers que l'on peut constituer,
- $\binom{N-N\pi}{n-k}$ est le nombre de groupes de $(n-k)$ individus non-particuliers que l'on peut constituer.

La dispersion des valeurs possibles pour k suit alors une variable aléatoire hypergéométrique $H(N, n, \pi)$. Ainsi, toute estimation de π est entachée d'une incertitude directement liée au fait qu'il faille prélever un échantillon, forcément de taille limitée. Cet échantillonnage, processus purement aléatoire, est donc la source d'une dispersion incompressible des estimateurs possibles de π.

Transposition de cette approche à l'analyse quantitative. L'échantillonnage des volumes élémentaires introduits précédemment relève de la même démarche. L'estimation du pourcentage de volumes élémentaires contenant l'analyte est donc inévitablement perturbé par cette opération. La seule différence réside dans la taille de ce qui est échantillonné. En effet, si une population déjà conséquente contient plusieurs milliers d'individus, un millilitre d'eau regroupe pas moins de plusieurs dizaines de millions de volumes élémentaires de la taille d'une bactérie. Or, on peut démontrer que lorsque $N \to \infty$, $H(N, n, \pi)$ converge vers une variable aléatoire binomiale $B(n, \pi)$ [Saporta, 2011], dont la loi de probabilité est :

$$P(X = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}.$$

En pratique, on peut raisonnablement remplacer la loi hypergéométrique par la loi binomiale dès lors que le taux de sondage n/N est inférieur à 10%. La simplification peut être poursuivie avec le

théorème de Polya sur l'uniformité de la convergence en loi d'une fonction de répartition continue. Ce théorème permet d'établir qu'une variable aléatoire de distribution binomiale $B(n, \pi)$ converge en loi vers une variable aléatoire de Poisson $P(\lambda)$, où $\lambda = n\pi$, lorsque $n \to \infty$ et $\pi \to 0$. La loi de Poisson peut raisonnablement être employée dès que $\pi < 0.1$ et $n > 50$. Ajoutons finalement que $P(\lambda)$ converge en loi vers la loi normale $N(\lambda, \sqrt{\lambda})$ lorsque $\lambda \to \infty$ [Neuilly and CEA, 1998]. Cela se traduit par le fait que les extrémités des bâtons du diagramme de la loi de Poisson sont voisines de la courbe de densité de la loi $N(\lambda, \sqrt{\lambda})$. On obtient donc une valeur approchée de $P(X = k)$ par la surface sous la courbe de densité comprise entre les droites d'abscisse $k - \frac{1}{2}$ et $k + \frac{1}{2}$. Ceci s'exprime par la formule ci-dessous, dite "avec correction de continuité" :

$$P(X = k) = P\left(\frac{k - \frac{1}{2} - \lambda}{\sqrt{\lambda}} < U < \frac{k + \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right).$$

Il est d'usage en pratique de substituer $N(\lambda, \sqrt{\lambda})$ à $P(\lambda)$ de cette façon dès lors que $\lambda > 18$. Lorsque λ devient vraiment élevé, il peut être considéré que $k - \frac{1}{2} - \lambda \approx k + \frac{1}{2} - \lambda$. La formule avec correction de continuité peut alors être simplifiée. La nature statistique de la dispersion liée à l'échantillonnage varie donc d'un processus analytique à l'autre. Les éléments déterminant la variable aléatoire adaptée sont le taux de sondage (n/N), la proportion de volumes élémentaires contenant l'analyte (π) et le nombre moyen de volumes élémentaires contenant l'analyte par prise d'essai $(n\pi)$.

1.3.2 Variables aléatoires utilisées

Précisons tout d'abord que pour les analyses de routine, il existe essentiellement deux techniques différentes de dénombrement en microbiologie :

- les techniques de type *énumératives*, qui reposent sur la mise en culture d'une aliquote de l'échantillon sur milieu solide. Après incubation, on considère que chaque colonie observée correspond à une bactérie.

- les techniques de type *quantiques*, qui reposent sur la mise en culture de dilutions successives de l'échantillon dans un milieu liquide. C'est la méthode du nombre le plus probable, dite "NPP".

Il nous faut évaluer le taux de sondage de la proportion de volumes élémentaires contenant l'analyte, et du nombre moyen de volumes élémentaires contenant l'analyte par prise d'essai, pour déterminer la variable aléatoire adaptée à la situation. C'est l'objet de l'Annexe A, issue de la procédure [AGLAE,] d'AGLAE.

On peut constater en Annexe A que n/N est toujours inférieur à 0.1. On peut donc approcher la loi hypergéométrique par la loi binomiale. De même, π est toujours largement inférieur à 0.1. La loi binomiale peut donc être remplacée par la loi de Poisson. En revanche, les valeurs de $n\pi$ sont variables. Certaines sont inférieures à 18, d'autres sont supérieures à 30. On ne doit donc surtout pas passer systématiquement à la loi normale.
En conclusion, la dispersion liée à l'échantillonnage pour une analyse microbiologique suivrait une loi de Poisson. Ce résultat est communément admis dans la littérature métier [BCR, 1993], [Bliss and Fisher, 1953]. L'usage d'une variable aléatoire normale doit donc être proscrit.

1.3.3 Notion de surdispersion

Introduction de la notion de surdispersion. En pratique, il arrive souvent que la dispersion des résultats de dénombrement de germes soit supérieure à celle attendue d'après le modèle de Poisson comme en témoignent notamment [ISO, 2000] et [Tillett and Lightfoot, 1995].
Il est admis que la loi normale surdispersée reste une loi normale (de moyenne inchangée m et d'écart-type σ' plus élevé que celui attendu σ). En revanche, la loi de Poisson surdispersée n'est plus une loi de Poisson. La question de sa représentation doit donc être posée.
Il existe plusieurs explications au phénomène de surdispersion. Tout d'abord, l'hypothèse d'homogénéité du milieu échantillonné peut ne pas être vérifiée. En effet, l'agglomération sur les parois du contenant (flacon ou pipette) ou sur les particules présentes, ainsi que la formation de chaînes ou d'amas bactériens susceptibles de se disloquer induisent une dérive par rapport à la loi de Poisson.
D'autre part, la variabilité des conditions de réalisation du dénombrement indirect des volumes élémentaires contenant l'analyte recherché constitue une source d'erreur. Il existe en effet toujours une incertitude admise sur le plan métrologique (volumétrie, température). Mais ce sont surtout l'expérience du manipulateur et le rendement de récupération (qui n'est jamais de 100%) qui expliquent cette surdispersion.
La question de la représentation mathématique de cette réalité a été beaucoup étudiée. De nombreux modèles ont fait l'objet de publications. Pour la microbiologie, nous retenons en particulier la loi binomiale négative [BCR, 1993], [El-Shaarawi et al., 1981]. Mais ces modèles sophistiqués ne sont pas les plus usités. Aussi, lorsqu'il s'agit d'effectuer un contrôle statistique de procédé analytique ou une certification de matériaux de référence, une représentation plus simple est généralement préférée. Cette représentation globalise l'effet des sources d'erreur, sous la forme d'un terme de surdispersion par rapport à la loi d'échantillonnage. Il suffit alors simplement de calculer le nombre de fois où il faut multiplier la variance attendue d'après la loi d'échantillonnage. On obtient alors la variance que l'on observe réellement sur la série de donnée. Concrètement, la variance σ^2 est multipliée par ce facteur K, tandis que les autres grandeurs caractéristiques de la loi restent inchangées.

Représentation mathématique de la répartition des données réelles. Pour une loi normale surdispersée, on gagne à transformer le produit $\sigma'^2 = \sigma^2 \cdot K$ en une somme :

$$\sigma'^2 = \sigma^2 \cdot K = \sigma^2 + \sigma''^2.$$

Une variable aléatoire N de loi normale surdispersée $\mathcal{N}(m, \sigma')$ peut alors être perçue comme la somme de 2 variables aléatoires normales, N_1 et N_2 :

$$N = N_1 + N_2,$$

où
$$N_1 \sim \mathcal{N}(m, \sigma) \quad \text{et} \quad N_2 \sim \mathcal{N}(m, \sigma'').$$

N_2 correspond à l'erreur aléatoire générée par l'hétérogénéité du matériau et par la variabilité des conditions de réalisation du dénombrement indirect des volumes élémentaires contenant l'analyte.
Pour la loi de Poisson surdispersée, la situation est en revanche plus complexe puisque la loi de distribution n'est pas complètement définie. Il est d'usage d'utiliser une variable aléatoire log-normale pour représenter une loi de Poisson surdispersée [Jarvis, 1989], [ISO, 2010b] lorsque le nombre de germes par prise d'essai est suffisamment élevé (en pratique supérieur à 12). La transformation logarithmique s'avère effectivement efficace pour stabiliser la variance.
Lorsque le nombre de germes par prise d'essai est faible, il est généralement suffisant d'utiliser

une loi de Poisson standard. En effet, à bas niveau de charge, la surdispersion apparaît réduite et peut généralement être négligée.

Dans les parties 1.4 à 1.8, nous présentons la procédure actuellement utilisée à AGLAE pour le contrôle de qualité en microbiologie. Il s'agit de la méthode qui va être modifiée suite à ce travail de thèse.

1.4 Modalités d'exploitation des données

Dans ce paragraphe, nous allons voir comment est réalisée l'exploitation statistique de données d'essais interlaboratoires, à partir de cette représentation de la variabilité du résultat de mesure défini en section 1.3.

1.4.1 Spécificité des méthodes microbiologiques

Comme souligné en partie 1.3.2, les méthodes statistiques mises en œuvre dans le cadre de l'exploitation des essais microbiologiques sont de nature différente en fonction de la méthode de dénombrement utilisée. Nous distinguons les essais dont les méthodes de dénombrement correspondent à des méthodes énumératives (comptage sur boîte de Petri) des essais dont les méthodes sont de type quantique (Nombre le Plus Probable). Pour les méthodes énumératives, le traitement statistique effectué par AGLAE est détaillé dans l'Organigramme C.1 de l'Annexe C, issu de la procédure [AGLAE,], et établi suivant [BCR, 1993]. Pour les méthodes quantiques, il reprend le principe de la détermination du Nombre le Plus Probable (NPP). Le calcul repose sur la maximisation de la fonction de vraisemblance de la variable NPP (k-uplets formés par la juxtaposition des nombres de réponses positives observées pour les k niveaux de dilution successifs) avec résolution de l'équation de vraisemblance selon la méthode de Newton [Maul, 1991], la maximisation de la fonction de vraisemblance n'ayant pas de solution littérale. L'écart-type décrit dans [Haldane, 1939], basé sur la distribution logarithme du NPP, est utilisé pour le calcul des intervalles de confiance. L'organigramme C.2 de l'Annexe C, issu de la procédure [AGLAE,], reprend la démarche mise en place par AGLAE pour ce type de traitement.
Pour la grande majorité des essais de microbiologie faisant appel à des méthodes énumératives, le plan d'essai mis en œuvre correspond à l'envoi de deux flacons appartenant au même lot avec l'analyse de deux répliques sur chaque flacon, dans des conditions de répétabilité. Lorsque la variable aléatoire est "Normalisable" par transformation logarithmique, ce plan permet d'estimer tous les paramètres du modèle en se raccordant à une analyse de la variance, comme présenté dans [ISO, 2006].

1.4.2 Analyse préliminaire des données

En microbiologie, l'analyse des données est menée en trois étapes, présentées dans [BCR, 1993].

1ère étape : Détection de laboratoires à répétabilité suspecte. Dans un premier temps, la cohérence des résultats observés sur un même flacon pour l'ensemble des données de l'essai est vérifiée. Il s'agit de s'assurer que la dispersion des valeurs obtenues sur un même échantillon corresponde à celle attendue d'après le modèle de Poisson. Cela revient à un contrôle de la répétabilité.
Pour ce faire, le test mis en œuvre est un test de Cochran, présenté dans sa version classique (pour variables aléatoires normales) en Annexe B.1, mais adapté au modèle de Poisson. La statistique de test correspondante est notée "T1". Le principe de ce test est basé sur le fait que différents jeux

de données appliqués au modèle de Poisson ne varient pas systématiquement de la même manière. En effet, si les niveaux de contamination sont différents, les variances sont également différentes. Au lieu de comparer directement les variances, il paraît alors plus approprié d'observer le ratio entre les variances et les moyennes.
Ce test de Cochran pour modèle de Poisson est utilisé pour détecter les laboratoires ayant des écarts importants entre deux mesures répétées sur un même flacon. Pour chaque laboratoire $i = \{1 \dots a\}$, la statistique suivante est calculée :

$$T1 = \sum_{j=1}^{b} \sum_{k=1}^{n} \left(\frac{(y_{jk} - \frac{y_{j+}}{n})^2}{\frac{y_{j+}}{n}} \right),$$

où :

- n est le nombre de répliques (usuellement $n = 2$),
- b est le nombre de flacons (habituellement $b = 2$),
- y_{jk} est le $k^{\text{ième}}$ résultat du dénombrement effectué sur le $j^{\text{ième}}$ flacon,
- y_{j+} correspond au dénombrement total de toutes les répliques du $j^{\text{ième}}$ flacon.

Si les mesures répétées sont distribuées suivant une loi de Poisson, alors $T1$ suit une distribution du χ^2 avec $b(n-b)$ degrés de liberté :

$$T1 \sim \chi^2 b(n-b).$$

Une valeur de $T1$ plus grande que la valeur critique du χ^2 avec un risque $\alpha = 0.05$ traduit une mauvaise répétabilité entre les différents dénombrements réalisés à partir d'un même flacon.
D'autre part, il est également nécessaire d'inspecter les valeurs faibles de la statistique de test $T1$, qui peuvent indiquer une répétabilité anormalement bonne, ce qui est également suspect.
Parfois, le rejet de l'hypothèse nulle est causé seulement par quelques observations aberrantes. On peut détecter les observations aberrantes en considérant la quantité :

$$T1_j = \sum_{n=1}^{n} \frac{(y_{jk} - \frac{y_{j+}}{n})^2}{\frac{z_{j+}}{n}}.$$

Cette quantité donne un apport à $T1$ pour chaque flacon $j = \{1 \dots n\}$. Si les flacons suivent la distribution de Poisson alors :

$$T1_j \sim \chi^2(n-b).$$

$2^{\text{ème}}$ étape : Détection de laboratoires aux écarts entre flacons suspects. La cohérence des écarts entre flacons observés sur les différents laboratoires ayant participé à l'essai est ensuite vérifiée pour l'ensemble des données de l'essai. Il s'agit de s'assurer que les observations de l'hétérogénéité du lot de flacons, effectuées par les différents laboratoires, sont compatibles.
Pour ce faire, le test mis en œuvre est un test de Cochran pour modèle de Poisson. La statistique de test correspondante est notée "T2". Le test de Cochran peut également être utilisé pour mesurer les écarts entre flacons à partir de la somme des comptages répétés pour chaque laboratoire :

$$T2 = \sum_{j=1}^{b} \frac{(y_{j+} - \frac{y_{++}}{b})^2}{\frac{y_{++}}{b}}.$$

En supposant le même nombre de répliques par flacon,

- b représente le nombre de flacons pour chaque laboratoire,

- y_{++} représente le nombre total des comptages effectués sur tous les flacons et sur toutes les répliques par le laboratoire considéré.

La statistique de test $T2$ est distribuée suivant une distribution χ^2 à $(b-1)$ degrés de liberté, multipliée par la constante K. Cette constante dénommée *facteur de dispersion* est calculée de la façon suivante :

$$K = \frac{T2}{b-1}.$$

L'estimation de K peut être obtenue par une méthode graphique ou par une méthode plus formelle.

- La **méthode graphique** utilisée afin de détecter les laboratoires aberrants, par rapport à une trop grande dispersion de leurs propres résultats, ou pour identifier une relation possible entre K et le nombre moyen de germes, est la suivante :
  Pour tous les laboratoires, le ratio $\frac{T2}{b-1}$ est calculé, puis tracé en fonction du dénombrement moyen de chaque laboratoire. Si la dispersion est égale pour tous les laboratoires, le graphe obtenu sera une droite. Dans le cas d'une distribution de Poisson, K est proche de 1. Les laboratoires suspects affichent une grande valeur de $\frac{T2}{b-1}$.

- Une **méthode plus formelle**, détaillée ci-dessous, peut également être utilisée.
  Soit T_{hom}/N la même mesure à partir d'un consensus de laboratoires, où N est le nombre de degrés de liberté de T_{hom}.
  Dans notre cas, T_{hom}/N est calculé à partir des observations $T2/b-1$ de l'essai considéré, et la moyenne arithmétique représente la meilleure estimation du consensus qui se dégage de l'ensemble des données.
  Les deux mesures de dispersion sont distribuées suivant un χ^2 multiplié par une constante, le facteur de surdispersion.
  Sur le plan théorique, il est reconnu que, sous l'hypothèse $\mathcal{H}_0$ de facteurs de surdispersion égaux, le quotient F_{hom}, défini ci-dessous, suit une distribution de Fisher avec $(b-1)$ et N degrés de liberté :

$$F_{hom} = \frac{\frac{T2}{b-1}}{\frac{T_{hom}}{N}}.$$

  L'approximation asymptotique de $F^{0,95}_{b-1,N}$ par $\frac{\chi^2_{b-1,0,95}}{b-1}$ est retenue dans le calcul correspondant.
  Nous pouvons alors calculer un seuil critique pour lequel une valeur de $\frac{T2}{b-1}$ devrait être rejetée avec une probabilité de 95% (ou de 99%).
  Soit $\frac{\chi^2_{(b-1,0,95)}}{I-1}$, le 95$^{\text{ème}}$ percentile de la distribution du χ^2 avec $(b-1)$ degrés de liberté.
  Le seuil critique pour rejeter une grande valeur de $\frac{T2}{b-1}$ est donc $\frac{\chi^2(b-1,0,95)}{b-1}\frac{T_{hom}}{N}$, qui peut être tracé sur la graphique représentant $\frac{T2}{b-1}$ avec la moyenne de chaque laboratoire (à la condition que chacun des laboratoires ait le même nombre de flacons).

Si $K \leq 1$, un ajustement au modèle de Poisson est alors effectué (voir annexe C.1). En revanche, pour $K \geq 1$, on effectue une analyse de la variance sur données préalablement log-transformées. Nous présenterons l'analyse de variance mise en œuvre en section 1.5.

3ème étape : Détection de laboratoires suspects par rapport au panel de laboratoires participant à l'essai. Finalement, la cohérence des valeurs moyennes observées par les différents laboratoires est vérifiée pour l'ensemble des données de l'essai. Il s'agit de statuer sur la comparabilité des résultats des différents participants, de pointer les valeurs aberrantes, et d'ajuster les données au modèle postulé.
Pour ce faire, un test de Grubbs et une analyse de la variance sont utilisés, après transformation logarithmique des données. Le principe du test de Grubbs est de comparer les moyennes des dénombrements des différents laboratoires. Ce test est présenté en Annexe B.2. Les comparaisons des valeurs des moyennes des résultats des laboratoires seront effectuées en utilisant une analyse de la variance sur les données logarithmiques des réplications. L'analyse de la variance est détaillée dans la section suivante.

1.5 Analyse de variance sur données normalisées

Le coeur de l'exploitation est l'estimation des paramètres décrivant la dispersion des données observées. Cette exploitation est classiquement menée par analyse de variance sur les données préalablement log-transformées.

1.5.1 Modèle à 2 facteurs aléatoires imbriqués

Le modèle employé est un modèle d'analyse de la variance à deux facteurs aléatoires imbriqués [ISO, 2006]. Pour pouvoir utiliser la méthode d'analyse de la variance, une transformation logarithmique est préalablement appliquée aux données, afin de les normaliser.
Nous considérons un essai interlaboratoires auquel participent a laboratoires. Chaque laboratoire participant à l'essai reçoit b flacons qu'il analyse chacun n fois (n réplications).
Notons y_{ijk} le résultat du dénombrement effectué pour la réplique k du flacon j du laboratoire i. Le modèle est le suivant :

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{(ij)k} \quad \begin{cases} i = 1, ..., a \\ j = 1, ..., b \\ k = 1, ..., n \end{cases} \tag{1.1}$$

où :

- μ est la valeur vraie du paramètre,
- α_i est l'erreur due au laboratoire,
- $\beta_{j(i)}$ est l'erreur due au $j^{\text{ème}}$ flacon du laboratoire i,
- $\epsilon_{(ij)k}$ est l'erreur de mesure.

Dans ce modèle, le facteur Flacon est imbriqué dans le facteur Laboratoire. Aucune interaction entre les deux facteurs n'est considérée. Ces deux facteurs sont aléatoires, normalement distribués, et d'espérances nulles. Les variances des variables aléatoires s'écrivent :

$$\mathbb{V}[\alpha_i] = \sigma_L^2, \qquad \mathbb{V}[\beta_{j(i)}] = \sigma_u^2, \qquad \text{et } \mathbb{V}[\epsilon_{(ij)k}] = \sigma_r^2.$$

Les paramètres à évaluer sont alors :

- La valeur de consensus m,
- La variance σ_L^2 de l'erreur inter laboratoire (α_i),

- La variance σ_u^2 due à l'hétérogénéité de lot ($\beta_{j(i)}$),
- La variance σ_r^2 de l'erreur de mesure (répétabilité ϵ_{ijk}).

Nous pouvons estimer ces paramètres par la méthode d'analyse de la variance, comme présenté ci-dessous. Nous considérons ici le même nombre de mesures répétées pour chaque laboratoire et le même nombre de flacons par réplications.
Suivant le type de plan d'essai mis en œuvre nous pouvons estimer tout ou partie de ces paramètres. Le plan d'essai par défaut correspond à l'envoi de plusieurs flacons appartenant au même lot avec des répliques sur chaque flacon dans des conditions de répétabilité. Ce plan pyramidal permet d'estimer tous les paramètres du modèle linéaire. Toutefois, il n'est pas toujours possible de mettre en place ce type de plan d'essai en raison de contraintes techniques ou économiques. En effet, il n'est pas possible pour certains essais d'envoyer plusieurs flacons appartenant à un même lot. Le plan se résume alors à plusieurs répliques sur un seul flacon. Ce plan ne permet pas d'avoir une estimation de l'hétérogénéité de lot β. L'erreur de mesure due à l'hétérogénéité de lot est alors intégrée dans l'erreur interlaboratoires α.

1.5.2 Vérification des hypothèses

La technique d'analyse de la variance étant une technique paramétrique, sa robustesse peut constituer une limite à son utilisation. Il convient donc de vérifier les deux conditions requises pour sa mise en œuvre :

- l'homogénéité de la variance résiduelle,
- la normalité de la distribution des données pour chaque variable aléatoire considérée.

Pour vérifier ces deux hypothèses, plusieurs tests sont mis en œuvre :

- le test de Cochran, présenté en Annexe B.1, a pour objectif de vérifier l'homogénéité des variances résiduelles d'un bloc de mesures répétées à l'autre,
- les tests de Grubbs et Dixon explicités en Annexes B.2 et B.3, visent à repérer les données singulières (singulièrement élevées ou singulièrement basses, c'est-à-dire potentiellement anormales),
- les tests de Shapiro-Wilk et de Kolmogorov-Smirnov, présentés en Annexe B.4 et B.5, ont pour objet de statuer sur la normalité de la distribution. L'utilisation conjointe des deux tests est intéressante car leur fonctionnement est complémentaire, le test de Shapiro étant davantage basé sur l'asymétrie, et le test de Kolmogorov portant plus sur les écarts à la courbe.

1.5.3 Mise en œuvre de l'analyse de variance

Décomposition de la variabilité totale. Notons $\bar{y}_{...}$ la moyenne globale des dénombrements sur tous les laboratoires :

$$\bar{y}_{...} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}}{abn},$$

$\bar{y}_{i..}$ la moyenne des résultats du laboratoire i :

$$\bar{y}_{i..} = \frac{\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}}{bn},$$

$\bar{y}_{ij.}$ la moyenne des résultats du flacon j du laboratoire i :

$$\bar{y}_{ij.} = \frac{\sum_{k=1}^{n} y_{ijk}}{n}.$$

L'analyse de la variance consiste en la décomposition de la variabilité totale des résultats de dénombrement en plusieurs composantes. La somme des carrés totale corrigée s'écrit :

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left((\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..}) + (y_{ijk} - \bar{y}_{ij.})\right)^2.$$

Cette équation peut s'écrire :

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2. \tag{1.2}$$

L'équation (1.2) indique que la somme des carrés des écarts totale des résultats de mesures peut être décomposée en une somme des carrées des écarts due au facteur Laboratoire, SC_L, une somme des carrés due au facteur Flacon sous les différents niveaux du facteur Laboratoire, $SC_{F(L)}$, et une somme des carrés des écarts due à l'erreur résiduelle, SC_E. Elle peut alors s'écrire sous la forme :

$$SC_T = SC_L + SC_{F(L)} + SC_E,$$

avec

$$SC_L = \frac{1}{bn}\sum_{i=1}^{a} y_{i..}^2 - \frac{y_{...}^2}{abn},$$

$$SC_{F(L)} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij.}^2 - \frac{1}{bn}\sum_{i=1}^{a} y_{i..}^2,$$

$$SC_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{abn}.$$

SC_T a $abn-1$ degrés de liberté, SC_L a $a-1$ degrés de liberté, $SC_{F(L)}$ a $a(b-1)$ degrés de liberté et SC_E a $ab(n-1)$ degrés de liberté. Les erreurs étant normalement distribuées suivant une même loi normale $N(0,\sigma^2)$, on peut alors diviser chaque somme des carrés de l'équation (1.2) par son degré de liberté afin d'obtenir des carrés moyens indépendamment distribués, de sorte que le quotient de deux de ces carrés moyens soit distribué suivant une loi de Fisher.
Les carrés moyens s'écrivent alors :

$$CM_L = \frac{SC_L}{a-1}, \qquad CM_{F(L)} = \frac{SC_{F(L)}}{a(b-1)}, \qquad CM_E = \frac{SC_E}{ab(n-1)}.$$

Tests de significativité. L'analyse de variance nous permet de réaliser les deux tests suivants :

- Test de significativité de la variance interlaboratoires : Pour effectuer ce test, on considère les hypothèses suivantes :

$$\begin{cases} \mathcal{H}_0 : \sigma_L^2 = 0 \\ \mathcal{H}_1 : \sigma_L^2 \neq 0. \end{cases}$$

Le test est effectué grâce à la statistique suivante :

$$F_{L/F(L)} = \frac{CM_L}{CM_{F(L)}}.$$

Si $F_{L/F(L)} > F_{(1-\alpha);(a-1);a(b-1)}$, on peut affirmer avec un risque α que le facteur interlaboratoires a une influence significative sur le résultat de mesure.

- Test de significativité de l'hétérogénéité de lot : Pour effectuer ce test, on considère les hypothèses suivantes :

$$\begin{cases} \mathcal{H}_0 : \sigma_u^2 = 0 \\ \mathcal{H}_1 : \sigma_u^2 \neq 0. \end{cases}$$

Le test est effectué grâce à la statistique suivante :

$$F_{L/E} = \frac{CM_{F(L)}}{CM_E}.$$

Si $F_{L/E} > F_{(1-\alpha);a(b-1);ab(n-1)}$, on peut affirmer avec un risque α que le facteur hétérogénéité du lot a une influence significative sur le résultat de mesure.

Table d'analyse de variance et distinction des cas. La table d'analyse de variance s'écrit :

Source	Somme des carrés	Degré de liberté	Moyenne quadratique	Espérance de la moyenne quadratique	Test de significativité
Interlaboratoires	SC_L	$df_1 = a - 1$	CM_L	$\sigma_r^2 + n\sigma_u^2 + bn\sigma_L^2$	$F_{L/F(L)}$
Interflacons	$SC_{F(L)}$	$df_2 = a(b-1)$	$CM_{F(L)}$	$\sigma_r^2 + n\sigma_u^2$	$F_{F(L)/E}$
Erreur de mesure (résiduelle)	SC_E	$df_3 = ab(n-1)$	CM_E	σ_r^2	
Totale	SC_T	$abn - 1$			

TABLE 1.1 – Table d'ANOVA

Pour estimer les paramètres σ_r^2, σ_u^2, σ_L^2, et σ_z^2, nous distinguons les quatre cas suivants, en fonction des résultats obtenus par les deux tests présentés ci-dessus :

		$F_{L/F(L)}$	
		Significatif	Non significatif
$F_{L/E}$	Significatif	1^{er} cas	$3^{\text{ème}}$ cas
	Non significatif	$2^{\text{ème}}$ cas	$4^{\text{ème}}$ cas

1^{er} cas : Les deux facteurs ont des influences significatives. La valeur vraie du paramètre μ est estimée par :

$$\hat{\mu} = m = \bar{x}.$$

L'intervalle de confiance sur m est :

$$\left[m - t_{1-\frac{\alpha}{2},(a-1)} \sqrt{\frac{CM_L}{abn}} \, ; \, m + t_{1-\frac{\alpha}{2},(a-1)} \sqrt{\frac{CM_L}{abn}} \right],$$

où $t_{1-\frac{\alpha}{2},(a-1)}$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student avec $(a-1)$ degrés de liberté.
L'incertitude-type sur m est :

$$u_m = \sqrt{\frac{CM_L}{abn}}.$$

Les écarts-types σ_L, σ_u et σ_r sont respectivement estimés par :

$$S_L = \sqrt{\frac{CM_L - CM_{F(L)}}{bn}}, \qquad S_u = \sqrt{\frac{CM_{F(L)} - CM_E}{n}}, \qquad \text{et } S_r = \sqrt{CM_E}.$$

On en déduit l'écart-type de dispersion des résultats moyens des laboratoires, S_Z :

$$S_Z = \sqrt{S_L^2 + \frac{S_u^2}{b} + \frac{S_r^2}{nb}}.$$

On en déduit aussi l'écart-type de reproductibilité S_R :

$$S_R = \sqrt{S_L^2 + S_r^2}.$$

$2^{\text{ème}}$ cas : Seul le facteur interlaboratoires a une influence significative. Dans le cas où la variance interflacons n'est pas significative ($S_u = 0$), nous ne considérons alors plus de variance interflacons, en prenant bn réplications sur un seul flacon.

La table d'analyse de variance devient :

Source	Somme des carrés	Degré de liberté	Moyenne quadratique	Espérance de la moyenne quadratique	Test de significativité
Interlaboratoires	SC_L	a-1	CM_L	$\sigma_r^2 + bn\sigma_u^2$	$F_{F(L)/E}$
Interflacons	-	-	-	-	
Erreur de mesure (résiduelle)	SC_E	$a(bn-1)$	CM_E	σ_r^2	
Totale	SC_T	$abn-1$			

TABLE 1.2 – Table d'ANOVA dans le cas où seul le facteur interlaboratoires a une influence significative

avec

$$SC_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{i..})^2$$

Le test de significativité de la variance interlaboratoires peut être repris.
Si $F_{F(L)/E} > F_{(1-\alpha);(a-1);a(bn-1)}$, le facteur interlaboratoires a une influence significative sur le résultat de mesure.

$F_{F(L)/E}$	
Significatif	Non significatif
$S_L = \hat{\sigma_L} = \sqrt{\frac{CM_L - CM_E}{bn}}$	$S_L \approx 0$ aller au 4ème cas

3ème cas : Seul le facteur hétérogénéité de lot est significatif. Dans le cas où la variance interlaboratoires n'est pas significative, nous ne considérons plus de variance interlaboratoires ($S_L = 0$).

La table d'analyse de variance s'écrit alors :

Source	Somme des carrés	Degré de liberté	Moyenne quadratique	Espérance de la moyenne quadratique	Test de significativité
Interlaboratoires	-	-	-	-	-
Interflacons	$SC_{F(L)}$	$ab-1$	$CM_{F(L)}$	$\sigma_r^2 + n\sigma_u^2$	$F_{F(L)/E}$
Erreur de mesure (résiduelle)	SC_E	$ab(n-1)$	CM_E	σ_r^2	-
Totale	SC_T	$abn-1$	-		-

TABLE 1.3 – Table d'analyse de variance dans le cas où seul le facteur hétérogénéité de lot est significatif

avec

$$SC_{F(L)} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2,$$

ce qui correspond à :

$$SC_{F(L)} = n\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2.$$

Le test de significativité de l'hétérogéneité de lot peut être repris. Si $F_{F(L)/E} = \frac{CM_{F(L)}}{CM_E} > F_{(1-\alpha);b(a-1);ab(n-1)}$, le facteur hétérogénéité du lot a une influence significative sur le résultat de mesure.

$F_{F(L)/E}$	
Significatif	Non significatif
$S_u = \hat{\sigma}_u = \sqrt{\frac{CM_{F(L)} - CM_E}{n}}$	$S_u \approx 0$ Aller au $4^{\text{ème}}$ cas

TABLE 1.4 – Distinction des cas lorsque seul le facteur hétérogénéité de lot est significatif

$4^{\text{ème}}$ cas : Aucun des deux facteurs n'est significatif. Nous ne considérons alors qu'un seul laboratoire. $S_u = 0$ et $S_L = 0$. La table d'analyse de variance se simplifie encore (Table 1.5). Dans cette table,

$$SC_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2.$$

Source	Somme des carrés	Degré de liberté	Moyenne quadratique	Espérance de la moyenne quadratique
Interlaboratoires	-	-	-	-
Interflacons	-	-	-	-
Erreur de mesure (résiduelle)	SC_E	$abn - 1$	CM_E	σ_r^2
Totale	SC_T	$abn - 1$	CM_T	

TABLE 1.5 – Table d'analyse de variance pour le cas où aucun des deux facteurs n'est significatif

1.5.4 Calcul de la reproductibilité et de la répétabilité

Les valeurs de fidélité sont calculées tel que décrit ci-dessus. En se basant sur [ISO, 1994a], nous en déduisons alors la limite de répétabilité r et la limite de reproductibilité R :

- Limite de répétabilité : $r = 2.83 \cdot S_r$.
  C'est la différence maximale à laquelle on doit s'attendre entre deux résultats de mesures indépendantes réalisées sur un matériau, par un même laboratoire et dans les mêmes conditions.
- Limite de reproductibilité : $R = 2.83 \cdot S_R$, où $S_R = \sqrt{S_L^2 + S_r^2}$.
  C'est la différence maximale à laquelle on doit s'attendre, entre deux résultats de mesures indépendantes réalisées par deux laboratoires sur un matériau identique.

Calcul additionnel de répétabilité spécifique aux dénombrements bactérien. Dans le cadre des méthodes énumératives, la répétabilité est également exprimée sous la forme du niveau de signification de l'écart par rapport à la dispersion incompressible d'échantillonnage (établie à l'aide du test T1 de Cochran présenté en section 1.4.2). De plus, le niveau de signification de l'écart entre unités du lot analysées peut aussi être indiqué. Le test T2, reposant également sur les bases de la loi de Poisson permet d'appréhender la dispersion des unités de lot. Pour les

méthodes quantiques, une approche par intervalles de confiance de la dispersion des écarts entre les mesures répétées et entre les unités de lot est effectuée. Le niveau de signification de l'écart est indiqué aux participants.

1.5.5 Calcul des coefficients de variation

Les coefficients de variation expriment les valeurs de fidélité r et R en pourcentage par rapport à la moyenne $\bar{y}_{...}$.
Les coefficients de variation de répétabilité CV_r, de reproductibilité CV_R, et d'hétérogénéité de lot CV_u (utilisé dans le cadre de l'évaluation de la qualité des échantillons) s'écrivent :

$$CV_r = \frac{S_r}{\bar{y}_{...}} \cdot 100, \qquad CV_R = \frac{S_R}{\bar{y}_{...}} \cdot 100, \qquad CV_U = \frac{S_U}{\bar{y}_{...}} \cdot 100.$$

1.6 Cas particulier du modèle Poissonien

Dans le cas de l'ajustement au modèle de Poisson, la dispersion totale observée est réduite à la variabilité aléatoire également appelée *dispersion incompressible d'échantillonnage*. Seul un classement qualitatif (présenté section 1.8.2), reflet du niveau de signification de l'écart à la variable aléatoire ajustée, est attribué aux laboratoires. Il permet de distinguer les participants qui retrouvent la valeur du matériau (homogènes entre eux), de ceux qui ne la retrouvent pas (présentant un biais jugé anormal).

1.7 Calcul des incertitudes de mesure

Les estimations d'incertitude sont exprimées par les *coefficients de surdispersion* notés u^2. Ils correspondent à l'écart-type-relatif au carré qui relate la surdispersion par rapport à la loi de Poisson. Ces estimateurs sont calculés d'après les normes [AFNOR, 1991] et [ISO, 2012b].
Dans l'application que les laboratoires vont faire de ces estimateurs, il leur est indiqué que :

- un u^2 est d'autant plus grand que l'incertitude générée par la mise en œuvre de l'analyse est importante (le u^2 maximal calculé est > 0.4),
- à l'inverse un u^2 égal ou proche de zéro indique une incertitude opérationnelle réduite au minimum (pas ou peu de surdispersion par rapport à la loi de Poisson).

Pour une interprétation plus concrète, le coefficient de surdispersion u^2 peut être exprimé en pourcentage, à l'aide de la formule suivante :

$$u\% = \sqrt{u^2} \cdot 100.$$

Calcul des estimations spécifiques aux laboratoires. Pour chaque laboratoire, il s'agit de calculer l'incertitude en répétabilité u_r^2 et l'incertitude en reproductibilité interlaboratoires u_R^2.

Calcul des estimations générales de la profession. Pour l'ensemble des laboratoires, il s'agit de calculer l'incertitude en répétabilité u_r^2 et l'incertitude en reproductibilité u_R^2.
Les u^2 présentés sont des estimations moyennées à partir des valeurs disponibles depuis la mise en œuvre du paramètre dans les essais, valeurs singulières d'origine identifiée exclues. La transformation des écart-types (échelle logarithmique) en u^2, ceci en répétabilité et en reproductibilité, est effectuée à partir de la formule figurant dans la norme [ISO, 2012b].
Dans les cas suivants, nous ne procédons pas au calcul des incertitudes pour la profession :

- Plus d'un quart des estimations spécifiques non calculables pour un paramètre,
- Problèmes majeurs de biais systématiques. C'est par exemple le cas des *staphylocoques* pathogènes, lorsque des écarts notables selon le type de milieu de culture sélectif employé conduisent à des traitements statistiques par sous-populations de résultats.

1.8 Évaluation de la performance analytique des laboratoires

Outre l'évaluation des valeurs de fidelité de la mesure (évaluation des écarts-type S_r et S_L), l'objectif des essais interlaboratoires est d'évaluer les performances analytiques des laboratoires ayant participé. A noter que l'évaluation des résultats des participants est effectuée sur tous les résultats, y-compris ceux pouvant être *a priori* considérés comme aberrants. L'objectif de cette section est de présenter ce dispositif de scoring.

1.8.1 Calcul du z-score

La composante laboratoire du biais de chaque participant est exprimée en terme de z-score lorsque la variable aléatoire est "Normalisable" par transformation logarithmique.

