Revue de l'analyse statistique de la réponse immunitaire HI liée aux vaccinations anti-grippe

Jérémy Magnanensi

GlaxoSmithKline Biologicals
Influenza Vaccines Department
Avenue Fleming, 20
B-1300 Wavre, Belgium

Université de Strasbourg
U.F.R. de Mathématique
7, rue René Descartes
67084 Strasbourg

Rapport de Stage
Master de Statistique 2ème Année
07 Février 2011 - 29 Juillet 2011

Intitulé du stage : Revue de l'analyse statistique de la réponse immunitaire HI liée aux vaccinations anti-grippe.

Stagiaire:

MAGNANENSI Jérémy

Stagiaire bio-statisticien

Maître de Stage :

DEWE Walthere P.

Senior Manager Bio-Statistician – Clinical Biometrics

Table des Matières :

Remerciements :

En premier lieu, je tenais à remercier Mr Dewe Walthere, mon responsable de stage, pour son écoute, ses aides et la confiance qu'il m'a porté dans l'exécution quotidienne de mes tâches tout au long du stage. Il a pris tout particulièrement soin à effectuer un suivi régulier de mon travail et a toujours été disponible en cas de problèmes lors de la résolution de mes objectifs.

Je remercie notamment Mr Tibaldi Fabian, Senior Manager in Clinical Biometrics, ainsi que l'ensemble des statisticiens du service pour l'aide et les conseils qu'ils nous ont apporté sur ce sujet ainsi que Mr Fourneau Marc, Directeur GRCD Biometrics pour la confiance qu'il m'a porté en validant mon stage au sein de GSK.

Enfin, je tiens naturellement à remercier ma famille, mon père, ma mère, ma soeur et mon frère pour leur soutien qu'ils m'ont apporté afin que je puisse réaliser mon stage dans les meilleures conditions possibles.

I. Préface

Lors de mon premier semestre de Master 2, j'ai décidé d'axer mes recherches de stage dans le domaine de la bio-statistique. En effet, depuis le début de ce Master, je savais que le domaine qui serait susceptible de me plaire le plus au sein des multiples applications des statistiques était celui de la bio-statistique. De plus, ayant déjà effectué mon stage de Master 1^{ere} année dans ce domaine, je n'avais donc plus aucun doute quant à mon intérêt pour ce domaine. J'ai donc recherché un stage et envoyé des lettres aux principaux laboratoires pharmaceutiques ainsi qu'aux C.R.O. (Contract Research Organisation), entreprises spécialisées dans les études cliniques pour le compte de l'industrie pharmaceutique. J'ai ainsi eu un retour positif de l'entreprise pharmaceutique GlaxoSmithKline basée à Wavre en Belgique, en périphérie de Bruxelles, entreprise au sein de laquelle j'avais déjà effectué mon stage de Master1. C'est donc sans aucune hésitation que j'ai accepté cette offre de stage. Cela m'a ainsi permit de rester dans une certaine continuité et cohérence vis à vis de mon précédent stage et je pense qu'il s'agit d'un réel avantage pour le futur. De plus, cela montre une certaine confiance de l'entreprise à mon égard, ce qui n'est pas négligeable qui plus est de la part du numéro deux mondial.

Ce stage étant d'une durée de 6 mois, le sujet qui m'a été proposé a donc été beaucoup plus ouvert et a naturellement demandé une quantité de travail bien plus importante que lors de mon stage de Master1. Cependant, j'ai retrouvé les avantages de l'année dernière, à savoir un sujet qui laisse place non seulement à de l'application directe de méthodes statistiques sur certaines données, mais également, et en grande partie, un travail conséquent de recherche à travers des lectures d'articles scientifiques, de publications etc... En effet, mon objectif pour ce stage a été d'élaborer de nouvelles méthodes d'analyse des titres obtenus lors d'essais cliniques sur les vaccins grippaux saisonniers. Le cadre de travail était donc très large puisqu'il m'a été possible d'envisager toutes sortes de méthodes afin d'améliorer le système actuel. Ceci fut d'ailleurs parfois compliqué puisqu'il m'a fallu faire des choix lors de la sélection des possibilités envisagées afin de ne pas partir dans plusieurs directions et ainsi avoir le temps de traiter une méthode clairement et correctement. Ce travail se révèle être très important pour GSK puisque, avec leur méthode actuelle d'analyse, beaucoup d'hypothèses et de suppositions ont été faites, sans que celles-ci ne soient réalisées ou vérifiées réellement et sans connaitre l'impact de telles suppositions non réalisées sur les résultats des analyses. Il s'agissait donc d'un sujet d'autant plus intéressant que les résultats allaient avoir un impact immédiat sur le travail des statisticiens au sein de l'équipe Flu. De plus, le fait de refaire une partie de mes études à l'étranger ne pouvait être que bénéfique pour moi et pour mon avenir.

Pour ma part, j'ai pu approfondir mes connaissances sur le fonctionnement de laboratoires pharmaceutiques dans le sens où l'année dernière mon stage s'est déroulé lors des mois estivaaux, mois lors desquels beaucoup de personnes sont en congés, le rythme de travail est différent et la majorité du travail statistique sur les études liées aux vaccins grippaux est bouclé. Or cette année, j'ai réellement pu assister au travail réel et quotidien que les statisticiens effectuent au sein d'un grand laboratoire pharmaceutique comme l'est GSK et ce fut réellement très intéressant et enrichissant. GSK est une entreprise très dynamique, composée d'un personnel, du moins au sein du service statistique dans lequel j'ai effectué ce stage, très ouvert et rigoureux dans leur travail. De plus, comme l'année dernière, je désirais effectuer mon stage dans une entreprise où je pourrai côtoyer des personnes spécialistes des statistiques et non une entreprise où l'on aurait compté que sur moi afin de résoudre un problème sans pouvoir apprendre d'une personne spécialisé dans le domaine. Tout ceci je l'ai trouvé chez GlaxoSmithKline, avec des personnes très accessibles et qui ont su prendre le temps de m'aider ou de m'expliquer certains points lorsque j'avais des difficultés.

II. Présentation de l'entreprise

2.1. Glaxo Smith Kline (GSK), un leader mondial

GSK est l'un des leaders mondiaux dans le domaine pharmaceutique. En effet, il s'agit du deuxième laboratoire pharmaceutique mondial, en terme de chiffre d'affaire, en affichant une forte présence dans une vingtaine de domaines thérapeutiques tel que l'oncologie, la neurologie, le VIH sans oublier la vaccinologie dont les vaccins antigrippaux, secteur dans lequel j'ai effectué mon stage.

Il s'agit d'une entreprise possédant un très grand passé. En effet, son historique remonte à 1715 avec la création, à Londres, de la pharmacie de Plough Court par Silvanus Bevan. En 1830, John K. Smith ouvrira sa première pharmacie à Philadelphie qui, rejoint par son frère en 1841, formeront John K Smith & Co. S'en suivront de multiples fusions et associations qui feront naître en 2000 GSK à travers la fusion de Glaxo Welcome et SmithKline Beecham.

GSK emploie aujourd'hui quelques 100.000 personnes à travers le monde en étant présent dans près de 120 pays. Elle est constituée de deux parties à savoir GSK Santé Grand Public, spécialisée dans les médicaments d'automédication vendus sans prescription en pharmacie et non remboursables ainsi que dans le domaine buccodentaire, et GSK Pharma, spécialisée dans les médicaments vendus sur prescription médicale, et à l'intérieur de laquelle se situe GSK Biologicals focalisée sur les vaccins. C'est donc au sein de GSK Bio que j'ai effectué mon stage, plus particulièrement dans le secteur s'occupant des virus anti-grippaux. GSK Bio c'est plus de 9000 personnes dans le monde dont 1600 scientifiques qui développent des vaccins pour le monde entier et notamment des vaccins contre les trois principaux fléaux planétaires, considérés comme prioritaires par l'OMS, à savoir le VIH/SIDA, la tuberculose et enfin le paludisme. GSK Biologicals abrite trois sites de production belges, le premier et plus ancien à Rixensart, puis un site à Gembloux et un dernier à Wavre, site sur lequel s'est déroulé mon stage. Ces trois sites réunis constituent l'un des plus importants pôles industriels de vaccins en Europe et accueillent également l'ensemble des services qui coordonnent l'activité de recherche, de développement et de production de GSK Biologicals dans le monde. GSK Biologicals, c'est aussi 14 sites de production sur 3 continents, qui chaque année produisent, formulent, conditionnent, contrôlent et expédient plus d'un milliard de doses d'une trentaine de vaccins différents dans 169 pays en sachant que près de 80% de la production est destinée aux pays en développement. Ces vaccins permettent d'éviter plus de trois millions de décès chaque année et des séquelles liées à des maladies infectieuses à plus de 750.000 enfants. Enfin, GSK possède un chiffre d'affaire de 34.1 milliards d'euros et en consacre 17% au seul secteur de Recherche & Développement. GSK Pharma représente 29.5 milliards d'euros et GSK Santé Grand Public 4.6 milliards. A noter que 445 millions d'euros sont consacrés à des programmes humanitaires soutenus par GSK et sa fondation.

En effet, GSK est une entreprise très engagée dans le suivi des patients, notamment des enfants et des populations du tiers-monde. Grâce, entre autre, à sa fondation, GSK développe des actions dans plusieurs domaines. En effet, ils sont présents dans l'accompagnement des enfants malades, le développement de la recherche dans le domaine de la santé des Femmes, l'amélioration de l'accès à l'information pour les malades ou encore l'amélioration de la santé dans les pays en développement. Pour cela, ils se fixent des domaines de Recherche & Développement sur des maladies frappant majoritairement le tiers-monde, fournissent à ces pays les vaccins, antirétroviraux et antipaludéens

à tarifs préférentiels ou encore soutiennent des programmes de proximité concernant des maladies qui sévissent plus particulièrement dans ces pays comme la filariose lymphatique, le VIH/SIDA, le paludisme ainsi que la diarrhée de l'enfant en procédant à des dons en médicaments, financiers ou technique. Afin de faciliter la mise à disposition de ses vaccins, GSK a également étendu ses partenariats avec l'Unicef, la Global Alliance for Vaccine Immunization (GAVI), l'Organisation mondiale de la Santé (OMS) et l'Organisation Panaméricaine de la Santé (PAHO). Lors de ce stage, j'ai ainsi pu assister à la remise d'un chèque de 50.000 euros à l'Unicef afin de financer le projet "Des écoles pour Haïti".

2.2. GSK Biologicals en Belgique

Tout d'abord, débutons par un petit historique. L'histoire commence par la création en 1945 de l'entreprise belge Recherche et Industrie Thérapeutiques (RIT) à Genval dans la banlieue de Bruxelles par le Docteur Pieter De Sommer. En 1956, débute la production du vaccin contre la polio qui va connaître un très grand succès d'où l'acquisition dès 1958 du site de Rixensart. En 1968, RIT est racheté par SmithKline & French et devient ainsi SmithKline-RIT avant de s'appeler en 1989 SmithKline Beecham Biologicals suite à la fusion de SmithKline et Beecham. En 1995, se produit l'extension aux deux sites supplémentaires de Wavre et Gembloux avant de finir par s'appeler GSK Biologicals suite à la dernière fusion en date de SmithKline et Glaxo en 2000.

Le site de Wavre est très récent et n'est même pas encore complètement achevé. Ce site emploie, au quotidien, aux alentours de 4000 personnes et est constitué de plus d'une quinzaine de bâtiments dont certains, comme celui dans lequel je travaillais, ont été inaugurés au courant des années 2009/2010. Il s'agit d'un site de production mais également de recherche regroupant beaucoup de domaines d'études ainsi que des services commerciales, financiers... En cumulant les 3 sites présents en Belgique, GSK Bio emploi, au mois de Septembre 2009, plus de 6800 personnes.

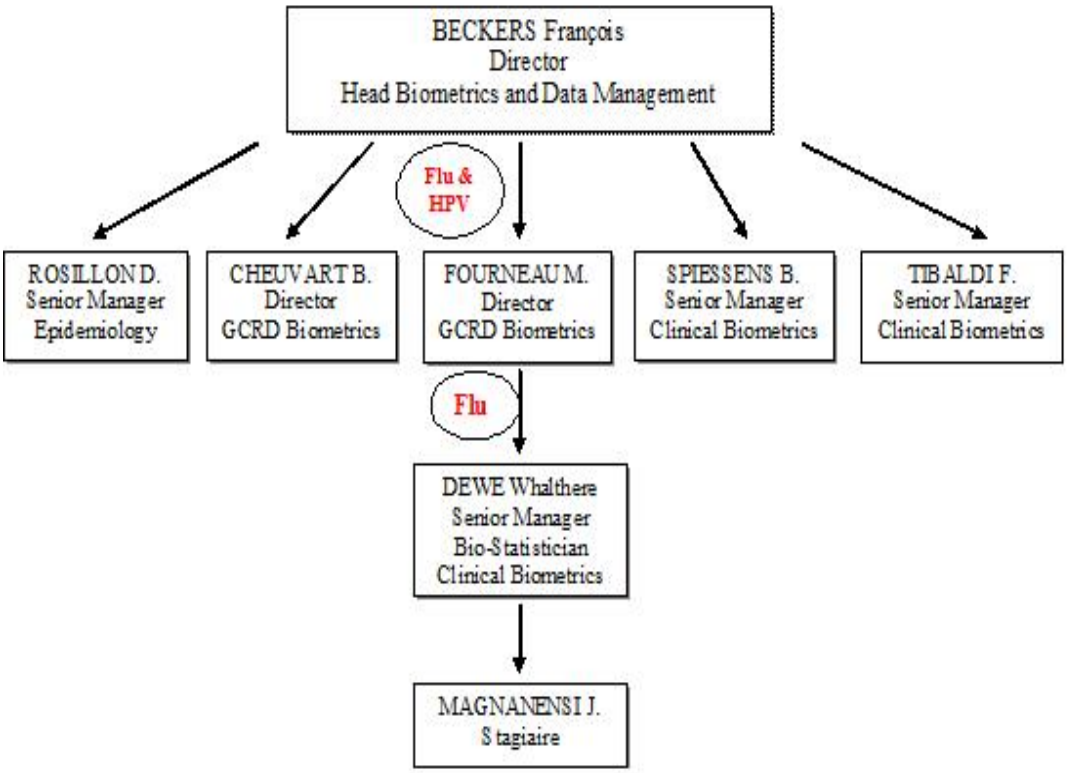
En ce qui concerne le service des statistiques plus précisément, il est constitué de quatre parties majeures :

- Epidémiologie
- Oncologie
- Pédiatrie
- Adult and Early Development

Le service statistique, regroupant ces quatre parties plus d'autres projets, est dirigé par Mr Beckers François.

Pour ma part, je travaillais au sein du groupe Flu (Grippe), groupe qui se trouve sous la direction de Mr Fourneau Marc et managé par mon maître de Stage, Mr Dewe Walthere. Ce groupe travail plus particulièrement sur l'analyse des données cliniques relatives aux vaccins contre les différentes grippes.

Diagramme de l'Organisation du Service Statistique :



III. Introduction

3.1. Développement clinique et questionnement

Le but d'un développement clinique d'un vaccin est de démontrer son efficacité en le comparant soit à un vaccin de référence soit à un placébo. En ce qui concerne la grippe, on sait que trois types principaux de virus circulent chez l'Homme à savoir les types A/H1N1, A/H3N2 et B. Chacun de ces types de virus possède une protéine, l'haemagglutinin (HA), sur son enveloppe. Dans le cadre de la vaccination anti-grippe, le marqueur d'efficacité est un titre appelé HI pour Heamagglutination-inhibition. Ces titres et la façon dont ils sont obtenus seront expliqués et définis dans la section 4.1. Ces mesures, une fois relevées sur l'ensemble des sujets participant à l'étude, sont analysées statistiquement à l'aide d'une ANCOVA, qui elle sera détaillée et définie dans la section 4.2. Une fois les résultats obtenus, on conclut à une efficacité significative ou non du vaccin testé à l'aide de moyennes géométriques des titres, appelés GMT, et de leurs rapports par groupe, appelés GMTRatio. Cependant, certaines hypothèses relatives à la méthode de l'ANCOVA et qui garantissent la régularité des résultats ne sont pas vérifiées, ou simplement supposées, et les conséquences sur les résultats obtenus restent inconnues.

3.2. Objectif du stage

3.2.1. Objectif

Malgré le fait que les méthodes d'analyse statistique actuellement utilisées soient validées, GSK s'est donc posé la question de savoir si ces approximations avaient un réel impact sur les conclusions de ces analyses et s'il était simplement possible d'améliorer ces méthodes. Ce travail devrait déboucher sur une publication en interne pour une modification des macros SAS si de nouvelles méthodes, apportant un gain significatif pour les études, étaient mis à jour et élaborées, suivi d'une publication officielle.

L'objectif de mon stage, qui se place dans un cadre inférentiel d'étude et de recherche, a donc été de revoir entièrement l'analyse de ces titres dans le cadre des données relatives à la vaccination contre la grippe saisonnière. Notons que la seule restriction qui m'a été faite fut de ne pas s'engager dans des modèles ou des méthodes trop compliqués. En effet, s'agissant d'analyses dans le cadre du vaccin grippal saisonnier, ces analyses sont réalisées très fréquemment, et ne doivent donc pas nécessiter un temps d'exécution trop long. Il s'agit donc trouver un compromis entre puissance des nouveaux modèles et complexité de ceux-ci.

Les résultats des travaux de recherche théorique ont été testés à l'aide de simulations réalisées avec le logiciel SAS version9.2 et les conclusions quant à l'efficacité des modèles développés ont été directement tirées des résultats de ces simulations.

3.2.2. Supports de travail

L'ensemble du travail informatique a été effectué à l'aide du logiciel statistique et de gestion de bases de données SAS version9.2. Ce rapport de stage a quant à lui été réalisé avec le logiciel LyX

ainsi que Microsoft Office Word 2007.

Un travail conséquent de recherche de la littérature a été entrepris et la majorité des références présentes dans ce rapport en sont le fruit. Unepartie de ces références est située en Annexe 11.

Enfin, ce sujet avait déjà été abordé lors d'un stage antérieur au sein de GSK. J'ai ainsi pu consulter le rapport de stage correspondant même si les méthodes entreprises ne sont pas identiques. En effet, cette étudiante était restée dans le cadre de l'ANCOVA et a montré qu'avec une analyse réalisée à l'aide d'une ANCOVA avec erreurs hétérogènes, on optimisait clairement la puissance de l'analyse, que l'on diminuait nettement le biais de l'estimateur de la différence entre les deux vaccins et que le précision de l'estimateur de cette différence était bien meilleure qu'avec une ANCOVA classique.

3.2.3. Confidentialité

Toutes les données brutes et résultats d'analyses cliniques relatifs aux études sur lesquelles j'ai pu travailler sont confidentiels et n'apparaissent donc pas dans ce rapport de stage. Seuls les formules mathématiques théoriques et les résultats de ces formules utilisant ces données sont autorisées à être révélés dans ce rapport.

IV. Matériel et méthodes

4.1. Caractérisation des données HI

4.1.1. Description du test HI

Les études considérées ont pour but de comparer la réponse immunitaire entre différents vaccins. Pour ce faire, une prise de sang est effectuée chez les sujets participants à l'étude juste avant la vaccination ainsi que 21 jours après celle-ci (le jour 21 post-vaccination représentant théoriquement le délai nécessaire à l'obtention du pic de la réponse immunitaire). Ces échantillons sont ensuite analysés afin d'obtenir, pour chaque souche vaccinale, une "quantité" d'anticorps présents dans le sang. Le test biologique se base sur une protéine, l'haemagglutinin (HA), présente sur l'enveloppe du virus de la grippe. Cette protéine HA adhère aux globules rouges causant ainsi une agglutination, appelée l'hémaglutination. A partir de l'échantillon de sang de base, on extrait un sérum comportant les anticorps (s'il y en a), sérum qui est dilué de façon successive (premier facteur de dilution 10x associé au seuil de séro-négativité, les suivants 2x) et exposé à l'antigène et aux globules rouges. En fonction de la quantité d'anticorps dans le sérum, ceux-ci inhiberont plus ou moins vite l'hémaglutination et le titre attribué sera le facteur de dilution le plus élevé où il y a inhibition complète.

4.1.2. Caractérisation des données

Ces titres HI prennent ainsi des valeurs discrètes de la forme :

$$HI = 5.2^k \ , \ k \in \mathbb{N}$$

De plus, ces analyses sont effectuées deux fois sur chaque échantillon de sang. On possède ainsi pour chaque jour et pour chaque individus deux titres que l'on ne tolère pas de différer de plus d'un facteur de dilution. Puis ils effectuaient la moyenne géométrique (MG) de ces deux titres obtenus et cette moyenne apparaissait ainsi comme titre unique dans la base de données.

Enfin, dans le cadre de ce stage et afin de simplifier nos recherches, nous avons décidé de se limiter à des études comportant deux groupes et deux temps. Ceci avec pour objectif futur une généralisation à plusieurs groupes et plusieurs temps de relevés.

Définition :

L'ensemble des titres au jour 0 obtenus constitue *la baseline*.

Sujet	Jour	MG(HI)
1	0	...
1	21	...
2	0	...
2	21	...
...	...	...
n	0	...
n	21	...

Fig1 : Exemple type de base de données

4.1.3. Données censurées

L'une des premières caractéristiques que l'on remarque est que toutes ces données sont censurées par intervalle. En effet si HI=5, en réalité cela équivaut à savoir que HI$\in [0;10[$ et si $HI = 5.2^n$ avec $n \geq 1$, en réalité cela veut dire que $HI \in \left[5.2^n; 5.2^{n+1}\right[$. De plus, on est également en présence d'une censure à droite dans le cas où toutes les dilutions successives possibles n'ont pas réussi à complètement stopper l'inhibition de ce phénomène d'agglutination. Dans ce cas là, le titre alloué à cet individu sera le titre maximum testé.

Un deuxième caractéristique réside dans le fait que l'on prenne les moyennes géométriques de deux titres à chaque jour pour chaque individus. En effet, le plus souvent les titres sont identiques et donc en prendre la moyenne géométrique ne change rien. Cependant, il arrive que ces deux titres diffèrent, d'au maximum un facteur de dilution puisque l'on n'autorise pas plus. Ainsi en faisant la moyenne géométrique de ces valeurs, on en obtient une nouvelle mais qui sera présente en moindre mesure dans la base de données. Ainsi, il est compliqué d'ajuster une courbe à un histogramme des données par exemple, puisque celui-ci va osciller en permanence ou encore cela va fausser les résultats des tests d'adéquation à une loi pour la même raison.

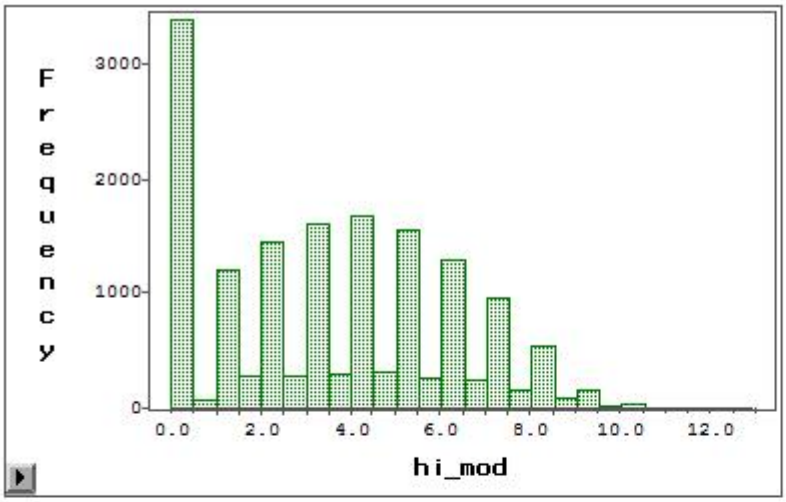


Fig2 : Histogramme représentant l'oscillation des données.

4.1.4. Excès potentiel de séro-négatifs

En traçant un histogramme des données, on se rend très vite compte qu'une caractéristique gênante de ces données va souvent nous empêcher d'y ajuster une distribution classique, celle-ci étant une proportion excessive de zéro c'est à dire une proportion excessive de titre HI égal à 5 dans notre cas. Ceci est logique et risque de se reproduire sur les études à venir. En effet, il n'est pas rare que le vaccin, lors de tests, ne réagisse pas suffisamment et que l'on obtienne ainsi des données égales à zéro. De plus, cela va être d'autant plus problématique que nous allons décider d'intégrer dans la variable réponse les titres au jour 0 qui eux sont majoritairement nuls puisque correspondant à des individus sains non encore vaccinés. Ceci restant naturellement à vérifier puisqu'un tel phénomène n'a pu être confirmé sur un grand nombre d'étude. Les explications et modèles retenus afin de tenir compte de ces caractéristiques seront détaillés dans la cinquième partie.

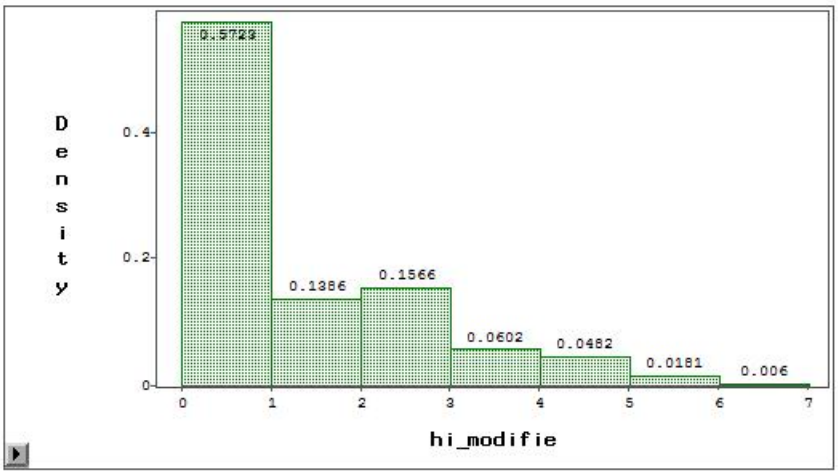


Fig3 : Histogramme type des titres montrant l'excès en zéros

4.2. Méthode d'analyse actuelle

4.2.1. Jeux de données et études

Lors de ce stage, j'ai été amené à étudier des jeux de données suivants un plan bien précis. On s'est focalisé plus principalement sur deux études dont une que nous avons eu le temps de réanalyser. Pour chacune, nous avons sélectionné deux groupes à comparer et pris les données relatives au Jour0 et Jour21 bien que d'autres données correspondantes à d'autres relevés étaient disponibles.

	Groupe 1	Groupe 2
Nombre de sujet	2447	2456

Répartition Effectif Etude 1

	Groupe 1	Groupe 2
Nombre de sujet	166	166

Répartition Effectif Etude 2

Dans la première base de données, on pouvait trouver plusieurs groupes de traitement différents. Nous en avons donc sélectionné que deux, afin de rester cohérent avec le modèle posé sous SAS et de rester dans un cadre le plus simple possible.

La deuxième base de données est celle que nous avons eu le temps de réanalyser et dont les résultats seront décrits à la fin de ce rapport. Précisons que pour réanalyser cette étude, je n'ai pas été mis au courant des résultats attendus et obtenus par GSK afin de ne pas être influencé par ceux-ci.

4.2.2. Analyse actuelle des données

L'analyse statistique de ces données se faisait à l'aide d'une ANCOVA sans interaction. Le modèle utilisé était le suivant :

$$Y_{ij} = \mu + \alpha_j + \beta.X_{ij} + \varepsilon_{ij}$$

avec :
Y_{ij} le titre au jour 21 de l'individu i ayant reçu le j^{eme} traitement
α_j l'effet fixe du j^{eme} traitement
X_{ij} le titre au jour 0 de l'individu i ayant reçu le j^{eme} traitement
ε_{ij} l'erreur résiduelle

et comme contraintes que : $\varepsilon_{ij} \sim N(O, \sigma^2)$; $\alpha_1 = 0$

L'analyse était effectuée de façon gaussienne, en supposant la log-normalité des titres HI. Ainsi, ils procédaient à une ANCOVA classique sur les $log_{10}(HI)$.

4.2.3. Hypothèses et problèmes liés à cette analyse

Pour effectuer une ANCOVA et pouvoir se fier aux résultats obtenus, plusieurs conditions doivent être remplies. Cependant certaines de ces conditions ne sont pas vérifiées ou sont simplement supposées. Certains problèmes et questionnements liés à ce type d'analyse se sont donc posés à moi, dû au type de données que l'on tente d'analyser :

1. S'agissant de données discrètes, peut-on réellement supposer une distribution continue sur ces mêmes titres ?
2. Comment traiter les problèmes de censure autant sur la variable réponse que sur la covariable X_{ij} qui se retrouve ainsi entachée d'une erreur de mesure ce qui n'est habituellement pas toléré dans ce type de modèle.
3. Un examen graphique montre l'hétéroscédasticité des résidus. En effet, plus le nombre d'anticorps présent au jour 0 (Baseline) est élevé, plus les résidus, en valeur absolue, diminuent.
4. Est-ce raisonnable de supposer le parallélisme des droites de régression entre les différents traitements ? Autrement dit, serait-il utile d'apporter dans le modèle une interaction entre la baseline et l'effet groupe ie poser le même modèle avec β_j au lieu de β ?
5. Envisager d'autres méthodes d'analyse qu'une ANCOVA ?

4.3. Orientation choisie

Après réflexion sur les différentes possibilités d'aborder le problème, on a décidé de tenter d'adapter une autre méthode statistique et de sortir de ce cadre de l'ANCOVA. On s'est ainsi dirigé vers des modèles linéaires généralisés mixtes à mesures répétées afin d'analyser ces données, en s'inspirant principalement des travaux de J.Nauta [1] et de Kaifeng Lu [2], le premier traitant de l'analyse de ce type de données et le deuxième introduisant la notion des modèles cLDA que j'expliquerai dans la section 5.1.2.

1. Jos NAUTA. Statistics in Clinical Vaccine Trials, Springer, October 25, 2010
2. Kaifeng Lu. On efficiency of Constrained Longitudinal Data Analysis versus Longitudinal Analysis of Covariance, Biometrics 66, 891-896, September 2010

V. Analyse des titres de base

5.1. Base de travail

5.1.1. Modification de la base de données

Afin de pouvoir analyser correctement les données issues des analyses cliniques, deux transformations des titres sont considérées :

1. La première a été de séparer le titre d'un individu par jour en deux afin de retrouver les deux titres de base ayant servis à la confection de la moyenne géométrique, valeur affichée dans notre base de donnée brute. L'algorithme, assez simple, fut le suivant :

Si $2.\left(\frac{ln(HI)-ln(5)}{ln(2)}\right) \equiv 0[2]$ alors $HI_1 = HI_2 = HI$

Sinon $\begin{cases} HI_1 = \frac{HI}{\sqrt{2}} \\ HI_2 = \sqrt{2}.HI \end{cases}$

Le calcul permettant de trouver cet algorithme est donné en annexe (Voir Annexe1).

On se retrouve ainsi avec deux titres, par jour et par individus, de deux formes possibles :

$$\begin{cases} HI_1 = 5.2^k \\ HI_2 = 5.2^k \end{cases} \quad \text{ou} \quad \begin{cases} HI_1 = 5.2^k \\ HI_2 = 5.2^{k+1} \end{cases}$$

2. La deuxième modification a été de créer une nouvelle variable, issue des HI_1 et HI_2, afin d'obtenir une variable de comptage allant de 0 à n correspondant au facteur de dilution associé au titre. Notons que cette transformation est utilisée par J.Nauta dans son livre "Statistics in Clinical Vaccine Trials". On a donc ajouté à notre base de données la variable suivante :

$$Y = log_2\left(\frac{HI_i}{5}\right) = k$$

Cela va nous permettre de pouvoir poser par la suite une distribution discrète de comptage par exemple et analyser en réalité Y et non plus les HI, sans pour autant fausser les résultats.

Sujet	Jour	Mesure	Y
1	0	1	...
1	0	2	...
1	21	1	...
1	21	2	...
2	0	1	...
2	0	2	...
2	21	1	...
2	21	2	...
...	...	...	...
n	0	1	...
n	0	2	...
n	21	1	...
n	21	2	...

Fig4 : Exemple type d'une base de données modifiée

Après lecture d'un certain nombre de publications et articles (voir Annexe 11), deux méthodes nous ont paru particulièrement intéressantes et adaptables à nos données. Nous nous sommes donc fortement inspirés de ces types de modèles afin de contrôler les interrogations que l'on se posait avec l'analyse des données faite par une ANCOVA.

5.1.2. cLDA model

Les Constrained Longitudinal Data Analysis (cLDA) models donnent une alternative intéressante à l'analyse de Covariance classique. Cette méthode est en réalité très utilisée lorsque la valeur de baseline influe sur la probabilité d'avoir un sujet censuré ou exclu de l'étude et ainsi amener un biais dans les résultats si cette baseline est prise comme covariable. Par exemple, si le fait d'avoir une baseline élevée augmente la probabilité pour le sujet de ne pas se représenter aux rendez-vous futurs et ainsi de sortir de l'étude, puisqu'il faut pour un sujet au moins une valeur en baseline et une valeur post-baseline, alors on obtiendra un biais puisque l'on va sous-estimer les valeurs de baseline. Ainsi, il a été proposé d'inclure cette baseline dans le vecteur réponse en imposant comme seule condition que la moyenne de cette baseline soit la même pour chaque groupe ce qui est le cas dans nos études puisque les sujets sont répartis aléatoirement parmi un échantillon d'individus sains. Il n'y a donc pas de raison, théoriquement, de croire qu'un groupe possèdera une baseline plus élevé qu'un autre. Ceci permettant ainsi d'inclure tous les sujets dans l'étude, et dans notre cas, d'éviter les erreurs de mesure liées à la censure sur une covariable.

Un théorème a été démontré, nous assurant que le fait d'adopter ce type de modèle ne peut qu'améliorer les résultats.

Théorème :

Le modèle cLDA est plus efficace que le modèle ANCOVA dans l'estimation des différences en post-baseline, i.e.,

$$Var_{cLDA}(\hat{\eta}_1 - \hat{\eta}_2) \leq Var_{ANCOVA}(\hat{\beta}_1 - \hat{\beta}_2)$$

avec : η_j la moyenne des titres post-baseline pour le groupe j
β_j l'effet groupe du modèle ANCOVA associé. (équivalent à $\mu + \alpha_j$ du modèle décrit en 4.2.2.)

Démonstration :

La démonstration est disponible en ligne à l'adresse http ://www.biometrics.tibs.org dans la section Paper Information.

Remarque :

1. On a que $\beta_j = \eta_j + \alpha.\gamma_0$, où γ_0 est la moyenne de la baseline supposée identique pour chaque groupe.
2. Les deux modèles sont de performance identique lorsque nous sommes en présence d'un plan où aucune valeur n'est manquante. Cependant, dans notre cadre, il sera toujours préférable d'adopter le cLDA modèle afin d'éviter d'avoir une covariable entachée d'erreurs de mesure.
3. Comme mentionné dans la section 4.3, tous ces résultats sont extraits d'une publication de Keifang Lu [3].

5.1.3. Zero inflated model

Un des problèmes liés à ces données, d'autant plus présent lorsque l'on ajoute les données liées au jour 0 dans la variable réponse, est que l'on se retrouve avec une forte proportion de zéro comme expliqué ci-dessus. L'idée est donc de se dire qu'il ne faut pas étudier ces données en un seul bloc mais plutôt les voir comme appartenant à deux groupes différents. Ainsi, la loi de probabilité P correspondante à ces modèles n'est en fait qu'un mélange de loi entre une mesure de dirac en 0 et une loi, de densité notée f, connue de préférence. Ainsi les modèles modifiés en zéro relève d'un mélange de loi et s'écrivent donc, de façon générale :

$$\mathrm{P(Y{=}y)} = \begin{cases} p_0 + (1-p_0).f(0) & ,y = 0 \\ (1-p_0).f(y) & ,y > 0 \end{cases}$$

Il existe alors plusieurs méthodes afin d'estimer les différents paramètres de ce type de distribution. La méthode retenue sera détaillée au paragraphe suivant.

3. Kaifeng Lu. On efficiency of Constrained Longitudinal Data Analysis versus Longitudinal Analysis of Covariance, Biometrics 66, 891-896, September 2010

5.2. Elaboration du nouveau modèle

5.2.1. Nouveau modèle retenu

A partir des deux méthodes étudiées ci-dessus, on a donc décidé de retenir deux distributions, l'une discrète et l'autre continue, basées sur un modèle linéaire généralisé à effets mixtes et à mesures répétées unique. L'idée était ensuite de comparer ces deux méthodes, à l'aide de simulations, et de déterminer laquelle de ces deux distributions proposées est la meilleure. Ces études et les résultats sont décris dans la prochaine partie.

En suivant l'idée des modèles cLDA, on a donc décidé d'inclure les données au jour 0 dans la variable réponse ainsi qu'un effet jour dans le modèle. Voici donc le nouveau modèle retenu :

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \delta_i + (\delta\beta)_{ik} + \varepsilon_{ijkl}$$

avec :
Y_{ijkl} le titre lié à l'individu i appartenant au groupe k, au jour j dans sa l^{eme}mesure
α_j l'effet fixe du j^{eme} groupe
β_k l'effet fixe du k^{eme} jour
$(\alpha\beta)_{jk}$ l'interaction entre le j^{eme} groupe et le k^{eme} jour
δ_i l'effet aléatoire du i^{eme} sujet
$(\delta\beta)_{ik}$ l'effet aléatoire de l'interaction entre le k^{eme} jour et le i^{eme} sujet
ε_{ijkl} l'erreur résiduelle

Les contraintes imposées sur ce modèle sont les suivantes :
$\alpha_2 = 0$; $\beta_2 = 0$; $(\alpha\beta)_{21} = (\alpha\beta)_{12} = (\alpha\beta)_{22} = 0$; $\delta_i \sim \mathrm{N}(0,\sigma^2_\delta)$; $(\delta\beta)_{ik} \sim \mathrm{N}(0,\sigma^{2^{(k)}}_{\delta\beta})$; $\varepsilon_{ijkl} \sim \mathrm{N}(0,\sigma^2)$

Remarque :

La variance de l'effet aléatoire sujet*jour est différente au Jour0 et Jour21. En effet, comme expliqué lors de l'introduction, on est en face d'un phénomène de saturation, ce qui implique une variabilité des titres non nécessairement égale au Jour0 et au Jour21.