Calcul du z-score pour le laboratoire i dans une échelle logarithmique. Le z-score du laboratoire i dans une échelle logarithmique se calcule de la façon suivante :

$$z_i = \frac{\bar{y}_{i..} - \bar{y}_{...}}{S_z}.$$

Cet indicateur s'interprète de la façon suivante :

- Si $|z| < 2$: Résultat non différent de l'ensemble des autres,
- Si $2 \leq |z| < 3$: Résultat potentiellement différent de l'ensemble des autres laboratoires,
- Si $3 \leq |z|$: Résultat notablement différent de l'ensemble des autres laboratoires.

Le signe du z-score s'interprète de la façon suivante :

- Si le z-score est positif : le laboratoire a tendance à majorer le résultat,
- Si le z-score est négatif : le laboratoire a tendance à minorer le résultat.

Remarque 1.8.1.
- *En cas de doute sur la qualité des matériaux, en cas de particularité méthodologique ou d'ajustement, des z-scores de valeur absolue supérieure à 2 pourront être jugés satisfaisants a priori.*
- *Dans le cas où la variance interlaboratoires n'est pas significative, nous ne calculons pas de z-score mais seul le classement qualitatif ci-dessous est proposé. De plus, un classement qualitatif exprime sous la forme d'une "note" la qualité de l'analyse fournie par chaque participant.*

1.8.2 Classement qualitatif et graphique de dispersion

En microbiologie, le classement qualitatif relativise la portée du z-score en intégrant l'incertitude sur sa détermination. L'intervalle de confiance sur l'estimation de la moyenne de chaque participant est calculé. Si cet intervalle chevauche la zone d'acceptation (intervalle statistique de dispersion de 95% des résultats), on procède à un reclassement du laboratoire. Il n'est toutefois pas effectué de reclassement de plus d'une "zone" ; à savoir un laboratoire ayant un z-score $3 < |z|$ ne pourra pas être reclassé en +/- A, ceci dans le but de ne pas masquer une dérive potentielle du système analytique. Par ailleurs, aucun reclassement n'est effectué lorsque l'ampleur de l'intervalle de confiance est anormalement élevée, à cause d'un défaut de répétabilité interne du laboratoire.

Un reclassement peut également être opéré suite à la détection d'une anomalie relative à la qualité des échantillons. Ce type de reclassement peut porter :

- soit sur la détection de laboratoires potentiellement pénalisés à la suite du constat d'un problème de stabilité des échantillons dans la période raisonnable,
- soit sur la détection de laboratoires potentiellement pénalisés à la suite du constat d'un problème d'hétérogénéité de lot lors de l'essai.
  Dans ce cas de figure, on procède généralement à un examen des données brutes des participants, et en particulier des répercussions des disparités inter-échantillons (d'après le test T2 section 1.6) sur l'évaluation des performances analytiques des participants. Le reclassement porte classiquement sur les participants potentiellement pénalisés retrouvant un résultat significativement élevé ou bas sur l'un de ses deux flacons, sans présenter par ailleurs de problème de répétabilité.

La Figure 1.3, issue de la procédure interne d'AGLAE [AGLAE,], présente l'interprétation de la performance analytique des laboratoires.

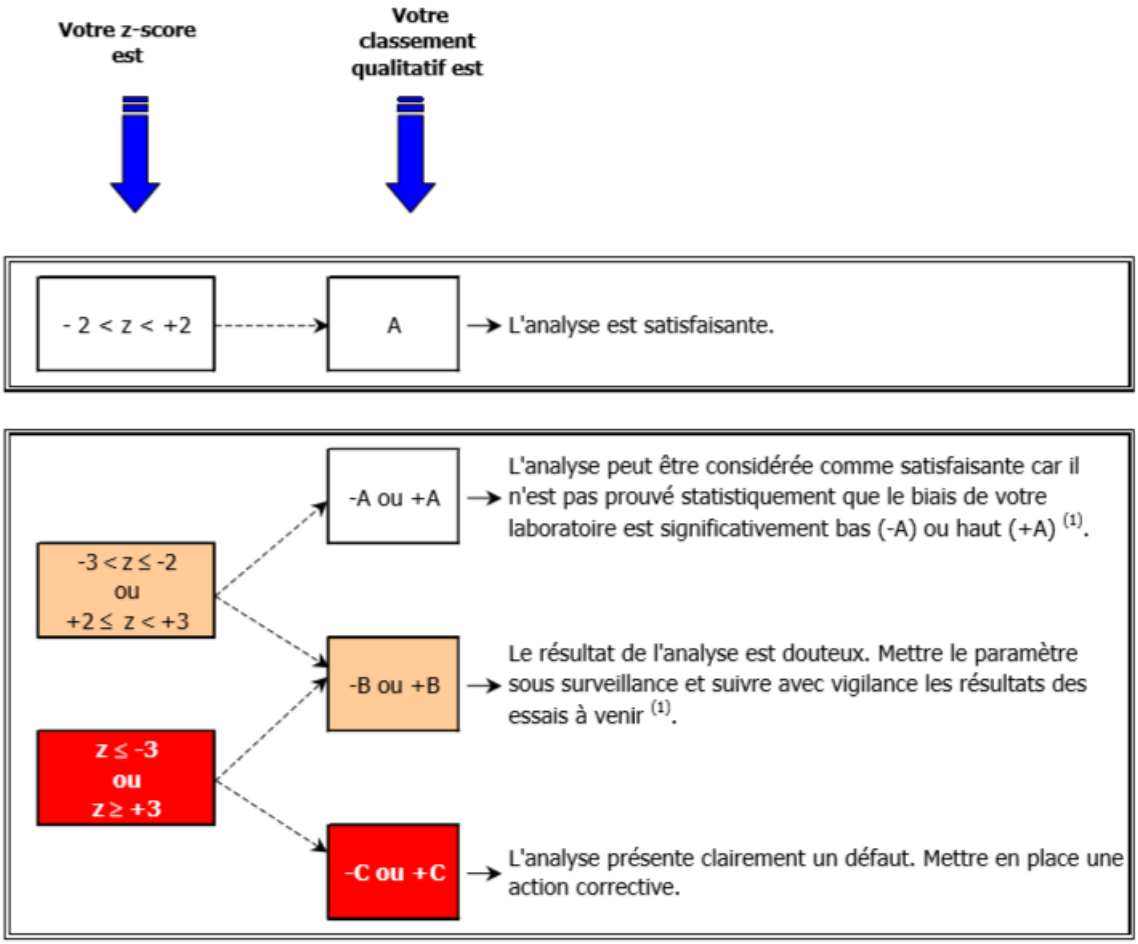

FIGURE 1.3 – Interprétation du z-score et du classement qualitatif d'un laboratoire

Cas des variables aléatoires non normalisables. Lorsque la variable aléatoire n'est pas normalisable (ajustement au modèle de Poisson), il est simplement fourni un classement qualitatif qui est le reflet du niveau de signification de l'écart à la variable aléatoire ajustée :

- "A" : pas d'écart significatif : le résultat appartient à la population,
- "B" : écart significatif (à 5%) : le résultat apparaît décalé par rapport à la moyenne, mais sans que l'on puisse affirmer que le résultat n'appartienne pas à la population,
- "C" : écart très significatif (à 1%) : le résultat n'appartient pas à la population.

1.8.3 Visualisation des résultats des participants

En microbiologie, des schémas sous forme de cartes de contrôle permettent de visualiser les résultats des participants par rapport à la valeur de consensus établie et aux bornes des zones d'acceptation calculées pour l'essai. Deux types de schémas peuvent être utilisés selon le niveau de charge bactérienne effectivement réalisé pour l'essai.

- Si l'exploitation statistique a été menée en utilisant le modèle de Poisson comme distribution théorique d'ajustement des données, la représentation de la dispersion sera effectuée selon ce même modèle (voir Figure 1.4, issue de [AGLAE,]).

Les bornes des zones d'acceptation à 95% et à 99% sont alors calculées compte-tenu du modèle de Poisson et du plan d'essai mis en œuvre. Concrètement, les valeurs d'une variable aléatoire suivant une loi de Poisson de paramètre λ (valeur de consensus) pris pour 2x2 répliques sont calculées pour les probabilités de 95% et 99%. Ce calcul repose donc sur la somme des 4 dénombrements réalisés. En ce qui concerne les intervalles de confiance sur la moyenne de chaque participant, le calcul est effectué selon le modèle de Poisson, d'après la loi du χ^2 [Brownlee, 1965].

- Si l'exploitation statistique a été menée en utilisant le modèle log-normal comme distribution théorique d'ajustement des données, la représentation de la dispersion sera effectuée selon ce même modèle (Figure 1.5). Pour ce type de schéma, les bornes des zones d'acceptation à 95% et à 99% sont calculées en considérant $2S_z$ et $3S_z$ de part et d'autre de la moyenne générale $\bar{y}_{...}$.

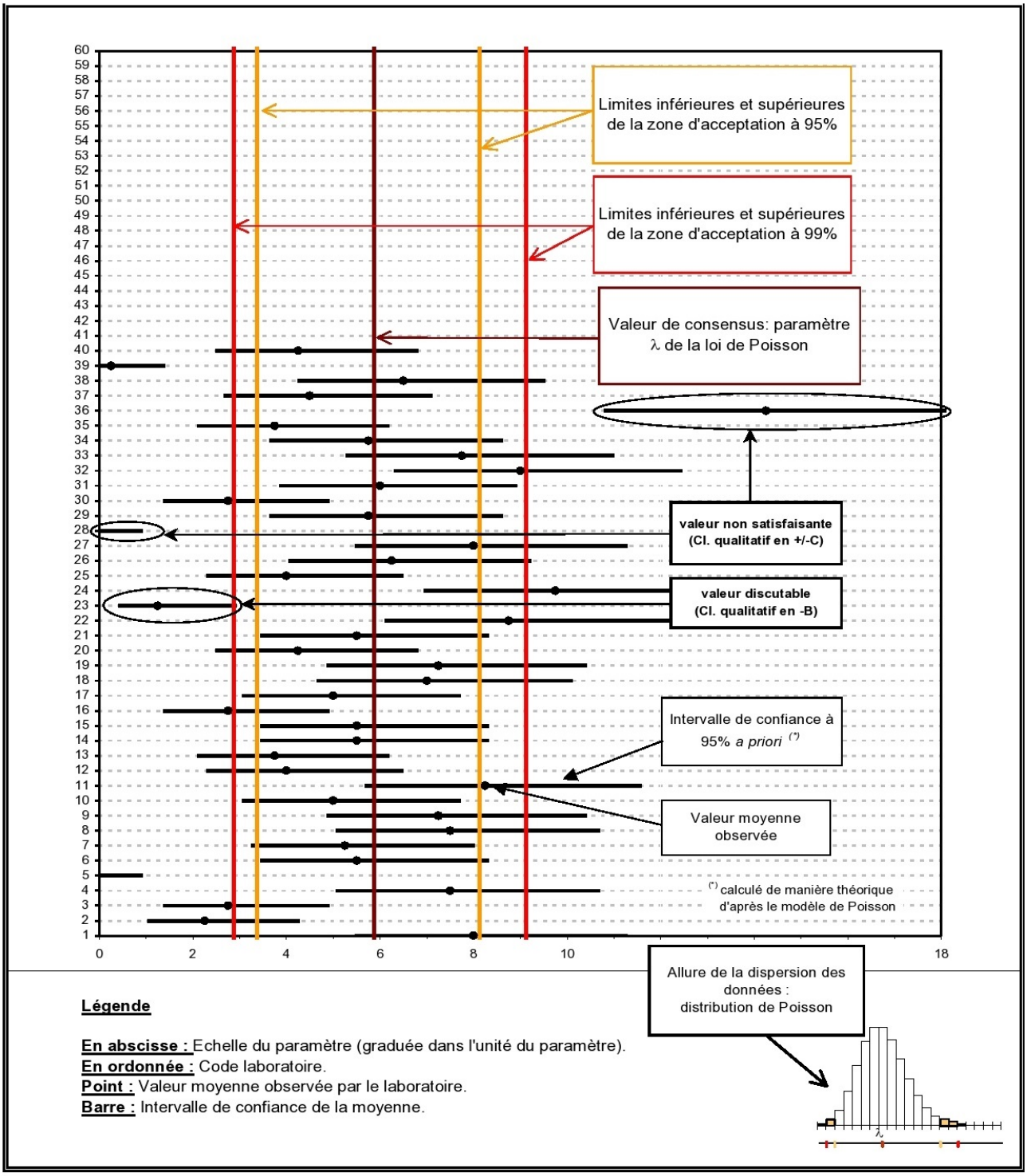

FIGURE 1.4 – Représentation de la dispersion des laboratoires selon un ajustement au modèle de Poisson

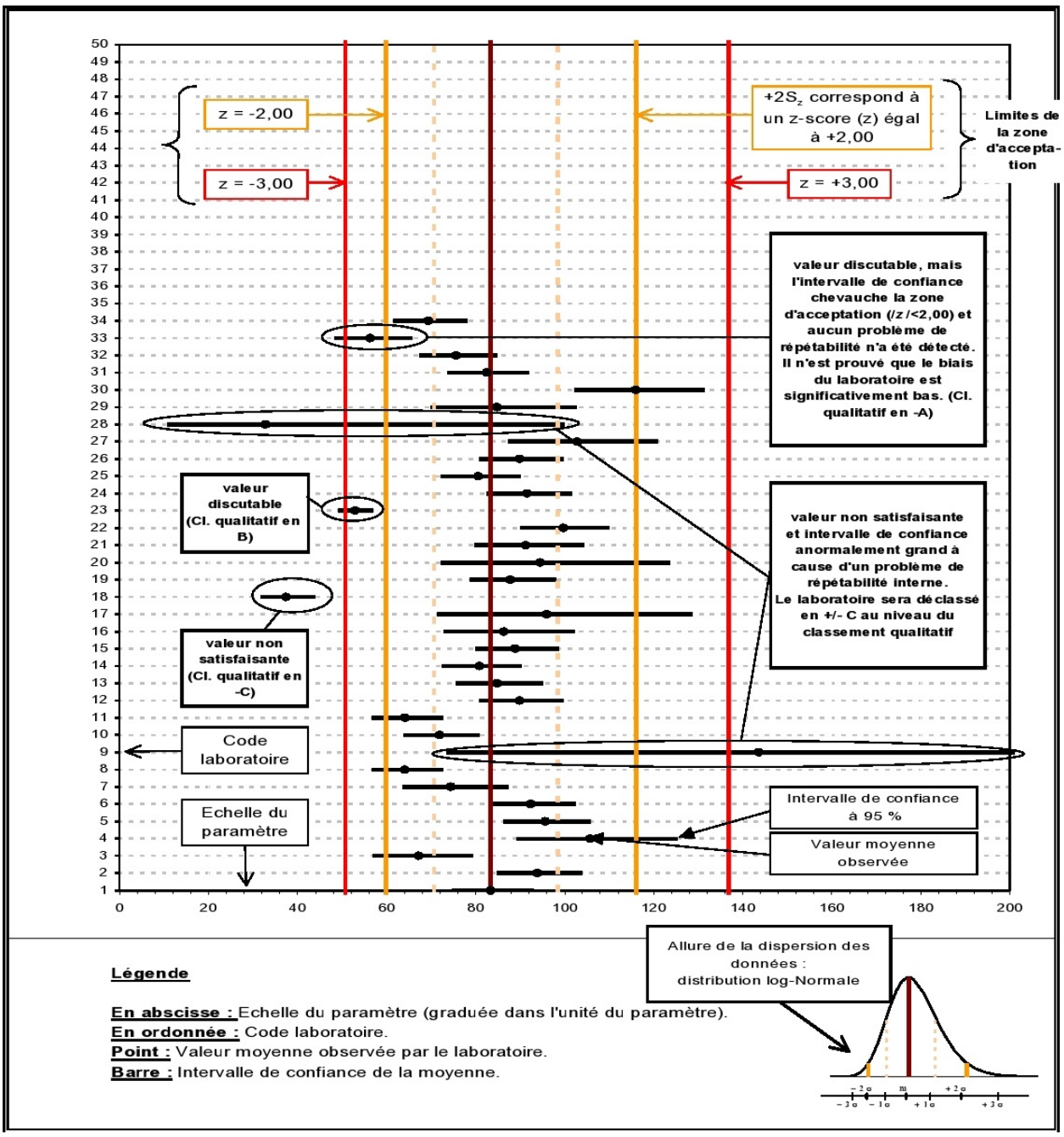

FIGURE 1.5 – Représentation de la dispersion des laboratoires selon un ajustement au modèle de log-normal

1.9 Inconvénients et limites de la méthode

Les données étudiées dans le domaine de la biologie sont souvent de nature discrète. Bien que l'utilisation de la transformation racine carrée soit parfois préconisée pour traiter de telles données, notamment dans [Jarvis, 1989, Lightfoot and Maier, 1998], elles sont le plus souvent log-transformées, puis étudiées à l'aide d'une méthode paramétrique, telle que l'analyse de variance, comme c'est le cas dans le domaine du contrôle de qualité en microbiologie.
La raison de l'utilisation de cette transformation logarithmique des données en microbiologie est liée à une pratique commune en termes de traitement des données de comptage bactériens, comme l'avancent [ISO, 2010b, Niemelä, 2002]. En pratique, on constate que les dénombrements bactériens sont souvent caractérisés par une distribution asymétrique, avec relativement plus de valeurs basses que de valeurs élevées. La loi log-normale décrit bien ce type de distribution. D'autre part, classiquement, les variables réponses sont transformées pour améliorer à la fois leur linéarité, et l'homogénéité de la variance (homoscédasticité). En microbiologie, la log-transformation des résultats de dénombrement permet d'obtenir des données qui sont normalement distribuées, et de variance homogène. Les hypothèses fondamentales nécessaires à la mise en œuvre d'une analyse de variance sont alors vérifiées. Pourtant, il n'y a aucune raison pour qu'une transformation soit optimale à la fois en terme de linéarité et en terme d'homogénéité. Aussi, l'utilisation d'une transformation logarithmique implique que les effets soient multiplicatifs à l'échelle des données brutes, ce qui n'est pas nécessairement le cas.

En pratique, lorsqu'on utilise la méthode présentée dans ce chapitre, on observe des ajustements insatisfaisants. La Figure 1.6 présente un exemple d'ajustement de données d'un essai interlaboratoires à la loi de Poisson. La zone d'acceptation des laboratoires est trop étroite.

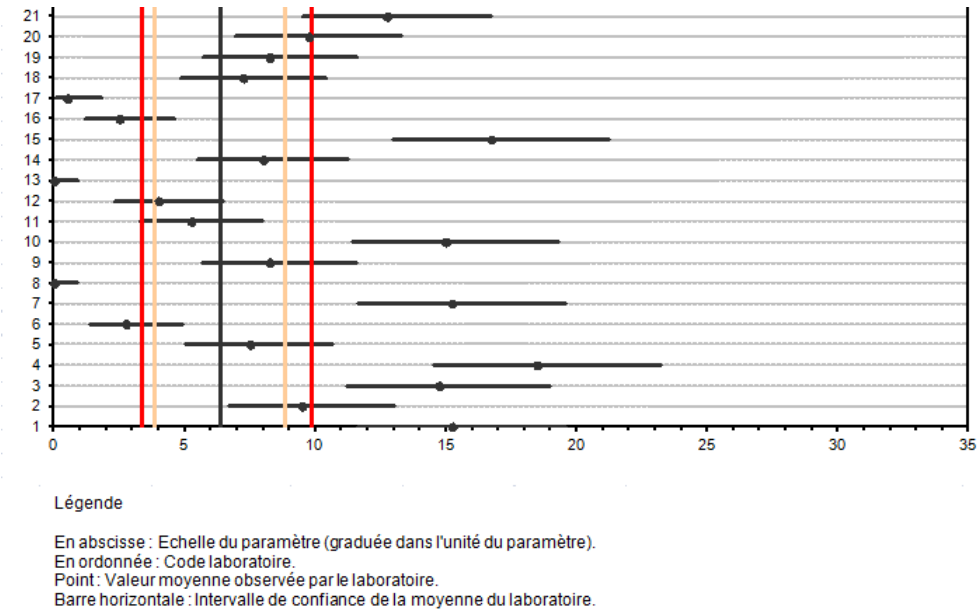

FIGURE 1.6 – Exemple d'un ajustement insatisfaisant au modèle de Poisson

La Figure 1.7 présente l'ajustement des données de l'essai interlaboratoires considéré dans la Figure 1.6, selon un modèle log-normal. La zone d'acceptation des laboratoires est cette fois trop large, à cause de la dispersion prise en compte dans l'échelle logarithmique. L'ajustement n'est donc satisfaisant ni suivant la loi de Poisson, ni suivant le modèle log-normal.

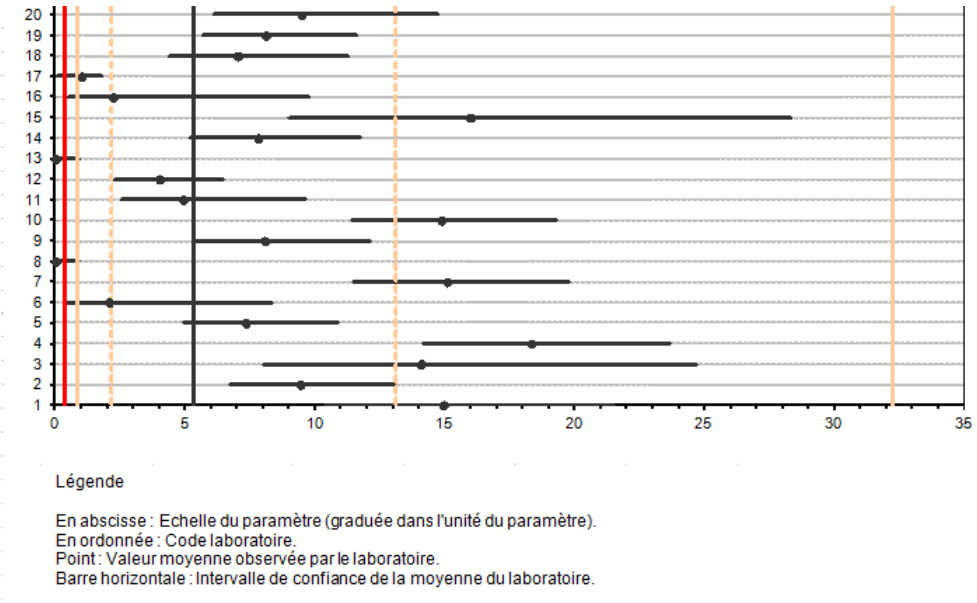

FIGURE 1.7 – Exemple d'un ajustement insatisfaisant au modèle log-normal

Or, les valeurs réglementaires (les spécifications qui permettent par exemple de classer une eau "potable" ou "non potable") sont presque toutes à des bas niveaux de charge bactérienne (voire à l'absence) dans le domaine du contrôle sanitaire des eaux et de l'environnement. Par exemple, pour le dénombrement des deux témoins de contamination fécale que sont *Escherichia coli* et les entérocoques, la valeur réglementaire est à l'absence de germes dans 100 millilitres. Les niveaux de charge qui intéressent la profession sont donc les bas niveaux, là où, comme présenté en Figures 1.6 et 1.7, ni le modèle de Poisson ni le modèle log-normal ne convient parfaitement.

Par ailleurs, la généralisation des estimations d'écarts-types en log-normal est difficile. En effet, S_R n'est pas constant sur la gamme de charges bactériennes intéressantes.

L'utilisation d'une transformation logarithmique des résultats de dénombrements pour permettre leur exploitation au moyen d'une analyse de variance peut donc être perçue comme un palliatif à l'absence de modèle adapté. Ce palliatif semble peut satisfaisant puisqu'il n'est pas homogène (valable seulement pour les niveaux de charge bactérienne m supérieurs à 30, ou supérieurs à 15 en extrême limite). D'autre part, à bas niveaux, cette méthode introduit des distorsions dans la représentation de la dispersion des résultats de mesures observés.

Or, la littérature scientifique dans le domaine de la microbiologie considère que la loi de Poisson surdispersée correspond à la loi binomiale négative [Niemelä, 2002]. **Il semble alors judicieux de réfléchir à un nouveau modèle basé sur cette loi de binomiale négative pour exploiter les données issues des essais interlaboratoires en microbiologie.**

1.10 Synthèse de la problématique et organisation de la thèse

Dans ce chapitre, nous avons introduit la problématique et les enjeux du travail présenté dans ce manuscrit. Les fondements du dénombrement en microbiologie, et notamment sa formalisation statistique sont abordés en section 1.3. En section 1.4, nous avons présenté la méthode utilisée par AGLAE pour l'étude de la variabilité des résultats de dénombrement d'un essai interlaboratoires en microbiologie. L'analyse de la variance est détaillée en section 1.5, tandis que l'ajustement

au modèle de Poisson est présenté en section 1.6. L'utilisation de la méthode présentée conduit pourtant à des ajustements insatisfaisants, imputables notamment à la transformation logarithmique qui est préalablement appliquée aux données (section 1.9).
Dans le chapitre 2, nous proposons un modèle linéaire semblable à celui qui est présenté dans ce chapitre, mais à facteurs fixes, permettant d'exploiter les données brutes (non transformées). Nous developperons des tests de significativité des facteurs sur la base de ce modèle.
Au chapitre 3, nous proposerons un modèle linéaire généralisé hiérarchique Gamma-Poisson pour expliquer la variabilité des résultats de dénombrement des essais interlaboratoires en microbiologie. Nous présenterons les méthodes d'estimation, de tests des effets, d'ajustement, et d'évaluation des laboratoires développées à partir de ce modèle.
Au chapitre 4, nous proposons une méthode de détection de résultats de mesure aberrants à partir d'une distribution presque exacte du produit de variables aléatoires indépendantes de lois Gamma.

Chapitre 2

Analyse de déviance à 2 facteurs fixes imbriqués sur données de Poisson

Sommaire

2.1 Introduction

Dans le chapitre 1, nous avons mis en évidence les limites de l'utilisation de la méthode d'analyse de variance sur données log-transformées lorsque les données sont de nature discrète, comme c'est le cas en microbiologie. Comme souligné par [O'Hara and Kotze, 2010], grâce aux modèles linéaires généralisés (GLM, que nous présenterons plus en détails à la section 3.2.3 du chapitre 3), il existe aujourd'hui des méthodes alternatives qui s'avèrent plus efficaces qu'une analyse sur données préalablement transformées.

Nous avons montré au chapitre 1, section 1.3.2, que les résultats de dénombrement d'un essai interlaboratoires en microbiologie suivent idéalement une loi de Poisson. D'autre part, le modèle utilisé d'un point de vue normatif pour le contrôle de qualité en chimie comme en microbiologie correspond à un modèle à facteurs imbriqués (1.5.1).

C'est pourquoi, dans ce chapitre, nous considérons un modèle à facteurs fixes imbriqués, sur

données de Poisson. Nous proposons une extension de la méthode d'analyse de variance à un modèle linéaire généralisé de Poisson. Ce travail a fait l'objet d'une publication pour les 46$^{\text{èmes}}$ Journées de Statistique [Loingeville et al., 2014].

Dans en premier temps, nous introduisons le modèle considéré. Nous présentons, dans un deuxième temps, une généralisation de l'analyse de variance aux modèles linéaires généralisés. Nous appliquons la méthode proposée au modèle (1.1) précédemment introduit. Dans un troisième temps, nous présentons les résultats fournis par cette méthode sur notre modèle, et nous soulignons ses avantages et ses limites.

2.2 Modèle à deux facteurs fixes imbriqués sur données de Poisson

Dans le cadre d'un essai interlaboratoires en microbiologie, nous cherchons à expliquer la variabilité de résultats de dénombrement. Comme nous l'avons vu au chapitre 1, la variabilité des résultats d'un essai interlaboratoires peut s'expliquer par le facteur *Laboratoire* et par le facteur *Flacon* (hétérogénéité de lot). Notons y_{ijk} le résultat de la mesure réalisée sur la $k^{\text{ième}}$ réplication du flacon j par le laboratoire i. Les résultats de dénombrement sont décrits par le modèle linéaire (1.1) introduit au chapitre 1 :

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \epsilon_{(ij)k} \quad \begin{cases} i = 1, ..., a \\ j = 1, ..., b \\ k = 1, ..., n \end{cases} \tag{2.1}$$

où

- μ est la moyenne globale,
- τ_i est l'effet du laboratoire i,
- $\beta_{j(i)}$ est l'effet du flacon j imbriqué dans le laboratoire i,
- $\epsilon_{(ij)k}$ est l'erreur de mesure entre les réplications d'un flacon pour un laboratoire.

Nous considérons dans ce chapitre un plan équilibré comportant le même nombre de niveaux du facteur *Flacon* ($b = 2$) sous chaque niveau du facteur *Laboratoire*, et un même nombre de réplications par flacon ($n = 2$). Le caractère imbriqué des facteurs Laboratoire et Flacon suffisant à justifier les différences entre les résultats des mesures effectuées par un même laboratoire sur une même réplication, nous considérerons dans ce chapitre les facteurs Laboratoire et Flacon fixes. Nous reviendrons en détails sur la distinction entre facteurs fixes et aléatoires au chapitre 3 en section 3.2.2.

Nous cherchons ici à identifier les facteurs qui contribuent à la variabilité des résultats de dénombrement d'un essai interlaboratoires considéré. Nous avons effectivement vu au chapitre 1, section 1.5.3, que, dans certains cas de figure, le facteur *Laboratoire* et/ou le facteur *Flacon* n'est pas significatif. Pour le modèle à facteurs fixes (2.1), les tests des effets *Laboratoire* et *Flacon* s'écrivent de la façon suivante :

- Test de significativité du facteur *Laboratoire* :

$$\begin{cases} \mathcal{H}_0 : \forall\, i,\ 1 \leq i \leq a\ ,\ \tau_i = 0 \text{ contre} \\ \mathcal{H}_1 : \exists\, i,\ \tau_i \neq 0. \end{cases}$$

- Test de significativité du facteur *Flacon* imbriqué dans le facteur *Laboratoire* :

$$\begin{cases} \mathcal{H}_0 : \forall (i,j) \ / 1 \leq i \leq a, 1 \leq j \leq b, \ \beta_{j(i)} = 0 \text{ contre} \\ \mathcal{H}_1 : \exists (i,j) \ / \beta_{j(i)} \neq 0. \end{cases}$$

Lorsque les données sont gaussiennes, la méthode d'analyse de variance est utilisée pour tester la significativité des effets. Les erreurs $\epsilon_{(ij)k}$ sont alors considérées 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, comme présenté dans [Montgomery, 2008].
Nous considérons dans ce chapitre que les résultats de dénombrement y_{ijk} sont distribués suivant une loi de Poisson (1.3.2), de paramètre que nous noterons λ_{ijk}. Le modèle 2.1 devient alors :

$$y_{ijk} \sim \mathcal{P}(\lambda_{ijk}) \quad \begin{cases} i = 1, ..., a \\ j = 1, ..., b \\ k = 1, ..., n \end{cases} \tag{2.2}$$

Pour transposer la méthode d'analyse de variance à la microbiologie, nous souhaitons alors définir des méthodes de test applicables sur données de comptage. Les données étant ici supposées suivre une loi de Poisson, l'hypothèse de distribution normale des erreurs n'est alors pas vérifiée.

Notons L et F les propositions *"Il y a un effet Laboratoire"*, et *"Il y a un effet Flacon"*, et $\bar{L}$ et $\bar{F}$ les propositions inverses *"Il n'y a pas d'effet Laboratoire"*, et *"Il n'y a pas d'effet Flacon"*. **Pour tester la significativité des facteurs Laboratoire et Flacon, nous proposons dans ce chapitre de comparer des modèles incluant ou non ces effets.** Nous distinguons ici les quatre modèles possibles suivants :

- $\bar{L}\,\bar{F}$: Il n'y a aucun effet,
- $L\,\bar{F}$: Il y a un effet *Laboratoire* mais pas d'effet *Flacon*,
- $\bar{L}\,F$: Il y a un effet *Flacon* mais pas d'effet *Laboratoire*,
- $L\,F$: Il y a un effet *Laboratoire* et un effet *Flacon*.

En fonction du modèle choisi, y_{ijk} suivra une loi de Poisson de paramètre λ, λ_i, λ_{ij} ou λ_j.

2.3 Ajustement du modèle

L'ajustement d'un modèle à un jeu de données permet de remplacer un jeu de données y par un jeu de données *ajustées* $\hat{\mu}$, obtenu à partir du modèle considéré. En général, les valeurs $\hat{\mu}$ ne correspondront pas exactement aux valeurs y. L'étude de l'écart entre les valeurs $\hat{\mu}$ et y est crucial car, si une légère différence est acceptable, une différence trop large ne l'est en revanche pas.

Un jeu de données de N observations, pourra être ajusté par des modèles contenant jusqu'à N paramètres. Le modèle le plus simple, noté dans la littérature modèle *nul*, a un seul paramètre, qui correspond à une moyenne globale μ. Dans le cadre de notre problématique, μ correspond à la moyenne de tous les résultats de dénombrement y_{ijk}. Pour le modèle *nul*, la variabilité des résultats de dénombrement ne s'explique que par le caractère aléatoire de la loi de Poisson. Au contraire, le modèle *complet* comporte autant de paramètres qu'il y a de données dans l'échantillon, c'est à dire *abn* dans le cas du modèle (2.2). Le modèle complet attribue alors la variation

entre les résultats de dénombrement au laboratoire qui réalise la mesure, au flacon sur lequel elle est réalisée (hétérogénéité de lot), ainsi qu'à la réplication (répétabilité). L'adéquation d'un modèle à un échantillon peut être étudiée de plusieurs façons. Dans le cas de modèles gaussiens, la méthode la plus courante consiste à étudier la distribution des résidus, notamment les résidus de Pearson et de Anscombe, présentés dans [McCullagh and Nelder, 1989]. Nous nous intéressons ici à la déviance, qui permet d'étudier l'adéquation de modèles à des données non gaussiennes.

La déviance est une statistique basée sur le rapport entre les vraisemblances de deux modèles. Elle permet de comparer un modèle donné au modèle de vraisemblance maximale, le modèle complet, noté ici *MC*. Pour le modèle (2.2), la déviance $D(y;\hat{\mu})$ s'écrit :

$$D(y;\hat{\mu}) = -2\{\log(L(y_{111},\ldots,y_{abn},\lambda)) - \log(L_{MC}(L(y_{111},\ldots,y_{abn},\lambda_{MC}))\}$$

Lorsque l'on compare des modèles, le modèle nul est généralement trop simple, tandis que le modèle complet n'est que rarement retenu car ne permet pas de synthétiser l'information fournie par les données. Cependant, le modèle complet a l'avantage de fournir une base de comparaison pour mesurer l'écart d'un modèle intermédiaire plus parcimonieux.
La comparaison de modèles basée sur la déviance est appelée analyse de déviance.

2.3.1 Généralisation de l'analyse de variance en terme de distribution

Dans le cas d'une distribution normale, la déviance correspond à la somme des carrés des résidus. On peut ainsi aisément considérer l'analyse de déviance comme une généralisation, en terme de distribution, de l'analyse de variance. Dans [McCullagh and Nelder, 1989], l'analyse de déviance est présentée comme une extension de la méthode d'analyse de variance à la classe plus large des modèles linéaires généralisés. Cette généralisation repose sur deux points. D'une part, contrairement aux facteurs d'un modèle d'analyse de variance, les termes d'un GLM ne sont généralement pas orthogonaux. D'autre part, les sommes des carrés utilisées pour l'analyse de variance ne sont alors plus appropriées pour mesurer la contribution d'un terme à la variabilité totale. En effet, les sommes des carrés d'une analyse de variance permettent d'expliquer la variabilité induite par les différents facteurs. Elles peuvent alors être considérées comme des différences entre des mesures d'adéquation de modèles aux données, chaque modèle incluant un terme de plus que le précédent. Par exemple, un modèle simple à deux facteurs A et B conduit à une analyse de variance à trois termes, A, B et l'interaction entre ces deux facteurs, notée AB. Les sommes des carrés des modèles que l'on peut ajuster à ce problème sont les différences entre les sommes des carrés résiduelles (somme des carrés de l'erreur) obtenues en ajustant successivement les modèles *Nul*, A, $A+B$, $A+B+AB$ aux données, où *Nul* correspond au modèle *Nul* qui ne comporte qu'un paramètre, la moyenne μ. La table d'analyse de variance 2.1 ci-dessous rend la généralisation explicite.

Modèle	Degré de liberté	Somme des carrés résiduelle	Somme des carrés du facteur	Degré de liberté	Facteur
Nul	df_{Null}	SC_E	SC_A	df_A	A
A	df_A	SC_E	SC_B	df_B	B
A+B	df_{A+B}	SC_E	SC_{AB}	df_{A+B}	$A+B$
A+B+AB	0	0			

TABLE 2.1 – Table d'Analyse de déviance à 2 facteurs

En considérant une séquence de modèles emboîtés, nous pouvons alors utiliser la déviance comme mesure générale de l'écart entre les modèles, et former ainsi une table d'analyse de déviance en calculant les différences entre les déviances associées à chaque modèle, comme fait dans la table 2.1. L'interprétation de la table d'analyse de déviance est cependant compliquée par le fait que les termes ne sont pas nécessairement orthogonaux. Chaque somme des carrés des facteurs de la table 2.1 représente la variation due au facteur correspondant, en ayant éliminé les effets des facteurs considérés au dessus et en ignorant les effets des facteurs mentionnés plus bas. Dans certains problèmes, on est alors amené à considérer plusieurs séquences de modèles, chacune possédant sa propre table d'analyse de déviance.
Le but de l'analyse de déviance est de développer des modèles parcimonieux, dans lesquels les facteurs qui ne sont pas nécessaires sont exclus. Notons que, lorsque les données sont complexes, il est peu probable qu'un unique modèle soit largement meilleur que les autres, c'est pourquoi on ne parle généralement pas de "meilleur" modèle, puisque d'autres modèles peuvent en être très proches.

2.3.2 Analyse de déviance pour le contrôle de qualité en microbiologie

Dans un modèle à facteurs imbriqués, le facteur imbriqué n'existe pas seul. En effet, il est possible de montrer qu'une analyse de variance à deux facteurs A et B, avec B imbriqué dans A est équivalente à une analyse de variance classique à deux facteurs A et B, sans effet du facteur B seul, mais avec effet du facteur A et effet d'interaction AB (Annexe D). Par conséquent, le modèle (2.2) ne permet pas de mettre en évidence l'effet du facteur *Flacon* seul. Cela signifie que nous ne pouvons pas détecter une hétérogénéité de lot à partir du modèle (2.2). Ce modèle permet en revanche de détecter l'effet du facteur *Flacon* imbriqué dans le facteur *Laboratoire* (test 2 détaillé en section 2.4.2).