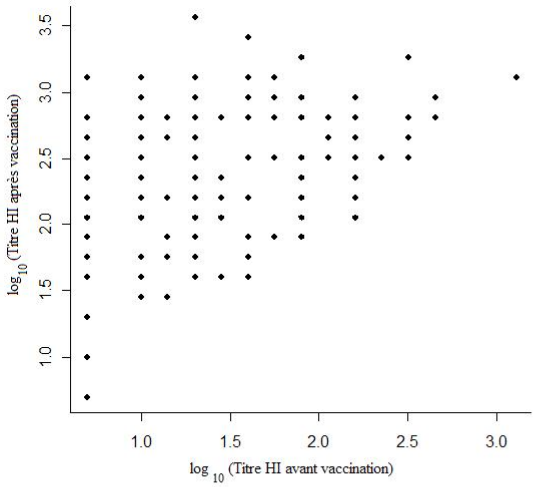


Fig5 : Réponse immunitaire en fonction de la baseline, effet de saturation.

Ce modèle va donc nous permettre d'obtenir des estimations des paramètres de nos vraisemblances pour chaque jour et chaque groupe grâce aux effets fixes présents dans ce modèle. L'effet aléatoire sujet va permettre de tenir compte de la variabilité liée à la répétition sur les jours et l'effet aléatoire représentant l'interaction jour-sujet va permettre de tenir compte de la variabilité liée aux répétitions au sein d'un même jour. On a tout de même décidé de vérifier la structure de la matrice de variance-covariance pour chacune des deux distributions afin de s'assurer de la qualité et la concordance du modèle posé avec notre plan d'étude. Le calcul a été effectué pour les deux distributions étudiées, explicitées ci après. Ces développements sont situés en Annexe (voir Annexe2), et les résultats sont donnés ci-dessous pour chacune des distributions.

Pour effectuer ces calculs, on a posé le modèle sous forme matricielle :

$$Y = X\beta + ZU + \varepsilon$$

Le détail des matrices est donné en Annexe (voir Annexe2), dans le développement des matrices de variance-covariance. Cette forme va également servir à l'écriture des estimateurs des différents paramètres des vraisemblances décrites ci-dessous.

Remarque :

On se situe en réalité dans un modèle à données longitudinales. On a ainsi choisi de poser un modèle dit "subject-specific" c'est-à-dire en intégrant des effets aléatoires dans le modèle afin de modéliser les sources de corrélation supposées. Un autre type de modèle, dit "population averaged", aurait consisté à ne pas mettre d'effets aléatoires dans le modèle mais d'imposer une structure à la working correlation matrix. Dans le type de modèle que l'on a choisi à savoir "subject-specific", il est important de noter que l'interprétation "effets fixes influent sur le niveau moyen de Y" et "effets aléatoires influent sur la dispersion de Y" ne tient plus à cause de la relation induite entre espérance et variance pour toute loi autre que la loi gaussienne.

5.2.2. Cas discret

On a tout d'abord pensé à étudier ces données à l'aide d'une distribution discrète, ceci permettant de traiter les données brutes directement sans considérer les effets de censure à l'exception de la censure à droite pour les titres les plus élevés. En effet, lorsque y=max(Y), cela peut vouloir dire qu'en réalité le test n'a pas abouti et que techniquement il n'était pas possible de diluer plus les échantillons par manque de matériel. Le titre rendu n'est donc à ce moment là qu'une borne inférieure. On peut donc reformuler la probabilité que Y=max(Y) sous la forme suivante :

Soit A_i l'évènement Y=i alors,

$$\mathrm{P}(\mathrm{Y} = \max(\mathrm{Y})) = \mathrm{P}(\bigcup_{i=y}^{+\infty} A_i) = \sum_{i=y}^{+\infty} P(A_i)$$

Ensuite, en ce qui concerne la loi, on a décidé d'opter pour la loi Binomiale Négative BN(k,p), de moyenne $\lambda = k.\frac{1-p}{p}$ et de variance $\sigma^2 = k.\frac{1-p}{p^2} = \frac{\lambda}{p}$. En pratique, pour une variable de comptage, la loi la plus fréquemment utilisée est la loi de Poisson. Cependant, cela implique, d'après ses

propriétés, que la moyenne de nos données doit être égale à leur variance or ceci n'est en réalité jamais le cas sur ce type de données. Il est donc préférable d'opter pour la loi binomiale négative qui peut ainsi être vu comme une généralisation de la loi de Poisson. On se situe donc dans le cadre d'un modèle linéaire généralisé mixte et comme fonction de lien j'ai choisi la fonction log, en accord avec mon responsable de stage. De plus, afin de traiter convenablement le problème des zéros en excès, on a adapté cette distribution pour obtenir un zero inflated model. Voici la distribution choisie :

$$\mathrm{P(Y{=}y)} = \begin{cases} p_0 + \frac{(1-p_0)}{(1+\frac{\lambda}{k})^k} & , y = 0 \\ (1-p_0).\left[\frac{\lambda^y}{y!}.\frac{\Gamma(y+k)}{\Gamma(k).(\lambda+k)^y}.(1+\frac{\lambda}{k})^{-k}\right] & , y > 0 \\ (1-p_0).\sum\limits_{i=y}^{+\infty}\left[\frac{\lambda^i}{i!}.\frac{\Gamma(i+k)}{\Gamma(k).(\lambda+k)^i}.(1+\frac{\lambda}{k})^{-k}\right] & , y = max(Y) \end{cases}$$

En ce qui concerne l'estimation des paramètres de cette vraisemblance, je me suis inspiré du travail et d'une présentation de Fabian Tibaldi basé sur le travail de plusieurs statisticiens [4]. Ainsi, afin d'estimer p_0, on va utiliser une fonction logistique. En effet, ces fonctions sont particulièrement adaptées à la modélisation de probabilité puisqu'elles prennent leurs valeurs entre 0 et 1 selon une courbe en S.

$$p_0 = \frac{exp(\mu+\alpha_j+\beta_k+(\alpha\beta)_{jk}+\delta_i+(\beta\delta)_{ik})}{1+esp(\mu+\alpha_j+\beta_k+(\alpha\beta)_{jk}+\delta_i+(\beta\delta)_{ik})}$$

Ensuite, on va estimer λ à l'aide de la fonction de lien. En effet, en ayant choisi la fonction log comme fonction de lien, on obtient comme égalité :

$$log(E(Y \mid X,\, Z,\, U)) = X\beta + ZU$$

$$\Rightarrow E(Y \mid X,\, Z,\, U) = exp(X\beta + ZU)$$

Or $\lambda = E(Y \mid X,\, Z,\, U)$, donc on obtient comme égalité nous servant à l'estimation de λ :

$$\lambda = exp(X\beta + ZU)$$

Puis on estime k de la même façon, en accord avec l'article de Mr Tibaldi et mon responsable de stage, et on pourra ainsi en déduire une estimation de p puisque l'on a l'égalité suivante :

$$p = \frac{k}{k+\lambda}$$

4. Peter L. Bonate, Crystal Sung, Karen Welch and Susan Richard ; Conditional modeling of Antibody titers using a zero-inflated random effects model : Application to Fabrazyme, Journal of Pharmacokinetics and Pharmacodynamics, 2009, 36 :443-459.

Cette forme parait bien adaptée puisqu'en développant, on retrouve une fonction logistique qui est donc tout à fait convenable pour la modélisation de cette probabilité. En effet, posons :

$\lambda = exp(X\beta_\lambda + ZU_\lambda)\, et\, k = exp(X\beta_k + ZU_k)$

Afin de simplifier le développement, posons $t = X\beta_\lambda + ZU_\lambda$ et $v = X\beta_k + ZU_k$

$$\begin{aligned} \text{D'où } p &= \frac{exp(v)}{exp(v)+exp(t)} \\ &= \frac{1}{exp(t)} \cdot \frac{exp(v)}{1+\frac{exp(v)}{exp(t)}} \\ &= \frac{exp(v-t)}{1+exp(v-t)} \end{aligned}$$

On reconnait bien ici une forme logistique.

Remarque :

Afin de déterminer la matrice de variance-covariance qui suit, on a effectué les calculs (Voir Annexe 2) en supposant que les variances liées aux effets aléatoires sujet*jour étaient égaux que l'on soit au jour0 ou 21 et ceci dans un but de simplification de l'écriture. La matrice propre à nos données est donc quelque peu différente puisque les coefficients a et b ne sont en réalité pas les mêmes entre ceux du Jour0 et ceux du Jour21. Cependant, cela suffit à démontrer que l'on a bien la structure de corrélation que l'on cherchait à mettre en évidence.

La matrice de variance-covariance obtenue est de la forme suivante :

$$Var(Y \mid X,\, Z) = \begin{bmatrix} a & b & c & c & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ b & a & c & c & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ c & c & a & b & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ c & c & b & a & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \end{bmatrix}$$

avec :

$$\begin{aligned} \text{a} &= Var(Y_i \mid X_i,\, Z_i), \text{ avec } Y_i = Y_{ijkl},\, X_i = X_{ijkl},\, Z_i = Z_{ijkl}\ \forall\{i,j,k,l\} \\ &= exp\left(2X_i\beta_{\lambda_i}\right).\left(exp\left(2.\left(Z_i^{(1)^2}.\sigma^2_{\lambda_i,\delta} + Z_i^{(2)^2}.\sigma^2_{\lambda_i,\delta\beta}\right)\right) - exp\left(Z_i^{(1)^2}.\sigma^2_{\lambda_i,\delta} + Z_i^{(2)^2}.\sigma^2_{\lambda_i,\delta\beta}\right)\right) \\ &\quad + exp\left(X_i(\beta_{\lambda_i} - \beta_p)\right).exp\left(\tfrac{1}{2}.\left(Z_i^{(1)^2}.(\sigma^2_{\lambda_i,\delta} - \sigma^2_{p,\delta}) + Z_i^{(2)^2}.(\sigma^2_{\lambda_i,\delta\beta} - \sigma^2_{p,\delta\beta})\right)\right) \\ &\quad + exp(X_i\beta_{\lambda_i}).exp\left(\tfrac{1}{2}.\left(Z_i^{(1)^2}.\sigma^2_{\lambda_i,\delta} + Z_i^{(2)^2}.\sigma^2_{\lambda_i,\delta\beta}\right)\right) \end{aligned}$$

$$
\begin{aligned}
\text{b} &= \text{Cov}(Y_{ijkl}, Y_{ijkm} \mid X_{ijkl},\, X_{ijkm},\, Z_{ijkl},\, Z_{ijkm}) \\
&= exp\left((X_{ijkl} + X_{ijkm}).\beta\right).\Big[exp\left(\tfrac{1}{2}.\left((Z^{(1)}_{ijkl} + Z^{(1)}_{ijkm})^2.\sigma^2_\delta + (Z^{(2)}_{ijkl} + Z^{(2)}_{ijkm})^2.\sigma^2_{\delta\beta}\right)\right) \\
&\qquad -exp\left(\tfrac{1}{2}.\left((Z^{(1)^2}_{ijkl} + Z^{(1)^2}_{ijkm}).\sigma^2_\delta + (Z^{(2)^2}_{ijkl} + Z^{(2)^2}_{ijkm}).\sigma^2_{\delta\beta}\right)\right)\Big],\ l \neq m
\end{aligned}
$$

$$
\begin{aligned}
\text{c} &= \text{Cov}(Y_{ijkl}, Y_{ijqm} \mid X_{ijkl},\, X_{ijqm},\, Z_{ijkl},\, Z_{ijqm}) \\
&= exp\left((X_{ijkl} + X_{ijqm}).\beta\right).\Big[exp\left(\tfrac{1}{2}.\left((Z^{(1)}_{ijkl} + Z^{(1)}_{ijqm})^2.\sigma^2_\delta + (Z^{(2)}_{ijkl} + Z^{(2)}_{ijqm})^2.\sigma^2_{\delta\beta}\right)\right) \\
&\qquad -exp\left(\tfrac{1}{2}.\left((Z^{(1)^2}_{ijkl} + Z^{(1)^2}_{ijqm}).\sigma^2_\delta + (Z^{(2)^2}_{ijkl} + Z^{(2)^2}_{ijqm}).\sigma^2_{\delta\beta}\right)\right)\Big],\ k \neq q,\ l, m \in \{1, 2\}
\end{aligned}
$$

Cette matrice, diagonale par bloc, où chaque bloc représente les liens intra-sujet, correspond bien aux liens que l'on voulait mettre en évidence au sein d'un même individu (représenté en bleu dans la matrice).

Tous les détails de ces formules sont situés en Annexe (voir Annexe2).

5.2.3. Cas continu

Comme je l'ai déjà expliqué, tous les titres que nous avons sont en réalité censurés. Si l'on avait les moyens techniques de déterminer exactement le taux d'anticorps présent dans le sang suite à la vaccination, cela parait évident qu'il s'agirait alors d'une variable continue. On a donc également pensé à la possibilité d'adapter un modèle de zéro en excès avec une loi Normal $\text{N}(\mu, \sigma^2)$. Pour que cela ait un sens, il a fallu adapter cette vraisemblance de telle sorte à tenir compte des différentes censures. Le choix de la loi Normale s'est imposé de lui-même puisqu'il est plus que courant dans le milieu de supposer la log-normalité des titres. Or, à travers la transformation des données effectuée, on a divisé les titres bruts par 5 puis on a pris le log en base deux de ces valeurs. On peut donc supposer qu'ils suivent une loi normale bien que les tests classiques de normalité ne le confirmeront pas puisque travaillant sur une base de données à valeurs discrètes. Nous avons donc procédé un examen graphique afin de vérifier si cette hypothèse pouvait être cohérente, et il s'avère, comme attendu, que c'est bien le cas (voir Annexe 5). Ceci ayant été vérifié, voici la vraisemblance proposée :

$$
\text{P(Y=y)} = \begin{cases}
p_0 + (1-p_0).\int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx & ,y = 0 \\
(1-p_0).\int_{y}^{y+1} \frac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx & ,y > 0 \\
(1-p_0).\int_{y}^{+\infty} \frac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx & ,y = max(Y)
\end{cases}
$$

Les bornes des intégrales ont été choisies de telle sorte à tenir compte de la censure des titres modifiés. En effet, voici le résumé des censures sur ces titres :

$Y = 0 \Leftrightarrow HI = 5$ signifie que $Y \in]-\infty,\, 1[\Leftrightarrow HI \in]0, 10[$
$Y = y \Leftrightarrow HI = 5.2^y$ signifie que $Y \in [y,\, y+1[\Leftrightarrow HI \in \left[5.2^y,\, 5.2^{y+1}\right[$
$Y = max(Y) \Leftrightarrow HI = 5.2^{max(Y)}$ signifie que $Y \in [max(Y), +\infty[\Leftrightarrow \left[5.2^{max(Y)},\, +\infty\right[$

En ce qui concerne les estimations des différents paramètres, p_0 sera estimé de la même façon que dans le cas discret, à savoir à l'aide d'une fonction logistique dépendant du modèle, et pour les

autres, la fonction de lien étant l'identité, les estimations s'effectuent de façon classique à savoir que l'on va expliquer la moyenne et le paramètre de dispersion σ^2 à l'aide de l'ensemble des effets aléatoires et fixes.

Remarque :

De même que dans le cas discret, afin de déterminer la matrice de variance-covariance qui suit, on a effectué les calculs (Voir Annexe 2) en supposant que les variances liées aux effets aléatoires sujet*jour étaient égaux que l'on soit au Jour0 ou 21 et ceci dans un but de simplification de l'écriture. La matrice propre à nos données sont donc quelque peu différentes puisque le terme $\sigma^2_{\delta\beta}$ n'est en réalité pas le même au Jour0 et au Jour21. Cependant, cela suffit à démontrer que l'on a bien la structure de corrélation que l'on cherchait à mettre en évidence.

Voici la forme de la matrice de variance-covariance liée à cette distribution (voir Annexe2 pour le développement) :

$$\begin{bmatrix}
\sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta & \sigma^2_\delta & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta & \sigma^2_\delta & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
\sigma^2_\delta & \sigma^2_\delta & \sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta+\sigma^2_{\delta\beta} & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
\sigma^2_\delta & \sigma^2_\delta & \sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
. & . & . & . & . & . & . & . & . & . & . & . & . \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & *
\end{bmatrix}$$

Cette matrice correspond bien aux liens que l'on voulait mettre en évidence au sein d'un même sujet (représenté en bleu dans la matrice).

Remarque :

Les calculs de détermination des vraisemblances marginales et conditionnelles sont disponibles en Annexe (voir Annexe3).

5.3. Simulation et validation du modèle

5.3.1. Validation du modèle

Le premier objectif a été de valider la structure du modèle mis en place afin de vérifier que les effets aléatoires introduists correspondent bien aux différents liens entre les données que l'on voulait mettre en évidence. Pour cela, nous avons simulé des données suivant une loi normale. On a ainsi simulé quatre données par individu, deux au jour0 et deux au jour21, et ce pour mille sujets. On a décidé de fixer 2 comme moyenne des titres au jour0 et 6 au jour21. On voulait ensuite imposer 2 comme variance de l'effet aléatoire sujet et 3 comme variance de l'effet aléatoire sujet*jour au Jour0 et 1 au Jour21. Puis il a fallu s'assurer que les données générées correspondaient bien à ce que l'on attendait c'est-à-dire que le programme de simulation SAS était correct. Pour cela, on s'est servi de la définition de la corrélation :

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{var(X).var(Y)}}$$

On a ensuite déterminé les différentes variances et covariances théoriques en se servant des formes obtenues dans la matrice de variance-covariance et des paramètres entrés dans la simulation. Puis, en comparant les corrélations théoriques et les corrélations effectivement présentes au sein de l'échantillon simulé, on a obtenu la conclusion que les données simulées étaient correctes et correspondaient bien à ce que l'on voulait analyser. Certains résultats et les codes SAS sont disponibles en Annexe (Voir Annexe 6).

Il s'est ensuite posé la question de savoir quelle procédure SAS nous allions utiliser. En effet, vu le temps restant il paraissait plus raisonnable d'utiliser une procédure existante que de créer une macro spécifique. Etant donné la particularité des lois de probabilité utilisées, notre choix s'est tourné vers la procédure NLMIXED. En effet, cette procédure permet d'entrer à la main une loi non commune et d'utiliser des modèles mixtes. On a donc décidé, même pour cet exemple où l'on a simulé des données non censurées et sans excès en zéro, d'utiliser cette procédure. Le principal problème a été de coder correctement les différents effets aléatoires afin que chacun représente les différents effets introduits. Une forme a été trouvée et ce modèle ainsi écrit renvoie bien ce à quoi on s'attendait c'est à dire aux paramètres fixés à l'avance. Les codes, résultats et sorties SAS sont situés en annexe (Voir Annexe 6).

5.3.2. Validation des distributions

Une fois le modèle validé, il a été question de vérifier le code SAS correspondant aux distributions que l'on a mis théoriquement en place. Avant de passer aux simulations, on a d'abord préféré vérifier que, sur les jeux de données qu'il m'était donné de réanalyser, on était bien en présence d'une surdispersion de la variance. En effet, ceci est indispensable dans le cadre discret puisque l'on se sert d'une distribution binomiale négative et que l'une des caractéristiques fondamentales de cette loi est que la variance se doit d'être plus grande que la moyenne. Or, après modification des titres afin d'obtenir des données de comptage, on a constaté que, dans certains cas, c'est-à-dire pour des données de jours et groupes précis, la variance était bien plus faible que la moyenne. Un tel résultat obtenu à l'aide la procédure UNIVARIATE de SAS est disponible enAnnexe (voir

Annexe 7).La décision a donc été prise d'abandonner le cas discret et de se focaliser sur la méthode utilisant la lognormalité des titres censurés.

Cette fois-ci, aucune condition n'étant à vérifier pour une loi normale, on a décidé de créer les programmes de simulations de données. Pour cela, on a simulé des données discrètes suivant une loi binomiale négative de paramètres permettant que la répartition des titres suive une allure de densité d'une loi normale. Les codes SAS avec les paramètres choisis sont situés enAnnexe (voir Annexe 8). On a procédé en plusieurs étapes progressives dans les simulations. On a tout d'abord simulé des données répétées sur 2 jours mais sans répétitions par jour, sans excès en zéro et avec des résidus normalement distribués. Puis on a intégré les répétitions par jour. Ces étapes se sont bien déroulées et la procédure mise en place fonctionnait très bien. Puis on a décidé de ne plus ajouter de résidus normalement distribués mais d'ajouter à la deuxième répétition au sein de chaque jour une variable suivant une loi de Bernoulli de paramètre 0.05 afin de bien obtenir des titres discrets, comme dans nos bases de données réelles. Cependant, cette modification a conduit à l'impossibilité de faire converger la procédure. Le problème restait naturellement le même lorsque l'on a ajouté des zéros en excès. Après vérification de la base de données obtenue par simulation, on a tout de même décidé d'appliquer cette procédure à des données réelles. Ces données étaient issues de l'une des deux études que nous devions à terme réussir à réanalyser. Le problème fut le même, pas de convergence dans l'estimation des paramètres et il nous a donc été impossible de valider cette distribution.

5.4. Problème et explication

Il nous aura fallu beaucoup de temps avant de conclure au fait qu'il s'agissait d'un problème numérique, inhérent à SAS. En effet, j'ai commencé par lire l'aide SAS sur cette procédure NLMIXED, les techniques d'approximation qu'il utilisait etc...Sans que cela apporte grand chose. Puis, après quelques discussions avec mon responsable de stage et d'autres statisticiens, et de multiples essais de simulations différentes, on a réussi à déterminer l'origine du problème sans pour autant réussir à déterminer comment le résoudre ni même pourquoi il s'agissait d'un problème pour SAS.

En réalité, ce qui posait problème, était la faible valeur de la variance résiduelle combinée à la distribution spécifique que l'on a inséré. En effet, cette variance ne pose aucun problème si l'on ne tient pas compte de la censure et que l'on indique comme loi sur les données une simple loi normale. Or dès que l'on code la distribution tenant compte de la censure, il faut une variance résiduelle minimum afin que la procédure puisse converger. Cette faible variance est dû au fait qu'elle représente, au sein de notre modèle, la variabilité intra sujet*jour. Or au sein d'un même sujet et d'un même jour on sait que la différence entre les deux titres est soit nulle soit égale à 1 et ceci avec une faible probabilité. De plus, en simulant ce type de données à l'aide d'une variable suivant une loi de Bernoulli pour représenter l'écart entre les deux répétitions au sein d'un même jour, on en déduit très rapidement que le maximum de la variance d'une telle variable est obtenu lorsque le paramètre p de cette loi est égal à 0.5 et ainsi ce maximum vaut 0.25. Or ceci est bien trop petit pour arriver à faire converger notre procédure. A nouveau, il nous a fallu nous résigner à oublier cette méthode qui ne pourra donc pas s'appliquer à nos bases de données réelles bien que théoriquement, et ceci est confirmé par simulation si l'on suppose une variabilité résiduelle plus forte, cette méthode marche très bien. Nous avons tenté de déterminer où se situait le problème numérique lors de l'éxécution sous SAS, sans succès. Différents exemples de simulations qui nous ont permis de déterminer ce problème sont situés en Annexe (Voir Annexe 8).

VI. Analyse des moyennes

Etant donné l'échec des précédentes tentatives et ayant déterminé l'origine du problème, nous avons décidé de revenir à un seul point par sujet et par jour, ce point représentant la moyenne géométrique des deux titres obtenus au laboratoire. Ceci a été pensé dans le but d'éviter d'avoir une variance résiduelle trop faible et ainsi permettre à la procédure de converger.

6.1. Modification de la base de données

En ce qui concerne la base de données, l'unique modification par convenance a été d'extraire à nouveau les facteurs de dilution en prenant le log2 des titres HI divisés par 5. En effet, comme on ne se focalise plus que sur le cas continu il n'était pas nécessaire d'extraire des données de comptage, chose que nous avions fait afin de pouvoir appliquer une loi binomiale négative. Il aurait donc été possible de prendre uniquement le logarithme en base 10 de ces titres que l'on suppose suivre une loi log-normale, comme ce qui se fait actuellement chez GSK.

6.2. Modification du modèle

Etant donné la nouveau plan de données que nous avons, il a fallu modifier le modèle et en quelque sorte le simplifier. En effet, nous sommes désormais face à un plan à mesures répétées sur les jours, chaque sujet possédant deux titres, l'un au jour0 et l'autre au jour21. Le modèle choisi a donc été le suivant :

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \delta_i + \varepsilon_{ijk}$$

avec :
Y_{ijk} le titre lié à l'individu i appartenant au groupe k, au jour j
α_j l'effet fixe du j^{eme} groupe
β_k l'effet fixe du k^{eme} jour
$(\alpha\beta)_{jk}$ l'interaction entre le j^{eme} groupe et le k^{eme} jour
δ_i l'effet aléatoire du i^{eme} sujet
ε_{ijk} l'erreur résiduelle

Les contraintes imposées sur ce modèle sont les suivantes :
$\alpha_2 = 0$; $\beta_2 = 0$; $(\alpha\beta)_{21} = (\alpha\beta)_{12} = (\alpha\beta)_{22} = 0$; $\delta_i \sim \mathrm{N}(0,\sigma_\delta^2)$; $\varepsilon_{ijk} \sim \mathrm{N}(0,\sigma_k^2)$

Remarque :

Un point important à souligner réside dans le fait que la variance résiduelle est différente entre le jour0 et jour21 et il a fallu en tenir compte dans l'écriture du programme SAS.

En ce qui concerne la distribution, on a naturellement été obligé de l'adapter au nouveau type de donnée et de censure que l'on a. En effet, plusieurs situations sont possibles pour ces moyennes. S'agissant de moyennes arithmétiques sur les titres modifiés (facteur de dilution), ces moyennes peuvent prendre deux formes. Soit les deux titres à l'origine de cette moyenne sont identiques et égaux à k, auquel cas il s'agira d'un entier naturel k, soit les deux titres diffèrent d'un facteur de dilution et sont donc égaux à k et k+1 respectivement, auquel cas la moyenne sera égale à $k + 1/2$. Or, comme expliqué lors de ce rapport, un titre égal à k équivaut à dire que le titre réel $Y \in [k; k+1[$. Ainsi, on en déduit qu'une moyenne sous la forme d'un entier k appartient en réalité à l'intervalle $[k; k+1[$ et une moyenne sous la forme $k+\frac{1}{2}$ appartient en réalité à l'intervalle $\left[k+\frac{1}{2}; k+\frac{3}{2}\right[$. On a donc adapté la densité, en adaptant les bornes des intégrales de la même façon en finalité que ce que l'on avait fait lors de la première approche.

6.3. Résultats

6.3.1. Validation du nouveau modèle et comparaison par simulation

Afin de valider cette deuxième approche sur les moyennes et la comparer aux autres on a décidé de passer au stade de la simulation de données. A nouveau, nous faisons l'hypothèse de log-normalité de ces moyennes. Ceci restant un point à vérifier. En effet, on suppose au départ la log-normalité des titres c'est à dire la normalité des Y=log(HI). Or on procède à la moyenne entre les deux répétitions par jour et par sujet qui ne sont pas indépendant. La normalité de la moyenne n'est donc pas assuré mathématiquement mais devant le manque de temps, on a décidé de la supposer.

On a décidé de simuler 1000 moyennes, appartenant à un unique groupe, suivant une loi Normale et avec une corrélation entre les données du jour 0 et du jour 21. On a commencé simplement par vérifier le modèle en procédant à une analyse de ces données en posant comme distribution dans la procédure SAS une loi Normale. On a ainsi pu vérifier que les estimations se faisaient très correctement et que l'effet aléatoire introduit représentait bien la covariance entre les données du Jour0 et celles du Jour21.

On a ensuite généré des valeurs suivant une loi Normale centrée et de variance faible, 0.01 dans notre simulation. On a ensuite simplement retiré et ajouté cette valeur à nos moyennes afin d'obtenir deux titres par jour et ce sans changer la moyenne simulée. Le fait d'imposer une variance faible permet de limiter au maximum la probabilité d'avoir deux facteurs de dilution d'écart, ce qui n'est pas toléré au laboratoire. Puis nous avons procédé à la censure telle qu'elle se fait au laboratoire, avant de déduire la moyenne arithmétique des titres censurés obtenus.

On a ensuite procédé aux analyses avec nos 3 méthodes étudiés à savoir l'équivalent de ce qui se faisait jusqu'à présent au sein de GSK, celle que nous avons développé et une dernière méthode qui provient en réalité de travaux effectués par Jos Nauta et disponible dans une publication intitulée "Statistics in Clinical Vaccine Trials" datant de 2010 [5]. Il s'agit, lorsque l'on travaille sur des données censurées par intervalle, de prendre le milieu de l'intervalle comme titre et non la borne inférieur. On a donc décidé d'également coder ce modèle là, à titre informatif, afin de déterminer si notre méthode amène de meilleurs résultats que celle-ci. Le résultat est clair, en voici un récapitulatif :

5. Jos NAUTA. Statistics in Clinical Vaccine Trials, Springer, October 25, 2010

Modèle SAS : (Voir le programme et les sorties SAS en Annexe 9)

$Y \sim N(mu, k)$ avec k l'écart-type résiduel
$mu = b_0 + b_1.t + eta_1$, avec eta_1 représentant l'effet aléatoire sujet et $eta_1 \sim N(0, Var1)$
$k = sqrt(c_0 + c_1.t)$
t=0 au Jour0 et t=1 au Jour21

	Resultat théorique	Methode GSK	Methode Nauta	Methode développée
b_0	2.155	1.746	2.215	2.17
b_1	4.407	4.332	4.448	4.408
c_0	0.687	0.531	0.832	0.668
c_1	1.18	1.552	1.093	1.177
Var_1	1.323	1.197	1.354	1.345

Tableau récapitulatif des sorties SAS

Ce qui donne donc comme estimation finale :

	Resultat théorique	Methode GSK	Methode Nauta	Methode développée
mu	2.155	1.746	2.215	2.17
Var	2.02	1.728	2.186	2.013

Jour 0 avec des données suivant une N(mu,var)

	Resultat théorique	Methode GSK	Methode Nauta	Methode développée
mu	6.562	6.078	6.663	6.578
Var	3.2	3.279	3.279	3.19

Jour 21 avec des données suivant une N(mu,var)

Le résultat est flagrant sur le gain de la nouvelle méthode. On remarque ainsi que le fait de ne pas tenir compte de la censure va naturellement biaiser les résultats qui seront sous estimés, puisque prenant comme titre les différentes bornes inférieures des intervalles auxquelles appartiennent les titres réels. Au contraire, la méthode décrite par Nauta a tendance à surestimer les résultats. Précisons tout de même que cette simulation a été faite sans présence de zéros en excès. Nous avons donc, suite à ça, décidé de réanalyser une étude réelle afin de comparer les résultats obtenus dans le passé par GSK et ceux obtenus avec cette nouvelle méthode.

6.3.2. Réanalyse d'une étude

Comme expliqué précédemment dans ce rapport, il m'a été demandé de réanalyser une étude de façon aveugle. En effet, on m'a uniquement transmis la base de données sans m'indiquer ni la correspondance des groupes avec les vaccins ou placebo, ni les résultats obtenus par GSK lors de l'analyse de cette base. Ceci afin que je ne sois nullement influencé lors de mon analyse avec la

méthode que l'on a développé. Il s'agit de la deuxième étude que l'on a décrite au paragraphe 4.2.1. On a donc transformé et recodé certaines informations afin que la base de données soit compatible avec le code SAS développé. On a décidé d'avancer par étape. On a tout d'abord commencé par une analyse graphique des données par jour et par groupe, à l'aide de SAS/Insight. En effet, cette étape est importante afin de déterminer quand est-ce qu'il est nécessaire d'introduire un paramètre de zéro en excès. Si l'on demande une estimation d'un tel paramètre p_0 alors que celui-ci est en réalité nul, ce qui est souvent le cas pour les données au Jour21, la procédure n'arrive pas à converger. Ceci est dû à la forme logistique sous laquelle nous avons exprimé p_0. Si ce dernier est nul, le numérateur $\exp(a_0)$ va tendre vers 0, autrement dit a_0 va tendre vers -∞ et la procédure ne va donc pas réussir à converger. Ces analyses nous ont indiqué qu'il n'y avait présence de zéro en excès qu'aux données relatives au Jour0. Une autre méthode basée sur une statistique, appelée statistique de Vuong, existe afin de tester la présence ou non de zéro en excès. Cette statistique, développée dans le cadre de la loi binomiale négative et adaptable pour une loi normale, est décrite en Annexe (Voir Annexe 4) mais on ne l'a pas utilisé pour cette analyse. Ensuite, on a recherché le titre maximum présent dans cette base de données. Il s'agissait de 9.5, ceci nous a donc conduit à ne pas introduire de censure à droite sur ces titres. En effet, avoir une moyenne de 9.5 provient du fait d'avoir deux titres de base valant respectivement 9 et 10. Or si 10 était censuré et représentait donc le facteur maximum de dilution que le laboratoire puisse tester sans qu'il n'y ait eu inhibition de la réaction, cela voudrait dire qu'il y'a au moins deux titres d'écart et ils n'auraient donc pas rendu ce résultat. Ce maximum n'est donc pas censuré à droite. Puis on a commencé par analyser groupe par groupe avant de passer à l'analyse de la base de données entière afin de s'assurer qu'aucun problème ne survenait et de nous donner une idée des différents paramètres d'entrée à indiquer dans le procédure NLMIXED pour l'analyse de la base de données complète (Voir le programme et certaines sorties SAS en Annexe 10). Les données étant confidentielles, je ne peux que communiquer les résultats obtenus par le procédure mise en place. Voici ces résultats :

Remarque :

Méthode 1 : Méthode développée avec zéros en excès au Jour 0
Méthode 2 : Méthode de Nauta avec zéros en excès au Jour 0
Méthode 3 : Méthode équivalente GSK avec zéros en excès au Jour 0
Méthode 4 : Méthode équivalente GSK sans zéros en excès au Jour 0

	Méthode 1	Méthode 2	Méthode 3	Méthode 4
p_0	0.4563	0.6014	0.5793	0
mu	1.9118	2.7010	2.0480	0.8802
var résid	2.3605	1.3389	1.3351	1.42
var pid	1.2231	0.7931	0.7903	0.6286

Tab1 : Tableau récapitulatif Jour0 Groupe1

	Méthode 1	Méthode 2	Méthode 3	Méthode 4
p_0	0.2667	0.547	0.5212	0
mu	1.4751	2.7277	2.075	1.0512
var résid	3.0314	1.2625	1.2593	1.4828
var pid	1.2231	0.7931	0.7903	0.6286

Tab2 : Tableau récapitulatif Jour0 Groupe2

	Méthode 1	Méthode 2	Méthode 3	Méthode 4
mu	6.7584	6.8451	6.2599	6.2573
var résid	0.9241	1.3694	1.3645	1.6014
var pid	1.2231	0.7931	0.7903	0.6286

Tab3 : Tableau récapitulatif Jour21 Groupe1

	Méthode 1	Méthode 2	Méthode 3	Méthode 4
mu	6.5571	6.6422	6.0572	6.0572
var résid	1.043	1.6605	1.6412	1.7308
var pid	1.2231	0.7931	0.7903	0.6286

Tab4 : Tableau récapitulatif Jour21 Groupe2

Un point important, permettant une meilleure visualisation de la répartition des données, ressort de ces résultats. Il s'agit de l'importance de l'apport d'un modèle de zéro en excès. En effet, dans cette étude, si l'on compare la méthode 3 et 4 qui sont les même avec et sans le paramètre de zéro en excès, on remarque à l'aide des tableaux 1 et 2 que la répartition des titres est faite à plus de 50% de titres nuls en excès et, pour le reste, de titres suivant une loi normale de moyenne légérement supérieure à 2.

Déterminons maintenant les valeurs des différentes moyennes géométriques des titres initiaux, appelés GMT, ainsi que leur rapport entre Groupe, appelés GMTratio. Pour cela, il suffit de prendre les estimations des moyennes que l'on vient d'obtenir, multipliées par la proportion des données concernées, et leur appliquer la transformation inverse de celle que l'on a appliqué pour obtenir les facteurs de dilution. Cette transformation est donc la suivante :

$$GMT = 5.2^{mu.(1-p_0)}$$

On obtient alors les résultats suivants :

	Méthode 1	Méthode 2	Méthode 3	Méthode 4	Résultats GSK
Jour0 Groupe1	10.2772	10.5454	9.0852	9.2032	$\cdots$
Jour0 Groupe2	10.5826	11.7744	9.9551	10.3612	$\cdots$
Jour21 Groupe1	541.3163	574.8446	383.1666	382.4767	383.4
Jour21 Groupe2	470.8188	499.4263	332.9422	332.9422	332.9

Tab5 : Tableau récapitulatif des GMT

	Méthode 1	Méthode 2	Méthode 3	Méthode 4	Résultats GSK
$Jour0\,Groupe1/Jour0\,Groupe2$	0.9711	0.8956	0.9126	0.8882	$\cdots$
$Jour21\,Groupe1/Jour21\,Groupe2$	1.1497	1.1510	1.1509	1.1488	1.1517

Tab6 : Tableau récapitulatif des GMTRatio

Ce qui intéresse plus particulièrement GSK sont les résultats au Jour21 naturellement. En effet, n'oublions pas que le but d'un développement clinique est de démontrer l'efficacité d'un vaccin et ceci se démontre évidemment à l'aide des données après vaccination, ici celles du Jour21.

Or en ce qui concerne les GMTRatio, on remarque bien, à l'aide tableau 6, que pour les données du Jour21, ils sont préservés quelque soit la méthode utilisée. Donc si l'on se limite à la comparaison entre deux groupes, on peut en conclure, du moins sur cette étude, que quelque soit la méthode utilisée les résultats sont équivalents. La comparaison des GMTRatio suivant les méthodes et leur égalité serait également à tester encore à l'aide de simulation ou tout autre moyen afin de déterminer s'il s'agit là d'un cas particulier ou si l'on peut s'attendre à voir ce rapport conserver sur toutes les études possibles. Ceci dit, les GMT sont réellement différents comme nous l'indique le tableau 5 et cela a également beaucoup d'importance. Et c'est là que la nouvelle méthode prend toute son importance. En effet, ces résultats représentant un ordre de grandeur d'anticorps présent dans le sang, les GMT peuvent s'avérer très important d'un point de vue médical (seuil de protection atteint ou non).

De plus, on remarque à l'aide tu tableau 6, qu'au Jour0 la différence sur les GMTRatio est plus importante. On peut supposer que la présence du paramètre p_0 de zéros en excès en est la cause et que donc, si ce paramètre est une fois amené à être utilisé au sein des données collectées après vaccination, on peut s'imaginer que ces ratios ne seront plus nécessairement préservés. Tout ceci restant naturellement au conditionnel et demandant encore à être démontré.