Le caractère imbriqué du facteur *Flacon* ne permet pas la transposition directe de la table d'analyse de déviance 2.1 au modèle (2.2). Pour mettre en évidence les facteurs qui ont une influence significative sur les résultats de dénombrement d'un essai interlaboratoires en microbiologie, nous proposons alors d'effectuer successivement les deux tests suivants :

1. Test 1 : Test d'existence d'un effet quelconque :

$$\begin{cases} \mathcal{H}_0 : \text{Modèle sans effet } \bar{L}\bar{F} \text{ contre} \\ \mathcal{H}_1 : \text{Modèle complet } LF. \end{cases}$$

2. Test 2 : Test d'identification de l'effet (s'il y a un effet) :

$$\begin{cases} \mathcal{H}_0 : \text{Modèle à effet Laboratoire seul } L\bar{F} \text{ contre} \\ \mathcal{H}_1 : \text{Modèle complet } LF. \end{cases}$$

Le test 1 permet d'identifier un éventuel effet. Si on détecte un effet (rejet de $\mathcal{H}_0$), le test 2 permet alors de déterminer si cet effet est uniquement dû au facteur *Laboratoire*. Lorsque l'on rejette l'hypothèse $\mathcal{H}_0$ du test 2, 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 autre modèle pour le déterminer.

Pour certains essais interlaboratoires, des flacons présentant des niveaux de concentration différents, ou issus de deux cuves différentes, sont parfois envoyés aux laboratoires. Dans ce cas de figure, nous savons alors qu'il existe un effet Laboratoire. Nous cherchons en revanche à mettre en évidence un effet Laboratoire. Nous proposons pour cela le test suivant :

- Test 3 : Test d'existence d'un effet Laboratoire :

$$\begin{cases} \mathcal{H}_0 : \text{Modèle à effet Flacon sans effet Laboratoire : } \bar{L}F \\ \mathcal{H}_1 : \text{Modèle avec effet Flacon et effet Laboratoire : } LF. \end{cases}$$

Pour pouvoir calculer les déviances associées à chacun des modèles mentionnés ci-dessus ($\bar{L}\bar{F}$, $L\bar{F}$, LF), nous devons connaître le paramètre λ de la loi de Poisson associée à chacun de ces modèles. Suivant le modèle, le paramètre λ s'exprime en fonction de i (λ_i), de i et de j (λ_{ij}), ou indépendamment de i et de j (λ).
Les tests 1, 2 et 3, et les paramètres des lois de Poisson associées aux modèles correspondants sont résumés dans la table 2.2.

	Test	$\mathcal{H}_0/\mathcal{H}_1$	Loi de y_{ijk} sous $\mathcal{H}_0$	Loi de y_{ijk} sous $\mathcal{H}_1$
1	Test d'existence d'un effet quelconque	$\bar{L}\,\bar{F}$ / LF	$y_{ijk} \sim P(\lambda)$	$y_{ijk} \sim P(\lambda_{ij})$
2	Test modèle à deux effets contre modèle effet "Labo"	$L\bar{F}/LF$	$y_{ijk} \sim P(\lambda_i)$	$y_{ijk} \sim P(\lambda_{ij})$
3	Test modèle à deux effets contre modèle effet "Flacon"	$\bar{L}F$ /LF	$y_{ijk} \sim \frac{1}{2}P(\lambda_1) + \frac{1}{2}P(\lambda_2)$	$y_{ijk} \sim P(\lambda_{ij})$

TABLE 2.2 – Tests des effets Laboratoire et Flacon

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 $\bar{L}\bar{F}$ ($\mathcal{H}_0$ du test 1). 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 :

$$\hat{\lambda} = \bar{y}_{...} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{2}\sum_{k=1}^{2} y_{ijk}}{4a}. \tag{2.3}$$

Les estimateurs $\hat{\lambda}_i$ de λ_i (configuration $L\,\bar{F}$, hypothèse $\mathcal{H}_0$ du test 2) et $\hat{\lambda}_{ij}$ de λ_{ij} (configuration LF, hypothèses $\mathcal{H}_1$ des tests 1 et 2) sont :

$$\hat{\lambda}_i = \bar{y}_{i..} = \frac{\sum_{j=1}^{2}\sum_{k=1}^{2} y_{ijk}}{4}, \qquad (2.4) \qquad \hat{\lambda}_{ij} = \bar{y}_{ij.} = \frac{\sum_{k=1}^{2} y_{ijk}}{2}. \qquad (2.5)$$

Pour effectuer les tests de la Table 2.2, nous formons leurs rapports de vraisemblance, Λ :

$$\Lambda = \frac{L_{\mathcal{H}_0}}{L_{\mathcal{H}_1}},$$

où $L_{\mathcal{H}_0}$ et $L_{\mathcal{H}_1}$ correspondent aux vraisemblances du modèle (2.1) sous les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$. Nous calculons ainsi la déviance D associée à un test :

$$D = -2\log(\Lambda).$$

2.4 Tests de significativité basés sur la déviance

2.4.1 Test de détection d'un effet

Le premier test consiste à détecter un éventuel effet. Les hypothèses de ce test, noté test 1 plus haut, sont :

$$\begin{cases} \mathcal{H}_0 : \bar{L}\bar{F} : y_{ijk} \sim P(\lambda) \\ \mathcal{H}_1 : LF : y_{ijk} \sim P(\lambda_{ij}). \end{cases}$$

Le rapport de vraisemblance nous permet de comparer le modèle LF ($\mathcal{H}_1$), et le modèle complet, $\bar{L}\bar{F}$ ($\mathcal{H}_0$). L'estimateur de λ étant $\hat{\lambda}$ (expression (2.3)) sous $\mathcal{H}_0$, la vraisemblance du modèle $\bar{L}\bar{F}$ s'écrit :

$$L_{\bar{L}\bar{F}}(y_{111}, \ldots, y_{abn}, \lambda) = \prod_{i=1}^{a}\prod_{j=1}^{2}\prod_{k=1}^{2} \hat{\lambda}^{y_{ijk}} \frac{e^{-\hat{\lambda}}}{y_{ijk}!}. \qquad (2.6)$$

L'estimateur de λ_{ij} étant $\hat{\lambda}_{ij}$ (expression (2.5)) sous $\mathcal{H}_1$, la vraisemblance du modèle LF s'écrit :

$$L_{LF}(y_{111}, \ldots, y_{abn}, \lambda_{11}, \ldots, \lambda_{1a}, \lambda_{21}, \ldots, \lambda_{2a}) = \prod_{i=1}^{a}\prod_{j=1}^{2}\prod_{k=1}^{2} \hat{\lambda}_{ij}^{y_{ijk}} \frac{e^{-\hat{\lambda}_{ij}}}{y_{ijk}!}. \qquad (2.7)$$

Nous pouvons former le rapport de vraisemblance Λ_1 du test 1 à partir des expressions (2.6) et (2.7) :

$$\Lambda_1 = \frac{L_{\bar{L}\bar{F}}}{L_{LF}} = \prod_{i=1}^{a}\prod_{j=1}^{2} e^{-2(\lambda-\lambda_{ij})} \left(\frac{\lambda}{\lambda_{ij}}\right)^{\sum_{k=1}^{2} y_{ijk}}. \qquad (2.8)$$

Nous obtenons l'expression de la déviance D_1 du test 1 à partir de l'expression (2.8) :

$$D_1 = -4\sum_{i=1}^{a}\sum_{j=1}^{2} \bar{y}_{ij.} \log\left(\frac{\bar{y}_{...}}{\bar{y}_{ij.}}\right). \qquad (2.9)$$

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_1 du jeu de données est indéfinie. En observant que $\lim_{\bar{y}_{ij.}\to 0}\left(\bar{y}_{ij.}\ \log(\frac{\bar{y}_{...}}{\bar{y}_{ij.}})\right) = 0$, nous pouvons réécrire l'expression (2.9) :

$$D_1 = -2\log(\Lambda_0) = -4\sum_{i=1}^{a}\sum_{\substack{j=1 \\ \forall\, j\,/\,\bar{y}_{ij.}\neq 0}}^{2} \bar{y}_{ij.}\log\left(\frac{\bar{y}_{...}}{\bar{y}_{ij.}}\right). \tag{2.11}$$

D'après un théorème asymptotique de [Lehmann and Romano, 2006], sous $\mathcal{H}_0$, la distribution de la déviance D_1 tend vers une loi du χ^2 à $2a-1$ degrés de liberté lorsque le nombre de réplications n tend vers l'infini. Étant donné le petit nombre de réplications considérées ici ($n=2$), le théorème de Lehmann ne peut être appliqué, comme le soulignent [McCulloch and Neuhaus, 2001]. 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, comme indiqué dans [Droesbeke et al., 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 $\mathcal{H}_0$) par simulation.

2.4.2 Test d'identification de l'effet

Si le test 1 conduit à la détection d'un effet (rejet de $\mathcal{H}_0$), il nous faut alors identifier cet effet. Il peut s'agir d'un effet Laboratoire seul (modèle $L\bar{F}$), d'un effet Flacon seul (modèle F) ou d'un effet Laboratoire et un effet Flacon simultanément (modèle LF). Pour déterminer dans quel configuration nous nous trouvons, nous allons comparer un modèle avec effet Laboratoire seul avec un modèle avec effet Laboratoire et Flacon.

Les hypothèses du test 2 sont les suivantes :

$$\begin{cases} \mathcal{H}_0 : L\bar{F} : y_{ijk} \sim P(\lambda_i) \\ \mathcal{H}_1 : LF : y_{ijk} \sim P(\lambda_{ij}). \end{cases}$$

Sous $\mathcal{H}_0$, la vraisemblance $L_{L\bar{F}}$ du modèle $L\,\bar{F}$ s'écrit :

$$L_{L\bar{F}}(y_{111},\ldots,y_{abn},\lambda_1,\ldots,\lambda_a) = \prod_{i=1}^{a}\prod_{j=1}^{2}\prod_{k=1}^{2}\hat{\lambda}_i^{y_{ijk}}\frac{e^{-\hat{\lambda}_i}}{y_{ijk}!}. \tag{2.12}$$

Sous $\mathcal{H}_1$, la vraisemblance L_{LF} du modèle LF est donnée par l'expression (2.7).
D'après les expressions (2.12),(2.7), (2.4) et (2.5), le rapport de vraisemblance Λ_2 associé au test 2 est alors :

$$\Lambda_2 = \frac{L\bar{F}}{LF} = \prod_{i=1}^{a}\prod_{j=1}^{2} e^{-2(\bar{y}_{i..}-\bar{y}_{ij.})}\left(\frac{\bar{y}_{i..}}{\bar{y}_{ij.}}\right)^{2\bar{y}_{ij.}}. \tag{2.13}$$

A partir du rapport de vraisemblance (2.13), on obtient la déviance D_2 suivante :

$$D_2 = -2\log(\Lambda_3) = -4\sum_{i=1}^{a}\sum_{j=1}^{2}\bar{y}_{ij.}\log\left(\frac{\bar{y}_{i..}}{\bar{y}_{ij.}}\right). \tag{2.14}$$

Comme pour le test 1, 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_2 du jeu de données est indéfinie. La déviance D_2 peut alors s'écrire :

$$D_2 = -4\sum_{i=1}^{a}\sum_{\substack{j=1 \\ \forall\, j\,/\,\bar{y}_{ij.}\neq 0}}^{2}\bar{y}_{ij.}\log\left(\frac{\bar{y}_{i..}}{\bar{y}_{ij.}}\right). \tag{2.15}$$

La loi de la déviance D_2 pourra être approchée par une loi du χ^2 à a degrés de liberté pour $\lambda \geq 10$. Pour des valeurs de λ plus faibles, nous approcherons la loi de la déviance (sous $\mathcal{H}_0$) par simulation.

2.4.3 Cas de l'envoi de deux flacons différents

Lorsqu'AGLAE envoi deux flacons différents, les deux flacons seront identifiés. On peut alors estimer les paramètres λ_1 et λ_2 associés aux deux flacons en faisant une moyenne sur les mesures effectuées par les laboratoires sur le flacon 1 (ou le flacon 2) :

$$\hat{\lambda}_1 = \bar{y}_{.1.} = \bar{y}_{i1k} = \frac{\sum_{i=1}^{a}\sum_{k=1}^{2} y_{i1k}}{2a}, \quad (2.16) \qquad \hat{\lambda}_2 = \bar{y}_{.2.} = \bar{y}_{i2k} = \frac{\sum_{i=1}^{a}\sum_{k=1}^{2} y_{i2k}}{2a}. \quad (2.17)$$

Remarque : On considère les facteurs Laboratoire et Flacon imbriqués, mais ils ne le sont plus dans ce cas, car le flacon 1 est le même pour tous les laboratoires, et le flacon 2 aussi.
Les hypothèses du test 3 sont les suivantes :

$$\begin{cases} \mathcal{H}_0 : \bar{LF} : y_{ijk} \sim \frac{1}{2}P(\lambda_1) + \frac{1}{2}P(\lambda_2) \\ \mathcal{H}_1 : LF : y_{ijk} \sim P(\lambda_{ij}). \end{cases}$$

Sous $\mathcal{H}_0$, la vraisemblance $L_{\bar{LF}}$ du modèle $\bar{LF}$ s'écrit :

$$L_{\bar{LF}}(y_{111}, \ldots, y_{abn}, \lambda_1, \lambda_2) = \prod_{i=1}^{a}\prod_{j=1}^{2} \frac{e^{-2\hat{\lambda}_j}}{\prod_{k=1}^{2} y_{ijk}!} \hat{\lambda}_j^{\sum_{k=1}^{2} y_{ijk}}. \quad (2.18)$$

Sous $\mathcal{H}_1$, la vraisemblance L_{LF} du modèle LF est donnée par l'expression (2.7).
D'après les expressions (2.18),(2.7), (2.16) et (2.17), le rapport de vraisemblance Λ_3 associé au test 3 est alors :

$$\Lambda_3 = \frac{\bar{LF}}{LF} = \prod_{i=1}^{a}\prod_{j=1}^{2} e^{-2(\bar{y}_{.j.} - \bar{y}_{ij.})} \left(\frac{\bar{y}_{.j.}}{\bar{y}_{ij.}}\right)^{2\bar{y}_{ij.}}. \quad (2.19)$$

A partir du rapport de vraisemblance (2.19), on obtient la déviance D_3 suivante :

$$D_3 = -2\log(\Lambda_3) = -4\sum_{i=1}^{a}\sum_{j=1}^{2} \bar{y}_{ij.} \log\left(\frac{\bar{y}_{.j.}}{\bar{y}_{ij.}}\right). \quad (2.20)$$

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_3 du jeu de données est indéfinie. La déviance D_3 peut alors s'écrire :

$$D_3 = -4\sum_{i=1}^{a} \sum_{\substack{j=1 \\ \forall\, j \,/\, \bar{y}_{ij.} \neq 0}}^{2} \bar{y}_{ij.} \log\left(\frac{\bar{y}_{.j.}}{\bar{y}_{ij.}}\right). \quad (2.21)$$

La loi de la déviance D_3 pourra être approchée par une loi du χ^2 à $2a-2$ degrés de liberté pour $\lambda \geq 10$. Pour des valeurs de λ plus faibles, nous approcherons la loi de la déviance (sous $\mathcal{H}_0$) par simulation.

2.5 Applications sur données réelles

Nous avons appliqué les tests 1 et 2 proposés ci-dessus à des jeux de données d'essais inter-laboratoires en microbiologie. Nous présentons dans cette section les résultats obtenus sur deux essais.
Le premier jeu de données étudié correspond à un essai interlaboratoires de dénombrement de *Pseudomonas aeruginosa* dans les eaux. Ce jeu de données de 202 laboratoires présente une large dispersion interlaboratoires et peu de dispersion interflacons. En appliquant les tests 1 et 2 à ce jeu de données, nous obtenons les résultats suivants :

- Test 1 : Détection d'un effet
  p-valeur=0 : rejet de H_0 au risque $\alpha = 5\%$: Il existe un effet
- Test 2 : Identification de l'effet
  p-valeur=0.876 : non rejet de H_0 : l'effet n'est dû qu'au laboratoire

Nous comparons cette méthode avec la méthode de test présentée au chapitre 1, correspondant à une analyse de la variance sur les données ayant préalablement subi une transformation logarithmique. Les deux tests des effets Laboratoire et Flacon présentés en section 1.5.3 nous permettent d'aboutir aux conclusions suivantes :

- Test de l'effet Laboratoire :
  $F_{0.95,186,187} = 1,26$,
  $F_0 = 3,7 > 1,26$, donc rejet de H_0 : Il y a un effet Laboratoire
- Test de l'effet Flacon :
  $F_{0.95,187,374} = 1,22$,
  $F_0 = 1,82 > 1,22$, donc rejet de H_0 : Il y a un effet Flacon

Nous constatons que la méthode que nous proposons dans ce chapitre nous amène à conclure à l'absence d'hétérogénéité de lot, contrairement à la méthode utilisée en microbiologie.

Le deuxième jeu de données étudié ici correspond à un essai interlaboratoires de dénombrement de spores auquel ont participé 103 laboratoires. Chaque laboratoire participant à cet essai a reçu deux flacons différents. Nous allons donc ici appliquer le test 3 défini ci-dessus pour identifier un éventuel effet Laboratoire.

- Test 3 :
  p-valeur=$0,007$: rejet de H_0 : Il y a un effet Laboratoire

Nous allons maintenant utiliser le test de l'effet Flacon actuellement utilisé en microbiologie

- Test de détection de l'effet Laboratoire :
  $F_{0.95,102,103} = 1,39$
  $F_0 = 4,24 > 1,39$, donc rejet de H_0 : Il y a un effet Laboratoire

Pour le second jeu de données étudié, les deux méthodes nous conduisent à conclure à l'existence d'un effet Laboratoire.

La comparaison des méthodes de test proposées avec les méthodes de test présentées au chapitre 1 sur des jeux de données réels fait apparaitre des différences mais ne permet pas de déterminer quelle méthode est la plus puissante. Pour cela, nous allons travailler sur des jeux de données simulés.

2.6 Étude de la puissance des tests proposés

2.6.1 Test de détection d'un effet quelconque

Pour évaluer les tests proposés, nous comparons leurs puissances à celle de la méthode utilisée en microbiologie, qui correspond à des tests d'analyse de variance sur des données normalisées. Nous présentons ici les courbes de puissance du test 1 pour $\lambda = 1$ et $\lambda = 15$. Nous effectuons 1000 tests, sur des données simulées sous $\mathcal{H}_1$ ($L\,\bar{F}$). Pour introduire un effet Laboratoire, nous simulons les résultats des mesures de 14 laboratoires suivant $\mathrm{P}(\lambda)$, et celles d'un laboratoire suivant $\mathrm{P}(\lambda + \delta)$. Puis nous calculons la puissance du test 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 $\mathcal{H}_0$ (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 2.1*, 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.

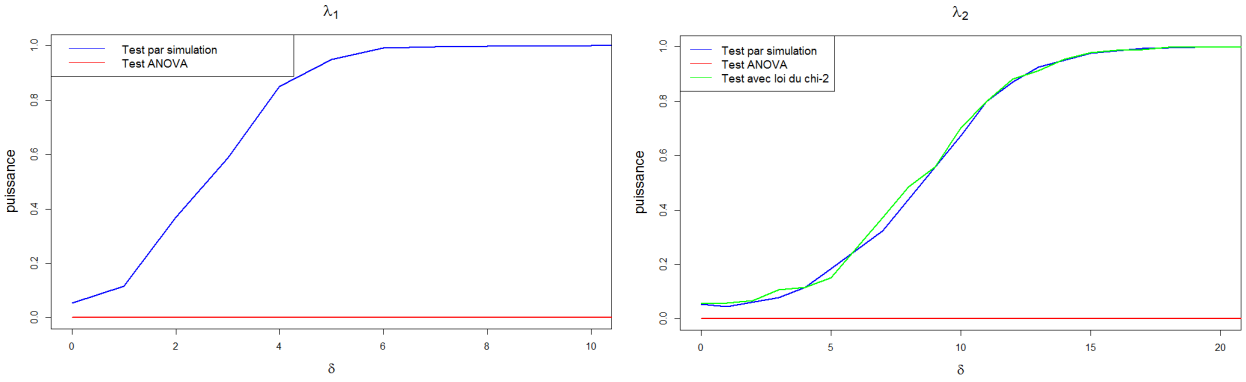

FIGURE 2.1 – Comparaison des puissances des trois méthodes pour le test 1

2.6.2 Test d'identification de l'effet

Pour évaluer le test 2, nous comparons sa puissance à celle de la méthode d'analyse de variance sur données log-transformées. Pour cela, nous simulons un jeu de données avec un effet Laboratoire et un effet Flacon. Nous considérons un jeu de données de 15 laboratoires. Nous simulons les résultats de dénombrements de 14 des laboratoires suivant un modèle à effet Laboratoire seul, dont le paramètres de la loi de Poisson est : $\lambda_{ij} = i$. Nous introduisons un effet Laboratoire ainsi qu'un effet Flacon sur le laboratoire restant. Les paramètres des lois de Poisson associées aux deux flacons de ce laboratoire sont alors : $\lambda_{i1} = i$ pour le laboratoire 1 et $\lambda_{i2} = i + \delta$ pour le laboratoire 2. Puis nous calculons la puissance du test 1 pour différentes valeurs de δ. Nous alors comparons la puissance du test 1 basé sur des simulations de la loi de la déviance sous $\mathcal{H}_0$ (courbe bleue, 10^4 simulations par p-valeur), avec celle du test d'ANOVA (courbe rouge), avec le test où la loi de la déviance est assimilée à une distribution du χ^2 à a degrés de liberté

(courbe verte). Sur la *Figure* 2.2, nous constatons que la méthode proposée ici est beaucoup plus puissante que la méthode d'ANOVA sur données log-transformées. En revanche, le test basé sur la loi du χ^2 donne des résultats très proches de ceux obtenus avec le test par simulation.

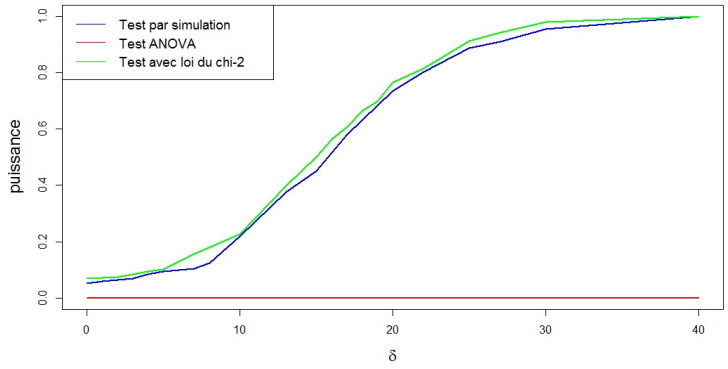

FIGURE 2.2 – Comparaison des puissances des trois méthodes pour le test 2

2.7 Conclusion

Dans ce chapitre, nous proposons un modèle linéaire à deux facteurs imbriqués, similaire à celui présenté au chapitre 1 section 1.5.1, pour expliquer la variabilité des résultats de dénombrement d'un essai interlaboratoires en microbiologie. La particularité de ce modèle par rapport à ce qui a été présenté au chapitre 1 est que nous nous attachons ici à travailler directement sur données discrètes non transformées.

Nous écrivons dans un premier temps les tests de significativité des facteurs pour ce modèle de façon générale, en soulignant la différence avec un modèle à données gaussiennes.

Dans un deuxième temps, nous présentons la méthode d'analyse de déviance, qui permet une généralisation de l'analyse de variance en termes de distribution. Nous proposons alors une stratégie de test des effets d'un essai interlaboratoires, basée sur l'analyse de déviance. Nous avons mis en évidence la pertinence de ces tests, en comparant leurs puissances à celles des tests d'analyse de variance sur données log-transformées.

Nous avons néanmoins souligné dans ce chapitre que le modèle à facteurs imbriqués proposé ne permet pas de tester l'effet Flacon seul (hétérogénéité de lot). Aussi, dans le cadre d'un essai interlaboratoires, nous souhaitons évaluer la dispersion des résultats de mesure attribuable à chacun des facteurs. Le modèle à facteurs fixes imbriqués semble alors présenter des limites. C'est pourquoi nous considérons à un modèle plus complexe, à facteurs aléatoires, au chapitre suivant.

Chapitre 3

Modèle Linéaire Généralisé Hiérarchique Gamma-Poisson

Sommaire

3.1 Introduction

Dans le chapitre 1, nous avons mis en évidence les limites de la méthode actuellement utilisée en microbiologie pour expliquer la variabilité des résultats de dénombrement d'un essai interlaboratoires. Ces limites sont notamment imputables à l'application préalable d'une transformation logarithmique aux données. La méthodologie d'analyse de variance n'est effectivement pas appropriée à l'analyse de données discrètes telles que les données de comptage. Certains auteurs, notamment [O'Hara and Kotze, 2010] et [Maindonald and Braun, 2006], affirment que les modèles linéaires généralisés ont fait disparaitre la nécessité d'appliquer une transformation aux données, et permettent des analyses directes et plus performantes de telles données.

Par ailleurs, bien que le dénombrement de particules dans une phase homogène soit idéalement représenté par la loi de Poisson, on observe régulièrement une surdispersion par rapport au modèle de Poisson [BCR, 1993]. Cette réalité confère à notre étude des contraintes sur la distribution de la variable à expliquer, mais aussi sur la nature des différents facteurs modélisant les sources de la surdispersion.

Nous proposons dans ce chapitre un modèle linéaire généralisé hiérarchique Gamma-Poisson afin d'expliquer la variabilité des résultats de dénombrement en microbiologie à l'aide de trois facteurs aléatoires. Une partie de ce travail a fait l'objet d'une publication lors des 47$^{\text{èmes}}$ Journées de Statistique [Loingeville et al., 2015].

Avant de présenter en détail le modèle adopté, nous revenons sur la classe des modèles que nous étudions. Les modèles linéaires généralisés hiérarchiques constituent en effet un ensemble de modèles s'inscrivant dans une démarche générale de modélisation, que nous retraçons en section 3.2. Nous présentons les hypothèses adoptées, ainsi que les termes employés dans la suite de ce document, et nous positionnons ce travail par rapport aux travaux réalisés dans le domaine. Nous adoptons la notation HGLM (Hierarchical Generalized Linear Models) pour désigner ces modèles. En section 3.3, nous présentons l'expression de la surdispersion en microbiologie en termes statistiques, et nous proposons finalement un modèle linéaire généralisé hiérarchique Gamma-Poisson permettant d'expliquer la variabilité des résultats de dénombrement, tout en prenant en compte cette surdispersion. Nous détaillons la procédure d'estimation des paramètres de ce modèle en section 3.4. Puis, nous étudions l'adéquation du modèle aux données issues des essais interlaboratoires en microbiologie en section 3.5. Nous proposons également des tests de significativité des effets des différents facteurs (section 3.6), ainsi qu'une méthode de scoring basée sur le modèle proposé, permettant d'évaluer la performance analytique des laboratoires qui participent à un essai dans le domaine de la microbiologie (section 3.7).

3.2 Les modèles linéaires généralisés hiérarchiques

3.2.1 Introduction

C'est au début du 19$^{\text{ème}}$ siècle, sous l'impulsion de Legendre et Gauss, que sont apparus les modèles linéaires classiques. Les données alors analysées sont issues du domaine de l'astronomie, et correspondent essentiellement à des quantités continues. La première extension de ce modèle simple fut à un modèle avec plusieurs termes d'erreurs. C'est ainsi que sont apparus les modèles linéaires mixtes (L2M), présentés dans [Henderson et al., 1959] . La seconde extension des modèles linéaires correspond aux modèles linéaires généralisés (GLM), introduits par [Nelder and Wedderburn, 1972]. Dans ces modèles, la classe de distribution est étendue aux distributions de la famille exponentielle à un paramètre, et l'additivité des effets peut être vérifiée par une transformation monotone de la moyenne (la fonction de lien), plutôt que sur la moyenne elle-même.

Ces deux extensions des modèles linéaires classiques furent ensuite combinées pour produire une classe de modèles linéaires généralisés mixtes (GL2M), dans lesquels le prédicteur linéaire introduit dans les GLM peut avoir, outre les effets fixes usuels, une ou plusieurs composantes aléatoires supposées gaussiennes.

Dans cette thèse, nous nous intéressons à la classe des modèles linéaires généralisés hiérarchiques introduits par [Lee and Nelder, 1996], dans lesquels la composante aléatoire provient d'une distribution de la famille exponentielle.

3.2.2 Les modèles linéaires mixtes

Nous nous intéressons dans un premier temps à la notion d'effets aléatoires comme outil de modélisation. Nous présentons notamment l'intérêt de l'introduction des tels effets pour la caractérisation de la variabilité des résultats de dénombrements d'un essai interlaboratoires en microbiologie. Puis nous verrons comment l'introduction d'effets aléatoires dans les modèles linéaires classiques a donné naissance aux modèles linéaires mixtes.

La modélisation avec effets aléatoires. Les résultats de dénombrement d'un essai interlaboratoire présentent une certaine variabilité. Le but de notre étude réside dans l'analyse de celle-ci. Nous souhaitons en effet caractériser cette variabilité, l'évaluer, et en déterminer les sources. Comme nous l'avons présenté dans le chapitre 1, la méthodologie d'analyse de la variance introduite par Fisher permet une telle étude dans le cas de données gaussiennes. Elle permet en effet de décomposer la variance totale des résultats de mesures en une variance interlaboratoire et une variance intralaboratoire, laquelle pouvant elle-même être décomposée en une variance interéchantillon (écart entre A et B sur la Figure 1.1) et intraéchantillon (écart A1, A2 ...). L'objectif est alors d'évaluer la significativité des différences entre les mesures sur la base de cette décomposition de la variance totale.

Comme souligné dans le chapitre 1, les facteurs du modèle d'analyse de variance utilisé sur données gaussiennes sont aléatoires. Les modèles à effets aléatoires constituent en effet un moyen plus élaboré d'étudier la variabilité des données. L'introduction de ces effets dans la modélisation nous permet de préciser les différentes sources de variation. Pour apporter une définition d'un effet aléatoire, nous allons opposer les deux natures possibles des effets : *fixe* ou *aléatoire*. Au cours d'une expérience, différents facteurs sont soupçonnés affecter les résultats de l'expérience, et donc les valeurs de la variable observée. Les données relevées peuvent alors être classées selon les différents niveaux de ces facteurs. On distingue deux types de facteurs :

- Les facteurs à *effets fixes*, qui ont un nombre fini de niveaux. Les données sont observées pour chacun de ces niveaux. Le but est alors de caractériser l'effet de chaque niveau sur la variable à expliquer.

- Les facteurs à *effets aléatoires*, présentant un nombre infini de niveaux. L'échantillon considéré étant de taille finie, les données observées ne peuvent se répartir sur tous les niveaux. Seul un échantillon de niveaux est considéré. Dans ce cas, l'influence des niveaux sur la variable à expliquer ne présente pas d'intérêt. Le but est de connaître la part de variabilité induite par cet effet à partir de l'échantillon de niveaux considéré.

La variabilité totale peut ainsi être séparée en une variabilité due aux effets aléatoires, et une variabilité attribuée aux erreurs. On est donc plus précis quant à l'origine de la variabilité puisqu'on l'explique à l'aide de différentes composantes.

Facteurs de variabilité des résultats de dénombrements pour les EIL en microbiologie. Dans le cas des essais interlaboratoires en microbiologie, plusieurs facteurs peuvent expliquer la dispersion des résultats de dénombrement.

- Le facteur *Laboratoire* : La technique employée par le laboratoire pour effectuer le dénombrement (méthode utilisée, milieu de culture, etc) peut influencer le résultat de la mesure. Mais ce sont aussi l'expérience du manipulateur et le rendement de récupération (qui n'est jamais de 100%) qui expliquent cette surdispersion.
- Le facteur *Flacon* : l'hypothèse d'homogénéité du milieu échantillonné peut ne pas être vérifiée. En effet, l'agglomération sur les parois du contenant (flacon ou pipette) ou sur les particules présentes, ainsi que la formation de chaînes ou d'amas bactériens susceptibles de se disloquer induisent une dérive par rapport à la loi de Poisson.
- Le facteur *Réplication* : la variabilité des conditions de réalisation du dénombrement indirect des volumes élémentaires contenant l'analyte recherché constitue une source d'erreur. Il existe en effet toujours une certaine incertitude sur le plan métrologique (volumétrie, température).

Remarque 3.2.1. *A noter que la nature de l'erreur associée au facteur Réplication est "aléatoire" au sens métrologique du terme, alors que celle de l'erreur associée au facteur Laboratoire est systématique.*

Certes, l'un des objectifs de la mise en place d'un essai interlaboratoire est l'évaluation de la performance individuelle de chacun des laboratoires participant à l'essai, mais, de façon plus générale, on souhaite aussi caractériser la part de variabilité induite par chacun des trois effets présentés ci-dessus sur l'ensemble des résultats de l'essai. C'est pourquoi il nous paraît plus approprié de considérer ces trois facteurs comme aléatoires dans la suite de ce travail.

Modèle à effets aléatoires et modèle avec surdispersion. Nous avons souligné en section 1.3.3 que la dispersion des résultats de mesure des essais interlaboratoires en microbiologie était fréquemment supérieure à celle attendue d'après le modèle de Poisson. Dans ce contexte, notre objectif est de construire un modèle qui permette de mettre en évidence les sources de variation des résultats de mesure, et d'identifier les variations induites par la présence des effets aléatoires identifiés. La variation supplémentaire par rapport à la loi de Poisson sera alors modélisée par des effets aléatoires. Il ne s'agit alors plus de surdispersion au sens statistique du terme.

Nous verrons dans le paragraphe suivant (3.2.3) que les modèles à effets aléatoires diffèrent des modèles avec surdispersion, apparus avec les GLM.

Les modèles linéaires mixtes. Les modèles linéaires classiques ne contenant que des effets fixes ont été élargis par l'ajout d'effets aléatoires. Il s'agit alors de modèles linéaires mixtes, notés L2M. Une partie aléatoire vient alors s'ajouter à la partie fixe de la façon suivante :

$$Y = X\psi + U\zeta + \epsilon,$$

où

- Y : vecteur aléatoire à expliquer de taille n,
- X : matrice de design des effets fixes, de taille $N \times p$, supposée connue,
- ψ : vecteur de paramètres inconnus des effets fixes, de taille p,

- ζ : vecteur d'effets aléatoires de taille q. Ce vecteur peut se décomposer en S parties $\zeta = (\zeta_1', \ldots, \zeta_S')'$ où S est le nombre d'effets aléatoires considérés dans le modèle. Chaque composante ζ_j est un vecteur aléatoire de dimension q_j. Il est constitué de q_j réalisations du $j^{\text{ième}}$ effet aléatoire, observées au sein des données ($\sum_{j=1}^{S} q_j = q$).

  Les effets aléatoires sont supposés gaussiens. Par ailleurs, $\forall i, j \in \{1, \ldots, S\}^2$, ζ_i et ζ_j sont indépendants.

- U : matrice de design connue, formée des différentes matrices de design U_j de chaque effet aléatoire : $U = [U_1 \ldots U_S]$

- ϵ : vecteur aléatoire d'erreurs de taille N. Puisque le modèle considéré est linéaire, la distribution de ϵ est gaussienne.

3.2.3 Les modèles linéaires généralisés

Au début du 20$^{\text{ème}}$ siècle, l'analyse de données discrètes, telles que les données de comptage, s'est développée. Au panel des distributions disponibles pour la modélisation sont alors venues s'ajouter les distributions de Poisson ou binomiale. La famille exponentielle permet de regrouper toutes ces lois et donne naissance à une nouvelle classe de modèles. Il s'agit des modèles linéaires généralisés (GLM). Cette terminologie est introduite par [Nelder and Wedderburn, 1972]. Comme son nom l'indique, cette classe de modèles généralise les modèles linéaires classiques. Il s'agit d'une généralisation en termes de loi de probabilité d'une part, mais aussi en termes de lien à la linéarité. L'hypothèse sur la distribution associée à chaque modélisation est alors remplacée par une propriété de linéarité commune à tous les modèles, et par une relation espérance-variance.

La classe des modèles linéaires généralisés permet l'analyse des données non gaussiennes, et notamment des données discrètes.

Un modèle linéaire généralisé est caractérisé par trois hypothèses : une hypothèse sur la distribution de la variable à expliquer, une hypothèse sur l'expression de la linéarité (faisant intervenir les variables explicatives), et une hypothèse sur le lien à la linéarité (c'est-à-dire le lien entre la variable réponse et les variables explicatives).

Distribution de la variable à expliquer. Soit Y la variable aléatoire que l'on cherche à expliquer et y le vecteur de taille N des observations. On suppose que les composantes Y_i ($i = 1, \ldots, N$) de Y sont indépendantes et identiquement distribuées selon une loi appartenant à la famille exponentielle [Nelder and Wedderburn, 1972]. La fonction densité de la variable aléatoire Y_i s'écrit :

$$f_{Y_i}(y_i, \theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{3.1}$$

où θ_i est un paramètre canonique et ϕ un paramètre de surdispersion. Les fonctions b et c sont spécifiques à chaque distribution et la fonction a_i s'écrit $a_i(\phi) = \frac{\phi}{\omega_i}$ où ω_i est un poids connu associé à l'observation i (différent de 1 lorsque les données ont été groupées).

La famille de lois exponentielle regroupe un certain nombre de lois dont les lois classiques telles que la loi binomiale, la loi de Poisson, la loi normale, la loi Gamma.

Un tableau présentant le paramètre ϕ ; l'expression du paramètre canonique θ_i en fonction des paramètres naturels de la loi, et les fonctions b et a_i associées est disponible dans [McCullagh and Nelder, 1989].

L'espérance et la variance de chacune de ces lois s'expriment à l'aide des fonctions a_i et b. En effet, notons $L(\theta; y) = \ln(f_Y(y\theta))$, la fonction de log-vraisemblance. Les relations classiques suivantes :

$$\begin{cases} \mathbb{E}(\frac{\partial L}{\partial \theta}) = 0 \\ \mathbb{E}(\frac{\partial^2 L}{\partial \theta^2}) + E((\frac{\partial L}{\partial \theta})^2) = 0 \end{cases},$$

nous permettent d'obtenir l'espérance de Y_i, notée m_i, et sa variance, comme dans [McCullagh and Nelder, 1989] :

$$\mathbb{E}(Y_i) = m_i = b'(\theta_i),$$
$$\mathbb{V}(Y_i) = b''(\theta_i)a_i(\phi).$$

Il existe donc une relation directe entre l'espérance de Y_i et sa variance :

$$\mathbb{V}(Y_i) = a_i(\phi)b''(b'^{-1}(m_i)) = \frac{\phi}{\omega_i}b''(b'^{-1}(m_i)).$$

On constate alors que dans le cas où ϕ est connu, la variance des observations est contrainte à être fonction de l'espérance, autrement dit, la connaissance de m_i implique celle de la variance de Y_i. Cette propriété essentielle des GLM n'était pas valable pour les modèles linéaires classiques.