VII. Discussion et perspective future

7.1. Approfondissement des recherches

Etant donné le temps dont je disposais, plusieurs points seraient encore à vérifier ou à développer, choses que je n'ai eu le temps de faire durant ce stage. Tout d'abord, il faudrait valider entièrement la méthode développée sur les moyennes en procédant à l'analyse des résidus, par exemple, ainsi qu'en vérifiant les différents points soulevés lors de la réanalyse de l'étude réelle. Ensuite, il serait peut-être possible de créer une macro SAS spécifique afin de pouvoir analyser les titres et non les moyennes sans que le problème de variance résiduelle faible ne pose problème. Pour cela, il faudrait tout d'abord déterminer ce qui pose problème au sein de la procédure NLMIXED et qui empêche la convergence dans les estimations des différents paramètres. Enfin, il faudrait établir une généralisation de la méthode à plusieurs jours. En effet, certaines études possèdent plus que deux relevés par individu. En somme, beaucoup de travail reste encore à faire sur ce sujet pour le traiter convenablement.

7.2. Développement futur

En plus des idées citées ci-avant qui sont en lien avec les méthodes développées, il serait intéressant de tenter d'analyser ces données à l'aide d'autres méthodes statistiques. Comme expliqué dans ce rapport, les données sont réparties suivant une loi de probabilité qui en réalité est un mélange de loi. Or il est bien connu que pour traiter les mélanges de lois, l'algorithme EM est une bonne méthode. Il serait donc intéressant d'analyser ce type de données à l'aide de cet algorithme et de comparer les résultats obtenus avec ceux obtenus à l'aide de la méthode que nous avons développé.

Une autre possibilité serait de traiter ces données à l'aide d'une analyse bayésienne. Une personne actuellement en doctorat et qui travaille au sein du même service que moi, a été confrontée au même problème et tente une approche bayésienne. Ne s'agissant pas de son sujet de thèse, il serait peut-être intéressant d'approfondir cette méthode en complétant son travail.

VIII. Epilogue

Revenir au sein de GSK cette année pour ce stage de 6 mois aura été un réel point positif pour moi sur plusieurs plans. Tout d'abord, d'un point de vue personnel, le fait que GSK m'ait à nouveau fait confiance est extrêmement gratifiant. Cela m'a apporté confiance en moi sur mes capacités au sein d'une telle structure.

D'un point de vue professionnel, ce stage m'a permit de découvrir ou redécouvrir plusieurs aspects propre au monde de l'entreprise. Déjà, le fait de travailler sur des données réelles est quelque chose qui complique énormément la tâche comparé à ce que l'on est amené à rencontrer à l'Université. En effet, tout ne se passe pas aussi bien que l'on imagine au début, et tout au long de mon stage, je suis allé de rebondissement en rebondissement. Une fois que j'avais pensé avoir résolu un problème, en voilà un autre, auquel je n'avais pas pensé, qui faisait surface. Cependant, un aspect très intéressant au sein d'une entreprise est la collaboration entre les personnes du même ou d'un différent service. Ceci m'a souvent aidé car cela permet de visualiser le problème auquel on est confronté sous un autre angle. Cela permet de prendre du recul et d'entrevoir une vision neuve du problème et ainsi s'orienter vers une solution à laquelle je n'aurai pas forcément pensé.

Un autre aspect auquel j'ai dû m'adapter est que parfois la théorie n'implique pas forcément la pratique. Je veux dire par là qu'une analyse théorique, bien que correcte, ne débouche pas forcément sur une application pratique qui va fonctionner. Ceci étant tout simplement dû aux limites de l'informatique, il a été compliqué de se dire que parfois il valait mieux revenir à une analyse moins précise mais qui fonctionne plutôt que de s'obstiner à faire fonctionner une méthode trop précise. Dans le monde de l'entreprise, on retrouve un enjeu économique qui n'est pas présent (ou moins) au sein des universités. Ceci vous oblige donc à avancer même si le résultat obtenu n'est pas forcément celui attendu.

D'un point de vue mathématique, j'ai beaucoup appris. En effet, le fait de se poser des questions sur des données si complexes, avec un modèle relativement spécial, m'a fait visualiser beaucoup de choses. La grande différence avec le monde de la faculté, à travers les cours que j'ai pu suivre, est qu'au sein de l'Université, la plus part du temps, tout se passe bien. On ne se pose donc que rarement des questions quant à la signification de ce que peuvent représenter les résidus par exemple. Or, en étudiant un certain type de données pendant six mois, et en tentant d'ajuster des modèles ou de trouver des distributions qui correspondent le mieux possible à ces données, on est amené à devoir réellement visualiser ce que l'on fait et où on veut aller.

Enfin, il s'agissait à mon avis d'un travail qui nécessite bien plus de temps que six mois. En effet, étant donné la complexité des données, de la planification, du nombre conséquent de méthodes différentes qu'il est possible d'imaginer et l'étendue des vérifications qu'il reste à faire, il n'était, je pense, pas possible de couvrir ce sujet intégralement en six mois.

Cela restera en tout cas une excellente expérience pour moi sur tout point de vue et, je le pense, une réelle chance pour mon futur.

Annexes :

Annexe 1 : Développement de l'algorithme de séparation des titres moyens

Les valeurs brutes de nos bases de données représentent un moyenne géométrique de deux titres qui peuvent soit être identiques soit différent d'un facteur de dilution près.

D'où $HI_{brut} = \sqrt{5.2^k.5.2^q}$, q=k ou q=k+1

$= 5.\sqrt{2^{k+q}}$

$= 5.2^{\frac{k+q}{2}}$

$\Rightarrow ln\left(\frac{HI_{brut}}{5}\right) = \frac{k+q}{2}.ln(2)$

$\Rightarrow 2.\left(\frac{ln(HI_{brut})-ln(5)}{ln(2)}\right) = k+q = \begin{cases} 2k & , si\, q = k \\ 2k+1 & , si\, q = k+1 \end{cases}$

Donc si $2.\left(\frac{ln(HI_{brut})-ln(5)}{ln(2)}\right) \equiv 0[2]$ alors q=k ie il s'agit de la moyenne géométrique de titres égaux

sinon $HI_1 = 5.2^k$, $HI_2 = 5.2^{k+1}$ et $HI_{brut} = 5.2^{k+\frac{1}{2}}$, d'où $HI_1 = {}^{HI_{brut}}/\sqrt{2}$ et $HI_2 = HI_{brut}.\sqrt{2}$

Annexe 2 : Développement des matrices de variance-covariance

– Cas continu :

Soient n_1 l'effectif du groupe 1 et n_2 l'effectif du groupe 2. Posons $n = n_1 + n_2$.
Soient :

$$Y = \begin{bmatrix} Y_{1111} \\ Y_{1112} \\ Y_{1121} \\ Y_{1122} \\ Y_{2111} \\ . \\ . \\ Y_{n_1122} \\ Y_{n_1+1,211} \\ Y_{n_1+1,212} \\ . \\ . \\ Y_{n222} \end{bmatrix}, X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ . & . & . & 0 & 1 \\ . & . & . & 1 & 0 \\ . & . & . & . & . \\ . & 1 & 0 & 0 & 1 \\ . & 0 & 1 & 1 & 0 \\ . & 0 & 1 & 1 & 0 \\ . & . & . & 0 & 1 \\ . & . & . & . & . \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & 0 & 0 & . & 0 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 \\ 1 & 0 & 0 & . & 0 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 \\ 1 & 0 & 0 & . & 0 & 0 & 1 & 0 & 0 & 0 & . & 0 & 0 \\ 1 & 0 & 0 & . & 0 & 0 & 1 & 0 & 0 & 0 & . & 0 & 0 \\ 0 & 1 & 0 & . & 0 & 0 & 0 & 1 & 0 & 0 & . & 0 & 0 \\ 0 & 1 & 0 & . & 0 & 0 & 0 & 1 & 0 & 0 & . & 0 & 0 \\ 0 & 1 & 0 & . & 0 & 0 & 0 & 0 & 1 & 0 & . & 0 & 0 \\ 0 & 1 & 0 & . & 0 & 0 & 0 & 0 & 1 & 0 & . & 0 & 0 \\ 0 & 0 & 1 & . & 0 & 0 & 0 & 0 & 0 & 1 & . & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & 1 & 0 & 0 & 0 & 0 & 0 & . & 1 & 0 \\ 0 & 0 & 0 & . & 1 & 0 & 0 & 0 & 0 & 0 & . & 0 & 1 \\ 0 & 0 & 0 & . & 1 & 0 & 0 & 0 & 0 & 0 & . & 0 & 1 \end{bmatrix}, U = \begin{bmatrix} \delta_1 \\ . \\ . \\ \delta_n \\ (\delta\beta)_{11} \\ (\delta\beta)_{12} \\ (\delta\beta)_{21} \\ (\delta\beta)_{22} \\ (\delta\beta)_{31} \\ . \\ . \\ (\delta\beta)_{n2} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_{1111} \\ \varepsilon_{1112} \\ \varepsilon_{1121} \\ \varepsilon_{1122} \\ \varepsilon_{2111} \\ . \\ . \\ \varepsilon_{n_1122} \\ \varepsilon_{n_1+1,211} \\ . \\ . \\ \varepsilon_{n222} \end{bmatrix}$$

Alors notre modèle peut s'écrire sous la forme : Y = X.β + Z.U + ε, avec $U \sim N(0, Var(U))$ et $\varepsilon \sim N(0, \sigma^2)$

Posons $Z^{(1)}$ la partie de la matrice Z liée aux effets aléatoires sujet (en bleue) et $Z^{(2)}$ la partie liée aux effets aléatoires de l'interaction sujet*jour (en verte).

$U^{(1)}$ la partie de la matrice U liée aux effets aléatoires sujet (en bleue) et $U^{(2)}$ la partie liée aux effets aléatoires de l'interaction sujet*jour (en verte).

Alors,

$$\begin{aligned} \text{Var(Y}|\ X,\ Z) &= Var(X.\beta + Z.U + \varepsilon \mid X,\ Z) \\ &= \text{Var(Z.U} + \varepsilon|\ X,\ Z) \\ &= \text{Var(Z.U}|\ Z) + Var(\varepsilon) \\ &= \text{Z.Var(U)}.Z^{'} + \sigma^2.Id \end{aligned}$$

On obtient donc que :

$$\text{Var(Y}|\ X,\ Z) = \text{Z.D.Z} + \sigma^2.Id$$

avec D = Var(U).

Or on suppose tous les effets aléatoires indépendants entre eux. On obtient donc :

$$D = \begin{bmatrix} \sigma^2_\delta.Id_n & 0 \\ 0 & \sigma^2_{\delta\beta}.Id_{2n} \end{bmatrix}$$

Ainsi on obtient que :

$$\begin{aligned} \text{Var(Y}|\ X,\ Z) &= \text{D.Z.Z} + \sigma^2.Id \\ &= \sigma^2_\delta.Z^{(1)}.Z^{(1)'} + \sigma^2_{\delta\beta}.Z^{(2)}.Z^{(2)'} + \sigma^2.Id \end{aligned}$$

Après avoir effectuer les calculs on obtient :

$$Z^{(1)}.Z^{(1)'} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & . & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$Z^{(2)}.Z^{(2)'} = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 1 & 1
\end{bmatrix}$$

D'où on obtient finalement comme matrice de variance-covariance pour Y, dans le cas d'une distribution continue que l'on suppose normale :

$$\begin{bmatrix}
\sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta & \sigma^2_\delta & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta & \sigma^2_\delta & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
\sigma^2_\delta & \sigma^2_\delta & \sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2_\delta+\sigma^2_{\delta\beta} & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
\sigma^2_\delta & \sigma^2_\delta & \sigma^2_\delta+\sigma^2_{\delta\beta} & \sigma^2+\sigma^2_\delta+\sigma^2_{\delta\beta} & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & . & 0 & 0 & 0 & 0 \\
. & . & . & . & . & . & . & . & . & . & . & . & . \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & * & * & * & *
\end{bmatrix}$$

– Cas discret :

Dans ce cadre discret, on a choisi comme fonction de lien g la fonction log et pour simplifier, supposons qu'il n'y a pas de zéro en excès.

Pour simplifier les écritures (hors covariance), posons $Y_i = Y_{ijkl}$, $X_i = X_{ijkl}$, $Z_i = Z_{ijkl}$.

Le modèle conditionnel est défini par les hypothèses suivantes :

- $L(Y_i \mid X_i,\, Z_i,\, U) = BN(k, p)$
- $E(Y_i \mid X_i,\, Z_i,\, U) = k.\frac{(1-p)}{p}$
- $Var(Y_i \mid X_i,\, Z_i,\, U) = k.\frac{(1-p)}{p^2}$
- $log(E(Y_i \mid X_i,\, Z_i,\, U)) = X_i.\beta + Z_i.U$
- Conditionnellement à U, les $(Y_i,\, X_i,\, Z_i)$ sont indépendants.
- $U \sim N(0,\, D) \Rightarrow \begin{cases} E(U) = 0 \\ Var(U) = D = E(U^2) \end{cases}$

Ensuite posons $\lambda_i = E(Y_i \mid X_i,\, Z_i,\, U)$, pour i = 1,...,n.

Alors on a l'égalité suivante :

$$\log(\lambda_i) = X_i.\beta + Z_i.U$$

$$\Rightarrow \lambda_i = exp(X_i.\beta + Z_i.U)$$

De plus, on sait que la variance, dans le cadre d'une loi binomiale négative BN(k,p), est égale à la moyenne divisé par p ie :

$$\mathrm{Var}(Y_i \mid X_i,\, Z_i,\, U) = \frac{\lambda_i}{p}$$

Déterminons désormais le modèle marginal :

- $\begin{aligned} E(Y_i \mid X_i,\, Z_i) &= E_U(E(Y_i \mid X_i,\, Z_i,\, U)) \\ &= E_U(\lambda_i) \\ &= E_U(exp(X_i.\beta + Z_i.U)) \\ &= exp(X_i.\beta).E_U(exp(Z_i.U)) \end{aligned}$

avec $\begin{aligned} E_U(exp(Z_i.U) &= E_U(exp(Z_i^{(1)}.U^{(1)} + Z_i^{(2)}.U^{(2)}) \\ &= E_U(exp(Z_i^{(1)}.U^{(1)}).exp(Z_i^{(2)}.U^{(2)})) \\ &= E_U(exp(Z_i^{(1)}.U^{(1)}).E_U(exp(Z_i^{(2)}.U^{(2)}) \end{aligned}$

par indépendance entre les deux types d'effets aléatoires.

Or $E_U(exp(Z_i^{(1)}.U^{(1)})) = \int_{\mathbb{R}} exp(Z_i^{(1)}.u).f_{U^{(1)}}(u).du$, avec $f_{U^{(1)}}(u)$ la densité d'une N(0, D).

$$\begin{aligned} &= \int_{\mathbb{R}} exp(Z_i^{(1)}.u).\frac{1}{\sqrt{2\pi.\sigma_\delta^2}}.exp\left(-\frac{u^2}{2.\sigma_\delta^2}\right).du \\ &= \int_{\mathbb{R}} exp\left(-\left(-\frac{2.\sigma_\delta^2.Z_i^{(1)}.u}{2.\sigma_\delta^2}\right)\right).\frac{1}{\sqrt{2\pi.\sigma_\delta^2}}.exp\left(-\frac{u^2}{2.\sigma_\delta^2}\right).du \end{aligned}$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi.\sigma_\delta^2}}.exp\left(-\frac{1}{2.\sigma_\delta^2}\left(\left(u - Z_i^{(1)^2}.\sigma_\delta^2\right)^2 - Z_i^{(1)^2}.\sigma_\delta^4\right)\right).du$$
$$= exp\left(-\frac{1}{2.\sigma_\delta^2}.\left(-Z_i^{(1)^2}.\sigma_\delta^4\right)\right).\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi.\sigma_\delta^2}}.exp\left(-\frac{1}{2.\sigma_\delta^2}\left(u - Z_i^{(1)}.\sigma_\delta^2\right)^2\right).du$$
$$= exp\left(\frac{1}{2}.Z_i^{(1)^2}.\sigma_\delta^2\right)$$

car $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi.\sigma_\delta^2}}.exp\left(-\frac{1}{2.\sigma_\delta^2}\left(u - Z_i^{(1)}.\sigma_\delta^2\right)^2\right).du = 1$ en tant qu'intégrale sur $\mathbb{R}$ d'une densité de $N(Z_i^{(1)}.\sigma_\delta^2, \sigma_\delta^2)$.

Le calcul est naturellement le même pour $E_U(exp(Z_i^{(2)}.U^{(2)}))$ et on obtient alors :
$E_U(exp(Z_i^{(2)}.U^{(2)}) = exp\left(\frac{1}{2}.Z_i^{(2)^2}.\sigma_{\delta\beta}^2\right)$

Donc, on obtient finalement que :

$$E(Y_i \mid X_i,\, Z_i) = exp(X_i.\beta).exp\left(\frac{1}{2}.Z_i^{(1)^2}.\sigma_\delta^2\right).exp\left(\frac{1}{2}.Z_i^{(2)^2}.\sigma_{\delta\beta}^2\right)$$
$$= exp\left(X_i.\beta + \frac{1}{2}.\left(Z_i^{(1)^2}.\sigma_\delta^2 + Z_i^{(2)^2}.\sigma_{\delta\beta}^2\right)\right)$$

Passons maintenant au calcul de la variance, en reprenant les notations introduites au 4.2.3.1. pour l'estimation des paramètres :

- $Var(Y_i \mid X_i,\, Z_i) = Var_U(E(Y_i \mid X_i,\, Z_i,\, U)) + E_U(Var(Y_i \mid X_i,\, Z_i,\, U))$
$= Var_U(\lambda_i) + E_U\left({}^{\lambda_i}/_{p}\right)$
$= Var_U\left(exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})\right) + E_U\left(\frac{exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})}{exp(X_i\beta_p + Z_iU_p)}\right)$

Or $Var_U\left(exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})\right) = exp\left(2X_i\beta_{\lambda_i}\right).Var_U\left(exp(Z_iU_{\lambda_i})\right)$
et $Var_U\left(exp(Z_iU_{\lambda_i})\right) = E\left(exp(2Z_iU_{\lambda_i})\right) - E\left(exp(Z_iU_{\lambda_i})\right)^2$
avec $2U \sim N(0, 4D)$

donc $E\left(exp(2Z_iU_{\lambda_i})\right) = exp\left(2.\left(Z_i^{(1)^2}.\sigma_{\lambda_i,\delta}^2 + Z_i^{(2)^2}.\sigma_{\lambda_i,\delta\beta}^2\right)\right)$
et $E\left(exp(Z_iU_{\lambda_i})\right)^2 = exp\left(\frac{1}{2}\left(Z_i^{(1)^2}.\sigma_{\lambda_i,\delta}^2 + Z_i^{(2)^2}.\sigma_{\lambda_i,\delta\beta}^2\right)\right)^2 = exp\left(Z_i^{(1)^2}.\sigma_{\lambda_i,\delta}^2 + Z_i^{(2)^2}.\sigma_{\lambda_i,\delta\beta}^2\right)$

d'où finalement :

$$Var_U\left(exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})\right) = exp\left(2X_i\beta_{\lambda_i}\right).\left(exp\left(2.\left(Z_i^{(1)^2}.\sigma_{\lambda_i,\delta}^2 + Z_i^{(2)^2}.\sigma_{\lambda_i,\delta\beta}^2\right)\right) - exp\left(Z_i^{(1)^2}.\sigma_{\lambda_i,\delta}^2 + Z_i^{(2)^2}.\sigma_{\lambda_i,\delta\beta}^2\right)\right)$$