Pour les familles exponentielles, le paramètre ϕ peut être connu (c'est le cas pour les lois binomiale, de Poisson et exponentielle), ou inconnu (lois normale et gamma).

Expression de la linéarité. Pour les GLM, les variables explicatives interviennent linéairement dans la modélisation, comme c'est le cas pour les modèles linéaires. On définit ainsi le prédicteur linéaire :

$$\eta = X\psi,$$

où ψ est un vecteur de paramètres inconnus de taille p, et X une matrice $N \times p$ connue.

Fonction de lien. Le lien entre la $i^{\text{ème}}$ composante de ce prédicteur linéaire et l'espérance de Y_i s'établit par l'intermédiaire d'une fonction g monotone et deux fois dérivables, appelée fonction de lien :

$$\eta_i = g(m_i). \tag{3.2}$$

La fonction de lien permettant d'égaler le prédicteur linéaire et le paramètre canonique est appelée fonction de lien canonique. Puisque $\eta_i = g(b'(\theta_i))$, la fonction de lien canonique associée à une distribution donnée sera $g = b'^{-1}$. Les fonctions de lien canoniques associées aux lois classiques sont indiquées dans [McCullagh and Nelder, 1989].

GLM avec surdipersion. Revenons sur la notion de surdispersion et la différence entre les modèles avec surdispersion et les modèles avec effets aléatoires. La surdispersion permet de modéliser une éventuelle variabilité supplémentaire des données par rapport à un choix de distribution. Elle intervient dans les GLM par l'intermédiaire du paramètre de surdispersion ϕ, qui permet de moduler la loi considérée. Si l'on souhaitait traduire la surdispersion en termes d'effet aléatoire, cela reviendrait à introduire un nouvel effet, présentant autant de réalisations que de données. Il est important de souligner la différence d'une telle modélisation avec les effets aléatoires classiques introduits dans la section 3.2.2. Deux arguments justifient cette différence. D'une part,

l'effet surdispersion ainsi introduit ne permet pas d'identifier la source de la variabilité supplémentaire. D'autre part, les composantes de ζ étant supposées indépendantes, le fait d'avoir une réalisation par donnée permet de conserver marginalement une indépendance des composantes Y_i de Y. Au contraire, les effets aléatoires que nous introduisons induisent une dépendance entre le vecteur de l'effet surdispersion introduit et les données.

En passant d'un modèle de Poisson surdispersé à un modèle à plusieurs effets aléatoires, nous caractérisons la surdispersion à l'aide de ces effets aléatoires.

Déviance. Si l'on note $\hat{\theta} = \theta(\hat{y})$ et $\tilde{\theta} = \theta(y)$ les estimateurs du paramètre canonique θ respectivement pour le GLM proposé et pour les données, en supposant que $a_i(\theta) = \phi/w_i$, la divergence entre le GLM (3.1) et les données auxquelles il est ajusté peut s'écrire :

$$\sum 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i)\}/\phi = D(y;\hat{y})/\phi, \tag{3.3}$$

où $\hat{y}$ correspond au vecteur de données ajustées suivant le modèle postulé, et $D(y;\hat{y})$ est la déviance du modèle et s'exprime simplement en fonction des données. On note

$$D^*(y;\hat{y}) = D(y;\hat{y})/\phi. \tag{3.4}$$

la déviance normalisée, qui correspond à la déviance exprimée comme un multiple du paramètre de dispersion ϕ.

3.2.4 Les modèles linéaires généralisés mixtes

Les GLM et les L2M ont ensuite été regroupés pour donner naissance à une nouvelle classe de modèles, les modèles linéaires généralisés mixtes, notés GL2M.

Hypothèses. De la même façon que des effets aléatoires gaussiens ont été introduits dans les LM (donnant naissance à la classe de modèles L2M), ils peuvent l'être au sein des GLM. La nouvelle classe de modèle ainsi obtenue est celle des GL2M. Une partie aléatoire vient alors s'ajouter à la partie fixe dans l'expression du prédicteur linéaire des GLM 3.2.
Le GL2M conserve toutes les propriétés du GLM.
En gardant les notations de la section précédente, le prédicteur s'exprime de la façon suivante :

$$\eta_\zeta = X\psi + U\zeta.$$

Dans cette nouvelle classe de modèles, les effets aléatoires ζ sont, comme pour les effets aléatoires introduits dans les L2M, supposés gaussiens. Le GL2M est défini dans un raisonnement conditionnel à ζ. Les composantes de Y sont, conditionnellement à ζ, indépendantes et de loi appartenant à la famille exponentielle.
De même que pour les GLM, nous pouvons définir pour les GL2M les fonctions de lien et de variance, mais cette fois ci dans un raisonnement conditionnel à ζ. L'espérance conditionnelle de Y est reliée au prédicteur linéaire par la fonction de lien :

$$\eta_\zeta = g(m_\zeta),$$

où

$$m_\zeta = E(Y|\zeta).$$

La fonction de variance V intervient quant à elle dans l'expression de la variance conditionnelle :

$$\forall i \in \{1, \ldots, N\}, \mathbb{V}(Y_i|\zeta) = a_i(\phi)V(m_{\zeta,i}).$$

Enfin, c'est sur la loi de Y conditionnelle à ζ qu'est formulée l'hypothèse de distribution. D'une part, on suppose que, conditionnellement à ζ, les composantes de Y sont indépendantes, on obtient donc la matrice de variance conditionnelle suivante : D'autre part, $\forall i \in \{1, \ldots, N\}, Y_i|\zeta$ est supposé distribué selon une loi issue de la famille exponentielle.
Ainsi, conditionnellement à ζ, le GL2M conserve toutes les propriétés du GLM. Par conséquent, le GL2M se trouve principalement défini dans un raisonnement conditionnel à ζ. Il peut être résumé de la façon suivante :

- Les composantes de Y sont, conditionnellement à ζ, indépendantes et de loi appartenant à la famille exponentielle,
- le prédicteur linéaire s'écrit : $\eta_\zeta = X\psi + U\zeta$,
- l'espérance conditionnelle de Y est reliée au prédicteur linéaire par la fonction de lien $\eta_\zeta = g(m_\zeta)$.

3.2.5 Les modèles linéaires généralisés hiérarchiques

La classe des GL2M a finalement été étendue par [Lee and Nelder, 1996] à une classe de modèles linéaires généralisés hiérarchiques, notés HGLM, dans lesquels les effets aléatoires ne sont pas nécessairement gaussiens, mais proviennent d'une distribution de la famille exponentielle. Nous conservons ici l'ensemble des notations définies dans les sections précédentes. Notons également u le vecteur des effets aléatoires considérés. Les HGLM vérifient les hypothèses suivantes :

- Conditionnellement aux effets aléatoires u, l'espérance du vecteur y de taille N des observations (réalisations de la variable aléatoire à expliquer) conditionnellement aux effets aléatoires u s'exprime de la façon suivante :

$$\begin{aligned} \mathbb{E}(y|u) &= m \\ \mathbb{V}(y|u) &= \phi V(m), \end{aligned}$$

  et la log-vraisemblance de y conditionnellement à u s'écrit :

$$l(\theta, \phi; y|u) = \frac{\{y\theta - b(\theta)\}}{a(\phi)} + c(y, \phi), \tag{3.5}$$

  où $\theta = \theta(u)$ est le paramètre canonique définissant la distribution. Le prédicteur linéaire prend alors la forme suivante :

$$\eta = g(m) = X\psi + Z\nu,$$

  où $\nu = \nu(u)$ est une fonction monotone de u, u correspondent aux effets aléatoires, et ψ correspond au vecteur d'effets fixes.

- Le vecteur d'effets aléatoires u suit une distribution conjuguée à celle de la réponse, de paramètre λ [Lee and Nelder, 1996].

La différence avec les GL2M réside dans l'hypothèse sur la distribution des effets aléatoires u, qui peuvent suivre une distribution autre que normale.

Le HGLM Gamma-Poisson On suppose que $y|\nu$ suit une distribution de Poisson dont l'espérance s'écrit :

$$m = \mathbb{E}(y|u) = \exp(X\psi)u,$$

En introduisant la fonction de lien logarithmique, on obtient :

$$\eta = \log m = X\psi + \nu,$$

où $\nu = \log u$. Si u suit une distribution Gamma, ν suit une distribution log-gamma, et le modèle correspond alors à un HGLM Gamma-Poisson. Les GL2M supposent en revanche une distribution gaussienne sur ν, autrement dit une distribution log-normale de u. Sous la paramétrisation $\nu = \log u$, le GL2M de Poisson correspondant serait alors un HGLM log-normal-Poisson. Pour un HGLM à un facteur aléatoire, la distribution normale pour u est la distribution conjuguée à la distribution normale de $y|m$, tandis que la distribution Gamma est la distribution conjuguée à la distribution de Poisson. Dans les HGLM, la distribution de u n'est pas nécessairement la distribution conjuguée à celle de $y|u$. Si c'est le cas, le modèle est un HGLM conjugué. Ainsi, le modèle Gamma-Poisson ainsi que le GL2M Poisson appartiennent tous les deux à la catégorie des HGLM, mais seul le modèle Gamma-Poisson est un HGLM conjugué.

3.3 Expression de la dispersion en microbiologie

Le dénombrement de particules dans une phase homogène est idéalement représenté par une loi de Poisson. En pratique, il s'avère que la dispersion des résultats de dénombrements de germes est fréquemment supérieure à celle attendue d'après le modèle de Poisson
[Tillett and Lightfoot, 1995].
Dans cette section, nous présentons la distribution statistique utilisée pour modéliser la surdispersion, ainsi que sa paramétrisation spécifique à la microbiologie. Nous écrivons finalement un modèle linéaire généralisé hiérarchique utilisant ces notations pour répondre au problème.

3.3.1 Représentation de la loi de Poisson surdispersée

En microbiologie, et notamment dans [ISO, 2000] et [Bliss and Fisher, 1953], la loi de Poisson surdispersée est représentée par la loi binomiale négative. La loi binomiale négative dépend de deux paramètres, et plusieurs paramétrisations sont envisageables. La paramétrisation classique introduit un entier naturel n non nul et un réel p compris entre 0 et 1. Il est également courant d'introduire la probabilité complémentaire $q = 1-p$. La loi de probabilité d'une variable aléatoire X distribuée selon une binomiale négative de paramètres n et p, notée $\mathcal{BN}(n,p)$, prend la forme suivante :

$$P(X = k; n, p) = \binom{k+n-1}{k} p^n q^k \qquad \text{pour } k \in \mathbb{N}.$$

La loi binomiale négative correspond à la loi de probabilité de la variable aléatoire X qui comptabilise le nombre d'échecs nécessaires avant obtention de n succès, où la probabilité d'un succès est p :

$$P(X = k) = \binom{k+n-1}{n-1} p^n q^k.$$

Les paramètres de la loi $\mathcal{BN}(n,p)$ sont :

	$\mathbb{E}(X)$	$\mathbb{V}(X)$	Coefficient d'asymétrie	Coefficient d'aplatissement
$\mathrm{BN}(n,p)$	$\frac{n(1-p)}{p}$	$\frac{n(1-p)}{p^2}$	$\frac{(2-p)^2}{n(1-p)}$	$3+\frac{\frac{p^2}{(1-p)}+6}{n}$

TABLE 3.1 – Paramètres de $\mathcal{BN}(n,p)$

Première re-paramétrisation. Pour rendre plus intelligible l'utilisation dans le domaine de la microbiologie de la loi binomiale négative telle que décrite dans la littérature, on opère un changement de variable en posant :

$$n=\frac{\lambda'}{K-1}, \qquad \text{et} \qquad p=\frac{1}{K}.$$

La loi de probabilité du nombre de germes x présents devient alors :

$$P(X=x,\lambda',K)=\binom{x+\frac{\lambda'}{K-1}-1}{x}\cdot\left(\frac{1}{K}\right)^{\frac{\lambda'}{K-1}}\cdot\left(1-\frac{1}{K}\right)^{x},$$

où

- λ' est un réel positif correspondant au niveau de charge bactérienne,
- K est un réel strictement supérieur à 1 correspondant à la surdispersion

Remarque 3.3.1. *K ne peut pas prendre ici n'importe quelle valeur car, pour pouvoir calculer le coefficient binomial, il faut que $\frac{\lambda'}{K-1}$ soit un entier naturel. Il faut donc que λ' soit un multiple de $K-1$.*

Les paramètres de $\mathcal{BN}(\lambda',K)$ sont :

	$\mathbb{E}(X)$	$\mathbb{V}(X)$	Coefficient d'asymétrie	Coefficient d'aplatissement
$\mathrm{BN}(\lambda',K)$	λ'	$\lambda'.K$	$\frac{(2.K-1)^2}{\lambda'.K}$	$3+\frac{\frac{1}{K}+6.(K-1)}{\lambda'}$

TABLE 3.2 – Paramètres de $\mathcal{BN}(\lambda',K)$

Deuxième re-paramétrisation. Lorsque l'on calcule les incertitudes de mesure, il est d'usage d'additionner l'effet des composantes. Dans le domaine du contrôle de qualité en chimie, la loi normale s'applique et on additionne ainsi les variances, comme détaillé en partie 1.3.3. En microbiologie, l'utilisation du ratio *variance/moyenne* est alors peu explicite.
Pour améliorer la formulation de la surdispersion, un coefficient de variation CV est introduit par [ISO, 2000].

Definition 3.3.1. On appelle **coefficient de variation**, noté CV, le rapport de l'écart-type à la moyenne. Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande. Il est généralement exprimé en pourcentage.

$$CV^2=\frac{\mathbb{V}(X)}{\mathbb{E}(X)^2}=\frac{1}{\lambda'}+u^2.$$

Definition 3.3.2. Il découle de ce coefficient de variation un **paramètre de surdispersion**, noté u^2, qui se calcule comme suit :

$$u^2 = \frac{K-1}{\lambda'} = \frac{1}{n}.$$

La paramétrisation de la loi binomiale négative en coefficient de variation CV^2 et paramètre de surdispersion u^2 correspond à :

$$K = 1 + \lambda' u^2, \qquad \mathbb{V}(X) = \lambda' + \lambda'^2 u^2.$$

Remarque 3.3.2. $\frac{1}{\lambda'}$ *est le carré du coefficient de variation pour la loi de Poisson.*

Nous obtenons ainsi un modèle additif en coefficients de variation. L'incertitude totale exprimée en coefficient de variation CV^2_{Total} est la somme de l'incertitude liée à la loi de Poisson (erreur d'aliquotage, exprimée en coefficient de variation $CV^2_{Poisson}$ et égal à l'inverse de la charge bactérienne), plus la somme des incertitudes liées aux différents facteurs source d'erreur exprimées en $CV^2_{facteur}$.

$$CV^2_{Total} = CV^2_{Poisson} + \sum CV^2_{facteurs} = \frac{1}{\lambda'} + \sum u^2_{facteurs}. \tag{3.6}$$

Remarque 3.3.3. u^2 *ne peut pas prendre ici n'importe quelle valeur car pour pouvoir calculer la combinaison il faut que* $\frac{1}{u^2}$ *soit un entier naturel.*

La loi de probabilité devient $\mathcal{BN}(\lambda', u^2)$ pour laquelle :

	$\mathbb{E}(X)$	$\mathbb{V}(X)$	Coefficient d'asymétrie	Coefficient d'aplatissement
BN(λ',u^2)	λ'	$\lambda' + \lambda'^2.u^2$	$\frac{(1+2.\lambda'.u^2)^2}{\lambda'+\lambda'^2.u^2}$	$3 + 6.u^2 + \frac{1}{\lambda'+\lambda'^2.u^2}$

TABLE 3.3 – Paramètres de $\mathcal{BN}(\lambda', u^2)$

3.3.2 Mélange Gamma-Poisson

Pour notre usage, la loi binomiale négative telle que décrite ci-dessus présente l'inconvénient majeur que n est un entier naturel non nul. Or, théoriquement $u^2 = \frac{1}{n}$ peut prendre n'importe quelle valeur réelle positive.
Nous allons donc travailler avec un généralisation de la loi binomiale négative à deux paramètres λ' et u^2, où u^2 peut prendre des valeurs réelles strictement positives. Il arrive en effet que l'occurrence de certains phénomènes suive une loi de Poisson d'intensité λ variant elle-même de manière aléatoire. C'est le cas lorsque le phénomène se produit dans des populations différentes dans lesquelles la loi de Poisson change de paramètre. C'est ce qu'on appelle des mélanges de lois de Poisson. Lorsque le paramètre λ suit une distribution Gamma, l'occurrence du phénomène suit une loi de mélange Gamma-Poisson [McCullagh and Nelder, 1989].

La loi Gamma. Une distribution Gamma est caractérisée par deux paramètres qui affectent respectivement la forme et l'échelle de sa représentation graphique. Les distributions Gamma sont utilisées pour modéliser une grande variété de phénomènes, et notamment les phénomènes se déroulant au cours du temps où, par essence, le temps écoulé est une grandeur réelle positive.

C'est le cas par exemple dans l'analyse de survie. Une distribution Gamma est caractérisée par deux paramètres. Usuellement un paramètre de forme k, et un paramètre d'échelle τ.
Une variable aléatoire X suit une loi Gamma de paramètres de forme $k > 0$ et d'échelle $\tau > 0$ si sa fonction de densité peut s'écrire sous la forme :

$$f(x; k, \tau) = \frac{x^{k-1} e^{-x/\tau}}{\Gamma(k)\tau^k}.$$

La loi Gamma peut également être décrite à l'aide du paramètre de forme k et d'un paramètre d'intensité ρ, qui correspond à l'inverse du paramètre d'échelle : $\rho = \frac{1}{\tau}$. La fonction de densité de la variable aléatoire X s'exprime alors :

$$f(x; k, \rho) = \frac{x^{k-1} \rho^k e^{-\rho x}}{\Gamma(k)}. \tag{3.7}$$

En utilisant cette paramétrisation, les moments de la variance aléatoire $X \sim \Gamma(k, \rho)$ sont :

	$\mathbb{E}(X)$	$\mathbb{V}(X)$	Coefficient d'asymétrie	Coefficient d'aplatissement
$\Gamma(k, \rho)$	k/ρ	k/ρ^2	$2/\sqrt{k}$	$6/k$

TABLE 3.4 – Paramètres de $\Gamma(k, \rho)$

Dans la suite de ce travail, nous utiliserons toujours cette paramétrisation $\Gamma(\text{forme}, \text{intensité})$ de la loi Gamma.

Paramétrisation du mélange Gamma-Poisson. Deux paramétrisations du mélange Gamma-Poisson sont particulièrement répandues. Il s'agit des paramétrisations NB1 et NB2 décrites par [McCullagh and Nelder, 1989] et [Cameron and Trivedi, 1998]. La paramétrisation NB2 peut s'exprimer de la façon suivante :
Soit X une variable aléatoire de loi de Poisson, dont l'intensité λ suit la loi Gamma suivante :

$$X \sim P(\lambda), \qquad \text{avec} \qquad \lambda \sim \Gamma(\frac{1}{u^2}, \frac{1}{u^2}).$$

L'espérance et la variance de λ s'expriment alors comme suit :

$$\mathbb{E}(\lambda) = 1 \qquad \text{et} \qquad \mathbb{V}(\lambda) = u^2.$$

La variable aléatoire X suit alors une loi Gamma-Poisson de paramètres λ' et u^2 :

$$X \sim \mathcal{GP}(\lambda', u^2).$$

Cette formulation de la loi Gamma est particulièrement explicite pour la microbiologie. En effet, les variances des différents facteurs sont égales à leurs paramètres de dispersion, donc d'après la formule (3.6), le calcul de l'incertitude total revient à sommer les coefficients de dispersion des différents facteurs.
Nous pouvons ainsi généraliser la loi binomiale négative à un paramètre u^2 réel positif. Les formules de la moyenne et de la variance indiquée dans la table 3.3 restent valables.
$\mathcal{BN}(\lambda', u^2)$ devient alors $\mathcal{GP}(\lambda', u^2)$ dont la loi de probabilité est :

$$P(X = k) = \frac{\Gamma(k + \frac{1}{u^2})}{\Gamma(\frac{1}{u^2 \cdot k!})} \cdot \left(\frac{1}{1 + \lambda' \cdot u^2}\right)^{\frac{1}{u^2}} \cdot \left(\frac{\lambda' \cdot u^2}{1 + \lambda' \cdot u^2}\right)^k,$$

où

- λ' est un réel positif correspondant au niveau de charge bactérienne,
- u^2 est un réel strictement positif correspondant à la surdispersion,
- k est un entier naturel correspondant au nombre de germes présents.

Les paramètres de $\mathcal{GP}(\lambda',u^2)$ sont les mêmes que ceux de $\mathcal{BN}(\lambda',u^2)$ (Table 3.3).

3.3.3 Modélisation de la variabilité des résultats d'un EIL

Différentes sources de dispersion. Trois facteurs, Laboratoire, Flacon, et Réplication, peuvent influencer la dispersion des résultats de dénombrement d'un essai interlaboratoires. Nous avons précisé l'influence de ces facteurs sur la variabilité des résultats de dénombrement dans la section 3.2.2. Nous associerons dans la suite de ce travail un coefficient de dispersion u_i^2, pour $i = \{1, 2, 3\}$ à chacun de ces facteurs.

- Le coefficient de dispersion u_1^2 mesure la dispersion induite par le facteur Laboratoire,
- Le coefficient de dispersion u_2^2 mesure la dispersion induite par le facteur Flacon,
- Le coefficient de dispersion u_3^2 mesure la dispersion induite par le facteur Réplication.

Remarque 3.3.4. *Nous abandonnons la notation "coefficient de surdispersion" de [ISO, 2000] au profit de celle de "coefficient de dispersion", puisque nous avons caractérisé la surdispersion à partir des trois facteurs Laboratoire, Flacon et Réplication. Il s'agit de la différence entre un modèle de Poisson surdispersé et un modèle à effets aléatoires, expliquée en section 3.2.3.*

HGLM Gamma-Poisson à 3 facteurs aléatoires. On note y_{ijk} le résultat du dénombrement effectué sur la réplication $k \in \{1, 2\}$ du flacon $j \in \{1, 2\}$ par le laboratoire $i \in \{1...a\}$. Considérons un modèle dont le prédicteur linéaire s'écrit :

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} \quad \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, b \\ k = 1, \ldots, n \end{cases} . \tag{3.8}$$

On suppose que la distribution de y_{ijk} conditionnellement aux effets aléatoires e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$ correspond à une loi de Poisson de paramètre λ_{ijk} :

$$\mathbb{E}(y_{ijk}|e^{\alpha_i}, e^{\beta_{ij}}, e^{\gamma_{ijk}}) = \lambda_{ijk} = e^{\mu_{ijk}}.e^{\alpha_i}.e^{\beta_{ij}}.e^{\gamma_{ijk}}, \tag{3.9}$$

où $e^{\mu} = \beta_0$ est un paramètre fixe correspondant à la moyenne globale des dénombrements de germes. D'autre part, on considère que e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$ sont distribuées suivant trois lois Gamma de paramètres de forme respectifs $\frac{1}{u_1^2}$, $\frac{1}{u_2^2}$ et $\frac{1}{u_3^2}$, et d'espérance 1 :

- $e^{\alpha_i} \sim \Gamma(\frac{1}{u_1^2}, \frac{1}{u_1^2})$ représente l'effet du laboratoire i,
- $e^{\beta_{ij}} \sim \Gamma(\frac{1}{u_2^2}, \frac{1}{u_2^2})$ représente l'effet du flacon j du laboratoire i,
- $e^{\gamma_{ijk}} \sim \Gamma(\frac{1}{u_3^2}, \frac{1}{u_3^2})$ représente l'erreur de mesure entre réplications d'un flacon.

Avec :

$$\mathbb{E}[e^{\alpha_i}] = \mathbb{E}[e^{\beta_{ij}}] = \mathbb{E}[e^{\gamma_{ijk}}] = 1.$$

Les paramètres d'échelle des variables aléatoires e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$ correspondent alors respectivement aux parts de dispersion induites par le laboratoire (u_1^2), le flacon (u_2^2), et la réplication (u_3^2).

$$\mathbb{V}[e^{\alpha_i}] = u_1^2, \qquad \mathbb{V}[e^{\beta_{ij}}] = u_2^2, \qquad \mathbb{V}[e^{\gamma_{ijk}}] = u_3^2.$$

La variable aléatoire λ_{ijk} correspond à un produit de trois variables aléatoires indépendantes de lois Gamma, dont la loi n'est pas connue. La caractérisation de la distribution de λ_{ijk} fera l'objet du chapitre 4.
Le modèle (3.9) correspond à un HGLM Gamma-Poisson à trois facteurs aléatoires.
Les composantes aléatoires α_i, β_{ij} et γ_{ijk} sont distribuées suivant des lois log-Gamma de paramètres respectifs u_1^2, u_2^2 et u_3^2 :

$$\begin{aligned}
\alpha_i &\sim \text{Log-Gamma}\left(\frac{1}{u_1^2}, \frac{1}{u_1^2}\right), \\
\beta_{ij} &\sim \text{Log-Gamma}\left(\frac{1}{u_2^2}, \frac{1}{u_2^2}\right), \\
\gamma_{ijk} &\sim \text{Log-Gamma}\left(\frac{1}{u_3^2}, \frac{1}{u_3^2}\right).
\end{aligned} \tag{3.10}$$

Dans la section suivante, nous nous intéressons à l'estimation des effets fixes et aléatoires, et des paramètres de dispersion.

3.4 Estimation de la dispersion induite par les différents facteurs

3.4.1 H-vraisemblance

[Lee and Nelder, 1996] définissent une vraisemblance spécifique pour l'inférence dans les HGLM, appelée *h-vraisemblance*. Notons α la variable aléatoire dont les réalisations pour une essai inter-laboratoires sont α_i, β la variable aléatoire dont les réalisations sont β_{ij}, γ la variable aléatoire dont les réalisations sont γ_{ijk}, et y la variable aléatoire dont les réalisations sont y_{ijk}. Pour le modèle (3.9), la log-h-vraisemblance s'écrit :

$$l(\mu, u_1^2, u_2^2, u_3^2, \alpha, \beta, \gamma) = l(y|\alpha, \beta, \gamma) + l(\alpha) + l(\beta) + l(\gamma), \tag{3.11}$$

où

- $l(y|\alpha, \beta, \gamma)$ correspond à la log-vraisemblance basée sur la distribution de y conditionnellement aux effets aléatoires α, β, et γ,
- $l(\alpha)$, $l(\beta)$ et $l(\gamma)$ sont les log-vraisemblances basées sur les distributions supposées des effets aléatoires, α, β, et γ.

Comme le soulignent [Lee et al., 2007], la h-vraisemblance (3.11) n'est pas une vraisemblance au sens classique du terme. Pourtant, en pratique, la h-vraisemblance $L(\mu, u_1^2, u_2^2, u_3^2, \alpha, \beta, \gamma)$ se calcule plus facilement que la vraisemblance marginale, $L(\mu, u_1^2, u_2^2, u_3^2)$. D'autre part, elle permet, contrairement à la vraisemblance marginale, d'estimer les effets aléatoires α_i, β_{ij}, et γ_{ijk}, ce qui s'avère nécessaire dans le cadre de notre problématique.
Finalement, la h-vraisemblance (3.11) est considérée comme une vraisemblance traditionnelle pour les paramètres μ et α_i, β_{ij}, γ_{ijk}, où α_i, β_{ij}, γ_{ijk} sont considérés comme étant des paramètres fixes pour des valeurs effectives mais non observables des effets aléatoires.

Pour les distributions de la famille exponentielle, la h-vraisemblance nous permet d'obtenir des estimateurs des effets aléatoires qui sont les meilleurs prédicteurs asymptotiques non biaisés, et un estimateur de l'effet fixe μ qui est aussi efficace asymptotiquement que l'estimateur

marginal du maximum de vraisemblance. Les bonnes propriétés des estimateurs du maximum de h-vraisemblance sont notamment illustrées dans [Commenges et al., 2011].

Remarque 3.4.1. *Nous avons vu en section 3.2.5 que l'expression de la log-vraisemblance basée sur la distribution de y conditionnellement aux effets aléatoires peut s'écrire sous la forme (3.5). Pour le modèle (3.9), les observations y_{ijk} suivent une distribution de Poisson conditionnellement aux effets aléatoires. Le paramètre de cette distribution, λ_{ijk} est lié aux effets aléatoires α_i, β_{ij} et γ_{ijk} par la relation (3.9). La log-vraisemblance basée sur la distribution de y conditionnellement à α, β et γ s'écrit alors :*

$$l(y|\alpha,\beta,\gamma) = y\log(\lambda) - \lambda - \log(y!). \tag{3.12}$$

Par identification des expressions (3.12) et (3.5), nous en déduisons que, pour le modèle (3.9), les paramètres ϕ, et θ définissant la famille exponentielle, ainsi que les fonctions b et c sont les suivants :

$$\phi = 1, \quad \theta = (u_1^2, u_2^2, u_3^2), \quad b(\theta) = \lambda_{ijk} = \exp(\mu + \alpha_i + \beta_{ij} + \gamma_{ijk}), \quad \textit{et} \;\; c(\theta,\phi) = -\log(y!).$$

[Lee et al., 2011, Lee and Nelder, 2002, Commenges et al., 2011] proposent des applications pratiques de l'utilisation de la h-vraisemblance.

3.4.2 Estimation du paramètre fixe et des effets aléatoires

[Pawitan, 2001, p466] souligne que, même dans le cas gaussien, la log-h-vraisemblance ne permet pas d'estimer à la fois les effets fixes et aléatoires, et les paramètres de dispersion. Par conséquent, pour le modèle 3.9, nous ne pouvons pas estimer à la fois μ, α, β, γ et les coefficients de dispersion u_1^2, u_2^2, et u_3^2 à partir de la seule log-h-vraisemblance $l(\mu, u_1^2, u_2^2, u_3^2, \alpha, \beta, \gamma)$.
Dans un premier temps, nous allons donc nous attacher à estimer le paramètre fixe μ, ainsi que les effets aléatoires α, β, γ, en fixant les paramètres u_1^2, u_2^2, et u_3^2.
Par ailleurs, l'expression de λ_{ijk} (3.9) rend l'écriture exacte de la vraisemblance basée sur la distribution de y conditionnellement à α, β, γ complexe. En considérant des valeurs fixes de u_1^2, u_2^2 et u_3^2, nous pouvons utiliser une approximation quadratique de la vraisemblance (3.11) pour estimer le paramètre μ, ainsi que les effets associés à chaque laboratoire, α_i, à chaque flacon, β_{ij}, et à chaque réplication, γ_{ijk}.
On peut écrire le prédicteur linéaire (3.8) sous forme matricielle :

$$\eta = \log(\lambda) = X\mu + Z\alpha + F\beta + R\gamma, \tag{3.13}$$

où :

- X est la matrice de design de l'effet fixe, de taille $(a \times b \times n) \times 1$, correspondant au vecteur : $X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$,

- Z est la matrice de design de l'effet Laboratoire de taille $(a \times b \times n) \times a$:

$$Z = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \end{pmatrix}}_{\text{a colonnes}} \Bigg\} (a \times b \times n) \text{ lignes,}$$

- F est la matrice de design de l'effet Flacon, de taille $(a \times b \times n) \times (b \times a)$:

$$F = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}}_{a \times b \text{ colonnes}} \Bigg\} (a \times b \times n) \text{ lignes,}$$

- R est la matrice de design de l'effet Réplication, de taille $(a \times b \times n) \times (a \times b \times n)$:

$$R = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix}}_{(a \times b \times n) \text{ colonnes}} \Bigg\} (a \times b \times n) \text{ lignes,}$$

En considérant des valeurs initiales μ^0, α^0, β^0 et γ^0 du paramètre fixe μ et des vecteurs α, β et γ, nous calculons λ^0, la valeur initiale de λ, en introduisant les valeurs α^0, β^0 et γ^0 dans l'expression 3.9. La log-vraisemblance basée sur la distribution de y conditionnellement aux effets aléatoires peut alors être approchée par l'expression suivante, valable pour les lois de la famille exponentielle [Pawitan, 2001, p464] :

$$l(y|\alpha,\beta,\gamma) \approx -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\tilde{Y} - X\mu - Z\alpha - F\beta - R\gamma)^{-1}\Sigma^{-1}(Y - X\mu - Z\alpha - F\beta - R\gamma), \tag{3.14}$$

où

- $\tilde{Y}$ est le vecteur dont les éléments sont les suivants :

$$\tilde{Y}_{ijk} = \mu^0 + \alpha^0 + \beta^0 + \gamma^0 + \frac{\partial \eta}{\partial \lambda_{ijk}}(y_{ijk} - \lambda^0_{ijk}),$$

- Σ est une matrice diagonale de taille abn de la variance du vecteur de travail. Ses termes diagonaux s'écrivent :

$$\Sigma_{ss} = \left(\frac{\partial \eta}{\partial \lambda_s}\right)^2 \lambda_s = \frac{1}{\lambda_s} \quad \text{pour } s = (1, \ldots, abn).$$

La valeur de λ_s est évaluée à partir de l'expression (3.9) en utilisant les valeurs courantes de μ et des réalisations α_i, β_{ij} et γ_{ijk} des variables aléatoires α, β, γ pour chaque mesure d'un essai interlaboratoires donné.

D'autre part, les effets aléatoires e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$ suivent des distributions Gamma de paramètres de forme respectifs $\frac{1}{u_1^2}$, $\frac{1}{u_2^2}$ et $\frac{1}{u_3^2}$. Les log-vraisemblances basées sur les variables aléatoires α_i, β_{ij} et γ_{ijk} s'écrivent donc :

$$\begin{aligned}
l(\alpha_i) &= \frac{1}{u_1^2}\alpha_i - \frac{1}{u_1^2}e^{\alpha_i} + \frac{1}{u_1^2}\log\frac{1}{u_1^2} - \log\Gamma\left(\frac{1}{u_1^2}\right), \\
l(\beta_{ij}) &= \frac{1}{u_2^2}\beta_{ij} - \frac{1}{u_2^2}e^{\beta_{ij}} + \frac{1}{u_2^2}\log\frac{1}{u_2^2} - \log\Gamma\left(\frac{1}{u_2^2}\right), \\
l(\gamma_{ijk}) &= \frac{1}{u_3^2}\gamma_{ijk} - \frac{1}{u_3^2}e^{\gamma_{ijk}} + \frac{1}{u_3^2}\log\frac{1}{u_3^2} - \log\Gamma\left(\frac{1}{u_3^2}\right).
\end{aligned} \tag{3.15}$$

Ces effets aléatoires n'étant pas gaussiens, [Pawitan, 2001, p465] suggère d'approcher les log-vraisemblances basées sur leurs distributions supposées par des formes quadratiques. Le théorème de Taylor permet d'obtenir les approximations suivantes :

$$\begin{aligned}
l(\alpha_i) &\approx l(\alpha_i^c) + \frac{1}{2}l''(\alpha_i^0)(\alpha_i - \alpha_i^c)^2, \\
l(\beta_{ij}) &\approx l(\beta_{ij}^c) + \frac{1}{2}l''(\beta_{ij}^0)(\beta_{ij} - \beta_{ij}^c)^2, \\
l(\gamma_{ijk}) &\approx l(\gamma_{ijk}^c) + \frac{1}{2}l''(\gamma_{ijk}^0)(\gamma_{ijk} - \gamma_{ijk}^c)^2.
\end{aligned}$$

où

$$\alpha_i^c = \alpha_i^0 - \frac{l'(\alpha_i^0)}{l''(\alpha_i^0)}, \qquad \beta_{ij}^c = \beta_{ij}^0 - \frac{l'(\beta_{ij}^0)}{l''(\beta_{ij}^0)}, \qquad \gamma_{ijk}^c = \gamma_{ijk}^0 - \frac{l'(\gamma_{ijk}^0)}{l''(\gamma_{ijk}^0)}.$$

Notons D_1^{-1}, D_2^{-1} et D_3^{-1} les matrices d'information de Fisher respectives de α, β et γ, et $l(\alpha^0)$, $l(\beta^0)$, $l(\gamma^0)$ les vecteurs respectifs de $l(\alpha_i^0)$, $l(\beta_{ij}^0)$ et $l(\gamma_{ijk}^0)$. On a :

$$\alpha^c = \alpha^0 + D_1 l'(\alpha^0), \qquad \beta^c = \beta^0 + D_2 l'(\beta^0), \qquad \gamma^c = \gamma^0 + D_3 l'(\gamma^0).$$

Nous pouvons alors écrire :

$$\begin{aligned} l(\alpha) &= l(\alpha^c) - \frac{1}{2}(\alpha-\alpha^c)' D_1^{-1}(\alpha-\alpha^c), \\ l(\beta) &= l(\beta^c) - \frac{1}{2}(\beta-\beta^c)' D_2^{-1}(\beta-\beta^c), \\ l(\gamma) &= l(\gamma^c) - \frac{1}{2}(\gamma-\gamma^c)' D_3^{-1}(\gamma-\gamma^c). \end{aligned} \tag{3.16}$$

Or, e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$ étant distribuées suivant des lois Gamma, les vraisemblances basées sur les distributions de ces variables aléatoires sont données par (3.15). Leurs dérivées premières et secondes sont les suivantes :

$$\left\{\begin{array}{ll} l'(\alpha_i) & = \frac{1}{u_1^2} - \frac{1}{u_1^2}e^{\alpha_i} \\ l''(\alpha_i) & = -\frac{1}{u_1^2}e^{\alpha_i} \end{array}\right. \quad \left\{\begin{array}{ll} l'(\beta_{ij}) & = \frac{1}{u_2^2} - \frac{1}{u_2^2}e^{\beta_{ij}} \\ l''(\beta_{ij}) & = -\frac{1}{u_2^2}e^{\beta_{ij}} \end{array}\right. \quad \left\{\begin{array}{ll} l'(\gamma_{ijk}) & = \frac{1}{u_3^2} - \frac{1}{u_3^2}e^{\gamma_{ijk}} \\ l''(\gamma_{ijk}) & = -\frac{1}{u_3^2}e^{\gamma_{ijk}} \end{array}\right.$$

D'après [Pawitan, 2001, p466], les expressions de D_1^{-1}, D_2^{-1} et D_3^{-1} pour le modèle (3.9) sont alors les suivantes :

$$D_1^{-1} = \text{diag}\left[\frac{1}{u_1^2}e^{\alpha_i}\right], \qquad D_2^{-1} = \text{diag}\left[\frac{1}{u_2^2}e^{\beta_{ij}}\right], \qquad D_3^{-1} = \text{diag}\left[\frac{1}{u_3^2}e^{\gamma_{ijk}}\right].$$

d'où :

$$\alpha^c = \alpha^0 + e^{-\alpha^0} - 1, \qquad \beta^c = \beta^0 + e^{-\beta^0} - 1, \qquad \gamma^c = \gamma^0 + e^{-\gamma^0} - 1.$$

En combinant les approximations des log-vraisemblances basées sur les effets aléatoires (3.16) avec l'approximation quadratique de $l(y|\alpha,\beta,\gamma)$ (3.14), nous pouvons approcher la log-h-vraisemblance (3.11) par :

$$\begin{aligned} h \approx & -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\tilde{Y} - X\mu - Z\alpha - F\beta - R\gamma)^T\Sigma^{-1}(\tilde{Y} - X\mu - Z\alpha - F\beta - R\gamma) \\ & + \log(L(\alpha^{(c)}) - \frac{1}{2}(\alpha-\alpha^{(c)})^T D_1^{-1}(\alpha-\alpha^{(c)}) + \log(L(\beta^{(c)}) - \frac{1}{2}(\beta-\beta^{(c)})^T D_2^{-1}(\beta-\beta^{(c)}) \\ & + \log(L(\gamma^{(c)}) - \frac{1}{2}(\gamma-\gamma^{(c)})^T D_3^{-1}(\gamma-\gamma^{(c)}). \end{aligned} \tag{3.17}$$

En dérivant (3.17) par rapport à μ, α, β et γ, à u_1^2, u_2^2 et u_3^2 fixés, on obtient le système d'équations suivant :

$$\begin{pmatrix} X^T\Sigma^{-1}X & X^T\Sigma^{-1}Z & X^T\Sigma^{-1}F & X^T\Sigma^{-1}R \\ Z^T\Sigma^{-1}X & Z^T\Sigma^{-1}Z + D_1^{-1} & Z^T\Sigma^{-1}F & Z^T\Sigma^{-1}R \\ F^T\Sigma^{-1}X & F^T\Sigma^{-1}Z & F^T\Sigma^{-1}F + D_2^{-1} & F^T\Sigma^{-1}R \\ R^T\Sigma^{-1}X & R^T\Sigma^{-1} & R^T\Sigma^{-1}F & R^T\Sigma^{-1}R + D_3^{-1} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} X^T\Sigma^{-1}\tilde{Y} \\ Z^T\Sigma^{-1}\tilde{Y} + D_1^{-1}\alpha^{(c)} \\ F^T\Sigma^{-1}\tilde{Y} + D_2^{-1}\beta^{(c)} \\ R^T\Sigma^{-1}\tilde{Y} + D_3^{-1}\gamma^{(c)}. \end{pmatrix}$$

Ces équations permettent d'estimer le paramètre fixe μ et les vecteurs d'effets aléatoires α, β, et γ, c'est à dire les effets α_i, β_{ij} et γ_{ijk} pour un essai interlaboratoires considéré. Nous utilisons pour cela la méthode dite de Jacobi ou de Gauss-Seidel en algèbre linéaire, également connue comme l'algorithme backfitting en statistique, introduit dans [Breiman and Friedman, 1985] et présenté dans [Pawitan, 2001]. La méthode est divisée en quatre étapes distinctes, présentée dans l'algorithme 1.