Il reste maintenant à déterminer $E_U\left(\frac{exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})}{exp(X_i\beta_p + Z_iU_p)}\right)$:

$$E_U\left(\frac{exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})}{exp(X_i\beta_p + Z_iU_p)}\right) = E_U\left(\frac{exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i}) + exp(X_i(\beta_p + \beta_{\lambda_i}) + Z_i(U_p + U_{\lambda_i})}{exp(X_i\beta_p + Z_iU_p)}\right)$$
$$= E_U\left(exp\left(X_i(\beta_{\lambda_i} - \beta_p) + Z_i(U_{\lambda_i} - U_p)\right) + exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})\right)$$
$$= E_U\left(exp\left(X_i(\beta_{\lambda_i} - \beta_p) + Z_i(U_{\lambda_i} - U_p)\right)\right) + E_U\left(exp(X_i\beta_{\lambda_i} + Z_iU_{\lambda_i})\right)$$
$$= exp\left(X_i(\beta_{\lambda_i} - \beta_p)\right).E_U\left(exp\left(Z_i(U_{\lambda_i} - U_p)\right)\right) + exp(X_i\beta_{\lambda_i}).E_U\left(exp(Z_iU_{\lambda_i})\right)$$

$$= exp\left(X_i(\beta_{\lambda_i} - \beta_p)\right).exp\left(\tfrac{1}{2}.\left(Z_i^{(1)^2}.(\sigma^2_{\lambda_i,\delta} - \sigma^2_{p,\delta}) + Z_i^{(2)^2}.(\sigma^2_{\lambda_i,\delta\beta} - \sigma^2_{p,\delta\beta})\right)\right)$$
$$+exp(X_i\beta_{\lambda_i}).exp\left(\tfrac{1}{2}.\left(Z_i^{(1)^2}.\sigma^2_{\lambda_i,\delta} + Z_i^{(2)^2}.\sigma^2_{\lambda_i,\delta\beta}\right)\right)$$

D'où finalement, on obtient que :

$$Var(Y_i \mid X_i,\, Z_i) = exp\left(2X_i\beta_{\lambda_i}\right).\left(exp\left(2.\left(Z_i^{(1)^2}.\sigma^2_{\lambda_i,\delta} + Z_i^{(2)^2}.\sigma^2_{\lambda_i,\delta\beta}\right)\right) - exp\left(Z_i^{(1)^2}.\sigma^2_{\lambda_i,\delta} + Z_i^{(2)^2}.\sigma^2_{\lambda_i,\delta\beta}\right)\right)$$
$$+exp\left(X_i(\beta_{\lambda_i} - \beta_p)\right).exp\left(\tfrac{1}{2}.\left(Z_i^{(1)^2}.(\sigma^2_{\lambda_i,\delta} - \sigma^2_{p,\delta}) + Z_i^{(2)^2}.(\sigma^2_{\lambda_i,\delta\beta} - \sigma^2_{p,\delta\beta})\right)\right)$$
$$+exp(X_i\beta_{\lambda_i}).exp\left(\tfrac{1}{2}.\left(Z_i^{(1)^2}.\sigma^2_{\lambda_i,\delta} + Z_i^{(2)^2}.\sigma^2_{\lambda_i,\delta\beta}\right)\right)$$

Enfin, il ne reste plus qu'à calculer les différents termes de covariance :

$$Cov(Y_{ijkl}, Y_{ijkm} \mid X_{ijkl},\, X_{ijkm},\, Z_{ijkl},\, Z_{ijkm})$$
$$= Cov_U\left(E(Y_{ijkl} \mid X_{ijkl},\, Z_{ijkl},\, U),\, E(Y_{ijkm} \mid X_{ijkm},\, Z_{ijkm},\, U)\right)$$
$$+E_U(Cov(Y_{ijkl}, Y_{ijkm} \mid X_{ijkl},\, X_{ijkm},\, Z_{ijkl},\, Z_{ijkm},\, U)$$

Or $Cov(Y_{ijkl}, Y_{ijkm} \mid X_{ijkl},\, X_{ijkm},\, Z_{ijkl},\, Z_{ijkm},\, U) = 0$ par indépendance conditionnelle
Donc, pour $l \neq m$:

$$Cov(Y_{ijkl}, Y_{ijkm} \mid X_{ijkl},\, X_{ijkm},\, Z_{ijkl},\, Z_{ijkm})$$
$$= exp(X_{ijkl}.\beta).exp(X_{ijkm}.\beta).Cov_U\left(exp(Z^{(1)}_{ijkl}.U^{(1)} + Z^{(2)}_{ijkl}.U^{(2)}),\, exp(Z^{(1)}_{ijkm}.U^{(1)} + Z^{(2)}_{ijkm}.U^{(2)})\right)$$
$$= exp\left((X_{ijkl} + X_{ijkm}).\beta\right).Cov_U\left(exp(Z^{(1)}_{ijkl}.U^{(1)} + Z^{(2)}_{ijkl}.U^{(2)}),\, exp(Z^{(1)}_{ijkm}.U^{(1)} + Z^{(2)}_{ijkm}.U^{(2)})\right)$$
$$= exp\left((X_{ijkl} + X_{ijkm}).\beta\right).\Big[E_U\left(exp\left((Z^{(1)}_{ijkl} + Z^{(1)}_{ijkm}).U^{(1)} + (Z^{(2)}_{ijkl} + Z^{(2)}_{ijkm}).U^{(2)}\right)\right)$$
$$-E_U\left(exp(Z^{(1)}_{ijkl}.U^{(1)} + Z^{(2)}_{ijkl}.U^{(2)})\right).E_U\left(exp(Z^{(1)}_{ijkm}.U^{(1)} + Z^{(2)}_{ijkm}.U^{(2)})\right)\Big]$$
$$= exp\left((X_{ijkl} + X_{ijkm}).\beta\right).\Big[E_U\left(exp\left((Z^{(1)}_{ijkl} + Z^{(1)}_{ijkm}).U^{(1)}\right).exp\left((Z^{(2)}_{ijkl} + Z^{(2)}_{ijkm}).U^{(2)}\right)\right)$$
$$-E_U\left(exp(Z^{(1)}_{ijkl}.U^{(1)}).exp\left(Z^{(2)}_{ijkl}.U^{(2)}\right)\right).E_U\left(exp(Z^{(1)}_{ijkm}.U^{(1)}).exp\left(Z^{(2)}_{ijkm}.U^{(2)}\right)\right)\Big]$$
$$= exp\left((X_{ijkl} + X_{ijkm}).\beta\right).\Big[E_U\left(exp\left((Z^{(1)}_{ijkl} + Z^{(1)}_{ijkm}).U^{(1)}\right)\right).E_U\left(exp\left((Z^{(2)}_{ijkl} + Z^{(2)}_{ijkm}).U^{(2)}\right)\right)$$
$$-E_U\left(exp(Z^{(1)}_{ijkl}.U^{(1)})\right).E_U\left(exp\left(Z^{(2)}_{ijkl}.U^{(2)}\right)\right).E_U\left(exp(Z^{(1)}_{ijkm}.U^{(1)})\right).E_U\left(exp\left(Z^{(2)}_{ijkm}.U^{(2)}\right)\right)\Big]$$
, par indépendance entre les deux types d'effets aléatoires.
$$= exp\left((X_{ijkl} + X_{ijkm}).\beta\right).\Big[exp\left(\tfrac{1}{2}.(Z^{(1)}_{ijkl} + Z^{(1)}_{ijkm})^2.\sigma^2_\delta\right).exp\left(\tfrac{1}{2}.(Z^{(2)}_{ijkl} + Z^{(2)}_{ijkm})^2.\sigma^2_{\delta\beta}\right)$$
$$-exp\left(\tfrac{1}{2}.Z^{(1)^2}_{ijkl}.\sigma^2_\delta\right).exp\left(\tfrac{1}{2}.Z^{(2)^2}_{ijkl}.\sigma^2_{\delta\beta}\right).exp\left(\tfrac{1}{2}.Z^{(1)^2}_{ijkm}.\sigma^2_\delta\right).exp\left(\tfrac{1}{2}.Z^{(2)^2}_{ijkm}.\sigma^2_{\delta\beta}\right)\Big]$$
$$= exp\left((X_{ijkl} + X_{ijkm}).\beta\right).\Big[exp\left(\tfrac{1}{2}.\left((Z^{(1)}_{ijkl} + Z^{(1)}_{ijkm})^2.\sigma^2_\delta + (Z^{(2)}_{ijkl} + Z^{(2)}_{ijkm})^2.\sigma^2_{\delta\beta}\right)\right)$$
$$-exp\left(\tfrac{1}{2}.\left((Z^{(1)^2}_{ijkl} + Z^{(1)^2}_{ijkm}).\sigma^2_\delta + (Z^{(2)^2}_{ijkl} + Z^{(2)^2}_{ijkm}).\sigma^2_{\delta\beta}\right)\right)\Big]$$

A l'aide du même développement, on trouve que :

$$Cov(Y_{ijkl}, Y_{ijqm} \mid X_{ijkl},\, X_{ijqm},\, Z_{ijkl},\, Z_{ijqm})$$
$$= exp\left((X_{ijkl} + X_{ijqm}).\beta\right).\Big[exp\left(\tfrac{1}{2}.\left((Z^{(1)}_{ijkl} + Z^{(1)}_{ijqm})^2.\sigma^2_\delta + (Z^{(2)}_{ijkl} + Z^{(2)}_{ijqm})^2.\sigma^2_{\delta\beta}\right)\right)$$
$$-exp\left(\tfrac{1}{2}.\left((Z^{(1)^2}_{ijkl} + Z^{(1)^2}_{ijqm}).\sigma^2_\delta + (Z^{(2)^2}_{ijkl} + Z^{(2)^2}_{ijqm}).\sigma^2_{\delta\beta}\right)\right)\Big],\ k \neq q,\ l, m \in \{1, 2\}$$

Annexe 3 : Vraisemblances

Dans notre cas, l'estimation des différents paramètres se fait par maximisation de la log-vraisemblance. J'ai donc essayé de calculer la vraisemblance marginale dans les deux situations décrites auparavant. Cependant, étant donné la forme obtenue, il m'a semblé très compliqué voir impossible de les calculer. En accord, avec mon maitre de stage, j'ai donc décidé de ne pas m'acharner plus longtemps sur ce calcul. En effet, des approximations de ces fonctions sont réalisées par la logiciel SAS grâce à de multiples algorithmes implantés notamment au sein de la procédure NLMIXED. Je vais donc me limiter à présenter la forme initiale de ces vraisemblances ainsi que la forme des vraisemblances conditionnelles.

– Vraisemblance conditionnelle :

Afin de simplifier l'écriture, posons $Y_p = Y_{ijkl}$, $\forall i, j, k, l$, avec p=1,...,n où n est le nombre total d'observations. Cela ne pose pas de problème puisque dans le cadre conditionnel, on a que les (Y_p, X_p, Z_p) sont indépendants pour p=1,...,n, avec X_p le vecteur ligne de la matrice X correspondant à l'observation p et Z_p le vecteur ligne de la matrice Z correspondant à l'observation p. On peut donc voir chacune des mesures comme étant une mesure à part entière, sans aucun lien de dépendance avec aucune des autres. On peut donc se passer de la distinction entre les mesures appartenant au même individu, ou au même individu et au même jour, ou encore celles qui sont indépendantes puisque n'appartenant pas au même sujet. La vraisemblance n'est ainsi rien d'autre que le produit des densités. On obtient donc que :

$$f_{Y|X,Z,U}(y) = \prod_{p=1}^{n} f_{Y_p|X_p,\,Z_p,\,U_p}(y_p)$$

Posons n_1 le nombre d'observations notées y_p^0 égal à 0, n_2 le nombre d'observations notées y_p^{int} censuré par intervalle différent de 0 et n_3 le nombre d'observations notées y_p^{dr} censurée à droite. Alors on obtient que :

$$f_{Y|X,Z,U}(y) = \prod_{p=1}^{n_1} f_{Y_p|X_p,\,Z_p,\,U_p}(y_p^0).\prod_{p=n_1+1}^{n_1+n_2} f_{Y_p|X_p,\,Z_p,\,U_p}(y_p^{int}).\prod_{p=n_1+n_2+1}^{n} f_{Y_p|X_p,\,Z_p,\,U_p}(y_p^{dr})$$

Appliquons donc ceci aux deux cas envisagés.

Cas discret :

Comme explicité ci-dessus, voilà la distribution choisie dans le cadre discret :

$$\mathrm{P}(Y_p{=}y_p) = \begin{cases} p_0 + \frac{(1-p_0)}{(1+\frac{\lambda}{k})^k} & , y_p = 0 \\ (1-p_0).\left[\frac{\lambda^{y_p}}{y_p!}\cdot\frac{\Gamma(y_p+k)}{\Gamma(k).(\lambda+k)^{y_p}}.(1+\frac{\lambda}{k})^{-k}\right] & , y_p > 0 \\ (1-p_0).\sum\limits_{i=y_p}^{+\infty}\left[\frac{\lambda^i}{i!}\cdot\frac{\Gamma(i+k)}{\Gamma(k).(\lambda+k)^i}.(1+\frac{\lambda}{k})^{-k}\right] & , y_p = \max\limits_{p}(Y_p) \end{cases}$$

Ainsi, on obtient comme vraisemblance conditionnelle :

$$\begin{aligned}
f_{Y|X,Z,U}(y) &= \prod_{p=1}^{n_1}\left[p_0+\tfrac{(1-p_0)}{(1+\frac{\lambda}{k})^k}\right].\prod_{p=n+1_1}^{n_1+n_2}(1-p_0).\left[\tfrac{\lambda^{y_p}}{y_p!}.\tfrac{\Gamma(y_p+k)}{\Gamma(k).(\lambda+k)^{y_p}}.(1+\tfrac{\lambda}{k})^{-k}\right] \\
&\quad .\prod_{p=n_1+n_2+1}^{n}(1-p_0).\sum_{i=y_p}^{+\infty}\left[\tfrac{\lambda^i}{i!}.\tfrac{\Gamma(i+k)}{\Gamma(k).(\lambda+k)^i}.(1+\tfrac{\lambda}{k})^{-k}\right] \\
&= \left(p_0+\tfrac{(1-p_0)}{(1+\frac{\lambda}{k})^k}\right)^{n_1}.(1-p_0)^{n_2}.\prod_{p=n_1+1}^{n_1+n_2}\left[\tfrac{\lambda^{y_p}}{y_p!}.\tfrac{\Gamma(y_p+k)}{\Gamma(k).(\lambda+k)^{y_p}}.(1+\tfrac{\lambda}{k})^{-k}\right].(1-p_0)^{n_3} \\
&\quad .\prod_{p=n_1+n_2+1}^{n}\left(\sum_{i=y_p}^{+\infty}\left[\tfrac{\lambda^i}{i!}.\tfrac{\Gamma(i+k)}{\Gamma(k).(\lambda+k)^i}.(1+\tfrac{\lambda}{k})^{-k}\right]\right) \\
&= (1-p_0)^{n_2+n_3}.\left(p_0+\tfrac{(1-p_0)}{(1+\frac{\lambda}{k})^k}\right)^{n_1}.\prod_{p=n+1_1}^{n_1+n_2}\left[\tfrac{\lambda^{y_p}}{y_p!}.\tfrac{\Gamma(y_p+k)}{\Gamma(k).(\lambda+k)^{y_p}}.(1+\tfrac{\lambda}{k})^{-k}\right] \\
&\quad .\prod_{p=n_1+n_2+1}^{n}\left(\sum_{i=y_p}^{+\infty}\left[\tfrac{\lambda^i}{i!}.\tfrac{\Gamma(i+k)}{\Gamma(k).(\lambda+k)^i}.(1+\tfrac{\lambda}{k})^{-k}\right]\right)
\end{aligned}$$

Soit F la fonction de répartition liée à cette distribution et f la fonction de densité liée. Alors on peut écrire cette vraisemblance sous la forme :

$$f_{Y|X,Z,U}(y)= (1-p_0)^{n_2+n_3}.(p_0+(1-p_0).f(0))^{n_1}.\prod_{p=n_1+1}^{n_1+n_2}f(y_p).\prod_{p=n_1+n_2+1}^{n}(1-F(y_p))$$

On en déduit donc une forme pour la log-vraisemblance, notée L :

$$\begin{aligned}
\mathrm{L(y|X,\ Z,\ U)} &= log\left((1-p_0)^{n_2+n_3}\right)+log\left((p_0+(1-p_0).f(0))^{n_1}\right)+\sum_{p=n_1+1}^{n_1+n_2}log\left(f(y_p)\right) \\
&\quad +\sum_{p=n_1+n_2+1}^{n}log\left(1-F(y_p)\right) \\
&= \left(n_2+n_3\right).log\left(1-p_0\right)+n_1.log\left(p_0+(1-p_0).f(0)\right)+\sum_{p=n_1+1}^{n_1+n_2}log\left(f(y_p)\right) \\
&\quad +\sum_{p=n_1+n_2+1}^{n}log\left(1-F(y_p)\right)
\end{aligned}$$

<u>Cas continu :</u>

En ce qui concerne le cas continu, rappelons la distribution que l'on a posé :

$$\mathrm{P}(Y_p{=}y_p) = \begin{cases} p_0+(1-p_0).\int_{-\infty}^{1}\frac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx & ,y_p=0 \\ (1-p_0).\int_{y_p}^{y_p+1}\frac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx & ,y_p>0 \\ (1-p_0).\int_{y_p}^{+\infty}\frac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx & ,y_p=\max_p(Y_p) \end{cases}$$

Ainsi, on obtient comme vraisemblance conditionnelle :

$$
\begin{aligned}
f_{Y|X,Z,U}(y) = \prod_{p=1}^{n_1} \left[p_0 + (1-p_0).\int_{-\infty}^{1} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx\right] . \prod_{p=n_1+1}^{n_1+n_2} (1-p_0).\int_{y_p}^{y_p+1} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx \\
. \prod_{p=n_1+n_2+1}^{k} (1-p_0).\int_{y_p}^{+\infty} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx \\
= \left(p_0 + (1-p_0).\int_{-\infty}^{1} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx\right)^{n_1} .(1-p_0)^{n_2}. \prod_{p=n_1+1}^{n_1+n_2} \left[\int_{y_p}^{y_p+1} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx\right] \\
.(1-p_0)^{n_3}. \prod_{p=n_1+n_2+1}^{k} \left(\int_{y_p}^{+\infty} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx\right) \\
= (1-p_0)^{n_2+n_3}.\left(p_0 + (1-p_0).\int_{-\infty}^{1} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx\right)^{n_1} . \prod_{p=n_1+1}^{n_1+n_2} \left[\int_{y_p}^{y_p+1} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx\right] \\
. \prod_{p=n_1+n_2+1}^{k} \left(\int_{y_p}^{+\infty} \tfrac{1}{\sqrt{2\pi}.\sigma}.e^{(-\frac{1}{2\sigma^2}.(x-\mu)^2)}dx\right)
\end{aligned}
$$

Soit F la fonction de répartition liée à cette distribution. Alors on peut écrire cette vraisemblance sous la forme :

$$
f_{Y|X,Z,U}(y) = (1-p_0)^{n_2+n_3}.(p_0 + (1-p_0).F(1))^{n_1} . \prod_{p=n_1+1}^{n_1+n_2} [F(y_p+1) - F(y_p)] . \prod_{p=n_1+n_2+1}^{k} (1-F(y_p))
$$

On en déduit donc une forme pour la log-vraisemblance, notée L :

$$
\begin{aligned}
\mathrm{L(y|X,\,Z,\,U)} = log\left((1-p_0)^{n_2+n_3}\right) + log\left((p_0 + (1-p_0).F(1))^{n_1}\right) + \sum_{p=n_1+1}^{n_1+n_2} log\left(F(y_p+1) - F(y_p)\right) \\
+ \sum_{p=n_1+n_2+1}^{k} log\left(1-F(y_p)\right) \\
= (n_2+n_3).log\left(1-p_0\right) + n_1.log\left(p_0 + (1-p_0).F(1)\right) + \sum_{p=n_1+1}^{n_1+n_2} log\left(F(y_p+1) - F(y_p)\right) \\
+ \sum_{p=n_1+n_2+1}^{k} log\left(1-F(y_p)\right)
\end{aligned}
$$

Passons désormais à la détermination des vraisemblances marginales, vraisemblances qui permettent l'estimation des paramètres par maximisation.