A partir des estimateurs de α, β, et γ, nous pouvons obtenir les effets associés à chaque laboratoire, à chaque flacon, et à chaque réplication. En effet, d'après [Lee and Nelder, 1996, p635],

Algorithme 1 Estimation de μ, α, β et γ

Tant que La convergence n'est pas atteinte **Faire**

Étape 1 : Calculer le vecteur de donné corrigé : $y^c = y - X\mu - Z\alpha - F\beta$.
Résoudre l'équation suivante pour estimer γ : $(R^T\Sigma^{-1}R + D_3^{-1})\gamma = R^T\Sigma^{-1}y^c + D_3^{-1}\gamma^c$.

Étape 2 : Calculer le vecteur suivant : $y^c = y - X\mu - Z\alpha - R\gamma$.
Estimer β en résolvant l'équation : $(F^T\Sigma^{-1}F + D_2^{-1})\beta = F^T\Sigma^{-1}y^c + D_2^{-1}\beta^c$.

Étape 3 : Calculer le vecteur corrigé suivant : $y^c = y - X\mu - F\beta - R\gamma$.
Résoudre l'équation suivante pour estimer α : $(Z^T\Sigma^{-1}Z + D_1^{-1})\alpha = Z^T\Sigma^{-1}y^c + D_1^{-1}\alpha^c$.

Étape 4 : Calculer : $y^c = y - Z\alpha - F\beta - R\gamma$.
Résoudre l'équation suivante pour ajuster l'estimateur de μ : $y(X^T\Sigma^{-1}X)\mu = X^T\Sigma^{-1}y^c$.

fin Tant que

pour les lois de la famille exponentielle, la h-vraisemblance est invariante par transformation des effets aléatoires. Les effets laboratoire sont donc estimés par $\exp(\hat{\alpha}_i)$, les effets flacon par $\exp(\hat{\beta}_{ij})$, et les effets réplication par $\exp(\hat{\gamma}_{ijk})$.

Le choix du critère de convergence est arbitraire et il est difficile de savoir *a priori* combien d'itérations seront nécessaires pour atteindre un seuil de convergence donné. Aussi, le modèle final dépend de l'ordre d'ajustement des variables α, β, γ et μ dans l'algorithme 1. Par conséquent, les estimateurs fournis par cet algorithme ne sont pas uniques.
Nous utilisons ici le package *HGLM* du logiciel R [Rönnegård et al., 2010], qui met en œuvre la procédure d'estimation présentée dans la présente section. Ce package défini un nombre maximal d'itération de 50, et un critère de convergence de 1×10^{-6}. Ces paramètres sont ajustables. Pour notre problématique, nous conservons le seuil de convergence par défaut. D'autre part, nous arrêtons l'algorithme lorsque l'écart entre les vecteurs données corrigé de deux itérations successives est suffisamment faible. On calcule $|y^c(i+1) - y^c(i)| \leq \epsilon = 1 \cdot 10^{-6}$

3.4.3 Estimation des paramètres de dispersion

La h-vraisemblance permet d'étudier des modèles à plusieurs paramètres, fixes et aléatoires. Cette vraisemblance multidimensitonnelle peut alors être compliquée à décrire. Même lorsqu'on est intéressé par plusieurs paramètres, il est toujours plus simple de décrire un paramètre à la fois. Dans le cadre de notre problématique, la log-vraisemblance jointe $l(\mu, u_1^2, u_2^2, u_3^2, \alpha, \beta, \gamma)$ ne permet pas d'estimer à la fois μ, α_i, β_{ij}, γ_{ijk}, et les paramètres de dispersion u_1^2, u_2^2, et u_3^2. Une méthode est alors nécessaire pour "concentrer" la vraisemblance sur le paramètre d'intérêt en éliminant les paramètres excédentaires. Expliquer l'incertitude supplémentaire due à ces paramètres excédentaires inconnus est une considération essentielle, en particulier lorsque les échantillons sont petits. Une approche basée sur la vraisemblance pour éliminer les paramètres excédentaires consiste à les remplacer par leurs estimateurs du maximum de vraisemblance, à chaque valeur fixée du paramètre d'intérêt. La vraisemblance résultante, présentée dans [Lee and Nelder, 1996, Lee and Nelder, 2001], est dite *profil de vraisemblance*.

Pour estimer les paramètres de dispersion, nous allons donc utiliser le profil de vraisemblance de (u_1^2, u_2^2, u_3^2).

Definition 3.4.1. Soit la vraisemblance jointe $L(u_1^2, u_2^2, u_3^2, \alpha, \beta, \gamma, \mu)$, le profil de vraisemblance de u_1^2, u_2^2, u_3^2, noté h_A, est donné par :

$$h_A = L(u_1^2, u_2^2, u_3^2) = max_{\alpha,\beta,\gamma,\mu} L(u_1^2, u_2^2, u_3^2, \alpha, \beta, \gamma, \mu), \tag{3.26}$$

où la maximisation est effectuée à μ, α, β, γ fixés.

Notons (3.13) sous la forme d'un modèle à une seule composante aléatoire, regroupant les effets laboratoire, flacon, et réplication :

$$\eta = \log(\lambda) = X\mu + \tilde{Z}t, \tag{3.27}$$

où

- $t = (\alpha, \beta, \gamma)$,
- $\tilde{Z}$ est la matrice de design $\tilde{Z} = [Z\ F\ R]$.

Propriété 3.4.1. *Il est montré dans [Pawitan, 2001, p446] que le profil de vraisemblance de (u_1^2, u_2^2, u_3^2) est équivalent à une vraisemblance modifiée, basée sur la vraisemblance jointe $L(\hat{\mu}, u_1^2, u_2^2, u_3^2, \hat{t})$ (expression 3.27) :*

$$L(u_1^2, u_2^2, u_3^2) = L(\hat{\mu}, u_1^2, u_2^2, u_3^2) = L(\hat{\mu}, u_1^2, u_2^2, u_3^2, \hat{t}) - \frac{1}{2}\log|\tilde{Z}^T\Sigma^{-1}\tilde{Z} + D^{-1}|, \tag{3.28}$$

où

- *D est la matrice diagonale suivante, de taille abn :*

$$D = \underbrace{\begin{pmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & D_3 \end{pmatrix}}_{(a\times b\times n)\ colonnes} \Bigg\} (a \times b \times n)\ lignes,$$

- *$\hat{t}$ correspond à la matrice des effets laboratoire, flacon et réplication pour un essai interlaboratoires donné, estimés par la méthode présentée ci-dessus à u_1^2, u_2^2, u_3^2 fixés.*

Les paramètres u_1^2, u_2^2 et u_3^2 interviennent dans l'expression (3.28) par l'intermédiaire des matrices Σ et D, et des estimateurs $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$, et $\hat{\gamma}$. Pour estimer les composantes de variance d'un HGLM, [Pawitan, 2001] propose d'estimer u_1^2, u_2^2 et u_3^2 en maximisant le profil de vraisemblance de (u_1^2, u_2^2, u_3^2).

3.4.4 Estimation jointe de l'ensemble des paramètres

Il est possible d'estimer simultanément le paramètre fixe μ, les vecteurs de réalisations des facteurs aléatoires $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$ et les paramètres de dispersion u_1^2, u_2^2 et u_3^2, en itérant l'algorithme 2, qui va calculer tour à tour $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ et $\hat{u}_1^2$, $\hat{u}_2^2$ et $\hat{u}_3^2$.

Algorithme 2 Estimation jointe de μ, α, β, γ et u_1^2 ,u_2^2, u_3^2

Tant que La convergence n'est pas atteinte **Faire**
Étape 1 : Fixer u_1^2, u_2^2 et u_3^2.
Calculer $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ en résolvant (3.4.2).
Étape 2 : Fixer μ, α, β, γ aux valeurs $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ calculées à l'étape 1. Calculer $\hat{u_1^2}$, $\hat{u_2^2}$ et $\hat{u_3^2}$ en maximisant 3.28.
fin Tant que

3.4.5 Estimation par quasi-h-vraisemblance

Comme le soulignent [Lee and Nelder, 2000], il existe deux façons de modéliser la surdispersion :

- par des modèles à effets aléatoires,
- à l'aide d'une approche par quasi-vraisemblance, introduite par [Wedderburn, 1974].

Ces deux approches conduisent à des variables réponses dont les fonctions de variance sont différentes.

Dans le cadre du modèle 3.9, la quasi-vraisemblance permet d'obtenir un estimateur de μ. En revanche, elle ne permet d'estimer ni les effets aléatoires, ni les paramètres de dispersion. [Nelder and Pregibon, 1987] ont développé une quasi-vraisemblance étendue Q définie par :

$$Q = -\sum \frac{d}{2\phi} - \sum \frac{1}{2}\log\{2\pi\phi V(y)\},$$

où

$$d = 2\int_y^{\lambda} \frac{y-x}{V(x)}\mathrm{d}x$$

est la composante de déviance, et $V()$ est la fonction de variance.
Pour le modèle 3.9, $V(y) = y$. Par conséquent, la déviance d s'écrit alors :

$$d = y\ln(\frac{\lambda}{y}) - (\lambda - y) \tag{3.29}$$

[Nelder and Lee, 1992] ont montré par simulations que les estimateurs de quasi-vraisemblance étendue des paramètres de dispersion sont efficaces sur des échantillons de taille finie et peuvent même s'avérer meilleurs que les estimateurs de maximum de vraisemblance.
[Lee and Nelder, 1996] définissent la quasi-h-vraisemblance, qui, pour le modèle 3.9, s'écrit :

$$h_E = Q + l(\alpha) + l(\beta) + l(\gamma).$$

Les équations $\partial h_E/\partial\mu = 0$, $\partial h_E/\partial\alpha = 0$, $\partial h_E/\partial\beta = 0$, et $\partial h_E/\partial\gamma = 0$ permettent d'obtenir les estimateurs du maximum de quasi-h-vraisemblance. D'autre part, en utilisant le profil de vraisemblance h_A, les equations $\partial h_E/\partial u_1^2 = 0$, $\partial h_E/\partial u_2^2 = 0$, $\partial h_E/\partial u_3^2 = 0$ permettent d'obtenir les estimateurs de quasi-h-vraisemblance des coefficients de dispersion, u_1^2, u_2^2, u_3^2.
[Wedderburn, 1974], [Lee et al., 2006, p66] soulignent que la quasi-vraisemblance est équivalente à la vraisemblance classique si et seulement si la variable réponse provient d'une distribution exponentielle à un paramètre. Autrement dit, pour le modèle 3.9, l'estimation par maximum de quasi-h-vraisemblance donne des résultats identiques à l'estimation par maximum de vraisemblance.

3.4.6 Estimation des variances du paramètre fixe et des effets aléatoires

Les variances des estimateurs peuvent être calculées en utilisant leurs matrices d'information de Fisher. D'après [Pawitan, 2001], par analogie avec le GL2M à facteurs gaussiens, l'information de Fisher pour le paramètre μ est :

$$I(\hat{\mu}) = X^T V^{-1} X,$$

où $V = \Sigma + \tilde{Z}D\tilde{Z}^T$, et la matrice d'information de Fisher pour t (modèle 3.27) est :

$$I(\hat{t}) = (\tilde{Z}^T\Sigma^{-1}\tilde{Z} + D^{-1}).$$

En pratique, ces matrices sont évaluées aux valeurs estimées de u_1^2, u_2^2, et u_3^2. Ces formules impliquent que :

- l'inférence sur μ repose sur le fait que t soit inconnu, mais pas u_1^2, u_2^2, et u_3^2,
- l'inférence sur t ne repose pas sur le fait qu'à la fois μ et u_1^2 ,u_2^2, et u_3^2 soient estimés.

3.4.7 Estimation des variances des paramètres de dispersion

La procédure d'estimation présentée ci-dessus nous permet d'obtenir des variances sur les estimateurs de l'effet fixe β_0 ainsi que des effets aléatoires α, β et γ. Elle ne permet en revanche pas de connaître la variance des estimateurs des paramètres de dispersion, $\hat{u}_1^2$, $\hat{u}_2^2$ et $\hat{u}_3^2$. Pour cela, nous allons avoir recours à une méthode par rééchantillonage. Deux méthodes de rééchantillonage peuvent être utilisées pour estimer l'erreur standard d'un estimateur, la jackknife et le bootstrap. Ces méthodes sont basées sur des simulations, et ont la particularité de ne pas nécessiter d'autre information que celle disponible dans l'échantillon : au contraire, elles vont même en éliminer des parties entières par roulement.

Le jackknife. En 1949, la méthode du Jackknife est introduite par [Quenouille, 1949] dans le but de réduire le biais d'un estimateur obtenu à partir d'un échantillon. Cette méthode fut ensuite étendue par [Tukey, 1958] qui montra que, si les observations d'un échantillon sont indépendantes et identiquement distribuées, il est alors possible de calculer la variance d'un estimateur obtenu sur cet échantillon, et qu'elle peut être approchée par une distribution de Student à $N-1$ degrés de liberté, où N est la taille de l'échantillon.
Notons r le paramètre à estimer à partir de l'échantillon $C = (c_1, \ldots, c_N)$. Une estimation de r est $\hat{r}$. L'algorithme 3 permet de calculer la variance de l'estimateur $\hat{r}$ ainsi qu'un intervalle de confiance sur cet estimateur à l'aide de la méthode du jackknife :

Remarque 3.4.2. *La méthode du Jackknife peut être étendue au retrait d'un groupe de k individus à chaque itération*

Algorithme 3 Jackknife

Pour i allant de 1 à N **Faire**
Calculer l'estimateur $\hat{r}_i$ de r sur l'échantillon privé de sa $i^{\text{ème}}$ observation : $C_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, c_N)$.
Calculer ensuite la $i^{\text{ème}}$ pseudo-valeur : $\hat{r}_i^* = N\hat{r} - (N-1)\hat{r}_i$.
Ces estimations partielles correspondent à des variables indépendantes et d'espérance r.
fin Pour
Calculer l'estimateur jackknife de r, qui correspond à la moyenne empirique suivante :

$$\hat{r}^* = \frac{1}{N}\sum_i \hat{r}_i^*.$$

Calculer la variance de cet estimateur :

$$\hat{\sigma}^2(\hat{r}^*) = \frac{1}{N}\hat{\sigma}^2(\hat{r}_i^*) = \frac{1}{N(N-1)}\sum_i (\hat{r}_i^* - \hat{r}^*)^2.$$

Calculer un intervalle de confiance sur $\hat{r}^*$ au seuil $1-\alpha$:

$$\hat{r}^* + -t_{\alpha/2;N-1}\sqrt{\hat{\sigma}^2(\hat{r}^*)},$$

où $t_{\alpha/2;N-1}$ est le quantile approprié d'une loi de Student.

Le bootstrap. Le bootstrap est une autre méthode de rééchantillonage permettant d'estimer la distribution d'un estimateur sur un échantillon [Efron, 1979]. Elle est apparue dans les années 1970, époque où devenaient abordables des calculs informatiques intensifs. Elle constitue une évolution de la méthode du jackknife. Le bootstrap consiste à effectuer un échantillonnage avec remise à partir de l'échantillon de départ, notamment dans le but d'obtenir des estimateurs robustes des erreurs standards ainsi, que des intervalles de confiance sur un paramètre de la population.
La méthode du bootstrap est souvent utilisée comme une alternative robuste à l'inférence basée sur des hypothèses paramétriques lorsque ces hypothèses ne sont pas fiables, ou lorsque l'inférence paramétrique est impossible, ou complexe à mettre en œuvre.

Remarque 3.4.3. *La méthode du bootstrap peut également être utilisée pour développer des tests d'hypothèses, comme nous le verrons dans la section 3.5.*

Le principe du bootstrap est de calculer un nombre n_i de fois la statistique étudiée sur des échantillons obtenus en sélectionnant systématiquement avec remise N_e individus parmi les N de l'échantillon initial, considérés indépendants et identiquement distribués. La série de n_i estimateurs obtenus permet alors de caractériser la distribution de la statistique étudiée, et notamment d'estimer sa variance. L'algorithme du boostrap est le suivant :

Algorithme 4 Bootstrap

Pour i allant de 1 à n_i **Faire**
Créer un échantillon de N_e individus avec remise à partir de l'échantillon initial
Calculer la statistique sur le $i^{\text{ème}}$ échantillon
fin Pour
Calculer un intervalle de confiance au risque $\alpha\%$ à partir des n_i valeurs de la statistique ainsi obtenues :
Calculer pour cela les quantiles à $\alpha/2\%$ et à $100 - \alpha/2\%$

Choix de la méthode. Le jackknife correspond en fait à une approximation linéaire du boostrap, comme le soulignent [Efron, 1979]. Il ne fournit un estimateur consistant que dans le cas où la statistique est "lisse" (ce qui n'est pas le cas de la médiane par exemple). Lorsque ce n'est pas le cas, l'estimateur fournit par le jackknife n'est pas consistant et on préfère utiliser le bootstrap. D'autre part, la différence entre les deux méthodes réside dans le fait que le jackknife permet d'estimer la variance d'un estimateur tandis que le bootstrap permet dans un premier temps de caractériser la distribution de l'estimateur, et ainsi d'en déduire sa variance. Le jackknife peut en revanche être utilisé pour la détection d'observations aberrantes. Nous préférerons donc ici utiliser le bootstrap préférentiellement au jackknife.

Calcul d'intervalles de confiance sur les coefficients de dispersion. Pour estimer les variances des estimateurs des coefficients de dispersion que nous avons obtenus à l'aide de la méthode explicitée en section 3.4.3, nous utilisons l'algorithme 5 :

Algorithme 5 Estimation des intervalles de confiance sur u_1^2, u_2^2 et u_3^2

Pour i allant de 1 à n_i **Faire**
Sélectionner nb_{lab} parmi les a laboratoires,
Estimer les paramètres u_1^2, u_2^2 et u_3^2 sur l'échantillon i.
fin Pour
Calculer des intervalles de confiance à 95% pour les valeurs estimées $\hat{u}_1^2$, $\hat{u}_2^2$, et $\hat{u}_3^2$ en calculant les quantiles à $\alpha/2\%$ et à $100 - \alpha/2\%$ à partir des distributions obtenues pour chacun de ces coefficients.

Nous choisissons de réaliser $n_i = 10000$ itérations de l'algorithme, et de rééchantilloner $nb_{lab} = \lfloor (3a)/4 \rfloor$ laboratoires parmi l'ensemble des laboratoires participants à un essai. Nous pouvons ainsi obtenir les variances des estimateurs des coefficients de dispersion.

3.5 Ajustement des données réelles au modèle

Dans les sections précédentes, nous avons défini un modèle pour notre problématique, et nous avons présenté une méthode permettant d'en estimer les paramètres. Il est maintenant nécessaire de s'intéresser à l'ajustement du modèle aux données issues du contrôle de qualité en microbiologie, c'est à dire d'étudier dans quelle mesure le modèle proposé s'ajuste à de tels échantillons. Comme souligné en section 2.3 du chapitre 2, les critères d'adéquation d'un modèle à un échantillon mesurent l'écart entre les valeurs observées sur cet échantillon et les valeurs attendues d'après le modèle considéré.

Il a été souligné dans [Lee and Nelder, 2002] que l'adéquation de modèles avec effets aléatoires était encore trop peu étudiée. Toutefois, la procédure d'estimation dans les HGLM présentée en

section 3.4 revient à ajuster les données à deux GLM interconnectés. [Lee and Nelder, 2002] proposent par conséquent de transposer les tests et procédures d'étude d'ajustement des GLM aux HGLM.

3.5.1 Test de déviance normalisée

Un test d'ajustement d'un HGLM aux données est proposé dans [Lee and Nelder, 1996]. Les hypothèses du test sont les suivantes :

$$\begin{cases} \mathcal{H}_0 : \text{Les données s'ajustent au modèle postulé} \\ \mathcal{H}_1 : \text{Les données ne s'ajustent pas au modèle postulé.} \end{cases}$$

Ce test est basé sur la déviance normalisée (*scaled deviance* en anglais), précédemment introduite pour les GLM en section 3.2.3. [Lee et al., 2006] soulignent que ce test, usuellement utilisé pour les GLM, peut être étendu aux HGLM. Notons $\hat{y}_{ijk}$ les valeurs ajustées suivant le HGLM (3.9). Pour ce modèle, $\phi = 1$, par conséquent la déviance normalisée correspond alors à la déviance classique, qui s'exprime uniquement à partir des données y_{ijk}. Elle peut alors s'écrire de la façon suivante :

$$D(y_{ijk}, \hat{y}_{ijk}) = -2\{l(\hat{y}_{ijk}; y_{ijk}|\alpha_i, \beta_{ij}, \gamma_{ijk}) - l(y_{ijk}; y_{ijk}|\alpha_i, \beta_{ij}, \gamma_{ijk})\} \tag{3.30}$$

La littérature ne propose toutefois pas de critère de comparaison strict pour le test de déviance normalisée. Une étude par simulation de la distribution de la déviance du modèle 3.9 nous a permis d'aboutir à la conclusion que la déviance normalisée $D(y_{ijk}, \hat{y}_{ijk})$ ne suit ni une loi du χ^2, ni une autre loi connue. Par conséquent, nous allons effectuer un test par bootstrap paramétrique pour statuer sur l'adéquation du modèle 3.9 aux données. Cette méthode a notamment été utilisée par [Li and Wang, 1998]. Nous allons procéder par simulation de la loi de la déviance sous l'hypothèse d'ajustement du modèle aux données.

3.5.2 Exemples de données issues d'essais interlaboratoires en microbiologie

Nous proposons ici d'étudier deux jeux de données d'AGLAE correspondant à deux essais interlaboratoires aux caractéristiques bien distinctes.

Jeu de données de *Pseudomonas aeruginosa*. Le premier jeu de données que nous étudions ici correspond à des dénombrements des *Pseudomonas aeruginosa* dans les eaux. Les caractéristiques de ce jeu de données sont les suivantes :

- Une méthode relativement bien décrite d'un point de vue normatif,
- Un type de colonies assez bien retrouvé par les participants.

Ce jeu de données correspond à un type de dénombrement de difficulté intermédiaire en termes d'exploitation des résultats d'essais interlaboratoires. Il présente les résultats de 202 laboratoires.

Dans un premier temps, nous estimons le paramètre fixe μ du modèle 3.9, dont l'exponentielle correspond à la moyenne des dénombrements bactériens sur l'essai, ainsi que les coefficients de dispersion u_1^2, u_2^2 et u_3^2. Nous utilisons la méthode décrite dans la section 3.4.2, implémentée dans le package R *HGLM*. Les estimateurs obtenus sont donnés dans la Table 3.5.

paramètre	valeur estimée	Intervalle de confiance au risque de 5%
μ	4.00294	$[3.979037; 4.026242]$
u_1^2	0.0269029	$[0.01823827; 0.03528619]$
u_2^2	0.0008897	$[0.000362266; 0.001962858]$
u_3^2	0.0004882	$[0.0003050309; 0.0009033621]$

TABLE 3.5 – Intervalles de confiance des estimateurs sur le jeu de données de *Pseudomonas aeruginosa*

Remarque 3.5.1. *L'estimateur du coefficient de dispersion du facteur Laboratoire, $\hat{u}_1^2$, est relativement petit, ce qui est cohérent avec le fait que les colonies de Pseudomonas aeruginosa sont bien retrouvées par les participants.*

Nous pouvons obtenir des intervalles de confiance sur ces valeurs estimées à l'aide d'un bootstrap paramétrique, comme présenté en section 3.4.7. Nous effectuons $N = 10000$ rééchantillonage de 140 laboratoires parmi les 202 laboratoires du jeu de données de *Pseudomonas aeruginosa*. Les intervalles de confiance au risque de 5% obtenus sont données dans la Table 3.5.

Pour étudier l'adéquation du modèle 3.9 au jeu de données de *Pseudomonas aeruginosa*, nous effectuons le test de déviance normalisée proposé ci-dessus. Nous obtenons une p-valeur $pv > 0.9$, avec $N = 1000$ simulations, qui nous amène à la conclusion que le modèle s'ajuste bien aux données.

Les diagrammes quantiles-quantiles représentés en Figure 3.1 montrent que les facteurs Laboratoire, Flacon, et Réplication, sont effectivement distribués suivant des lois Gamma.

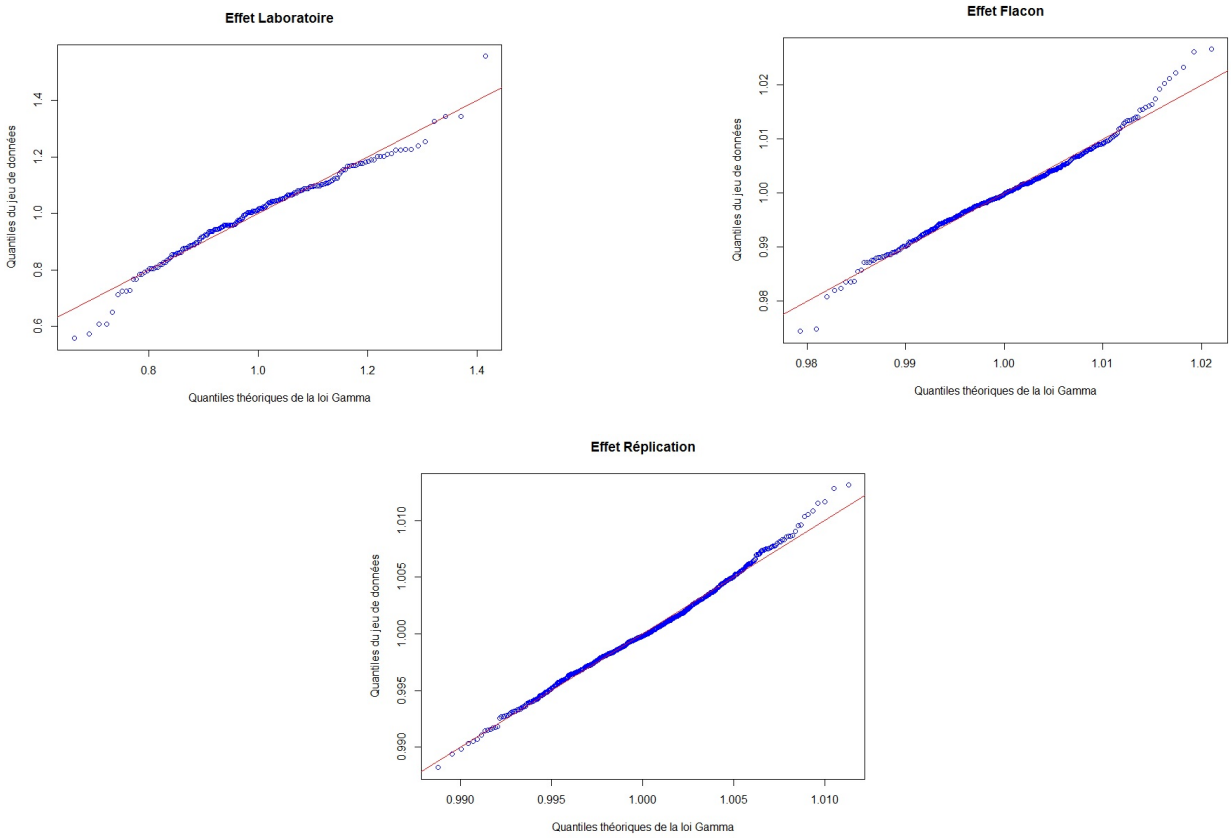

FIGURE 3.1 – Ajustement des facteurs à des distributions Gamma pour l'essai sur *Pseudomonas aeruginosa*

D'autre part, le graphique des résidus présenté en Figure 3.2 ne présente aucune tendance particulière, et les valeurs ajustées obtenues suivant le modèle 3.9 permettent de prédire correctement les mesures effectuées par les laboratoires participant à l'essai.

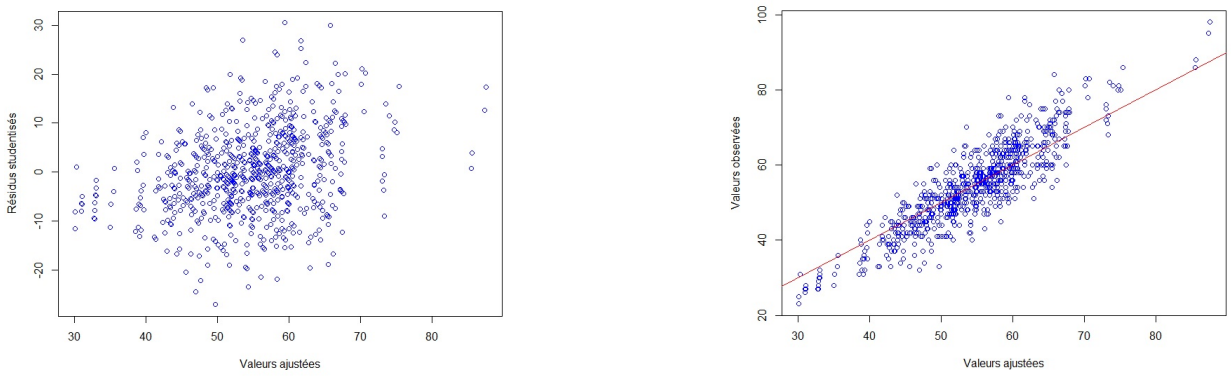

FIGURE 3.2 – Graphiques d'ajustement des données de l'essai sur *Pseudomonas aeruginosa* au modèle postulé

Jeu de données de *Staphylocoques* pathogènes. Nous étudions ici un second jeu de données, issu d'un essai interlaboratoires de dénombrement de *staphylocoques* pathogènes dans les eaux. Les caractéristiques de cet essai sont les suivantes :

- Des disparités interlaboratoires notoires liées à l'utilisation de méthodes différentes (milieu de culture différents) par le panel de participants,
- Une dispersion interlaboratoires relativement large, qui suscite des réflexions de la normalisation AFNOR, en vue d'une tentative d'harmonisation des pratiques.

Il s'agit d'un type de dénombrement plutôt délicat à exploiter. Le jeu de données présente les résultats de 167 laboratoires.
Nous ajustons les données de cet essai suivant le modèle 3.9. Nous estimons ensuite les paramètres correspondants à l'aide du package *HGLM* de R. Nous obtenons ainsi les estimateurs présentés en Table 3.6.

paramètre	valeur estimée	Intervalle de confiance au risque de 5%
μ	3.11840	$[3.053135; 3.178043]$
u_1^2	0.181982	$[0.1272418; 0.2380033]$
u_2^2	0.008177	$[0.001716398; 0.01889575]$
u_3^2	0.001525	$[0.0005353817; 0.003300832]$

TABLE 3.6 – Intervalles de confiance des estimateurs sur le jeu de données de *Staphylocoques* pathogènes

Remarque 3.5.2. *On constate que le coefficient de surdispersion de l'effet Laboratoire estimé,* $\hat{u}_1^2 = 0.181982$*, sur le jeu de données de Staphylocoques pathogènes est nettement supérieure à celui obtenu sur le jeu de données de Pseudomonas aeruginosa.*

Nous calculons ensuite des intervalles de confiance sur les valeurs estimées à l'aide d'un bootstrap paramétrique, comme présenté en section 3.4.7. Nous effectuons $N = 10000$ rééchantillonage de 117 laboratoires parmi les 167 laboratoires du jeu de données de *Staphylocoques* pathogènes étudié. Les intervalles de confiance au risque de 5% obtenus sont données dans la Table 3.6.

Afin de vérifier que le jeu de données de *Staphylocoques* pathogènes s'ajuste bien au modèle postulé 3.9, nous effectuons un test de déviance normalisée. La p-valeur obtenue, $pv > 0.9$, avec $N = 1000$ simulations nous conduit à conclure à l'ajustement des données de l'essai au modèle postulé.

Par ailleurs, les diagrammes quantile-quantile présentés en Figure 3.3 suggèrent que, pour le jeu de données de *Staphylocoques* pathogènes, les facteurs Laboratoire, Flacon, et Réplication suivent effectivement des distributions Gamma.

Pour étudier l'ajustement de données à un modèle, [Lee and Nelder, 1998] suggèrent de tracer les valeurs ajustées en fonctions des résidus studentisés. On constate sur la Figure 3.4 que ce graphique ne présente aucune tendance particulière. D'autre part, les valeurs prédites à l'aide du modèle 3.9 sont très proches des mesures effectivement réalisées par les laboratoires participant à l'essai. Les deux graphiques correspondant prouvent que les données du jeu de données de *Staphylocoques* pathogènes s'ajustent bien au modèle 3.9.

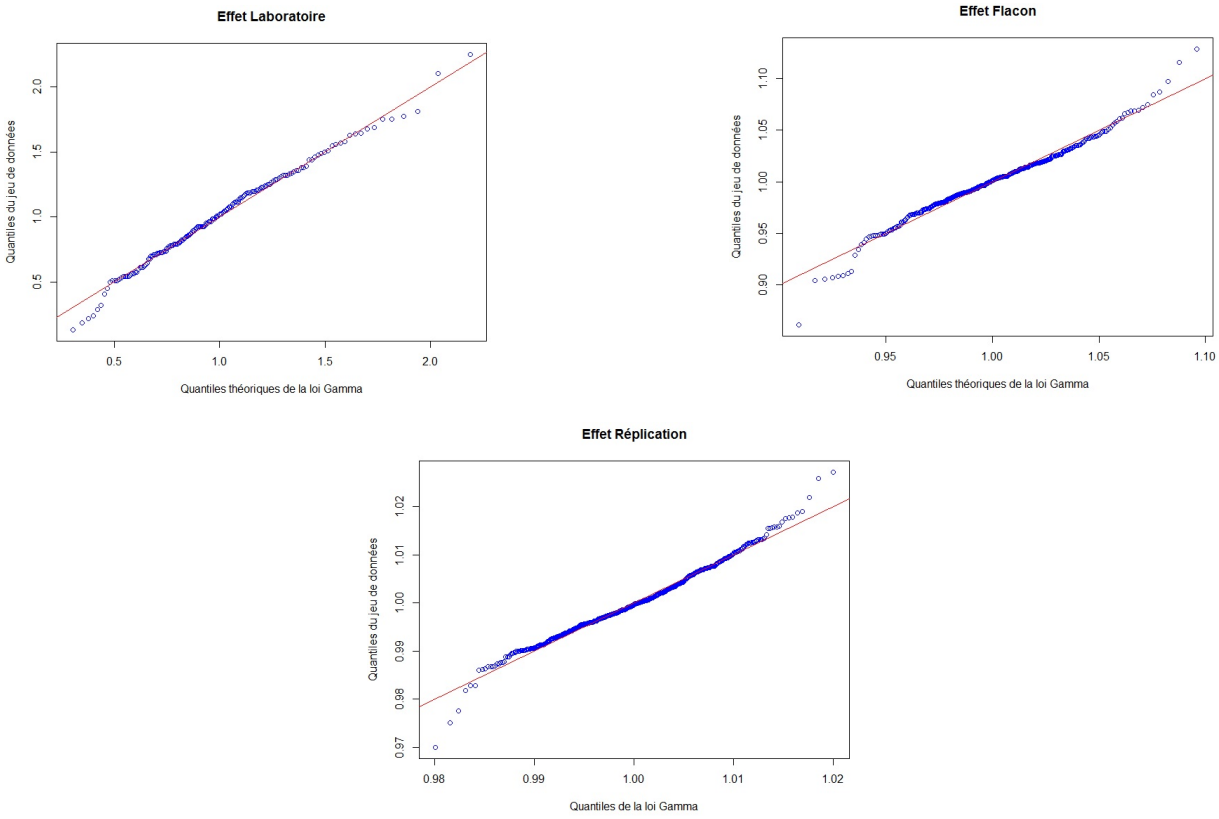

FIGURE 3.3 – Ajustement des facteurs à des distributions Gamma pour l'essai sur *Staphylocoques* pathogènes

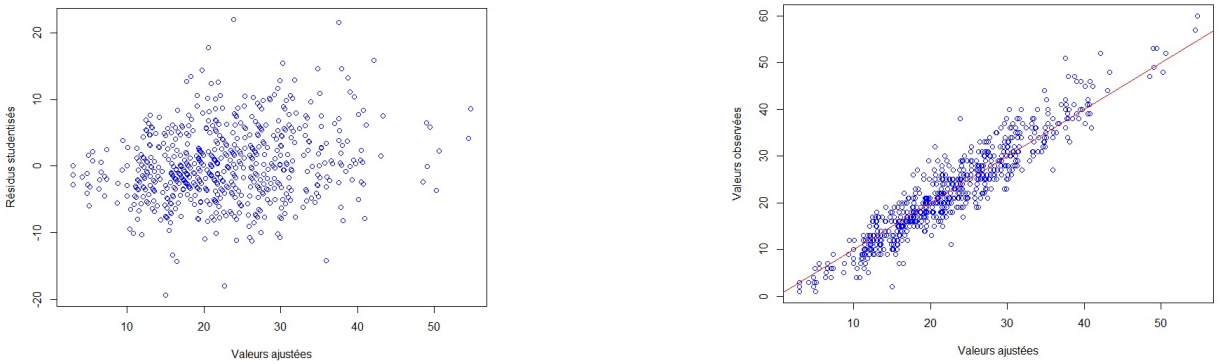

FIGURE 3.4 – Graphiques d'ajustement des données de l'essai sur *Staphylocoques* pathogènes au modèle postulé

3.6 Test de significativité des composantes aléatoires

Le modèle 3.9 comporte trois facteurs : Laboratoire, Flacon, et Réplication. Pour un essai interlaboratoire donné, nous souhaitons savoir si les effets de chacun de ces facteurs sont significatifs ou non. Notamment, nous souhaitons tester les hypothèses suivantes pour le facteur Laboratoire :

$$\begin{cases} \mathcal{H}_0 : u_1^2 = 0 : \text{L'effet Laboratoire n'est pas significatif} \\ \mathcal{H}_1 : u_1^2 \neq 0 : \text{L'effet Laboratoire est significatif.} \end{cases}$$

Remarque 3.6.1. *Nous pouvons formuler des tests similaires sur u_2^2 et u_3^2, pour les effets des*

facteurs Flacon et Réplication respectivement.

Lorsque le choix entre plusieurs modèles n'est pas clair, ou lorsqu'on hésite à prendre en compte certains effets dans le modèle retenu, [Besse,] propose de regarder les critères de choix de modèle tels que le AIC. Si l'un des modèles envisagés minimise ce critère, ce modèle pourra être retenu. En particulier, pour la sélection de modèle dans les HGLM, [Lee et al., 2006] recommandent l'utilisation du AIC conditionnel introduit par [Vaida and Blanchard, 2005], préférentiellement aux profils de vraisemblance $p_\mu = L(\mu)$ et $p_{\mu,\alpha,\beta,\gamma} = L(\mu, \alpha, \beta, \gamma)$. Cette méthode a notamment été utilisée par [Ha et al., 2007] pour la sélection de modèles de fragilité à plusieurs composantes.

Le AIC conditionnel est défini de la façon suivante :

$$AIC_c = abn \log \phi + \sum (Y - \tilde{Z}^T \hat{b})^2 / \phi + 2p_D,$$

où p_D intervient dans le calcul du degré de liberté de la déviance normalisée, et est spécifié dans [Lee et al., 2006, p193].

Dans cette section, nous proposons d'effectuer des tests de significativité des différents facteurs en comparant des modèles incluant ou non les différents effets sur la base du critère AIC conditionnel. Pour un essai interlaboratoires étudié, nous proposons alors de comparer les modèles suivants :

- Modèle 1 : Modèle sans effet
- Modèle 2 : Modèle à effet Laboratoire seul
- Modèle 3 : Modèle à effet Flacon seul
- Modèle 4 : Modèle à effet Réplication seul
- Modèle 5 : Modèle à effet Laboratoire et effet Flacon
- Modèle 6 : Modèle à effet Laboratoire et effet Réplication
- Modèle 7 : Modèle à effet Flacon et effet Réplication
- Modèle 8 : Modèle complet : avec effet Laboratoire, effet Flacon, et effet Réplication

Pour déterminer si cette stratégie est pertinente, nous proposons d'en étudier la puissance. Pour cela, nous allons simuler 1000 jeux de données d'essais interlaboratoires correspondant à chacune des 8 situations décrites ci-dessous. Pour chacun des 1000 jeux de données ainsi simulés, nous calculons les AIC conditionnels correspondants aux 8 modèles présentés ci-dessus. Nous comptons finalement le nombre de fois ou chacun des modèles a été choisi. Les 8 situations considérées sont les suivantes :

- Cas 1 : Jeu de données sans effet : $\mu = 5$, $u_1^2 = u_2^2 = u_3^2 = 0$
- Cas 2 : Jeu de données avec effet Laboratoire seul : $\mu = 5$, $u_1^2 = 0.2$, $u_2^2 = u_3^2 = 0$
- Cas 3 : Jeu de données avec effet Flacon seul : $\mu = 5$, $u_2^2 = 0.1$, $u_1^2 = u_3^2 = 0$
- Cas 4 : Jeu de données avec effet Réplication seul : $\mu = 5$, $u_3^2 = 0.1$, $u_1^2 = u_2^2 = 0$
- Cas 5 : Jeu de données avec effet Laboratoire et Flacon : $\mu = 5$, $u_1^2 = 0.2$, $u_2^2 = 0.1$, $u_3^2 = 0$

- Cas 6 : Jeu de données avec effet Laboratoire et Réplication : $\mu = 5$, $u_1^2 = 0.2$, $u_2^2 = 0$, $u_3^2 = 0.1$
- Cas 7 : Jeu de données avec effet Flacon et Réplication : $\mu = 5$, $u_1^2 = 0$, $u_2^2 = u_3^2 = 0.1$
- Cas 8 : Jeu de données avec effet Laboratoire, Flacon et Réplication (modèle complet) : $\mu = 5$, $u_1^2 = 0.2$, $u_2^2 = u_3^2 = 0.1$

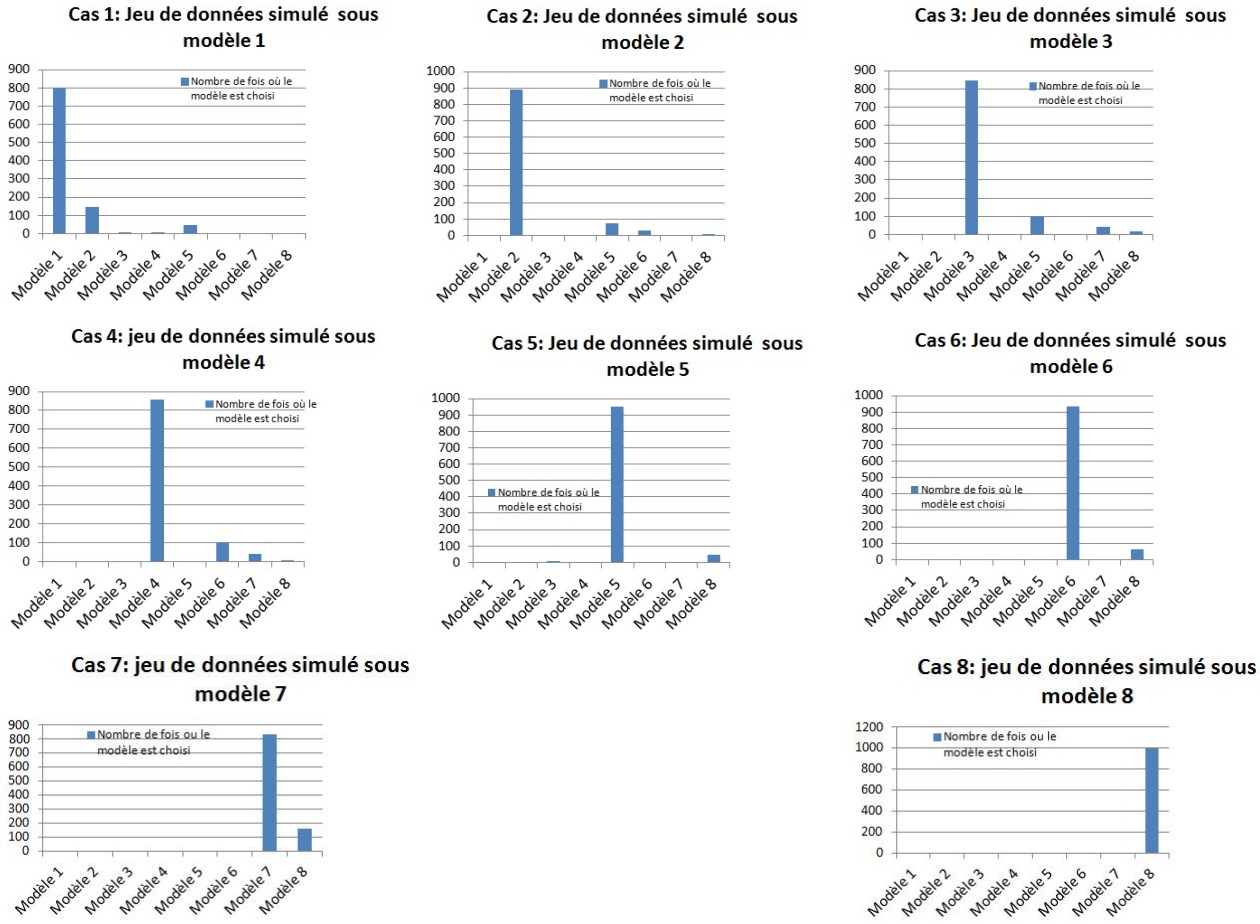

FIGURE 3.5 – Proportions de modèles choisis avec le AIC conditionnel dans les 8 cas étudiés

Pour chacune des 8 situations considérées, on constate (Figure 3.5 et Tables E.1, E.2, E.3, E.4, E.5, E.6, E.7, E.8 en Annexe E) que la grande majorité des modèles choisis correspond au modèle que nous avons simulé. En effet, la proportion de modèles choisis correspondant au modèle simulé est supérieur où égale à $80,2\%$. La stratégie de test proposée nous paraît donc pertinente.

3.6.1 Application à des données issues du contrôle de qualité en microbiologie

Jeu de données de *Pseudomonas aeruginosa*. Nous appliquons la stratégie proposée pour tester la significativité des trois effets sur le jeu de données de *Pseudomonas aeruginosa* introduit en section 3.5.2. La Table 3.6 présente les résultats de ce test.

Nous constatons que le modèle présentant le plus petit AIC conditionnel est le modèle 2, c'est-à-dire le modèle avec effet Laboratoire seul. Nous concluons donc à l'existence d'un effet Laboratoire, et à l'absence d'effets Flacon et Réplication pour cet essai interlaboratoires. Les résultats de l'estimation des coefficients de dispersion présentés en Table 3.5 corroborent les résultats de notre test. En effet, le coefficient de dispersion associé à l'effet Laboratoire, u_1^2 vaut 0.0269029, tandis que les coefficients de dispersion associées aux effet Flacon et Réplication sont tous deux très proches de zéro.

	AIC conditionnel
Modèle 1	6173.548
Modèle 2	**5526.092**
Modèle 3	5681.883
Modèle 4	5978.704
Modèle 5	5539.439
Modèle 6	5553.097
Modèle 7	5712.007
Modèle 8	5546.53

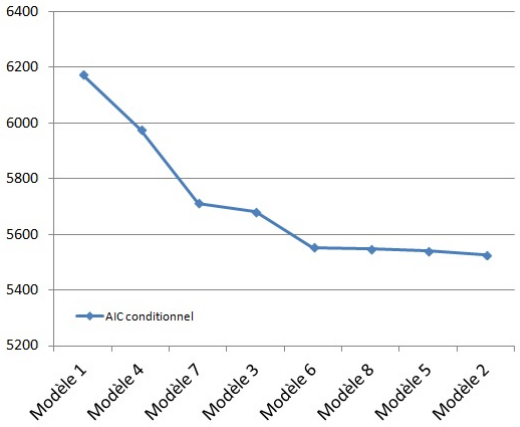

FIGURE 3.6 – Sélection de modèle par AIC conditionnel - Jeu de données de *Pseudomonas aeruginosa*

Jeu de données de *Staphylocoques* pathogènes. Nous appliquons ensuite la méthode de test proposée sur le jeu de données de *Staphylocoques* pathogènes introduit en section 3.5.2. La Table 3.7 présente les résultats obtenus avec cette stratégie.

	AIC conditionnel
Modèle 1	4988.098
Modèle 2	4077.47
Modèle 3	4142.1
Modèle 4	4481.155
Modèle 5	**4061.256**
Modèle 6	4086.676
Modèle 7	4152.863
Modèle 8	4068.263

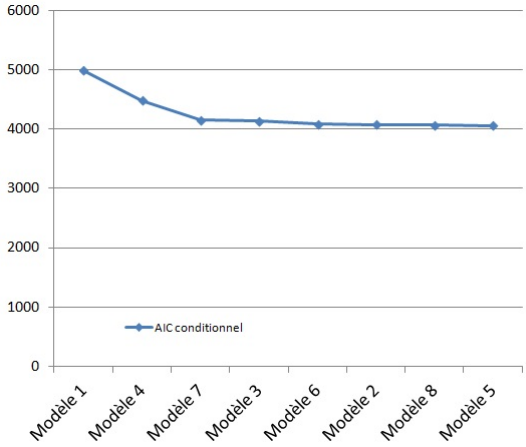

FIGURE 3.7 – Sélection de modèle par AIC conditionnel - Jeu de données de *Staphylocoques* pathogènes

Nous constatons que ces résultats corroborent les valeurs estimées des coefficients de disper-

sion, présentées en Table 3.6. En effet, l'estimateur du coefficient de dispersion de l'effet Laboratoire, u_1^2, présente une valeur bien plus élevée que les coefficients de dispersion associés aux deux autres facteurs. Lorsqu'on étudie les valeurs des AIC conditionnels obtenus pour chacun des modèles considérés, on constate que tous les modèles prenant en compte l'effet Laboratoire ont des AIC conditionnels bien plus petits que les autres modèles. Il nous paraît alors clair que l'effet Laboratoire est significatif. Le modèle présentant le plus petit AIC conditionnel est finalement le modèle 5, correspondant au modèle avec effet Laboratoire et effet Flacon. La significativité du facteur Flacon et la non- significativité du facteur Réplication sont d'autre part cohérentes avec les valeurs de $\hat{u}_2^2$ et $\hat{u}_3^2$.

3.7 Évaluation de la performance analytique des laboratoires

Nous avons vu au chapitre 1, section 1.8.1, que la composante laboratoire du biais de chaque participant est exprimée en terme de z-score. Ce z-score permet au laboratoire de tracer une carte de contrôle de son erreur de mesure. AGLAE est accréditée pour cette activité. Le référentiel régissant l'accréditation d'AGLAE, [ISO, 2010a], impose de ne pas faire peser sur le laboratoire les risques liés à l'hétérogénéité de lot et à l'instabilité du matériau, et par extension à l'erreur aléatoire résiduelle (la répétabilité). La finalité de cette contrainte est que les laboratoires un peu marginaux ne reçoivent pas un signal trop interféré par l'erreur aléatoire.
C'est pourquoi le z-score proposé en partie 1.8.1 prend en compte l'erreur introduite par l'effet Flacon, ainsi que l'erreur résiduelle. Ce mode de scoring a été audité et jugé conforme à la norme [ISO, 2010a].

Pour définir une méthode de scoring pour le modèle considéré dans ce chapitre, nous nous intéressons à la somme des résultats de dénombrements d'un laboratoire. Soit S_i la variable aléatoire correspondante :

$$S_i = y_{i11} + y_{i12} + y_{i21} + y_{i22} \qquad \text{avec} \quad i = \{1, \dots, a\}$$

Remarque 3.7.1. *$\forall (i,j,k)$, les variables α_i, β_{ij}, et γ_{ijk} sont indépendantes. En revanche, les variables y_{i11}, y_{i12}, y_{i21} et y_{i22} sont dépendantes car elles partagent une information commune, α_i.*

Calculons l'espérance $\mathbb{E}(S_i)$ et la variance $\mathbb{V}(S_i)$ de S_i, la somme des mesures du laboratoire i. Rappelons dans un premier temps les propriétés suivantes :

Propriété 3.7.1. *Soit X et Y deux variables aléatoires indépendantes. Alors*

- $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$
- $\mathbb{V}(X \cdot Y) = \mathbb{V}(X) \cdot \mathbb{V}(Y) + \mathbb{V}(X) \cdot \mathbb{E}^2(Y) + \mathbb{V}(Y) \cdot \mathbb{E}^2(X)$

Propriété 3.7.2. *Soit X et Y deux variables. Alors*

- $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$
- $\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}(Y|X))$

La Propriété 3.7.2 intervient notamment dans le calcul des espérances et variances des variables aléatoires données par des lois composées, notamment le couple Gamma-Poisson.

Proposition 3.7.1. *Soient y_{i11}, y_{i12}, y_{i21} et y_{i22} telles que*

$$y_{ijk} \sim \mathcal{P}(e^{\mu}e^{\alpha_i}e^{\beta_{ij}}e^{\gamma_{ijk}})$$

avec

$$\begin{cases} \mu \in \mathbb{R} \\ e^{\alpha_i} \sim \Gamma(\frac{1}{u_1^2}, \frac{1}{u_1^2}) \\ e^{\beta_{ij}} \sim \Gamma(\frac{1}{u_2^2}, \frac{1}{u_2^2}) \\ e^{\gamma_{ijk}} \sim \Gamma(\frac{1}{u_3^2}, \frac{1}{u_3^2}) \end{cases}$$

Alors

- $\mathbb{E}(y_{i11} + y_{i12} + y_{i21} + y_{i22}) = 4e^{\mu}$
- $\mathbb{V}(y_{i11}+y_{i12}+y_{i21}+y_{i22}) = 4e^{\mu}+e^{2\mu}(16u_1^2+8u_2^2+4u_3^2+8u_1^2u_2^2+4u_1^2u_3^2+4u_2^2u_3^2+4u_1^2u_2^2u_3^2)$

Remarque 3.7.2. *Dans la Proposition 3.7.1, on observe le terme de surdispersion de la variance par rapport à la moyenne.*

Preuve : Calculons l'espérance $\mathbb{E}(y_{i11} + y_{i12} + y_{i21} + y_{i22})$:

$$\begin{aligned} \mathbb{E}(y_{i11} + y_{i12} + y_{i21} + y_{i22}) &= \mathbb{E}(y_{i11}) + \mathbb{E}(y_{i12}) + \mathbb{E}(y_{i21}) + \mathbb{E}(y_{i22}) \\ &= e^{\mu} + e^{\mu} + e^{\mu} + e^{\mu} \\ &= 4e^{\mu} \end{aligned}$$

Notons $X = (\alpha_i, \beta_{i1}, \beta_{i2}, \gamma_{i21}, \gamma_{i22})$. La variance $\mathbb{V}(y_{i11} + y_{i12} + y_{i21} + y_{i22})$ s'écrit alors :

$$\begin{aligned} \mathbb{V}(y_{i11} + y_{i12} + y_{i21} + y_{i22}) =& \mathbb{V}(\mathbb{E}(y_{i11} + y_{i12} + y_{i21} + y_{i22}|X)) \\ &+ \mathbb{E}(\mathbb{V}(y_{i11} + y_{i12} + y_{i21} + y_{i22}|X)) \end{aligned} \tag{3.31}$$

Or, la linéarité de l'espérance conditionnelle donne :

$$\mathbb{E}(y_{i11} + y_{i12} + y_{i21} + y_{i22}|X) = \mathbb{E}(y_{i11}|X) + \mathbb{E}(y_{i12}|X) + \mathbb{E}(y_{i21}|X) + \mathbb{E}(y_{i22}|X)$$

avec

$$\mathbb{E}(y_{ijk}|X) = e^{\mu}e^{\alpha_i}e^{\beta_{ij}}e^{\gamma_{ijk}}$$

Le premier terme de 3.31 peut alors s'écrire

$$\begin{aligned} \mathbb{V}(\mathbb{E}(y_{i11} + y_{i12} + y_{i21} + y_{i22}|X)) =& \mathbb{V}\left(e^{\mu}e^{\alpha_i}[e^{\beta_{i1}}(e^{\gamma_{i11}} + e^{\gamma_{i12}}) + e^{\beta_{i2}}(e^{\gamma_{i21}} + e^{\gamma_{i22}})]|X\right) \\ =& e^{2\mu}\{u_1^2 \cdot \mathbb{V}[e^{\beta_{i1}}(e^{\gamma_{i11}} + e^{\gamma_{i12}}) + e^{\beta_{i2}}(e^{\gamma_{i21}} + e^{\gamma_{i22}})] \\ &+ 1 \cdot \mathbb{V}[e^{\beta_{i1}}(e^{\gamma_{i11}} + e^{\gamma_{i12}}) + e^{\beta_{i2}}(e^{\gamma_{i21}} + e^{\gamma_{i22}})] + u_1^2(2+2)^2\} \\ =& e^{2\mu}\{(u_1^2+1)[\mathbb{V}(e^{\beta_{i1}}(e^{\gamma_{i11}} + e^{\gamma_{i12}})) + \mathbb{V}(e^{\beta_{i2}}(e^{\gamma_{i21}} + e^{\gamma_{i22}}))] + 16u_1^2\} \\ =& e^{2\mu}\left(16u_1^2 + 8u_2^2 + 4u_3^2 + 8u_1^2u_2^2 + 4u_1^2u_3^2 + 4u_2^2u_3^2 + 4u_1^2u_2^2u_3^2\right) \end{aligned} \tag{3.32}$$

Calculons maintenant le deuxième terme de 3.31 :

$$\begin{aligned} \mathbb{V}(y_{i11} + y_{i12} + y_{i21} + y_{i22}|X) &= e^{\mu}e^{\alpha_i}e^{\beta_{i1}}e^{\gamma_{i11}} + e^{\mu}e^{\alpha_i}e^{\beta_{i1}}e^{\gamma_{i12}} + e^{\mu}e^{\alpha_i}e^{\beta_{i2}}e^{\gamma_{i21}} + e^{\mu}e^{\alpha_i}e^{\beta_{i2}}e^{\gamma_{i22}} \\ &= e^{\mu}e^{\alpha_i}[e^{\beta_{i1}}(e^{\gamma_{i11}} + e^{\gamma_{i12}}) + e^{\beta_{i2}}(e^{\gamma_{i21}} + e^{\gamma_{i22}})] \end{aligned}$$

Donc

$$\begin{aligned}\mathbb{E}(\mathbb{V}(y_{i11}+y_{i12}+y_{i21}+y_{i22}|X)) =&\mathbb{E}[e^{\mu}e^{\alpha_i}[e^{\beta_{i1}}(e^{\gamma_{i11}}+e^{\gamma_{i12}})+e^{\beta_{i2}}(e^{\gamma_{i21}}+e^{\gamma_{i22}})]]\\ =&4e^{\mu}\end{aligned} \tag{3.33}$$

En combinant 3.32 et 3.33, la variance $\mathbb{V}(y_{i11}+y_{i12}+y_{i21}+y_{i22})$ s'écrit alors :

$$\mathbb{V}(y_{i11}+y_{i12}+y_{i21}+y_{i22}) = 4e^{\mu}+e^{2\mu}(16u_1^2+8u_2^2+4u_3^2+8u_1^2u_2^2+4u_1^2u_3^2+4u_2^2u_3^2+4u_1^2u_2^2u_3^2$$

□

Notons

$$\begin{cases}\mathbb{E}(y_{i11}+y_{i12}+y_{i21}+y_{i22}) = M_S\\ \mathbb{V}(y_{i11}+y_{i12}+y_{i21}+y_{i22}) = V_S.\end{cases}$$

En vue de la définition d'un score pour les laboratoires, approchons S_i par une loi binomiale négative $\mathcal{BN}(r,p)$, avec $r>0$, et $p\in(0,1)$, d'espérance M_S et de variance V_S. On devrait alors avoir :

$$\begin{cases}\frac{r(1-p)}{p} = M_S\\ \frac{r(1-p)}{p^2} = V_S,\end{cases}$$

autrement dit,

$$\begin{cases}p = \frac{M_S}{V_S^2}\\ r = \frac{M_S^2}{V_S^2-M_S}.\end{cases} \tag{3.34}$$

Pour vérifier la pertinence de l'approximation proposée, nous avons effectué des simulations en faisant varier les paramètres dans notre gamme de travail. la démarche proposée est la suivante :

- On fixe μ, u_1^2, u_2^2, u_3^2.
- On génère 10^3 laboratoires participant à un tel essai interlaboratoires avec ces paramètres. On a alors 10^3 jeux de quatre mesure $\{y_{i11}, y_{i12}, y_{i21}, y_{i22}\}$.
- On trace le diagramme quantile-quantile de l'échantillon $(S_i,\dots,S_N)$ par rapport à la loi binomiale négative de paramètres 3.34 correspondante.

Nous étudions ici plusieurs configurations :

- Cas 1 : $\mu=5$, $u1=0.2$, $u2=0.1$, et $u3=0.05$.
- Cas 2 : $\mu=5$, $u1=0$, $u2=0.1$, et $u3=0.05$,
- Cas 3 : $\mu=5$, $u1=0.1$, $u2=0$, et $u3=0.05$,
- Cas 4 : $\mu=5$, $u1=0.1$, $u2=0.1$, et $u3=0$,
- Cas 5 : $\mu=1$, $u1=0.2$, $u2=0.1$, et $u3=0.05$,
- Cas 6 : $\mu=7$, $u1=0.2$, $u2=0.1$, et $u3=0.05$.

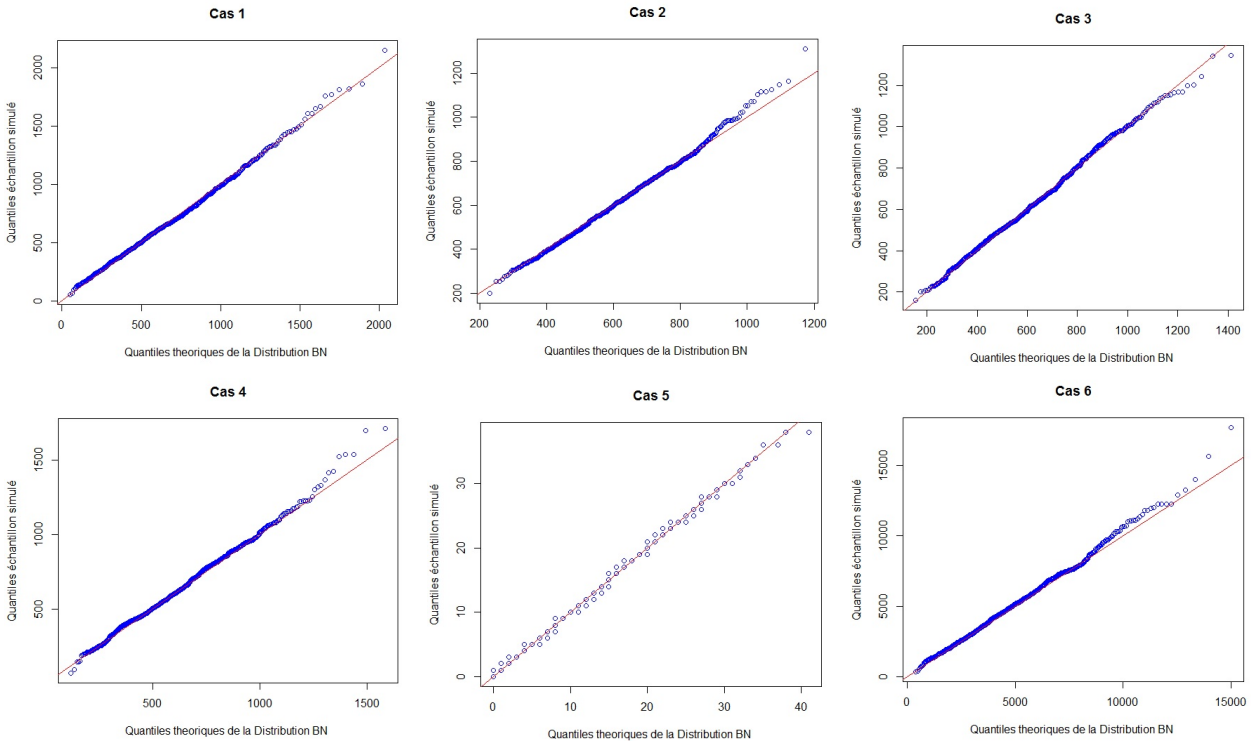

FIGURE 3.8 – Adéquation de S_i à une loi binomiale négative pour différentes configurations

Dans les huit cas présentés ci-dessus, nous considérons des modèles avec un ou plusieurs effets, à des niveaux de concentration $\exp(\mu)$ différents. Les diagrammes quantile-quantile présentés en Table 3.8 montrent que la variable aléatoire S_i semble effectivement suivre une loi binomiale négative de paramètres 3.34.

Pour calculer un score pour chaque laboratoire participant à l'essai, nous pouvons ensuite ramener la distribution binomiale négative à une distribution normale centrée réduite. Nous pouvons calculer le z-score à partir de l'approximation à la loi binomiale négative proposée. Le z-score est alors le quantile de la loi normale centrée réduite correspondant. Autrement dit, pour le laboratoire i, on calcule d'abord $\mathbb{P}(\mathcal{BN}(r,p) \geq S_i) = p_i$. Le z-score du laboratoire i correspond alors à $z_i = \mathbb{P}(\mathcal{N}(0,1) \leq p_i)$.

3.8 Conclusion

Dans ce chapitre, nous revenons tout d'abord sur les extensions du modèle linéaire qui ont été développées au cours du temps (Section 3.2), par l'introduction des effets aléatoires d'une part (L2M), et par la généralisation en terme de distribution de la variable à expliquer et de lien à la linéarité (GLM) d'autre part. La combinaison de ces deux extensions a donné naissance aux GL2M, puis aux HGLM. Dans ces derniers modèles, la contrainte sur la distribution des variables explicatives est moins stricte. Il suffit qu'elles appartiennent à la famille exponentielle.

Nous avons ensuite présenté le mélange Gamma-Poisson, qui permet de modéliser la surdispersion des résultats de dénombrement en microbiologie (Section 3.3). Nous en retenons finalement une paramétrisation faisant intervenir un coefficient de variation CV^2, et un coefficient de dispersion, u^2. L'avantage de cette paramétrisation est que l'incertitude totale exprimée en coefficient de variation CV^2 correspond à la somme de l'incertitude liée à la loi de Poisson, plus la somme des incertitudes liées aux différents facteurs sources d'erreur.

Nous proposons alors un HGLM Gamma-Poisson utilisant cette paramétrisation afin d'expliquer la variabilité des résultats de dénombrement en microbiologie. Nous présentons la méthode utilisée pour estimer les différents paramètres du modèle (Section 3.4). Cette méthode requière l'utilisation d'une vraisemblance particulière, la h-vraisemblance, et d'un algorithme backfitting, permettant d'estimer à la fois l'effet fixe, les effets aléatoires, et les coefficients de dispersion. L'ajustement du modèle aux données réelles est étudié grâce à un test basé sur la déviance normalisée (Section 3.5). Nous présentons également des graphiques d'ajustement pour deux essais interlaboratoires en microbiologie. A partir du modèle proposé, nous définissons de nouveaux tests de significativité des effets aléatoires (Section 3.6). Nous proposons aussi une méthode d'évaluation de la performance analytique des laboratoires, basée sur une approximation de la distribution de la variable somme des résultats de mesures des laboratoires par une loi binomiale négative (Section 3.7).

Chapitre 4

Produit de variables aléatoires indépendantes de lois Gamma

Sommaire

4.1 Introduction

Au chapitre 3, nous proposons un modèle linéaire généralisé hiérarchique (3.9) pour expliquer la variabilité des résultats de dénombrements d'un essai interlaboratoires en microbiologie. Ce modèle correspond à un processus de Poisson dont l'intensité est distribuée suivant un produit de trois variables aléatoires indépendantes de lois Gamma. Dans ce chapitre, nous nous intéressons à la distribution de l'intensité de ce processus de Poisson. Sa caractérisation a en effet plusieurs applications pratiques pour le contrôle de qualité en microbiologie.

Il n'existe pas de propriété sur le produit de variables aléatoires indépendantes de lois Gamma. Comme le soulignent [Withers and Nadarajah, 2013], nous connaissons la forme analytique de sa densité. Elle peut effectivement s'exprimer à l'aide de la fonction de Bessel. Toutefois, nous n'en connaissons une forme accessible que pour certains cas bien particuliers. Nous allons par conséquent nous intéresser au produit de variables aléatoires indépendantes de distributions Gamma généralisées.

La distribution Gamma généralisée correspond, comme son nom l'indique, à une généralisation de la distribution Gamma. Le produit de telles variables aléatoires a été étudié dès les années 1960. Nous reviendrons sur la distribution Gamma généralisée dans la section 4.2. Cette distribution a été utilisée pour la modélisation de nombreux problèmes statistiques, dans des domaines très divers, grâce à sa flexibilité. Ainsi, [Zaninetti, 2008] a utilisé une distribution Gamma

généralisée de paramètre d'intensité égal à 2 dans le cadre de problèmes liés aux fonctions de luminosité pour les galaxies. [Aalo et al., 2005] l'a utilisée pour caractériser les processus de zones d'ombres dans les systèmes de communication sans fil. Dans [Ali et al., 2008], les auteurs considèrent une application à des données sur la sécheresse au Nebraska. Ils ont utilisé le produit de deux variables aléatoires de lois Gamma généralisées pour caractériser la distribution de la magnitude de la sécheresse. On trouve d'autres applications de la distribution Gamma généralisée en physique. Dans [Karagiannidis et al., 2006], les auteurs considèrent notamment un produit de variables aléatoires de distributions de Rayleigh, qui correspondent à un cas particulier de distributions Gamma généralisées. Par ailleurs, [Coelho and Arnold, 2014] ont présenté d'autres cas particuliers du produit de distributions Gamma généralisées indépendantes : la distribution exacte de discriminants ou de déterminants de Vandermonde, ainsi que la distribution du produit d'une puissance de la valeur absolue de variables aléatoires gaussiennes indépendantes. Dans [Marques et al., 2014], plusieurs applications intéressantes sont illustrées pour une combinaison linéaire de variables aléatoires de Gumbel.

Bien que la distribution Gamma généralisée soit utilisée dans de nombreux domaines, la distribution exacte du produit de telles variables aléatoires n'a pas une expression simple, ce qui rend difficile son utilisation en pratique. La majorité des approches existantes est basée sur des fonctions Meijer-G ou H, qui, de nos jours, demeurent difficiles à utiliser, comme l'ont souligné [Burda et al., 2012], [Coelho and Arnold, 2014], et [Marques et al., 2014].
La distribution exacte du produit de variables indépendantes de lois Gamma généralisées a d'abord été étudiée par [Mathai, 1972], en utilisant la transformation inverse de Mellin et les fonctions H. Cependant, ces fonctions ont des expressions très complexes et ne sont pas implémentées dans la majorité des logiciels. Par ailleurs, une représentation de la distribution du produit de variables aléatoires indépendantes de lois Gamma généralisées en terme de fonction Meijer-G a été développée dans [Podolski, 1972], pour tous les coefficients $\tilde{\beta}_j$ dans l'expression (betaj) égaux. Dans [Ali et al., 2008] et [Malik, 1968], les expressions exactes de la fonction de densité et de la fonction de répartition sont obtenues, pour le cas du produit de deux variables aléatoires Gamma généralisées, en utilisant la fonction de Bessel modifiée. Plus récemment, de nouveaux résultats ont été obtenus dans [Karagiannidis et al., 2006] pour le produit de variables aléatoires Gamma généralisées et dans [Salo et al., 2004], [Salo and Vainikainen, 2006] pour le produit de distributions indépendantes de Rayleigh, qui correspondent à des cas particuliers de la distribution Gamma généralisée. Cependant ces résultats sont également basés sur la fonction Meijer-G, et sont par conséquent difficiles à utiliser.
[Coelho and Arnold, 2014] ont développé des distributions presque exactes pour le produit de variables aléatoires indépendantes de lois Gamma généralisées en utilisant une approche différente de celle que nous utilisons ici, basée sur des troncatures d'une représentation infinie de la fonction Gamma. Cependant, [Marques, 2012] a montré qu'il est possible d'améliorer les résultats obtenus par [Coelho and Arnold, 2014] avec une nouvelle approche générale que nous utilisons dans ce chapitre. Cette approche permet d'aborder des problèmes complexes avec une grande précision.

Dans ce chapitre, nous proposons d'utiliser une approximation développée par [Marques, 2012] afin de caractériser l'intensité du processus de Poisson du modèle développé au chapitre 3. Ce travail a fait l'objet d'un article coécrit avec Filipe Marques et soumis au journal *Computational statistics and data analysis* [Marques and Loingeville, 2015].

Nous rappelons dans un premier temps les définitions des distributions utilisées dans l'approximation (Section 4.2). Nous présentons ensuite l'approximation du produit de variables aléatoires indépendantes de lois Gamma généralisées développée par [Marques, 2012] (Section 4.3). Nous utilisons dans un troisième temps cette approximation pour caractériser l'intensité du processus de Poisson introduit au chapitre 3 (Section 4.4).

4.2 Définitions

Dans ce chapitre, nous proposons de caractériser la loi de l'intensité du processus de Poisson introduit au chapitre 3 à l'aide d'une approximation du produit de variables aléatoires indépendantes de lois Gamma généralisées. Dans cette section, nous revenons sur les définitions des lois utilisées dans le cadre de cette approximation.

Distribution Gamma généralisée (GG). La distribution Gamma généralisée a été introduite par [Stacy, 1962]. Soit X une variable aléatoire de distribution Gamma de paramètre d'intensité $\tilde{\lambda} > 0$ et de forme $r > 0$, c'est à dire $X \sim \Gamma(r, \tilde{\lambda})$. On dit que la variable aléatoire $Y = X^{1/\tilde{\beta}}$, pour $\tilde{\beta} \neq 0$, suit une distribution Gamma généralisée. On note cela de la façon suivante :

$$Y \sim G\Gamma(r, \tilde{\lambda}, \tilde{\beta})$$

La fonction densité de probabilité de cette variable Y s'exprime alors de la façon suivante :

$$f_Y(y) = |\tilde{\beta}| \frac{\tilde{\lambda}^r}{\Gamma(r)} y^{\tilde{\beta} r - 1} e^{-\tilde{\lambda} y^{\tilde{\beta}}}, \quad y > 0$$

et ses moments non centraux de la façon suivante :

$$E\left[Y^h\right] = \int_{-\infty}^{\infty} y^h f_Y(y) \mathrm{d}y = \frac{\Gamma\left(r + h/\tilde{\beta}\right)}{\Gamma(r)} \tilde{\lambda}^{-h/\tilde{\beta}}. \tag{4.1}$$

Remarque 4.2.1. *Si $\tilde{\beta} = 1$, la variable aléatoire X suit alors une loi Gamma.*

Distribution Gamma décalée. Un paramètre "de décalage" peut être ajouté à la distribution Gamma généralisée, de sorte à ce que le domaine de définition de x commence à une valeur autre que zéro. On notera cette distribution SGG pour Shifted Generalized Gamma.