– Vraisemblance marginale :

Soit y le vecteur réponse, X, Z et U comme défini en Annexe. Alors, on obtient que :

$$\begin{aligned} f_{Y|X,Z}(y) &= \int_q f_{Y|X,Z,U=u}(y).f_U(u).du \\ &= \int_q \prod_{p=1}^{n_1} f_{Y_p|X_p,\,Z_p,\,U_p=u_p}(y_p^0).\prod_{p=n_1+1}^{n_1+n_2} f_{Y_p|X_p,\,Z_p,\,U_p=u_p}(y_p^{int}) \\ &\quad .\prod_{p=n_1+n_2+1}^{k} f_{Y_p|X_p,\,Z_p,\,U_p=u_p}(y_p^{dr}).f_U(u).du \end{aligned}$$

avec :

$f_U(u)$ la densité liée à une loi normale N(0,D)

$f_{Y|X,Z,U=u}(y)$ les densités décrites ci-dessus suivant que l'on soit dans le cas discret ou continu

q la dimension du vecteur U

On obtient naturellement comme log-vraisemblance à maximiser :

$$\begin{aligned} \mathrm{L(y|X,Z)} &= log\left(\int_q f_{Y|X,Z,U=u}(y).f_U(u).du\right) \\ &= log\left(\int_q \prod_{p=1}^{n_1} f_{Y_p|X_p,\,Z_p,\,U_p=u_p}(y_p^0).\prod_{p=n_1+1}^{n_1+n_2} f_{Y_p|X_p,\,Z_p,\,U_p=u_p}(y_p^{int}).\prod_{p=n_1+n_2+1}^{k} f_{Y_p|X_p,\,Z_p,\,U_p=u_p}(y_p^{dr}).f_U(u).du\right) \end{aligned}$$

Je me passerai de présenter les formes explicites obtenues dans les deux cas. En effet, celles-ci sont très complexes et incalculables, selon moi, sans l'aide d'algorithme d'approximation. Elles seront donc approchées par l'intermédiaire d'algorithmes implantés au sein de la procédure NL-MIXED de SAS.

Annexe 4 : Statistique de Vuong

Une statistique a été créée, dite statistique de Vuong, permettant de comparer un modèle adapté pour une situation de zéro en excès et son homologue classique. Dans notre cas, il s'agira de comparer un zero-inflated binomial négative (ZINB) model avec un modèle classique où la variable réponse repose sur une loi binomiale négative. Cette statistique repose sur la comparaison de la probabilité d'observer y_{ijkl} étant donnée une loi ZINB et de la probabilité d'observer y_{ijkl} étant donnée une loi binomiale négative.

Notons $f_j(y_{ijkl} \mid X_{ijkl}, Z_{ijkl}, U)$ la probabilité prédite pour que la variable aléatoire Y soit égale à y_{ijkl} sous l'hypothèse que la densité est $f_j(y_{ijkl} \mid X_{ijkl}, Z_{ijkl}, U)$, pour j=1,2.
Notons $m_{ijkl} = log\left(\frac{f_1(y_{ijkl}|X_{ijkl}, Z_{ijkl}, U)}{f_2(y_{ijkl}|X_{ijkl}, Z_{ijkl}, U)}\right)$.

Cette statistique de Vuong est alors définie par l'équation suivante :

$$V = \frac{\sqrt{n}.\left(\frac{1}{n}.\sum_{i=1}^{n} m_{ijkl}\right)}{\sqrt{\frac{1}{n}.\sum_{i=1}^{n}(m_{ijkl}-\overline{m})^2}}$$

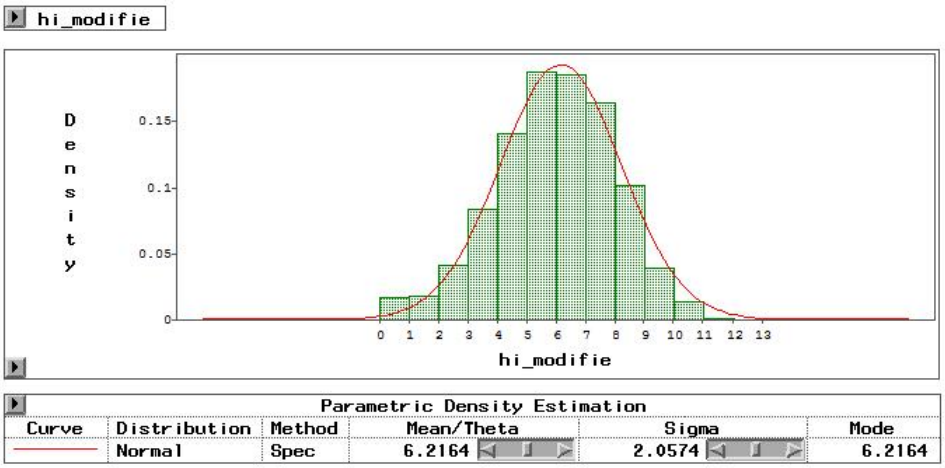
$$= \frac{\sqrt{n}.\overline{m}}{S_m}$$

V suit une loi de student à n degrés de liberté, loi que l'on peut approximer pour n grand, ce qui est souvent le cas dans ce type d'études, par une loi Normale. On peut alors se donner, en prenant comme seuil de confiance un seuil à 5%, la règle de décision suivante :

Si V $>$ 1,96 le modèle ZINB est préféré.
Si -1,96 $<$ V $<$ 1,96 aucun des deux modèles n'est préférable, le test est indécis.
Si V $<$ 1,96 le modèle binomial négatif est préféré.

Ce test pourrait ainsi être intéressant à appliquer lors de l'analyse des études afin de déterminer quelle méthode est préférable.

Annexe 5 : Exemple graphique de répartition des titres comparé à une densité de loi normale



Exemple de répartition des données au Jour 21.

Annexe 6 : Programme et sorties SAS simulation pour validation du 1er modèle.

```
options nodate nonumber nocenter;


/*****Simulation des données*****/

data one;
  do pid=1 to 1000;
      u_pid=sqrt(2)*rannor(13);
      u_pidday1=sqrt(3)*rannor(71);
      u_pidday21=sqrt(1)*rannor(72);
    do t=0 to 1;
      if t=0 then do;
        mu=2;
          do rep=1 to 2;
              eps=sqrt(0.75)*rannor(110);
              hi_modifie=mu+u_pid+u_pidday1+eps;
              output;
          end;
      end;
      if t=1 then do;
        mu=6;
          do rep=1 to 2;
              eps=sqrt(0.75)*rannor(112);
              hi_modifie=mu+u_pid+u_pidday21+eps;
              output;
          end;
       end;
      end;
    end;
  run;


/*****Création des différentes tables et vérification des corrélations*****/

proc sql;
 create table d0rep1 as
 select pid, hi_modifie as hi_d0rep1 from one
 where t=0 and rep=1;
quit;

proc sql;
 create table d0rep2 as
 select pid,hi_modifie as hi_d0rep2 from one
 where t=0 and rep=2;
quit;

proc sql;
 create table d21rep1 as
 select pid,hi_modifie as hi_d21rep1 from one
 where t=1 and rep=1;
quit;

proc sql;
 create table d21rep2 as
 select pid,hi_modifie as hi_d21rep2 from one
 where t=1 and rep=2;
quit;
```

```
proc sql;
 create table d21 as
 select d21rep1.*,hi_d21rep2 from d21rep1, d21rep2
 where d21rep1.pid=d21rep2.pid;
quit;

proc corr data=d21;
 var hi_d21rep1 hi_d21rep2;
run;
```

Corrélation théorique = 3/3.75 = 0.8

Coefficients de corrélation de Pearson, N = 1000 Proba > |r| sous H0: Rho=0		
	hi_d21rep1	**hi_d21rep2**
hi_d21rep1	1.00000	0.80514 <.0001
hi_d21rep2	0.80514 <.0001	1.00000

```
proc sql;
 create table d0 as
 select d0rep1.*,hi_d0rep2 from d0rep1, d0rep2
 where d0rep1.pid=d0rep2.pid;
quit;

proc corr data=d0;
 var hi_d0rep1 hi_d0rep2;
run;
```

Corrélation théorique = 5/5.75 = 0.8696

Coefficients de corrélation de Pearson, N = 1000 Proba > |r| sous H0: Rho=0		
	hi_d0rep1	**hi_d0rep2**
hi_d0rep1	1.00000	0.86986 <.0001
hi_d0rep2	0.86986 <.0001	1.00000

```
proc sql;
 create table rep1 as
 select d0rep1.*,hi_d21rep1 from d0rep1, d21rep1
 where d0rep1.pid=d21rep1.pid;
quit;

proc corr data=rep1;
 var hi_d0rep1 hi_d21rep1;
run;
```

Corrélation théorique = $2/(\sqrt{5.75}.\sqrt{3.75}) = 0.4307$

Coefficients de corrélation de Pearson, N = 1000 Proba > |r| sous H0: Rho=0		
	hi_d0rep1	**hi_d21rep1**
hi_d0rep1	1.00000	0.43749 <.0001
hi_d21rep1	0.43749 <.0001	1.00000

```
proc sql;
 create table rep2 as
 select d0rep2.*,hi_d21rep2 from d0rep2, d21rep2
 where d0rep2.pid=d21rep2.pid;
quit;

proc corr data=rep2;
 var hi_d0rep2 hi_d21rep2;
run;
```

Corrélation théorique = $2/(\sqrt{5.75}.\sqrt{3.75}) = 0.4307$

Coefficients de corrélation de Pearson, N = 1000 Proba > |r| sous H0: Rho=0		
	hi_d0rep2	**hi_d21rep2**
hi_d0rep2	1.00000	0.43357 <.0001
hi_d21rep2	0.43357 <.0001	1.00000

```
proc sql;
 create table rep2piddiff as
 select d0rep2.*,hi_d21rep2 from d0rep2, d21rep2
 where d0rep2.pid+2=d21rep2.pid;
quit;

proc corr data=rep2piddiff;
 var hi_d0rep2 hi_d21rep2;
run;
```

Corrélation théorique = 0

Coefficients de corrélation de Pearson, N = 998 Proba > |r| sous H0: Rho=0		
	hi_d0rep2	**hi_d21rep2**
hi_d0rep2	1.00000	-0.04004 0.2063
hi_d21rep2	-0.04004 0.2063	1.00000

```
/*****Détermination de l'écriture du modèle à l'aide de NLMIXED*****/

proc nlmixed data = one;
 parms b0=2 c0=0.8 b1=4 c1=0 var1=3 var2=1 var3=2;
 bounds var1>0, var2>0;
  mu = b0+b1*t + eta0 + (eta1-eta0)*t + eta2;
  k = c0 + c1*t;
 model hi_modifie ~ normal(mu,k);
 random eta0 eta1 eta2~ normal([0,0,0], [var1,0,var2,0,0,var3]) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	2.0232	0.07368	997	27.46	<.0001	0.05	1.8786	2.1678	0.000043
c0	0.7569	0.03385	997	22.36	<.0001	0.05	0.6905	0.8233	-0.00017
b1	4.0050	0.06912	997	57.94	<.0001	0.05	3.8693	4.1406	-0.00002
c1	0.01170	0.04824	997	0.24	0.8084	0.05	-0.08297	0.1064	-0.00014
var1	2.9500	0.1931	997	15.27	<.0001	0.05	2.5710	3.3290	-3.4E-6
var2	1.0653	0.1391	997	7.66	<.0001	0.05	0.7923	1.3384	-3.37E-6
var3	2.1008	0.1539	997	13.65	<.0001	0.05	1.7988	2.4028	-5.85E-7

Annexe 7 : Programme et sortie SAS exemple de sous-dispersion des données.

```
/********** Exemple concret de sous-dispersion des données **********/

options nodate nonumber nocenter;
libname BD 'C:\Users\Darkflow\Documents\cours\Master Stat\Stage M2
Stat\Analysis of HI data\Stage Jérémy';

data bd;
 set BD.cop_hi_t;
run;


/*****Transformation de la base de données*****/

data bd;
set bd;
  if mod(2*((log(HI) - log(5))/log(2)),2)=0 then do;
     hi1brut=HI;
       hi2brut=HI;
  end;
  else do;
     hi1brut=round(HI/sqrt(2),5);
       hi2brut=round(HI*sqrt(2),5);
  end;
hi1_modifie=log2(hi1brut/5);
hi2_modifie=log2(hi2brut/5);
run;

proc sql;
 create table bd_etude as
 select pid,trt,day,hi1_modifie as hi_modifie from bd
 where day<22 and trial="NG6";
quit;

proc sql;
 create table bd_etude2 as
 select pid,trt,day,hi2_modifie as hi_modifie from bd
 where day<22 and trial="NG6";
quit;

proc append base=bd_etude data=bd_etude2 force;
run;

proc sort data=bd_etude;
 by trt pid day;
run;

data bd_etude;
 set bd_etude;
  if trt="FLU-NG" then trt_etude=1;
  else trt_etude=0;
  if day=0 then day_etude=0;
  else day_etude=1;
run;
```

```
/*****Création de la table Groupe1 Jour21*****/

proc sql;
 create table tr1d21 as
 select * from bd_etude
 where  day_etude=1 and trt_etude=1;
run;
quit;


/*****Détermination de la moyenne et variance de ces données*****/

proc univariate data=tr1d21;
 var hi_modifie;
run;
```

Mesures statistiques de base			
Location		**Variability**	
Moyenne	5.835912	**Ecart-type**	1.95509
Médiane	6.000000	**Variance**	3.82239
Mode	6.000000	**Intervalle**	12.00000
		Ecart interquartile	2.00000

Sortie SAS prouvant la sous-dispersion des données.

Annexe 8 : Simulation et sorties SAS pour validation de la distribution continue dans la première approche.

```
options nodate nonumber nocenter;

/*****************Simulations données sans excès en zéro*****************/


/*Simulation données 1 jour 2 répétitions*/


data one;                                  /*Variance résid théorique=(0.25/2)*/
  do pid=1 to 1000;
    hi_modifie_pre= RAND('NEGBINOMIAL',0.9,40);          /*m=4.44 var=4.94*/
          do rep=0 to 1;
             if rep=0 then hi_modifie=hi_modifie_pre;
             else hi_modifie=hi_modifie_pre+sqrt(0.25)*rannor(112);
             output;
          end;
  end;
run;


data one_b;                                 /*Variance résid théorique=(2.5/2)*/
  do pid=1 to 1000;
    hi_modifie_pre= RAND('NEGBINOMIAL',0.9,40);          /*m=4.44 var=4.94*/
          do rep=0 to 1;
             if rep=0 then hi_modifie=hi_modifie_pre;
             else hi_modifie=hi_modifie_pre+sqrt(2.5)*rannor(112);
             output;
          end;
  end;
run;


/*Simulation données 2 jours et 2 répétitions avec erreur résiduelle
normale*/


data two;                                              /*Var résid théorique=0.25*/
  do pid=1 to 1000;
    hi_modifie_pre= RAND('NEGBINOMIAL',0.9,40);
    u_pidday1=RAND('NEGBINOMIAL',0.99,40);
    u_pidday21=RAND('NEGBINOMIAL',0.99,40);
      do t=0 to 1;
        if t=0 then do;
         do rep=0 to 1;
          if rep=0 then
                hi_modifie=hi_modifie_pre+u_pidday1+sqrt(0.25)*rannor(110);
          else hi_modifie=hi_modifie_pre+u_pidday1+sqrt(0.25)*rannor(112);
          output;
         end;
        end;
        else do;
         do rep=0 to 1;
          if rep=0 then
              hi_modifie=hi_modifie_pre+4+u_pidday21+sqrt(0.25)*rannor(71);
```

```
          else
              hi_modifie=hi_modifie_pre+4+u_pidday21+sqrt(0.25)*rannor(64);
          output;
         end;
        end;
      end;
  end;
run;


data two_b;                                        /*Var résid théorique=2.5*/
  do pid=1 to 1000;
    hi_modifie_pre= RAND('NEGBINOMIAL',0.9,40);
    u_pidday1=RAND('NEGBINOMIAL',0.99,40);
    u_pidday21=RAND('NEGBINOMIAL',0.99,40);
      do t=0 to 1;
        if t=0 then do;
          do rep=0 to 1;
             if rep=0 then
                 hi_modifie=hi_modifie_pre+u_pidday1+sqrt(2.5)*rannor(110);
             else
                 hi_modifie=hi_modifie_pre+u_pidday1+sqrt(2.5)*rannor(112);
             output;
          end;
        end;
        else do;
          do rep=0 to 1;
             if rep=0 then
               hi_modifie=hi_modifie_pre+4+u_pidday21+sqrt(2.5)*rannor(71);
             else
               hi_modifie=hi_modifie_pre+4+u_pidday21+sqrt(2.5)*rannor(64);
             output;
          end;
        end;
      end;
  end;
run;


 /*Simulation données étudiées 2 jours et 2 répétitions*/


data three;                                  /*var résid théorique=(0.0475/2)*/
   do pid=1 to 1000;
     hi_modifie_pre= RAND('NEGBINOMIAL',0.9,40);          /*var B-PID=4.94*/
     u_pidday1=RAND('NEGBINOMIAL',0.99,40);          /*var B-PID*day=0.408*/
     u_pidday21=RAND('NEGBINOMIAL',0.99,40);
      do t=0 to 1;
        if t=0 then do;
          do rep=0 to 1;
             if rep=0 then hi_modifie=hi_modifie_pre+u_pidday1;
             else
                hi_modifie=hi_modifie_pre+u_pidday1+RAND('BERNOULLI',0.05);
             output;
          end;
        end;
        else do;
          do rep=0 to 1;
             if rep=0 then hi_modifie=hi_modifie_pre+3+u_pidday21;
             else
             hi_modifie=hi_modifie_pre+3+u_pidday21+RAND('BERNOULLI',0.05);
             output;
```

```
          end;
        end;
      end;
    end;
run;



/*****************Simulations données avec excès en zéro*****************/


/*Simulation données étudiées 2 jours et 2 répétitions*/

data four;
 do pid=1 to 1200;              /* p0=1/6 */
  if pid<1001 then do;
    hi_modifie_pre= RAND('NEGBINOMIAL',0.9,40);
    u_pidday1=RAND('NEGBINOMIAL',0.99,40);
    u_pidday21=RAND('NEGBINOMIAL',0.99,40);
      do t=0 to 1;
        if t=0 then do;
           do rep=0 to 1;
              if rep=0 then hi_modifie=hi_modifie_pre+u_pidday1;
              else
                hi_modifie=hi_modifie_pre+u_pidday1+RAND('BERNOULLI',0.05);
              output;
           end;
        end;
        else do;
           do rep=0 to 1;
              if rep=0 then hi_modifie=hi_modifie_pre+3+u_pidday21;
              else
             hi_modifie=hi_modifie_pre+3+u_pidday21+RAND('BERNOULLI',0.05);
              output;
           end;
        end;
      end;
  end;
  else do;
    do t=0 to 1;
        do rep=0 to 1;
          hi_modifie=0;
          output;
        end;
    end;
  end;
 end;
run;
```