Distribution Gamma généralisée entière (GIG). La distribution Gamma généralisée entière (notée GIG pour Generalized Integer Gamma) correspond à la distribution de la somme de variables aléatoires indépendantes de lois Gamma qui ont toutes des paramètres de forme entiers, et des paramètres d'intensité différents.

Soit $X_j \sim \Gamma(r_j, \tilde{\lambda}_j)$, où $j = \{1, \ldots, p\}$, p variables aléatoires indépendantes, où tous les paramètres r_j sont des entiers positifs et tous les paramètres $\tilde{\lambda}_j$ sont différents. Alors la variable aléatoire Y définie par $Y = \sum_{j=1}^{p} X_j$ suit une distribution Gamma généralisée entière (GIG) de paramètre de profondeur p, de paramètre de forme r_j, et de paramètres d'intensité $\tilde{\lambda}_j$, où $j = \{1, \ldots, p\}$. Nous notons cela de la façon suivante : $Y \sim GIG(r_j, \tilde{\lambda}_j; p)$.

Distribution Gamma généralisée presque entière (GNIG). La distribution Gamma généralisée presque entière (notée GNIG pour Generalized Near Integer Gamma) correspond à la distribution de la somme de variables aléatoires indépendantes de lois Gamma généralisées qui ont toutes des paramètres de forme entiers, sauf une, dont le paramètre de forme est un réel, et des paramètres d'intensité différents.

Soit $Z = Y_1 + Y_2$, où $Y_1 \sim GIG(r_j, \tilde{\lambda}_j; p)$ et $Y_2 \sim \Gamma(r, \tilde{\lambda})$ sont deux variables aléatoires indépendantes, où r est un réel positif non entier et $\tilde{\lambda} \neq \tilde{\lambda}_j$, pour $j = \{1, \ldots, p\}$. La variable aléatoire Z suit une distribution Gamma généralisée presque entière (GNIG) de paramètre de profondeur $p + 1$, de paramètres de forme (r_j, r), et de paramètres d'intensité $(\tilde{\lambda}_j, \tilde{\lambda})$, où $j = \{1, \ldots, p\}$. Nous notons cela de la façon suivante : $Y \sim GNIG(r_j, r, \tilde{\lambda}_j, \tilde{\lambda}; p + 1)$.

4.3 Produit de variables aléatoires de lois Gamma généralisées

Dans cette section, nous présentons une approximation du produit de variables aléatoires de lois Gamma généralisées, proposée par [Marques, 2012].

Le principe de cette méthode consiste à factoriser l'expression de la fonction caractéristique du logarithme du produit de variables aléatoires indépendantes de lois Gamma généralisées. Nous pouvons ensuite approcher l'un des facteurs par une distribution connue. En se basant sur ce résultat, il est possible de développer des distributions presque exactes très précises, basées sur la distribution Gamma généralisée presque entière décalée (SGNIG), ou sur des mélanges de ces distributions. Un paramètre de précision, γ, peut être introduit dans ces représentations. En ajustant sa valeur, il est possible d'améliorer la qualité des distributions presque exactes obtenues. D'autre part, les distributions presque exactes développées auront leurs m^* premiers moments identiques à ceux de la distribution exacte. Par conséquent, augmenter la valeur de m^* permet également d'améliorer la précision de ces distributions.

4.3.1 Distribution exacte

Nous nous intéressons dans un premier temps à la distribution exacte du produit de variables aléatoires indépendantes de lois Gamma généralisées.

Soit X_j une variable aléatoire de distribution Gamma d'intensité $\tilde{\lambda}_j > 0$ et de paramètre de forme $r_j > 0$, c'est à dire $X_j \sim \Gamma(r_j, \tilde{\lambda}_j)$ avec $j = 1, \ldots, p$. On dit que la variable aléatoire $Y_j = X_j^{1/\tilde{\beta}_j}$, pour $\tilde{\beta}_j \neq 0$, suit une distribution Gamma généralisée $Y_j \sim G\Gamma(r_j, \tilde{\lambda}_j, \tilde{\beta}_j)$.
La fonction densité de probabilité de X_j est donnée par :

$$f_{Y_j}(y) = |\tilde{\beta}_j| \frac{\tilde{\lambda}_j^{r_j}}{\Gamma(r_j)} y^{\tilde{\beta}_j r_j - 1} e^{-\tilde{\lambda}_j y^{\tilde{\beta}_j}} \quad y > 0 \tag{4.2}$$

et les moments non centraux sont :

$$E\left[Y_j^h\right] = \int_{-\infty}^{\infty} y^h f_{Y_j}(y) \mathrm{d}y = \frac{\Gamma\left(r_j + h/\tilde{\beta}_j\right)}{\Gamma(r_j)} \tilde{\lambda}_j^{-h/\tilde{\beta}_j} \tag{4.3}$$

Nous souhaitons étudier la distribution du produit de variables aléatoires indépendantes de lois Gamma généralisées :

$$Z = \prod_{j=1}^{p} Y_j \tag{4.4}$$

avec $Y_j \overset{\text{ind.}}{\sim} G\Gamma(r_j, \tilde{\lambda}_j, \tilde{\beta}_j)$, $j = 1, \ldots, p$. la fonction caractéristique de la variable aléatoire $W = -\log Z$ est défini comme suit :

$$\Phi_W(t) = \int_{-\infty}^{\infty} e^{itw} f_W(w) \mathrm{d}w = E[e^{itW}]$$

En utilisant l'expression des moments non centraux dans (4.3), cette expression peut s'écrire de la façon suivante :

$$\Phi_W(t) = E[Z^{-it}] = \prod_{j=1}^{p} E[Y_j^{-it}] = \prod_{j=1}^{p} \frac{\Gamma\left(r_j - \frac{it}{\tilde{\beta}_j}\right)}{\Gamma(r_j)} \tilde{\lambda}_j^{\frac{it}{\tilde{\beta}_j}} \quad t \in \mathbb{R}. \tag{4.5}$$

Théorème 4.3.1 ([Marques, 2012]). *La fonction caractéristique de* $W = -\log \prod_{i=j}^{p} Y_j$ *avec* $Y_j \overset{ind.}{\sim} G\Gamma(r_j, \tilde{\lambda}_j, \tilde{\beta}_j)$ *pour* $r_j > 0$, $\tilde{\lambda}_j > 0$, $\tilde{\beta}_j \neq 0$, *et* $\gamma \in \mathbb{N}$ *peut s'écrire :*

$$\Phi_W(t) = \underbrace{\left\{ \prod_{j=1}^{p} \frac{\Gamma\left(r_j + \gamma - \frac{it}{\tilde{\beta}_j}\right)}{\Gamma(r_j + \gamma)} \right\}}_{\Phi_{W_1}(t)} \underbrace{\left\{ \prod_{j=1}^{p} \prod_{k=0}^{\gamma-1} ((r_j + k)\tilde{\beta}_j)((r_j + k)\tilde{\beta}_j - it)^{-1} \right\} e^{it \sum_{j=1}^{p} \log \tilde{\lambda}_j^{\frac{1}{\tilde{\beta}_j}}}}_{\Phi_{W_2}(t)} \tag{4.6}$$

A partir de l'expression (4.6) et des propriétés de la fonction caractéristique, [Marques, 2012] explique que, lorsque $\tilde{\beta}_j > 0$, la distribution exacte de $W = -\log Z$ correspond à la distribution de la somme de deux variables aléatoires indépendantes, W_1, dont la distribution correspond à celle de la somme de variables aléatoires indépendantes de distributions log-Gamma de paramètres $r_j + \gamma$ et 1, multiplié par le paramètre $\frac{1}{\tilde{\beta}_j}$ et W_2, dont la distribution correspond à celle d'une somme décalée de $p \times \gamma$ distributions exponentielles indépendantes de paramètres $(r_j + k)\tilde{\beta}_j$ pour $j = 1, \ldots, p$ et $k = 0, \ldots, \gamma - 1$, et de paramètre de décalage

$$\tilde{\theta} = \sum_{j=1}^{p} \log \tilde{\lambda}_j^{1/\tilde{\beta}_j}. \tag{4.7}$$

Si l'on somme les distributions exponentielles de même paramètre, on peut écrire $\Phi_{W_2}(t)$ dans l'expression (4.6) de la façon suivante :

$$\Phi_{W_2}(t) = \left\{ \prod_{j=1}^{\ell} \tilde{\alpha}_j^{\delta_j} (\tilde{\alpha}_j - it)^{-\delta_j} \right\} e^{it\tilde{\theta}}, \tag{4.8}$$

où ℓ est le nombre de distributions exponentielles de paramètres différents, $\tilde{\alpha}_j$ sont les paramètres de telles distributions exponentielles, et δ_j est le nombre de distributions exponentielles de même paramètre $\tilde{\alpha}_j$, for $j = 1, \ldots, \ell$. Ainsi, on peut dire que la distribution de W_2 correspond à une distribution Gamma généralisée entière décalée (SGIG), de paramètres de forme δ_j, d'intensité $\tilde{\alpha}_j$, pour $j = 1, \ldots, \ell$, de profondeur ℓ et de paramètre de décalage $\tilde{\theta}$ correspondant à (4.7). Nous notons cela de la façon suivante : $W_2 \sim \text{SGIG}(\delta_1, \ldots, \delta_\ell\,;\, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell\,;\, \ell\,;\, \tilde{\theta})$.

En utilisant la représentation proposée dans les expressions (4.6) and (4.8), [Marques, 2012] a formulé des approximations presque exactes très précises pour le produit de variables aléatoires indépendantes de lois Gamma généralisées. Nous présentons ces approximations ci-dessous.

4.3.2 Distribution presque exacte pour $W = -\log Z$ et pour Z

L'une des méthodes utilisées pour développer des distributions presque exactes consiste à utiliser une factorisation de la fonction caractéristique d'une variable aléatoire ou de son logarithme, puis d'approcher l'un des facteurs sans changer les autres, de sorte à ce que la fonction caractéristique obtenue corresponde à une distribution connue et facile à utiliser en pratique. En commençant par la factorisation de $\Phi_W(t)$ dans (4.6) avec $\Phi_{W_2}(t)$ dans (4.8), une première distribution presque exacte simple peut être obtenue en approchant la fonction caractéristique dans 4.6 par la fonction caractéristique d'une variable aléatoire de distribution Gamma décalée, sans changer $\Phi_{W_2}(t)$. [Marques, 2012] justifie cette stratégie par le fait qu'il est possible de prouver qu'une variable aléatoire de distribution log-Gamma peut être représentée comme une

somme infinie de distributions exponentielles décalées, par conséquent, on peut en déduire que la somme de variables aléatoires indépendantes de distributions log-Gamma peut être correctement approchée par une seule distribution Gamma décalée [Marques, 2012]. Ainsi, la fonction caractéristique presque exacte a la structure suivante :

$$\Phi_{W_1^*}(t)\Phi_{W_2}(t) \tag{4.9}$$

où

$$\Phi_{W_1^*}(t) = \omega^{\rho}(\omega - it)^{-\rho}e^{it\upsilon}, \quad t \in \mathbb{R}, \tag{4.10}$$

est la fonction caractéristique de W_1^*, de distribution Gamma décalée, de paramètre de forme ρ, d'intensité ω et de décalage υ. Nous notons cela de la façon suivante $W_1^* \sim \text{SGamma}(\rho, \omega, \upsilon)$. Les paramètres ρ, ω, et υ correspondent aux solutions numériques du système d'équations suivant :

$$\left.\frac{\partial^j \Phi_{W_1^*}(t)}{\partial t^j}\right|_{t=0} = \left.\frac{\partial^j \Phi_{W_1}(t)}{\partial t^j}\right|_{t=0}, \quad j = 1, 2, 3\,. \tag{4.11}$$

La distribution presque exacte ainsi obtenue aura alors ses trois premiers moments égaux à ceux de la distribution exacte de W.

Théorème 4.3.2 ([Marques, 2012]). *En utilisant comme approximation asymptotique pour $\Phi_{W_1}(t)$ dans (4.6) la fonction caractéristique $\Phi_{W_1^*}(t)$ dans (4.10), nous obtenons comme distribution presque exacte pour $W = -\log\prod_{j=1}^{p} Y_j$ avec $Y_j \overset{ind.}{\sim} G\Gamma(r_j, \tilde{\lambda}_j, \tilde{\beta}_j)$ pour $r_j > 0$, $\tilde{\lambda}_j > 0$ et $\tilde{\beta}_j > 0$ $(j = 1, \ldots, p)$, une distribution Gamma généralisée presque entière décalée (SGNIG) de paramètres de forme entiers $\delta_1, \ldots, \delta_\ell$ exprimés dans (4.8) et non entier ρ, de paramètres d'intensité $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell$ exprimés dans (4.8) et ω, de profondeur $\ell + 1$ et de paramètre de décalage $\tilde{\theta} + \upsilon$. Nous notons cela de la façon suivante : $SGNIG(\delta_1, \ldots, \delta_\ell\,, \rho\,;\, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell, \omega\,;\, \ell + 1\,;\, \tilde{\theta} + \upsilon)$. Les paramètres ρ, ω and υ correspondent aux solutions numériques du système (4.11) et $\tilde{\theta}$ est exprimé par (4.7). La fonction densité de probabilité et la fonction de répartition de W sont obtenues à partir des fonctions densité de probabilité et de répartition de la distribution GNIG correspondante, et exprimées respectivement par :*

$$f_W(w) \approx f^{\text{GNIG}}\Big(w - (\tilde{\theta} + \upsilon)|\ \delta_1, \ldots, \delta_\ell, \rho\,;\, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell, \omega\,;\, \ell + 1\Big) \tag{4.12}$$

et

$$F_W(w) \approx F^{\text{GNIG}}\Big(w - (\tilde{\theta} + \upsilon)\,|\,\delta_1, \ldots, \delta_\ell, \rho\,;\, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell, \omega\,;\, \ell + 1\Big)\,. \tag{4.13}$$

Des études numériques ainsi que des graphiques peuvent être trouvés pour cette première approximation dans [Marques, 2012].

En se basant sur cette première approche, [Marques and Loingeville, 2015] proposent de développer des distributions presque exactes encore plus précises. Pour cela, l'idée principale est d'approcher la fonction caractéristique $\Phi_{W_1}(t)$ dans 4.6 par la fonction caractéristique d'un mélange de distributions Gamma décalées, plutôt que par la fonction caractéristique d'une seule variable aléatoire Gamma décalée. Pour justifier cette nouvelle approximation, nous pouvons avancer les arguments suivants :

- La somme de variables aléatoires indépendantes de distributions log-Gamma peut s'exprimer comme une somme infinie de distributions exponentielles décalées indépendantes.
- La somme de variables aléatoires indépendantes de distributions exponentielles peut être représentée comme un mélange de variables aléatoires de distributions Gamma.

Par conséquent, un mélange de variables aléatoires de distributions Gamma décalées peut être considéré comme une approximation de la somme de variables aléatoires indépendantes de distributions log-Gamma. Comme nous l'avons vu ci-dessus, la partie cruciale correspond à l'estimation des paramètres.

La fonction caractéristique de W aura la structure suivante :

$$\Phi_{W_1^{**}}(t)\Phi_{W_2}(t) \tag{4.14}$$

où

$$\Phi_{W_1^{**}}(t) = \sum_{k=0}^{m^*} p_k \omega^{\rho+k}(\omega - it)^{-\rho-k} e^{it\upsilon}, \quad t \in \mathbb{R}, \tag{4.15}$$

où ρ, ω et υ sont les mêmes que dans (4.10), obtenus comme étant les solutions du système d'équations (4.11) et sont fixés et déterminés dans une étape initiale. Les poids p_k $(k = 0, \ldots, m^* - 1)$ sont ensuite déterminés de sorte que

$$\frac{\partial^j}{\partial t^j}\,\Phi_{W^{**}}(t)\bigg|_{t=0} = \frac{\partial^j}{\partial t^j}\,\Phi_{W_1}(t)\bigg|_{t=0}, \quad j = 1, \ldots, m^* \quad \text{avec} \quad p_{m^*} = 1 - \sum_{k=0}^{m^*-1} p_k \tag{4.16}$$

La particularité de cette approche est que les paramètres ρ, ω et υ sont fixés dès le départ tandis que les poids sont déterminés en résolvant le système d'équations (4.16). Ce système possède une unique solution et se résout facilement. Cette construction nous assure que la distribution presque exacte développée aura ses m^* premiers moments égaux à ceux de W, et aussi que nous aurons comme distribution presque exacte de W des mélanges de $m^* + 1$ distributions Gamma généralisées presque entières décalées (SGNIG).

Théorème 4.3.3 ([Marques and Loingeville, 2015]). *En utilisant comme approximation asymptotique de $\Phi_{W_1}(t)$ dans (4.6) la fonction caractéristique $\Phi_{W_1^{**}}(t)$ dans (4.15), nous obtenons une distribution presque exacte pour $W = -\log\prod_{j=1}^{p} Y_j$ avec $Y_j \overset{ind.}{\sim} G\Gamma(r_j, \tilde{\lambda}_j, \tilde{\beta}_j)$ pour $r_j > 0$, $\tilde{\lambda}_j > 0$ et $\tilde{\beta}_j > 0$ $(j = 1, \ldots, p)$ un mélange de $m^* + 1$ distributions Gamma généralisées presque entière décalées (SGNIG) de paramètres de forme entiers $\delta_1, \ldots, \delta_\ell$ exprimés dans (4.8) et non entiers $\rho + k$, $k = 0, \ldots, m^*$, de paramètres d'intensité $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell$ exprimés dans (4.8) et ω, de profondeur $\ell + 1$ et de paramètre de décalage $\tilde{\theta} + \upsilon$. les paramètres ρ, ω et υ sont obtenus en résolvant le système (4.16) et $\tilde{\theta}$ est fourni par (4.7). La fonction densité de probabilité, ainsi que la fonction de répartition correspondent à celles d'une distribution Gamma généralisée presque entière de tels paramètres, et s'expriment comme suit :*

$$f_W(w) \approx \sum_{k=0}^{m^*} p_k f^{\text{GNIG}}\Big(w - (\tilde{\theta} + \upsilon)|\ \delta_1, \ldots, \delta_\ell, \rho + k\,;\, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell, \omega\,;\, \ell + 1\Big) \tag{4.17}$$

et

$$F_W(w) = \sum_{k=0}^{m^*} p_k F^{\text{GNIG}}\Big(w - (\tilde{\theta} + \upsilon)\,|\,\delta_1, \ldots, \delta_\ell, \rho + k\,;\, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell, \omega\,;\, \ell + 1\Big) \tag{4.18}$$

où les poids p_k sont déterminés de la même façon que les solutions du système 4.16.

Preuve : Similaire à la preuve du Théorème 4.3.2.

Les expressions des fonctions densité de probabilité et de répartition de la distribution presque exacte de Z dans (4.4) peuvent être, cette fois encore, aisément obtenues par une transformation simple des expressions (4.17) et (4.18) et s'expriment comme suit :

$$\sum_{k=0}^{m^*} p_k f^{\mathrm{GNIG}}\Big(-\log(w) - (\tilde{\theta}+\upsilon)|\ \delta_1,\ldots,\delta_\ell,\rho+k\,;\,\tilde{\alpha}_1,\ldots,\tilde{\alpha}_\ell,\omega\,;\,\ell+1\Big)1/w \tag{4.19}$$

et

$$1-\sum_{k=0}^{m^*} p_k F^{\mathrm{GNIG}}\Big(-\log(w) - (\tilde{\theta}+\upsilon)\,|\,\delta_1,\ldots,\delta_\ell,\rho+k\,;\,\tilde{\alpha}_1,\ldots,\tilde{\alpha}_\ell,\omega\,;\,\ell+1\Big). \tag{4.20}$$

4.4 Caractérisation de l'intensité du processus de Poisson utilisé

Au chapitre 3, nous proposons un modèle permettant d'expliquer la variabilité des résultats de dénombrement en microbiologie par un HGLM Gamma-Poisson à trois facteurs aléatoires (3.9). Pour rappel, nous supposons que la distribution de y_{ijk} sachant les effets aléatoires e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$, correspond à une loi de Poisson de paramètre λ_{ijk} tel que :

$$E(y_{ijk}|e^{\alpha_i},e^{\beta_{ij}},e^{\gamma_{ijk}}) = \lambda_{ijk} = e^{\mu_{ijk}}.e^{\alpha_i}.e^{\beta_{ij}}.e^{\gamma_{ijk}}, \tag{4.21}$$

où $e^{\mu_{ijk}} = \beta_0$ est un paramètre fixe tandis que e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$ sont trois variables aléatoires indépendantes telles que :

- $e^{\alpha_i} \sim \Gamma(\frac{1}{u_1^2},\frac{1}{u_1^2})$
- $e^{\beta_{ij}} \sim \Gamma(\frac{1}{u_2^2},\frac{1}{u_2^2})$
- $e^{\gamma_{ijk}} \sim \Gamma(\frac{1}{u_3^2},\frac{1}{u_3^2})$

Autrement dit, la distribution des résultats de dénombrement y_{ijk} suit une loi de Poisson dont l'intensité correspond à un produit de trois variables aléatoires indépendantes de lois Gamma, multiplié par une constante :

$$y_{ijk} \sim P(e^{\mu_{ijk}}.e^{\alpha_i}.e^{\beta_{ij}}.e^{\gamma_{ijk}}). \tag{4.22}$$

L'intensité du processus de Poisson (4.22) correspond donc à un cas particulier du produit de variables aléatoires indépendantes de lois Gamma généralisées, que nous avons étudié ci-dessus. En effet, elle correspond au produit de trois variables aléatoires indépendantes de lois Gamma généralisées dont les paramètres β_j définis dans la section 4.2 sont tels que $\beta_j = 1 \quad \forall j = \{1,2,3\}$. Nous ne pouvons pas obtenir des expressions exactes simples des fonctions de densité et de répartition de λ_{ijk} dans (4.22). La distribution de λ_{ijk} peut en revanche être exprimée comme un mélange de distributions SGNIG, comme dans le théorème 4.3.3.

Intérêt de la caractérisation de l'intensité du processus de Poisson. La distribution de l'intensité du processus de Poisson peut être utilisée en pratique pour détecter des résultats de mesure aberrants. Une observation y_{ijk} est considérée aberrante si la valeur estimée du paramètre λ_{ijk} qui lui est associée (calculée à partir des valeurs estimées $\hat{\mu}$, $\hat{\alpha}_i, \hat{\beta}_{ij}$, et $\hat{\gamma}_{ijk}$), est située en dehors d'un intervalle de tolérance statistique sur l'intensité de la distribution de Poisson correspondante.

Applications à 2 jeux de données issus de la microbiologie. Nous présentons ci-dessous la caractérisation de l'intensité du processus de Poisson (4.22) pour les deux exemples étudiés en section 3.5.2.
A partir des valeurs estimées des paramètres, fournies en Tables 3.5 et 3.6, nous pouvons utiliser les distributions SGNIG pour étudier la distribution de l'intensité du processus de Poisson de chacun des essais considérés.
Nous pouvons obtenir les expressions des distributions presque exactes des fonctions de densité et des fonctions de répartition du paramètres λ_{ijk} pour les jeux de données à partir des expressions (4.19) et (4.20). La Figure 4.1 présente les graphiques des fonctions de densité et de répartition correspondantes.

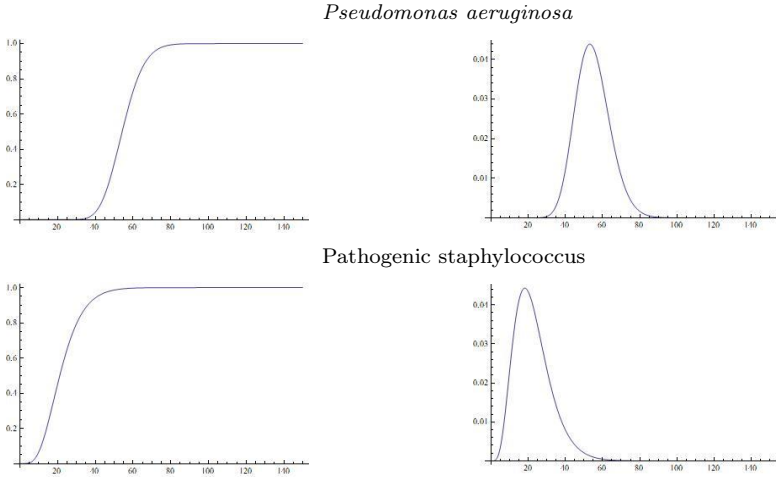

FIGURE 4.1 – Fonctions densité de probabilité et de répartition presque exactes de λ_{ijk} pour les 2 exemples pour $\gamma = 5$ et $m^* = 5$

Nous pouvons aussi calculer les quantiles presque exactes des distributions de l'intensité λ_{ijk} du processus de Poisson pour les deux exemples étudiés (Table 4.1).

	0.90	0.95	0.99
Pseudomonas aeruginosa			
Quantile presque exacte ($\gamma = 20$, $m^* = 1$)	66.57358	70.36831	77.85539
Staphylocoques pathogènes			
Quantile presque exacte ($\gamma = 20$, $m^* = 1$)	35.87963	41.09503	52.23848

TABLE 4.1 – Quantiles presque exactes pour λ_{ijk} dans les 2 exemples

L'adéquation des deux jeux de données étudiés au modèle (4.21) a été étudiée en section 3.5.2. En particulier, la Figure 4.2 montre que les valeurs $\hat{\lambda}_{ijk}$ associées à chaque résultat de mesure et estimées à partir de (4.21), s'ajustent bien à la distribution SGNIG du produit des trois variables aléatoires correspondantes.

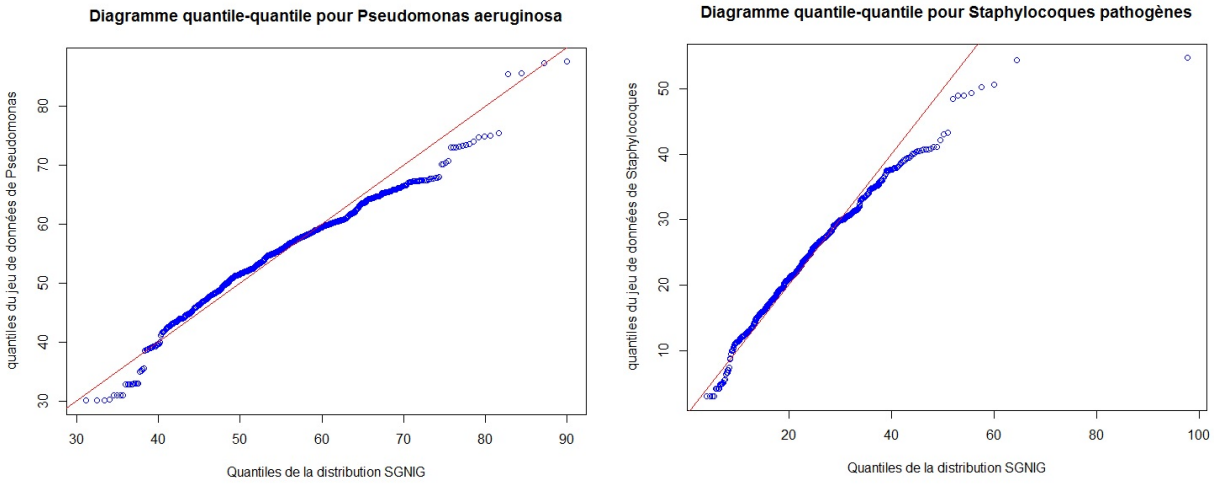

FIGURE 4.2 – Adéquation de la distribution des valeurs de λ_{ijk} ajustées par le modèle à la distribution SGNIG correspondante pour $\gamma = 5$ et $m^* = 5$

Comme $\mathbb{E}(e^{\alpha_i}) = \mathbb{E}(e^{\beta_{ij}}) = \mathbb{E}(e^{\gamma_{ijk}}) = 1$, nous avons $\mathbb{E}(\lambda_{ijk}) = \exp(\mu)$ pour les deux exemples étudiés. Nous vérifions dans la Table 4.1 que les médianes presque exactes sont effectivement proches des espérances associées ($E[\hat{\mu}] = 54.76$ pour l'essai sur *Pseudomonas aeruginosa* et $E[\hat{\mu}] = 22.61$ pour l'essai sur *Staphylocoques* pathogènes).

Comme mentionné ci-dessus, nous pouvons détecter des résultats de mesures aberrants dans un jeu de données d'essai interlaboratoires en utilisant la distribution de l'intensité du processus de Poisson. Pour cela, nous calculons dans un premier temps un intervalle de tolérance statistique sur λ_{ijk}, à partir de la distribution SGNIG correspondante. Pour le jeu de données de dénombrement de *Staphylocoques* pathogènes, l'intervalle de tolérance statistique sur λ_{ijk} au risque de 5% est $[7.665142; 46.02191]$, comme écrit dans la Table 4.1. Nous pouvons ensuite calculer les paramètres λ_{ijk} associés à chaque résultat de mesure du jeu de données, en remplaçant les effets e^{α_i}, $e^{\beta_{ij}}$, et $e^{\gamma_{ijk}}$ dans 3.9 par leurs estimateurs $\hat{e}^{\alpha_i}$, $\hat{e}^{\beta_{ij}}$ et $\hat{e}^{\gamma_{ijk}}$. Nous considérerons alors les dénombrements dont les valeurs de λ_{ijk} sont situées hors de l'intervalle de tolérance statistique comme potentiellement suspects. Par exemple, pour le jeu de données de *Staphylocoques* pathogènes, le laboratoire 5 a mesuré $y_{512} = 1$ sur la réplication 1 du flacon 2. Cette valeur est située hors de l'intervalle de tolérance statistique de λ_{ijk} au risque de 5%, mentionné ci-dessus. Nous pouvons alors considérer ce résultat de dénombrement suspect. La distribution presque exacte peut également permettre de détecter un problème d'hétérogénéité de lot. Pour cela, nous caractérisons alors la distribution de λ_{ij} par un produit de deux variables aléatoires de lois Gamma, et nous effectuons les tests de la même façon que présenté ici.

4.5 Conclusion

Nous avons proposé au chapitre 3 un modèle linéaire généralisé hiérarchique Gamma-Poisson, correspondant à un modèle de Poisson dont l'intensité est distribuée suivant un produit de 3 variables aléatoires indépendantes de lois Gamma. La distribution de cette intensité est donc complexe et difficilement caractérisable.
Dans ce chapitre, nous proposons d'utiliser une approximation du produit de variables aléatoires indépendantes de lois Gamma généralisées développée par [Marques, 2012] pour caractériser la distribution de λ_{ijk}. Cette approximation découle d'une factorisation de la fonction caracté-

ristique du produit considéré, et de l'approximation de l'un des facteurs par une distribution connue. Un paramètre de précision γ et un paramètre m^* correspondant au nombre des premiers moments égaux pour les distributions exactes et presque exactes, permettent d'obtenir des distributions presque exactes très précises.
L'utilisation de cette distribution presque exacte pour caractériser l'intensité du processus de Poisson du modèle (3.9) permet de détecter des résultats de mesure aberrants parmi les résultats d'un essai interlaboratoires. La méthode proposée peut également être appliquée pour détecter des problèmes d'hétérogénéité de lot.

Conclusion générale et perspectives

Conclusion générale

Dans le cadre de cette thèse, nous nous sommes attachés à développer une méthode statistique spécifique au contrôle de qualité dans le domaine de la microbiologie. Pour bien cerner les enjeux de ce travail, il a tout d'abord été nécessaire d'étudier la méthode d'analyse de variance actuellement utilisée, ses tenants et aboutissements, et d'en cerner les avantages, les inconvénients et les limites (chapitre 1). La méthode utilisée jusqu'alors en microbiologie est basée sur un modèle d'analyse de variance à deux facteurs aléatoires imbriqués. Pour que les données issues de la microbiologie puissent être exploitées à l'aide d'un tel modèle, il est alors nécessaire de les normaliser. Malgré cette transformation des données, il demeure tout de même des ajustements insatisfaisants pour certains essais interlaboratoires. Nous proposons dans les chapitres 2, 3 et 4 une nouvelle méthode d'exploitation des données des essais interlaboratoires en microbiologie.

Nous proposons au chapitre 2 une première modélisation du problème par un modèle linéaire à deux facteurs fixes imbriqués, sur données non-transformées. Nous avons alors utilisé la méthode d'analyse de déviance pour développer des tests de significativité des facteurs, qui s'avèrent efficaces sur des jeux de données d'essais interlaboratoires en microbiologie. Un modèle à facteurs fixes ne permet toutefois ni d'évaluer précisément les parts de dispersion induites par chacun des facteurs, ni d'évaluer la performances des laboratoires participant à un essai.

Nous avons alors utilisé des facteurs aléatoires dans une nouvelle modélisation, présentée au chapitre 3. Le caractère aléatoire des facteurs nous a permis d'évaluer et de caractériser la surdispersion des résultats de dénombrement en microbiologie, ce qui était l'un des objectifs principaux de la thèse. Le modèle développé correspond à un modèle linéaire généralisé hiérarchique Gamma-poisson à trois facteurs aléatoires. Des applications pratiques de cette méthode à des données d'essais interlaboratoires sur *Pseudomonas aeruginosa* et *Staphylocoques* pathogènes montrent que le modèle proposé s'ajuste bien aux données réelles. Ce modèle permet d'estimer les effets fixes, aléatoires, et les paramètres de dispersion associés aux différents facteurs. Nous utilisons pour cela une vraisemblance spécifique aux modèles hiérarchiques, la h-vraisemblance, et un algorithme backfitting. Nous pouvons également tester la significativité des facteurs à l'aide d'une méthode basée sur le critère AIC conditionnel, et évaluer la performance analytique des laboratoires participants à un essai.

Le modèle proposé suppose que les données sont distribuées suivant une loi de Poisson dont l'intensité est elle-même distribuée suivant un produit de trois variables aléatoires indépendantes de lois Gamma. La méthode d'estimation proposée au chapitre 3 n'est pas basée sur la distribution exacte de cette intensité, car celle-ci est difficilement caractérisable en pratique. Au chapitre 4, nous proposons de la caractériser à l'aide d'une distribution presque exacte développée par

[Marques, 2012]. Cette caractérisation permet notamment de détecter des éventuels résultats de mesure aberrants lors d'un essai interlaboratoires en microbiologie.

Perspectives

Le travail présenté dans ce manuscrit offre certaines perspectives, autant sur le plan théorique qu'appliqué. Cette section les présente de manière succincte.

Étude de la distribution de la déviance normalisée. Nous avons présenté en Section 3.5.1 un test d'ajustement des données au modèle proposé, basé sur la déviance normalisée. Ce test de déviance normalisée est communément utilisé pour l'étude de l'adéquation d'un HGLM à un jeu de données. Pourtant, aucun critère strict pour ce test n'est présenté dans la littérature. Les travaux effectués proposent une simple comparaison visuelle de la déviance normalisée avec le degré de liberté estimé pour conclure à l'adéquation ou à la non-adéquation du modèle aux données. C'est notamment le cas de [Lee and Nelder, 2002], qui ont travaillé sur un HGLM dont les effets aléatoires permettent de modéliser l'influence de différents hôpitaux sur un traitement administré en recherche clinique. Il semblerait alors intéressant de caractériser la distribution exacte de la déviance normalisée sous l'hypothèse d'adéquation d'un HGLM aux données. Pour notre application au contrôle de qualité en microbiologie, la connaissance de la distribution exacte de la déviance normalisée nous éviterait de réaliser un test par bootstrap, et nous permettrait alors de raccourcir la durée d'exécution du test.

Calcul des variances des paramètres de dispersion. Nous proposons, en section 3.4.3, de calculer des intervalles de confiance sur les paramètres de dispersion associés à chaque facteurs (u_1^2, u_2^2 et u_3^2) à l'aide d'un bootstrap. Cette méthode s'avère efficace mais pas forcément optimale dans le cadre d'un processus de routine, car elle revient à effectuer l'estimation un nombre de fois égal au nombre d'itérations du bootstrap. Une autre perspective du travail présenté dans cette thèse est alors de développer une méthode d'estimation de la variance des paramètres de dispersion d'un HGLM.

Analyse de variance pour variables aléatoires fonctionnelles. Les analyses écotoxicologiques ont pour objectif d'évaluer la concentration d'un composé chimique ou d'un échantillon environnemental qui est susceptible de causer des effets toxiques sur des organismes tests. Lors de la mise en œuvre des tests, les organismes sont exposés à différentes concentrations de la substance à analyser. Ces tests portent sur la mortalité, l'inhibition de croissance ou de reproduction des organismes soumis aux essais, exprimées sous la forme de proportions. La mesure réalisée par un laboratoire consiste donc en une série de proportions fonction de la concentration, et peut être apparentée à une courbe, observée en pratique qu'en un nombre fini de points. Pour les paramètres écotoxicologiques, l'état de l'art en matière de maîtrise statistique des résultats d'analyse de routine est à un stade bien plus précoce que pour la microbiologie. Les approches classiques consistent à modéliser de façon paramétrique les courbes obtenues [Maul, 1991], et à relier ces paramètres avec une quantité ayant un intérêt particulier dans le cas d'analyses écotoxicologiques : la concentration qui produit un certain niveau de toxicité fixé à l'avance (par exemple 50% de mortalité). L'une des perspectives de ce travail est donc de définir une méthodologie d'évaluation de l'exactitude des résultats de mesures des laboratoires lorsque ces dernières sont des courbes. Les travaux de relatifs à l'analyse de variance fonctionnelle de [Ramsay and Silverman, 2002] peuvent constituer un premier axe de recherche.

Annexe A

Calcul des éléments déterminants pour plusieurs cas de figure représentatifs des techniques microbiologiques

Taille de la bactérie recherchée	Volume de la prise d'essai	Volume total du matériau échantillonné	n	n/N	π			$n\pi$		
					pour 10 bactéries par litre	pour 10 bactéries par ml	pour 10.000 bactéries par ml	pour 10 bactéries par litre	pour 10 bactéries par ml	pour 10.000 bactéries par ml
1 µm	1 ml	0,1 m^3	1.10^{12}	1.10^{-5}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**1.10^{-2}**	**10**	*1.10^{4}*
		1 m^3	1.10^{12}	1.10^{-6}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**1.10^{-2}**	**10**	*1.10^{4}*
		10 m^3	1.10^{12}	1.10^{-7}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**1.10^{-2}**	**10**	*1.10^{4}*
	100 ml	0,1 m^3	1.10^{14}	1.10^{-3}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**1**	*1.10^{3}*	*1.10^{6}*
		1 m^3	1.10^{14}	1.10^{-4}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**1**	*1.10^{3}*	*1.10^{6}*
		10 m^3	1.10^{14}	1.10^{-5}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**1**	*1.10^{3}*	*1.10^{6}*
	1 litre	0,1 m^3	1.10^{15}	1.10^{-2}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**10**	*1.10^{4}*	*1.10^{7}*
		1 m^3	1.10^{15}	1.10^{-3}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**10**	*1.10^{4}*	*1.10^{7}*
		10 m^3	1.10^{15}	1.10^{-4}	1.10^{-14}	1.10^{-11}	1.10^{-8}	**10**	*1.10^{4}*	*1.10^{7}*
5 µm	1 ml	0,1 m^3	8.10^{9}	1.10^{-5}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**1.10^{-2}**	**10**	*1.10^{4}*
		1 m^3	8.10^{9}	1.10^{-6}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**1.10^{-2}**	**10**	*1.10^{4}*
		10 m^3	8.10^{9}	1.10^{-7}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**1.10^{-2}**	**10**	*1.10^{4}*
	100 ml	0,1 m^3	8.10^{11}	1.10^{-3}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**1**	*1.10^{3}*	*1.10^{6}*
		1 m^3	8.10^{11}	1.10^{-4}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**1**	*1.10^{3}*	*1.10^{6}*
		10 m^3	8.10^{11}	1.10^{-5}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**1**	*1.10^{3}*	*1.10^{6}*
	1 litre	0,1 m^3	8.10^{12}	1.10^{-2}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**10**	*1.10^{4}*	*1.10^{7}*
		1 m^3	8.10^{12}	1.10^{-3}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**10**	*1.10^{4}*	*1.10^{7}*
		10 m^3	8.10^{12}	1.10^{-4}	$1,25.10^{-12}$	$1,25.10^{-9}$	$1,25.10^{-6}$	**10**	*1.10^{4}*	*1.10^{7}*

FIGURE A.1 – Calcul des éléments déterminants pour plusieurs cas de figure représentatifs des techniques microbiologiques

Annexe B

Tests

B.1 Test de Cochran

Les tests de Cochran sont des tests de comparaison de variances utilisables sur des données distribuées suivant une loi normale. Comme le soulignent [Neuilly and CEA, 1998] et [ISO, 1994b], ces tests sont utilisables pour comparer entre elles des estimations de variance S_i^2 lorsque celles-ci sont obtenues sous forme de carrés moyens, c'est-à-dire en divisant une somme Q de carrés d'écarts par le nombre ν de degrés de liberté correspondant :

$$S_i^2 = \frac{Q_i}{\nu_i}$$

Ces tests supposent la normalité des variables X_i. Toutefois, leur robustesse fait qu'on peut tolérer des écarts à cette loi de probabilité.
Le test de Cochran est utilisable si tous les nombres de degrés de liberté ν_i sont égaux entre eux :

$$\nu_1 = \nu_2 = \nu_a = ddl$$

où :

- a est le nombre de laboratoires,
- S_i^2 est la variance de chaque laboratoire.

Ces tests sont utilisés pour tester l'homoscédasticité des variances de répétabilité. Si S_{max}^2 est la plus grande des estimations S_i^2, la fonction discriminante de Cochran est :

$$g = \frac{S_{max}^2}{\sum_i S_i^2}$$

Si $g > g_{1-\alpha}$, où $g_{1-\alpha}$ correspond à la valeur critique du χ^2 au risque α, le test est significatif.
Dans le cas des essais interlaboratoires, si le test est significatif, on ne considère plus le laboratoire qui a la S_{max}^2 et on refait tourner le test.

B.2 Test de Grubbs

D'après [ISO, 1994b], le test de Grubbs appartient à la famille des tests numériques des valeurs aberrantes.

- Test pour une observation aberrante

Soit un ensemble de données g_i avec $i = \{1, \ldots, a\}$ (a est le nombre de laboratoires), rangées en ordre croissant. Pour déterminer si la plus grande observation g_a est une valeur aberrante, il faut calculer la statistique suivante :

$$G_a = \frac{(g_a - \bar{g})}{s},$$

où :

$$\bar{g} = \frac{\sum_{i=1}^{a} g_i}{a}, \qquad \text{et} \qquad s = \sqrt{\sum_{i=1}^{a} \frac{(g_i - \bar{g})^2}{a-1}}.$$

Pour tester la significativité de la plus petite valeur g_1, on calcule :

$$G_1 = \frac{(g_1 - \bar{g})}{s}.$$

Si G_1 est plus grande que les valeurs critiques au risque α, on considère alors que la valeur est aberrante.

- Test pour deux observations aberrantes

Pour tester si les deux plus grandes observations sont aberrantes, on calcule :

$$G = \frac{s^2_{a-1,a}}{s^2_0}.$$

où :

$$s^2_0 = \sum_{i=1}^{a} (g_i - \bar{g})^2, \qquad \text{et} \qquad s^2_{a-1,a} = \sum_{i=1}^{a-2} (g_i - \bar{g}_{a-1,a})^2,$$

avec :

$$\bar{g}^2_{a-1,a} = \frac{\sum_{i=1}^{a-2} g_i}{a-2}.$$

Alternativement, pour tester les deux plus petites observations, on calcule :

$$G = \frac{s^2_{1,2}}{s^2_0},$$

où :

$$s^2_0 = \sum_{i=1}^{a} (g_i - \bar{g})^2 \qquad \text{et} \qquad s^2_{1,2} = \sum_{i=1}^{a} (g_i - \bar{g}_{1,2})^2$$

avec :

$$\bar{g}_{1,2} = \frac{\sum_{i=3}^{a} g_i}{a-2}.$$

Si G est plus grande que les valeurs critiques au risque α, on considère que les valeurs sont aberrantes.

B.3 Test de Dixon

Comme le souligne [Neuilly and CEA, 1998], ce test appartient à la famille des tests numériques des valeurs aberrantes.
Soit un ensemble de données x_i avec $i = \{1, \ldots, a\}$, (a est le nombre de laboratoires) rangées en ordre croissant. Ce test permet alors de tester si la première valeur x_1 ou la dernière valeur x_a est aberrante. Suivant que le résultat douteux est le plus faible x_1 ou le plus fort x_a et que le nombre de mesures est supérieur ou non à 10, la fonction discriminante est donnée par l'expression r_1 ou r_2 indiquée dans la Table B.1 :

	x_1 est douteux	x_a est douteux
$n \leq 10$	$r_1 = \frac{x_2 - x_1}{x_a - x_1}$	$r_1 = \frac{x_a - x_{a-1}}{x_a - x_1}$
$n > 10$	$r_2 = \frac{x_3 - x_1}{x_{a-2} - x_1}$	$r_2 = \frac{x_a - x_{a-2}}{x_a - x_{a-2}}$

TABLE B.1 – Calcul de la statistique de test - Test de Dixon

Lorsque la valeur expérimentale r_1 ou r_2 est supérieure à la valeur critique, il faut admettre l'une des hypothèses suivantes avec un risque d'erreur α :

- La valeur douteuse est aberrante,
- La distribution n'est pas normale.

B.4 Test de Shapiro-Wilks

Ce test appartient à la famille des tests de normalité. Il est fondé sur le rapport W de deux estimations de la variance de la population dont provient l'échantillon :

- l'une, fonction des étendues partielles $x_n - x_1$, $x_{n-1} - x_2, \cdots$, que l'on peut déduire de la suite ordonnée ($x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$) des n observations indépendantes d'un échantillon d'effectif n.
- l'autre Z^2, fonction des carrés des écarts à la moyenne x des observations.

Ce test est appliqué si le nombre n de résultats est inférieur à 50. Il est probablement applicable au dessus, mais nous n'avons pas les valeurs critiques. Les résultats sont classés par valeurs croissantes : $x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$ On calcule :

$$Z^2 = \sum_i (x_i - \bar{x})^2 = \frac{n \sum_i x_i^2 - (\sum_i x_i)^2}{n}.$$

Puis on calcule les différences successives :

$$\begin{aligned} d_1 &= x_n - x_1, \\ d_2 &= x_{n-1} - x_2, \\ d_3 &= x_{n-2} - x_3, \\ \ldots d_j &= x_{n-j+1} - x_j. \end{aligned}$$

Avec $j = \{1, \ldots, n/2\}$ si n est pair,
et $j = \{1, \ldots, (n-1)/2\}$ si n est impair.
La fonction discriminante du test est :

$$W = \frac{(\sum a_j d_j)^2}{Z^2}.$$

Les coefficients a_j pour $5 \leq n \leq 50$ en fonction de n et de i sont donnés dans une table de [AFNOR, 1991]. La distribution de W, lorsque l'hypothèse de normalité est vérifiée, a été tabulée et la table originale de Shapiro-Wilk fournie dans [AFNOR, 1991] donne en fonction de n ($5 \leq n \leq 50$), les valeurs W_α telles que :

$$Prob(W < W_{alpha}) = \alpha,$$

avec $\alpha = 5\%$ et $\alpha = 1\%$. L'hypothèse de normalité est rejetée si W est inférieur à la valeur W_α correspondant à n et à α.

B.5 Test de Kolmogorov-Smirnov

Le test présenté ici est une adaptation du test de Kolmogorov-Smirnov au cas de la loi normale dont les paramètres sont estimés [AFNOR, 1991]. Ce test est fondé sur la comparaison de la loi normale $F(x) = N(\mu, \sigma)$ (avec μ et σ estimés par x et s) avec la loi de probabilité $F_n(x)$ obtenue à partir d'un échantillon aléatoire de n observations :

$$x_1 \leq x_2 \leq ... \leq x_n$$

On centre et on réduit la loi normale en posant :

$$z_i = \frac{x_i - \bar{x}}{s}$$

où $\bar{x}$ est la moyenne de l'échantillon et :

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

On considère les effectifs cumulés de la façon suivante : Effcumul $= F(z_i)$ où F est la fonction de répartition de la loi normale centrée réduite. Ces valeurs sont lues dans la table de la loi normale centrée réduite. On définit :

$$\begin{aligned} D^+ &= \max\ (\frac{i}{n} - \text{Effcumul}) \text{ pour } 1 \leq i \leq n \\ D^- &= \max\ (\text{Effcumul} - \frac{i-1}{n}) \text{ pour } 1 \leq i \leq n \\ D &= \max\ (D^+, D^-) \end{aligned}$$

On calcule ensuite la statistique modifiée $T(D)$ en utilisant la formule suivante :

$$T(D) = D(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}})$$

On compare $T(D)$ avec les valeurs limites données dans [AFNOR, 1991]. Si $T(D) >$ vl, où vl est une valeur limite, le test est significatif.

Annexe C

Organigrammes récapitulatifs de l'exploitation des données des laboratoires

C.1 Méthodes énumératives

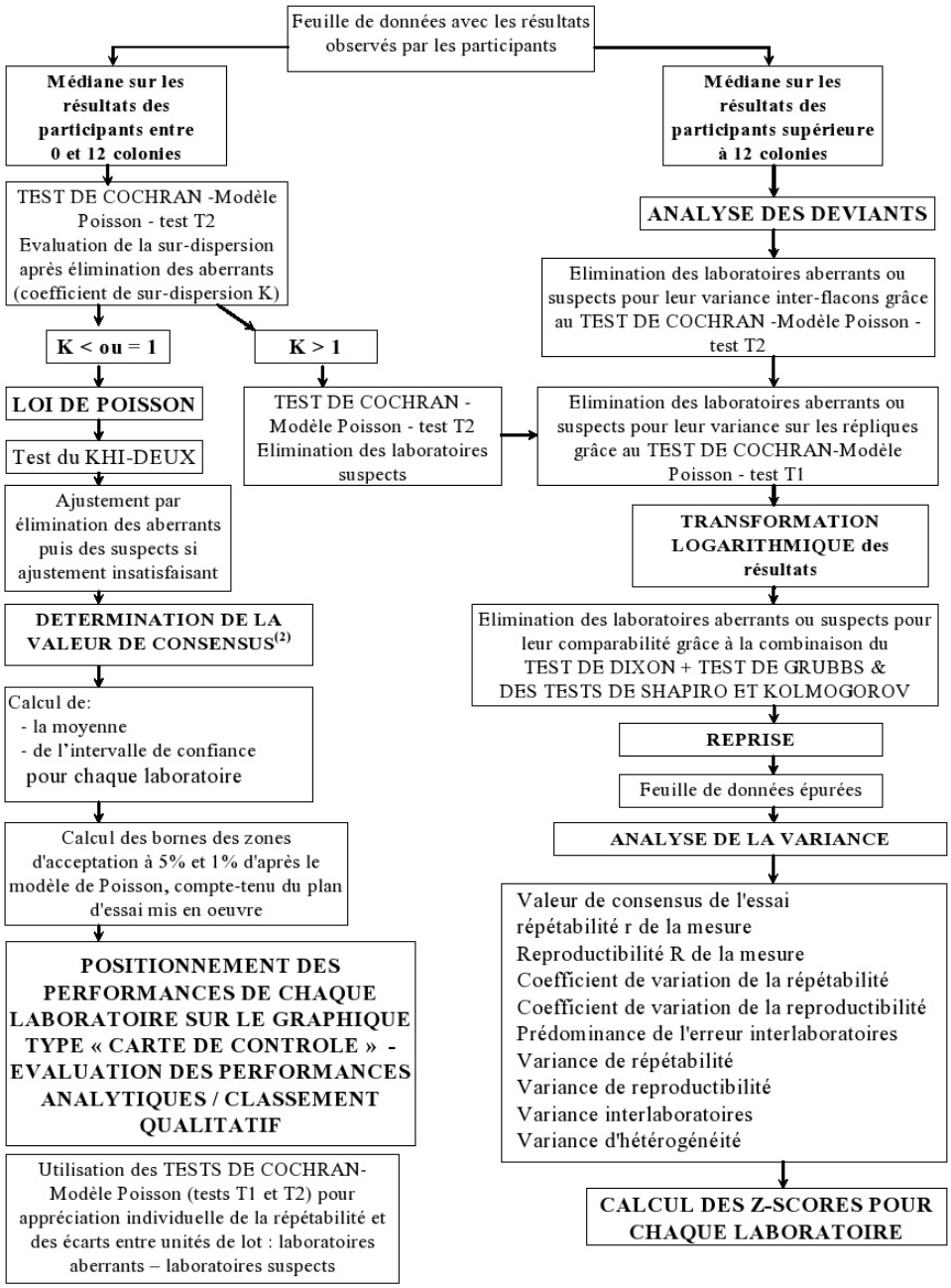

FIGURE C.1 – Organigramme relatif au traitement de données pour les paramètres microbiologiques - méthodes énumératives

C.2 Méthodes quantiques

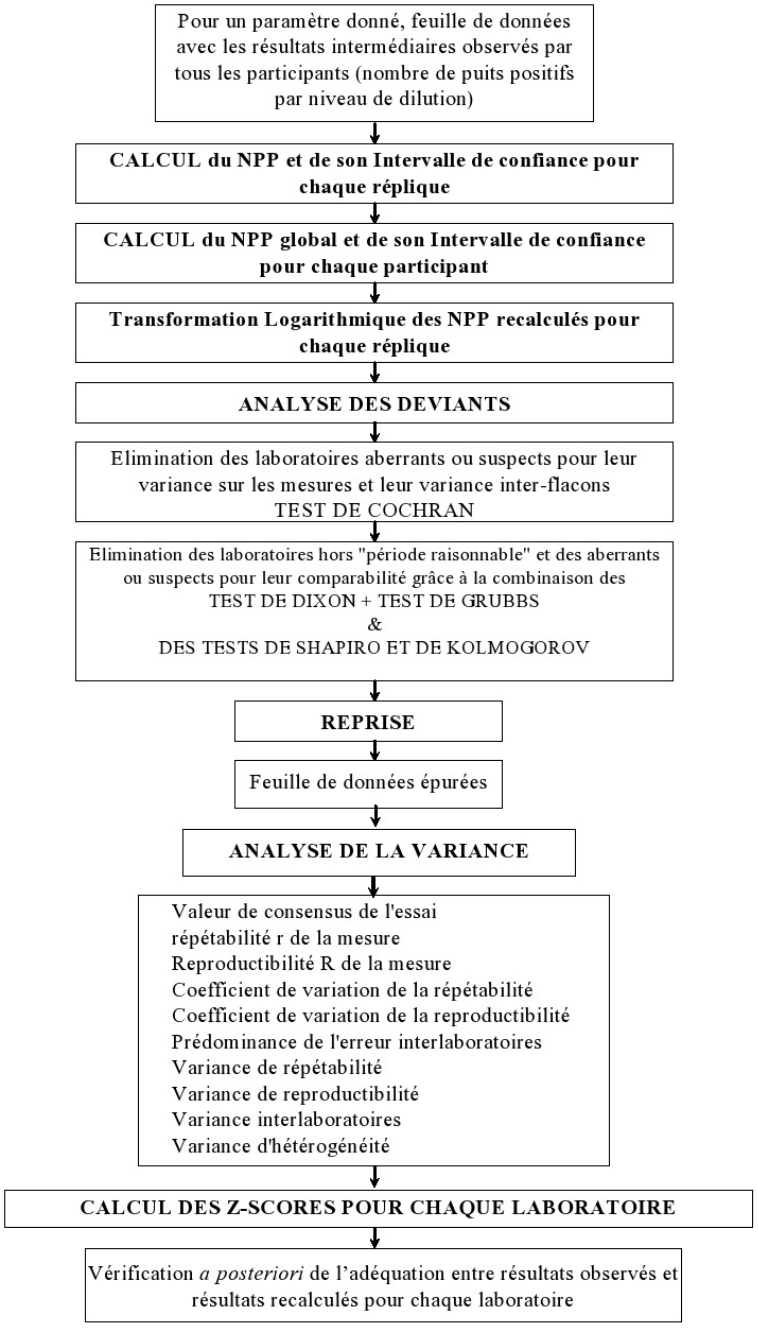

FIGURE C.2 – Organigramme relatif au traitement de données pour les paramètres microbiologiques - méthodes quantiques

Annexe D

Analyse de variance à 2 facteurs fixes imbriqués

Proposition D.0.1. *Une analyse de variance à deux facteurs A, et B imbriqué dans A est équivalente à une analyse de variance à deux facteurs A et B, avec effet du facteur A et effet d'interaction entre les deux facteurs A et B, mais sans effet du facteur B seul.*

Preuve : Soit le modèle d'analyse de variance à deux facteurs imbriqués suivant :

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

Dans ce modèle, le facteur A a a niveaux, le facteur B a b niveaux imbriqués dans chaque niveau du facteur A, et il y a n réplications par flacon. L'indice $j(i)$ indique que le $j^{\text{ème}}$ niveau du facteur B est imbriqué dans le $i^{\text{ème}}$ niveau du facteur A. μ est la moyenne globale, τ_i est l'effet du laboratoire i, $\beta_{j(i)}$ est l'effet du flacon j imbriqué dans le laboratoire i, et $\epsilon_{(ij)k}$ est l'erreur de mesure entre les réplications d'un flacon pour un laboratoire.
Il s'agit d'un plan imbriqué équilibré comportant un même nombre de niveaux du facteur B au sein de chaque niveau du facteur A, ainsi qu'un même nombre de réplications.
Pour ce modèle, la somme des carrés corrigée s'écrit :

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[(\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{ij.} - \overline{y}_{i..}) + (y_{ijk} - \overline{y}_{ij.})]^2 \tag{D.2}$$

Compte-tenu que les trois produits croisés valent zéro, en développant le membre droit de l'équation D.2, on obtient :

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2 = bn\sum_{i=1}^{a}(\overline{y}_{i..} - \overline{y}_{...})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.} - \overline{y}_{i..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2. \tag{D.3}$$

L'équation D.3 indique que la somme des carrés totale peut être divisée en une somme de carrés due au facteur A, une somme de carrés due au facteur B sous les différents niveaux du facteur A, et une somme de carrés attribuée l'erreur. On peut alors écrire l'équation D.3 de la façon suivante :

$$SC_T = SC_A + SC_{B(A)} + SC_E \tag{D.4}$$

SC_T correspond à la somme des carrés totale et a $abn-1$ degrés de liberté. SC_A correspond à la somme des carrés du facteur A et a $a-1$ degrés de liberté, et SC_B correspond à la somme des carrés du facteur B et a $a(b-1)$ degrés de liberté. L'erreur a $ab(n-1)$ degrés de liberté. On remarque que :

$$abn-1=(a-1)+a(b-1)+ab(n-1)$$

Comparons l'analyse de variance à deux facteurs imbriqués, dans laquelle le deuxième facteur prend ses différentes modalités pour chaque modalité du premier facteur, avec une analyse de variance classique à deux facteurs.

Considérons une analyse de variance classique à deux facteurs, A et B, où le facteur A a a niveaux, et le facteur B b niveaux. Ces niveaux sont arrangés suivant un plan factoriel, c'est à dire que chaque réplication de l'expérience comporte toutes les combinaisons ab possibles. On considère n réplications.
Soit y_{ijk} la réponse observée quand le facteur A est au niveau i, et le facteur B au niveau j pour la $k_{i^{\text{ème}}}$ réplication.
L'analyse de variance classique à deux facteurs s'écrit :

$$y_{ijk}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+\epsilon_{ijk} \quad \begin{cases} i & =1,\dots,a \\ j & =1,\dots,b \\ k & =1,\dots,n \end{cases} \tag{D.5}$$

où

- μ est la moyenne globale,
- τ_i est l'effet du $i^{\text{ème}}$ niveau du facteur A,
- β_j est l'effet du $j^{\text{ème}}$ niveau du facteur B,
- $(\tau\beta)_{ij}$ est l'effet de l'interaction entre τ_i et β_j,
- ϵ_{ijk} est la composante d'erreur aléatoire.

La décomposition de la somme des carrés totale s'écrit :

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[(\bar{y}_{i..}-\bar{y}_{...})+(\bar{y}_{.j.}-\bar{y}_{...})+(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})+(y_{ijk}-\bar{y}_{ij.})]^2 \\ &= bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2+an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2+n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 \\ &\quad+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2, \end{aligned} \tag{D.6}$$

soit

$$SC_T=SC_A+SC_B+SC_{AB}+SC_E$$

où SC_A est la somme des carrés due au facteur A :

$$SC_A=bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2.$$

SC_B est la somme des carrés due au facteur B :

$$SC_B = an \sum_{j=1}^{b} (\bar{y}_{.j.} - \bar{y}_{...})^2 .$$

SC_{AB} est la somme des carrés due à l'interaction des facteurs A et B :

$$SC_{AB} = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 .$$

SC_E est la somme des carrés résiduelle :

$$SC_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{ij.})^2 .$$

Si l'effet du facteur B est nul, alors les effets des j niveaux du facteur B, β_j, sont tous égaux :

$$\beta_1 = \beta_2 = ... = \beta_j \quad ,$$

et

$$\bar{y}_{.j.} = \bar{y}_{...}.$$

La décomposition de la variance totale devient alors :

$$\begin{aligned} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{...})^2 = bn \sum_{i=1}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2 \\ + n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..})^2 \\ + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{ij.})^2 \end{aligned} \tag{D.7}$$

On constate que l'équation D.7 ci-dessus est équivalente à l'équation D.3. La décomposition de la variance totale d'une analyse de variance à deux facteurs imbriqués est donc la même que celle d'une analyse de variance factorielle à deux facteurs, sans effet du facteur B.
On en conclue que l'analyse de variance à deux facteurs imbriqués est équivalente à une analyse de variance factorielle à deux facteurs, sans effet du facteur B seul.

□

Annexe E

Sélection de modèles par AIC conditionnel

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	3080.089	3136.978	3186.891	802
2	3089.995	3138.975	3187.623	147
3	3099.877	3146.407	3190.665	1
4	3104.586	3150.356	3189.36	4
5	3096.031	3142.1	3187.111	46
6	3105.313	3150.157	3189.249	0
7	3106.053	3150.961	3189.922	0
8	3105.944	3150.926	3190.65	0

TABLE E.1 – Sélection de modèles par AIC conditionnel - simulations - Cas 1 (modèle 1)

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	4361.817	4501.38	4638.798	0
2	3128.873	3192.742	3254.844	889
3	3237.738	3289.861	3348.504	0
4	3444.906	3483.484	3518.86	0
5	3136.444	3197.232	3259.702	75
6	3147.783	3204.815	3257.296	29
7	3251.774	3300.094	3350.886	0
8	3147.154	3204.851	3258.738	7

TABLE E.2 – Sélection de modèles par AIC conditionnel - simulations - Cas 2 (modèle 2)

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	4150.617	4239.86	4322.203	0
2	5306.989	6089.961	6920.797	0
3	3265.033	3307.995	3353.038	845
4	3472.436	3491.083	3509.157	0
5	3265.436	3308.307	3353.252	98
6	3465.311	3484.789	3504.813	0
7	3280.157	3317.67	3352.71	39
8	3282.298	3319.246	3353.988	18

TABLE E.3 – Sélection de modèles par AIC conditionnel - simulations - Cas 3 (modèle 3)

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	4176.785	4242.784	4302.943	0
2	6823.002	7551.052	8312.383	0
3	5606.079	6173.72	6755.928	0
4	3476.731	3491.03	3503.029	856
5	5606.492	6174.229	6756.524	0
6	3477.08	3491.416	3503.236	102
7	3477.37	3491.917	3503.406	40
8	3477.549	3492	3503.535	2

TABLE E.4 – Sélection de modèles par AIC conditionnel - simulations - Cas 4 (modèle 4)

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	4530.118	4680.446	4822.092	0
2	5173.563	6023.223	7015.072	0
3	3212.828	3269.056	3326.185	1
4	3420.21	3464.557	3504.733	0
5	3210.442	3266.709	3324.222	952
6	3392.292	3440.024	3483.01	0
7	3226.835	3280.296	3330.982	0
8	3223.182	3276.04	3328.751	47

TABLE E.5 – Sélection de modèles par AIC conditionnel - simulations - Cas 5 (modèle 5)

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	4606.129	4677.915	4750.568	0
2	6929.135	7695.479	8557.015	0
3	5681.744	6274.582	7012.921	0
4	3472.956	3487.688	3500.974	0
5	5677.217	6269.68	7007.98	0
6	3460.534	3475.98	3489.736	935
7	3465.819	3480.532	3494.022	0
8	3461.658	3476.871	3490.659	65

TABLE E.6 – Sélection de modèles par AIC conditionnel - simulations - Cas 6 (modèle 6)

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	4439.605	4521.404	4612.5	0
2	9078.44	10338.07	11930.86	0
3	5602.153	6192.789	6847.079	0
4	3461.131	3482.172	3504.965	0
5	5602.351	6192.877	6847.184	0
6	3459.734	3480.966	3504.119	0
7	3456.678	3477.417	3500.939	836
8	3456.719	3477.487	3501.049	164

TABLE E.7 – Sélection de modèles par AIC conditionnel - simulations - Cas 7 (modèle 7)

	Quantiles			
Modèle	0.025	0.5	0.975	Nombre de fois où le modèle est choisi
1	4675.962	4816.97	4970.614	0
2	8683.096	10328.08	12104.68	0
3	5438.495	6113.46	6918.305	0
4	3404.283	3449.122	3490.98	0
5	5344	6112.011	6917.058	0
6	3395.122	3441.181	3483.606	0
7	3391.782	3437.491	3482.093	3
8	3390.09	3436.378	3480.42	997

TABLE E.8 – Sélection de modèles par AIC conditionnel - simulations - Cas 8 (modèle 8)

Bibliographie

[Aalo et al., 2005] Aalo, V., Piboongungon, T., and Iskander, C. (2005). Bit-error rate of binary digital modulation schemes in generalized gamma fading channels. *Communications Letters, IEEE*, 9(2) :139–141.

[AFNOR, 1991] AFNOR (1991). X 06-050. *Application de la Statistique : Etude de la normalité d'une distribution.*

[AFNOR, 2000] AFNOR (2000). Iso/iec 17025 : Exigences générales concernant la compétence des laboratoires d'étalonnages et d'essais.

[AGLAE,] AGLAE. Pr-6-26b - traitement statistique des données biologie.

[Ali et al., 2008] Ali, M., Woo, J., and Nadarajah, S. (2008). Generalized gamma variables with drought application. *Journal of the Korean Statistical Society*, 37(1) :37–45.

[BCR, 1993] BCR (1993). Bcr information : Statistical analysis of certification trials for microbiological reference materials. *Report EUR 150008 EN.*

[Besse,] Besse, P. Modèles à effets aléatoires et modèles mixtes.

[Bliss and Fisher, 1953] Bliss, C. and Fisher, R. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2) :176–200.

[Breiman and Friedman, 1985] Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391) :580–598.

[Brownlee, 1965] Brownlee, K. (1965). *Statistical theory and methodology in science and engineering*, volume 150. Wiley New York.

[Burda et al., 2012] Burda, M., Harding, M., and Hausman, J. (2012). A poisson mixture model of discrete choice. *Journal of econometrics*, 166(2) :184–203.

[Cameron and Trivedi, 1998] Cameron, A. and Trivedi, P. (1998). Regression analysis of count data cambridge university press. *Nueva York.*

[Coelho and Arnold, 2014] Coelho, C. and Arnold, B. (2014). On the exact and near-exact distributions of the product of generalized gamma random variables and the generalized variance. *Communications in Statistics-Theory and Methods*, 43(10-12) :2007–2033.

[Commenges et al., 2011] Commenges, D., Jolly, D., Drylewicz, J., Putter, H., and Thiébaut, R. (2011). Inference in hiv dynamics models via hierarchical likelihood. *Computational Statistics & Data Analysis*, 55(1) :446–456.

[Droesbeke et al., 2005] Droesbeke, J., Lejeune, M., and Saporta, G.and Lejeune, M. (2005). *Modèles statistiques pour données qualitatives*. Technip.

[Efron, 1979] Efron, B. (1979). Bootstrap methods : another look at the jackknife. *The annals of Statistics*, pages 1–26.

[El-Shaarawi et al., 1981] El-Shaarawi, A., Esterby, S., and Dutka, B. (1981). Bacterial density in water determined by poisson or negative binomial distributions. *Applied and Environmental Microbiology*, 41(1) :107–116.

[Ha et al., 2007] Ha, I. D., Lee, Y., and MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statistics in medicine*, 26(26) :4790–4807.

[Haldane, 1939] Haldane, J. (1939). Sampling errors in the determination of bacterial or virus density by the dilution method. *Journal of Hygiene*, 39(03) :289–293.

[Henderson et al., 1959] Henderson, C., Kempthorne, O., Searle, S., and Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2) :192–218.

[ISO, 1994a] ISO (1994a). Iso 5725-1 : Exactitude (justesse et fidélité) des résultats et méthodes de mesure - partie 1 : Principes généraux et définitions.

[ISO, 1994b] ISO (1994b). Iso 5725-2. exactitude (justesse et fidélité) des résultats et méthodes de mesure - partie 2 : Méthode de base pour la détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée.

[ISO, 2000] ISO (2000). Iso/tr 13843 : Qualité de l'eau - lignes directrices pour la validation des méthodes microbiologiques.

[ISO, 2005] ISO (2005). Iso 13528 : Méthodes statistiques utilisées dans les essais d'aptitude par comparaisons interlaboratoires.

[ISO, 2006] ISO (2006). Iso 35 : Matériaux de référence - principes généraux et statistiques pour la certification.

[ISO, 2010a] ISO (2010a). Iso 17043 : Évaluation de la conformité – exigences générales concernant les essais d'aptitude.

[ISO, 2010b] ISO (2010b). Iso/ts 22117 : Microbiology of food and animal feeding stuffs - specific requirements and guidance for proficiency testing by interlaboratory comparison.

[ISO, 2012a] ISO (2012a). Iso 15189 : Laboratoires d'analyses de biologie médicale – exigences particulières concernant la qualité et la compétence.

[ISO, 2012b] ISO (2012b). Iso 29201 : Qualité de l'eau – variabilité des résultats d'essais et incertitude de mesure des méthodes d'énumération microbienne.

[Jarvis, 1989] Jarvis, B. (1989). *Statistical aspects of the microbiological analysis of foods*.

[Johnson and Nier, 1953] Johnson, E. and Nier, A. (1953). Angular aberrations in sector shaped electromagnetic lenses for focusing beams of charged particles. *Physical Review*, 91(1).

[Karagiannidis et al., 2006] Karagiannidis, G., Tsiftsis, T., and Mallik, R. (2006). Bounds for multihop relayed communications in nakagami-m fading. *Communications, IEEE Transactions on*, 54(1) :18–22.

[Lawrence et al., 2013] Lawrence, C., Attar, A., and Barbut, F. (2013). Accréditation pour la recherche des légionelles dans l'eau selon la norme nf en iso/cei 17025 : retour d'expérience de deux laboratoires hospitaliers. *Revue Francophone des Laboratoires*, 2013(453) :53–58.

[Lee and Nelder, 1996] Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678.

[Lee and Nelder, 1998] Lee, Y. and Nelder, J. (1998). Generalized linear models for the analysis of quality-improvement experiments. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 95–105.

[Lee and Nelder, 2000] Lee, Y. and Nelder, J. (2000). Two ways of modelling overdispersion in non-normal data. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 49(4) :591–598.

[Lee and Nelder, 2001] Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models : a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4) :987–1006.

[Lee and Nelder, 2002] Lee, Y. and Nelder, J. (2002). Analysis of ulcer data using hierarchical generalized linear models. *Statistics in medicine*, 21(2) :191–202.

[Lee et al., 2007] Lee, Y., Nelder, J., and Noh, M. (2007). H-likelihood : problems and solutions. *Statistics and Computing*, 17(1) :49–55.

[Lee et al., 2011] Lee, Y., Nelder, J., and Park, H. (2011). Hglms for quality improvement. *Applied Stochastic Models in Business and Industry*, 27(3) :315–328.

[Lee et al., 2006] Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized linear models with random effects : unified analysis via H-likelihood*. CRC Press.

[Lehmann and Romano, 2006] Lehmann, E. and Romano, J. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

[Li and Wang, 1998] Li, Q. and Wang, S. (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics*, 87(1) :145–165.

[Lightfoot and Maier, 1998] Lightfoot, N. and Maier, E. (1998). *Microbiological analysis of food and water : Guidelines for quality assurance*. Elsevier.

[Loingeville et al., 2014] Loingeville, F., Jacques, J., Preda, C., Guarini, P., and Molinier, O. (2014). Analyse de variance à 2 facteurs imbriqués sur données de comptage-application au contrôle de qualité. *46èmes Journées de Statistique*.

[Loingeville et al., 2015] Loingeville, F., Jacques, J., Preda, C., Guarini, P., and Molinier, O. (2015). Modèle linéaire généralisé hiérarchique gamma-poisson à 3 facteurs aléatoires-application au contrôle de qualité. *47èmes Journées de Statistique*.

[Maindonald and Braun, 2006] Maindonald, J. and Braun, J. (2006). *Data analysis and graphics using R : an example-based approach*, volume 10. Cambridge University Press.

[Malik, 1968] Malik, H. (1968). Exact distribution of the product of independent generalized gamma variables with the same shape parameter. *The Annals of Mathematical Statistics*, 39(5) :1751–1752.

[Marques et al., 2014] Marques, F., Coelho, C., and de Carvalho, M. (2014). On the distribution of linear combinations of independent gumbel random variables. *Statistics and Computing*, 25(3) :683–701.

[Marques and Loingeville, 2015] Marques, F. and Loingeville, F. (2015). On the distribution of the product of independent generalized gamma random variables - an application to quality control in microbiology. Technical report.

[Marques, 2012] Marques, F. J. (2012). On the product of independent generalized gamma random variables. Technical report, Discussion Paper 19–2012, CMA-FCT-Universidade Nova de Lisboa.

[Mathai, 1972] Mathai, A. (1972). Products and ratios of generalized gamma variates. *Scandinavian Actuarial Journal*, 1972(2) :193–198.

[Maul, 1991] Maul, A. (1991). Aspects statistiques des méthodes de quantification en virologie. *Virologie des milieux hydriques*, pages 143–171.

[McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. (1989). *Generalized linear models*, volume 37. CRC press.

[McCulloch and Neuhaus, 2001] McCulloch, C. and Neuhaus, J. (2001). *Generalized linear mixed models*. Wiley Online Library.

[Montgomery, 2008] Montgomery, D. (2008). *Design and analysis of experiments*. John Wiley & Sons.

[Nelder and Lee, 1992] Nelder, J. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood : some comparisons. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 273–284.

[Nelder and Pregibon, 1987] Nelder, J. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74(2) :221–232.

[Nelder and Wedderburn, 1972] Nelder, J. and Wedderburn, J. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*.

[Neuilly and CEA, 1998] Neuilly, M. and CEA (1998). Modélisation et estimation des erreurs de mesure.

[Niemelä, 2002] Niemelä, S. (2002). *Uncertainty of quantitative determinations derived by cultivation of microorganisms*. Centre for Metrology and Accreditation Helsinki.

[O'Hara and Kotze, 2010] O'Hara, R. and Kotze, D. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2) :118–122.

[Pawitan, 2001] Pawitan, Y. (2001). *In all likelihood : statistical modelling and inference using likelihood*. Oxford University Press.

[Podolski, 1972] Podolski, H. (1972). The distribution of a product of n independent random variables with generalized gamma distribution. *Demonstratio Math*, 4(2) :119–123.

[Quenouille, 1949] Quenouille, M. (1949). Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 483–484. Cambridge Univ Press.

[Ramsay and Silverman, 2002] Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis : methods and case studies*, volume 77. Springer New York.

[Rönnegård et al., 2010] Rönnegård, L., Shen, X., and Alam, M. (2010). hglm : A package for fitting hierarchical generalized linear models. *The R Journal*, 2(2) :20–28.

[Salo and Vainikainen, 2006] Salo, J., E.-S. H. and Vainikainen, P. (2006). The distribution of the product of independent rayleigh random variables. *Antennas and Propagation, IEEE Transactions on*, 54(2) :639–643.

[Salo et al., 2004] Salo, J., Suvikunnas, P., El-Sallabi, H., and Vainikainen, P. (2004). Approximate distribution of capacity of rayleigh fading mimo channels. *Electronics Letters*, 40(12) :741–742.

[Saporta, 2011] Saporta, G. (2011). *Probabilités, analyse des données et statistique*. Editions Technip.

[Stacy, 1962] Stacy, E. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, pages 1187–1192.

[Tillett and Lightfoot, 1995] Tillett, H. and Lightfoot, N. (1995). Quality control in environmental microbiology compared with chemistry : What is homogeneous and what is random ? *Water science and technology*, 31(5) :471–477.

[Tukey, 1958] Tukey, J. (1958). Bias and confidence in not-quite large samples. In *Annals of Mathematical Statistics*, volume 29, pages 614–614.

[Vaida and Blanchard, 2005] Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, 92(2) :351–370.

[Wedderburn, 1974] Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3) :439–447.

[Withers and Nadarajah, 2013] Withers, C. and Nadarajah, S. (2013). On the product of gamma random variables. *Quality & Quantity*, 47(1) :545–552.

[Zaninetti, 2008] Zaninetti, L. (2008). On the product of two gamma variate with argument 2 : Application to the luminosity function for galaxies. *arXiv preprint arXiv :0806.4086*.