```
/**********************Analyse des données simulées**********************/


/*****Analyse un jour 2 rep avec résidus suivant une loi normale*****/

/*Analyse sans censure*/

proc nlmixed data = one;          /*OK*/
parms  b0=4.5 c0=1.2 var1=5;

  mu = b0+eta0;
  k = sqrt(c0);

model hi_modifie ~ normal(mu,k*k);
random eta0~ normal(0,var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	4.4626	0.06960	999	64.12	<.0001	0.05	4.3260	4.5992	0.000016
c0	0.1307	0.005847	999	22.36	<.0001	0.05	0.1193	0.1422	0.000317
var1	4.7789	0.2167	999	22.06	<.0001	0.05	4.3537	5.2040	0.000028

```
/*Analyse avec censure*/

proc nlmixed data = one;                                    /*problème*/
 parms  b0=4.5036 c0=1.2 var1=4.949;
  mu = b0+eta0;
  k = sqrt(c0);
   if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
   else like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0,var1) subject=PID;
run;
```

```
WARNING: The final Hessian matrix is full rank but has at least one negative eigenvalue.
                    Second-order optimality condition violated.
```

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	4.5965	0.07555	999	60.84	<.0001	0.05	4.4482	4.7447	-67.1454
c0	0.7971	.	999	.	.	0.05	.	.	484.425
var1	4.9401	0.2658	999	18.59	<.0001	0.05	4.4186	5.4617	6.167015

```
proc nlmixed data = one;                                        /*problème*/
 parms  b0=4.5036 c0=0.2 var1=4.949;
  mu = b0+eta0;
  k = sqrt(c0);
   if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
   else like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0,var1) subject=PID;
run;
```

NOTE: Execution error for observation 37.

```
proc nlmixed data = one_b;                                      /*OK*/
 parms  b0=4.5036 c0=1.9 var1=4.949;
  mu = b0+eta0;
  k = sqrt(c0);
   if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
   else like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0,var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
b0	4.8474	0.07035	999	68.90	<.0001	0.05	4.7093	4.9854	0.001767
c0	1.2281	0.05887	999	20.86	<.0001	0.05	1.1125	1.3436	0.001021
var1	4.2908	0.2238	999	19.17	<.0001	0.05	3.8515	4.7300	-0.00005

```
/*****Analyse 2 jours 2 rep avec résidus suivant une loi normale*****/

/*Analyse sans censure*/

proc nlmixed data = two;                    /*OK*/
 parms b0=4.5 b1=4 c0=0.28 var1=0.5 var2=0.5 var3=5;
  mu = b0+b1*t+ eta0 + (eta1-eta0)*t + eta2;
  k = sqrt(c0);
 model hi_modifie ~ normal(mu,k*k);
 random eta0 eta1 eta2~ normal([0,0,0], [var1,0,var2,0,0,var3]) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	4.7689	0.07299	997	65.34	<.0001	0.05	4.6257	4.9122	8.55E-6
b1	4.0093	0.03255	997	123.18	<.0001	0.05	3.9454	4.0732	0.000067
c0	0.2479	0.007839	997	31.62	<.0001	0.05	0.2325	0.2633	-0.00018
var1	0.3363	0.07662	997	4.39	<.0001	0.05	0.1859	0.4866	0.000093
var2	0.4751	0.07851	997	6.05	<.0001	0.05	0.3211	0.6292	0.000075
var3	4.8667	0.2298	997	21.18	<.0001	0.05	4.4158	5.3176	-7.46E-6

```
/*Analyse avec censure*/

proc nlmixed data = two;                   /*Problème*/
 parms b0=4.5 b1=4 c0=2 var1=0.5 var2=0.5 var3=5;
  mu = b0+b1*t+ eta0 + (eta1-eta0)*t + eta2;
  k = sqrt(c0);
   if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
   else like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0 eta1 eta2~ normal([0,0,0], [var1,0,var2,0,0,var3]) subject=PID;
run;
```

```
WARNING: The final Hessian matrix is full rank but has at least one negative eigenvalue.
                    Second-order optimality condition violated.
```

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	4.8561	0.07990	997	60.78	<.0001	0.05	4.6994	5.0129	-69.162
b1	4.0719	0.04237	997	96.11	<.0001	0.05	3.9888	4.1551	3.063281
c0	1.1799	.	997	.	.	0.05	.	.	746.328

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
var1	0.1990	0.1506	997	1.32	0.1868	0.05	-0.09660	0.4946	112.1607
var2	0.2141	0.1412	997	1.52	0.1297	0.05	-0.06291	0.4912	107.7255
var3	5.1004	0.2689	997	18.97	<.0001	0.05	4.5727	5.6281	3.98617

```
proc nlmixed data = two_b;                    /*OK*/
 parms b0=4.5 b1=4 c0=2.5 var1=0.5 var2=0.5 var3=5;
  mu = b0+b1*t+ eta0 + (eta1-eta0)*t + eta2;
  k = sqrt(c0);
   if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
   else like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0 eta1 eta2~ normal([0,0,0], [var1,0,var2,0,0,var3]) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
b0	5.4319	0.08122	997	66.88	<.0001	0.05	5.2726	5.5913	-0.00015
b1	3.9953	0.05298	997	75.41	<.0001	0.05	3.8913	4.0993	0.000127
c0	2.3952	0.07838	997	30.56	<.0001	0.05	2.2414	2.5490	-0.00016
var1	0.2716	0.1495	997	1.82	0.0694	0.05	-0.02164	0.5649	0.000034
var2	0.05672	0.1454	997	0.39	0.6965	0.05	-0.2285	0.3420	-0.00026
var3	5.0853	0.2607	997	19.51	<.0001	0.05	4.5738	5.5969	-0.00005

```
/*****Analyse vrai type de base de données sans excès en zéro*****/

/*Analyse sans censure*/

proc nlmixed data = three;                /*OK*/
parms b0=4.5 b1=4 c0=0.025 var1=1.2 var2=2 var3=5;

mu = b0+b1*t+ eta0 + (eta1-eta0)*t + eta2;
k = sqrt(c0);

model hi_modifie ~ normal(mu,k*k);
random eta0 eta1 eta2~ normal([0,0,0], [var1,0,var2,0,0,var3]) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	4.9295	0.07401	997	66.61	<.0001	0.05	4.7843	5.0747	-0.00028
b1	2.9875	0.02840	997	105.19	<.0001	0.05	2.9318	3.0432	-0.00093
c0	0.02125	0.000672	997	31.62	<.0001	0.05	0.01993	0.02257	0.036037
var1	0.3923	0.06768	997	5.80	<.0001	0.05	0.2595	0.5251	-0.00038
var2	0.3931	0.06768	997	5.81	<.0001	0.05	0.2603	0.5259	-0.00225
var3	5.0739	0.2361	997	21.49	<.0001	0.05	4.6106	5.5372	-0.00007

```
/*Analyse avec censure*/

proc nlmixed data = three;                /*Problème*/
 parms b0=4.5 b1=4 c0=3 var1=1.2 var2=2 var3=5;
  mu = b0+b1*t+ eta0 + (eta1-eta0)*t + eta2;
  k = sqrt(c0);
   if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
   else like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0 eta1 eta2~ normal([0,0,0], [var1,0,var2,0,0,var3]) subject=PID;
run;
```

NOTE: FCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	4.8461	0.1045	997	46.36	<.0001	0.05	4.6410	5.0513	-69.7914
b1	3.4015	0.06908	997	49.24	<.0001	0.05	3.2659	3.5371	74.88427
c0	1.2186	.	997	.	.	0.05	.	.	849.4605

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
var1	0.9564	0.4855	997	1.97	0.0491	0.05	0.003746	1.9091	90.01689
var2	1.6965	0.4467	997	3.80	0.0002	0.05	0.8198	2.5731	100.7178
var3	4.9462	0.3156	997	15.67	<.0001	0.05	4.3269	5.5655	5.931378

```
/*****Analyse vrai type de base de données avec excès en zéro*****/

/*Analyse sans censure*/

proc nlmixed data = four;
parms a0=-2 b0=4.5 b1=4 c0=6;

p0= exp(a0)/(1+exp(a0));
mu = b0+b1*t;
k = sqrt(c0);

if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
else like=(1-p0)*PDF('NORMAL',hi_modifie,mu,k);
ll=log(like);

model hi_modifie ~ general(ll);
run;

proc nlmixed data = four;
parms b0=4.8568 b1=3.0271 c0=0.069 var1=0.9 var2=0.9 var3=5;

p0= 0.1636;
mu = b0+b1*t+ eta0 + (eta1-eta0)*t + eta2;
k = sqrt(c0);

if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
else like=(1-p0)*PDF('NORMAL',hi_modifie,mu,k);
ll=log(like);

model hi_modifie ~ general(ll);
random eta0 eta1 eta2~ normal([0,0,0], [var1,0,var2,0,0,var3]) subject=PID;
run;
```

Annexe 9 : Simulation et sorties SAS pour validation du modèle dans la deuxième approche et comparaison des méthodes.

```
options nodate nonumber nocenter;

/*******Simulation sans zéro en excès*******/

proc iml;
  call randseed(1);
  N=1000;
  Mean = {2.1 6.5};
  Corr = {1 0.5,0.5 1};
  Var = {1.9 3};
  Cov = Corr # sqrt(Var` * Var);            /*create the covariance matrix*/
  x = RANDNORMAL( N, Mean, Cov );
  SampleMean = x[:,];
  n = nrow(x);
  y = x - repeat( SampleMean, n );
  SampleCov = y`*y / (n-1);
  print SampleMean Mean, SampleCov Cov ;
  varnames='x1':'x2';
  create myhidata from x [colname=varnames];
  append from x;
quit;
```

SampleMean		Mean	
2.1551297	6.5620799	2.1	6.5

SampleCov		Cov	
2.0198202	1.322617	1.9	1.1937336
1.322617	3.199509	1.1937336	3

```
data one;
  set myhidata;
  array x(2) x1 x2;
  pid=_n_;
  do i=1 to 2;
    t=i-1;
      hi_modifie=x(i);
  output;
  end;
run;

data one;
  set one;
   res=sqrt(0.01)*rannor(17);
   hi_modifie_rep0=hi_modifie-res/2;
   hi_modifie_rep1=hi_modifie+res/2;
run;
```

```
/***Discrétisation***/

data bd;
set one;

 if hi_modifie_rep0<0 then hi_modifie_rep0=0;
 else
   if mod(hi_modifie_rep0,1)>0.5 then
        hi_modifie_rep0=round(hi_modifie_rep0,1)-1;
   else hi_modifie_rep0=round(hi_modifie_rep0,1);

 if hi_modifie_rep1<0 then hi_modifie_rep1=0;
 else
   if mod(hi_modifie_rep1,1)>0.5 then
        hi_modifie_rep1=round(hi_modifie_rep1,1)-1;
   else hi_modifie_rep1=round(hi_modifie_rep1,1);
run;

/***Vérification des données et moyenne géométrique***/

data bd_etude;
 set bd;

       /*Identifier les rep ayant + que 1 titres d'écart*/
   if abs(hi_modifie_rep0-hi_modifie_rep1)>1 then hi_modifie=-1;

       /*Moy Arithmétique des titres modifiés ~ Moy géo des HI*/
   else hi_modifie=(hi_modifie_rep0+hi_modifie_rep1)/2;
run;

/*Lors de cette simulation aucune donné ne possédait une répétition avec un
écart plus grand que 1 et le titre maximum présent est 12*/
```

```
/**********************Analyse des données simulées**********************/


/*Méthode intervalle*/

proc nlmixed data = bd_etude;
 parms b0=2 c0=1.8 b1=4.3 c1=0.8 var1=2 ;
  mu = b0+b1*t + eta0 ;
  k = sqrt(c0 + c1*t);
   if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
   else if hi_modifie=12 then like=1-CDF('NORMAL',12,mu,k);
        else
        like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	2.1704	0.04693	999	46.25	<.0001	0.05	2.0784	2.2625	-0.00038
c0	0.6680	0.08500	999	7.86	<.0001	0.05	0.5012	0.8348	-0.0009
b1	4.4077	0.05278	999	83.50	<.0001	0.05	4.3042	4.5113	7.474E-6
c1	1.1771	0.1592	999	7.39	<.0001	0.05	0.8647	1.4896	-0.00032
var1	1.3447	0.09576	999	14.04	<.0001	0.05	1.1568	1.5327	0.000313

```
/*Méthode des points milieux*/

proc nlmixed data = bd_etude;
 parms b0=2 c0=1.8 b1=4.3 c1=0.8 var1=2;
  mu = b0+b1*t + eta0 ;
  k = sqrt(c0 + c1*t);
   if hi_modifie=0 then like=PDF('NORMAL',0,mu,k);
   else like=PDF('NORMAL',hi_modifie+log2(1.5),mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	2.2151	0.04676	999	47.37	<.0001	0.05	2.1234	2.3069	-0.00085
c0	0.8324	0.08199	999	10.15	<.0001	0.05	0.6715	0.9933	-0.00115

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b1	4.4478	0.05252	999	84.69	<.0001	0.05	4.3448	4.5509	-0.00036
c1	1.0931	0.1541	999	7.10	<.0001	0.05	0.7908	1.3955	-0.00073
var1	1.3539	0.09488	999	14.27	<.0001	0.05	1.1677	1.5401	0.000492

```
/*Equivalent à méthode actuelle*/

proc nlmixed data = bd_etude;
 parms b0=2 c0=1.8 b1=4.3 c1=0.8 var1=2;
  mu = b0+b1*t + eta0 ;
  k = sqrt(c0 + c1*t);
  like=PDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
b0	1.7460	0.04156	999	42.01	<.0001	0.05	1.6644	1.8275	-0.0082
c0	0.5309	0.06926	999	7.67	<.0001	0.05	0.3950	0.6668	-0.00153
b1	4.3320	0.05112	999	84.73	<.0001	0.05	4.2317	4.4323	0.001095
c1	1.5519	0.1475	999	10.52	<.0001	0.05	1.2625	1.8413	-0.00211
var1	1.1966	0.08424	999	14.20	<.0001	0.05	1.0313	1.3619	0.000832

Annexe 10 : Programme et certaines sorties SAS pour réanalyse d'une base de données réelle.

```
/******Réanalyse d'une étude à l'aide de l'approche sur les moyennes******/


options nodate nonumber nocenter;
libname BD 'C:\Users\Darkflow\Documents\cours\Master Stat\Stage M2
Stat\Analysis of HI data\Stage Jérémy';

data bd;
 set BD./*Confidentiel*/;
run;


/*Transformation de la base de données*/

data bd;
 set bd;
  if NUM_RES>0 then do;
   hi_modifie=round(log2(NUM_RES/5),0.1);
  end;
  else do;
   hi_modifie=-2;
  end;
run;

proc sql;
 create table bd_etudes as
 select pid,group_nb,timing as day,hi_modifie from bd
 where timing='PRE' or timing='PI(D21)';
quit;

proc sql;
 create table bd_etude as
 select pid,group_nb,day,hi_modifie from bd_etudes
 where hi_modifie>-1 and pid ne 146;
quit;

data bd_etude;
 set bd_etude;
  if group_nb=2 then group_nb=0;
  if day='PRE' then day_etude=0;
  else day_etude=1;
run;

proc sort data=bd_etude;
 by group_nb pid day_etude;
run;


/*Création des différentes tables par groupe et jour*/

proc sql;
 create table gr1 as
 select * from bd_etude
 where group_nb=0;
quit;
```

```
proc sql;
 create table gr2 as
 select * from bd_etude
 where group_nb=1;
quit;

proc sql;
 create table d0tr1 as
 select * from bd_etude
 where day_etude=0 and group_nb=0;
quit;

proc sql;
 create table d21tr1 as
 select * from bd_etude
 where day_etude=1 and group_nb=0;
quit;

proc sql;
 create table d0tr2 as
 select * from bd_etude
 where day_etude=0 and group_nb=1;
quit;

proc sql;
 create table d21tr2 as
 select * from bd_etude
 where day_etude=1 and group_nb=1;
quit;


/*****Analyse du groupe1 suivant les 4 méthodes*****/

/*Méthode développée*/

proc nlmixed data = gr1;
 parms a0=-0.3 b0=2 c0=2 b1=4.5 c1=-1.5 var1=1 ;
  p0= exp(a0)/(1+exp(a0));
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
  if day_etude=0 then
    if hi_modifie=0 then like=p0+(1-p0)*CDF('NORMAL',1,mu,k);
    else
like=(1-p0)*(CDF('NORMAL',hi_modifie+1,mu,k)CDF('NORMAL',hi_modifie,mu,k));
  else
    if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
    else like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;


/*Méthode Nauta, point milieu*/

proc nlmixed data = gr1;
 parms a0=-1.0471 b0=1.2315 c0=0.6 b1=6.7608 c1=0 var1=2.1573 ;
  p0= exp(a0)/(1+exp(a0));
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
   if day_etude=0 then
```

```
     if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
     else if mod(hi_modifie,1)=0 then
               like=(1-p0)*PDF('NORMAL',hi_modifie+log2(1.5),mu,k);
          else like=(1-p0)*PDF('NORMAL',hi_modifie-0.5+log2(2.5),mu,k);
    else
     if hi_modifie=0 then like=PDF('NORMAL',0,mu,k);
     else if mod(hi_modifie,1)=0 then
                like=PDF('NORMAL',hi_modifie+log2(1.5),mu,k);
          else like=PDF('NORMAL',hi_modifie-0.5+log2(2.5),mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;


/*Méthode GSK avec excès en zéro*/

proc nlmixed data = gr1;
 parms a0=-1.0471 b0=1.2315 c0=0.6 b1=6.7608 c1=0 var1=2.1573 ;
  p0= exp(a0)/(1+exp(a0));
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
   if day_etude=0 then
     if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
     else like=(1-p0)*PDF('NORMAL',hi_modifie,mu,k);
   else like=PDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;


/*Méthode GSK sans excès en zéro*/

proc nlmixed data = gr1;
 parms b0=1.2315 c0=0.6 b1=6.7608 c1=0 var1=2.1573 ;
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
  like=PDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;


/*****Analyse du groupe2 suivant les 4 méthodes*****/

/*Méthode développée*/

proc nlmixed data = gr2;
 parms a0=-0.5 b0=2.2315 c0=0.6 b1=4.7608 c1=0 var1=1 ;
  p0= exp(a0)/(1+exp(a0));
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
   if day_etude=0 then
     if hi_modifie=0 then like=p0+(1-p0)*CDF('NORMAL',1,mu,k);
     else
like=(1-p0)*(CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k));
   else
     if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
     else
        like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
```

```
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;


/*Méthode Nauta, point milieu*/

proc nlmixed data = gr2;
 parms a0=-1.0471 b0=1.2315 c0=0.6 b1=6.7608 c1=0 var1=2.1573 ;
  p0= exp(a0)/(1+exp(a0));
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
   if day_etude=0 then
     if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
     else if mod(hi_modifie,1)=0 then
             like=(1-p0)*PDF('NORMAL',hi_modifie+log2(1.5),mu,k);
          else like=(1-p0)*PDF('NORMAL',hi_modifie-0.5+log2(2.5),mu,k);
   else
     if hi_modifie=0 then like=PDF('NORMAL',0,mu,k);
     else if mod(hi_modifie,1)=0 then
             like=PDF('NORMAL',hi_modifie+log2(1.5),mu,k);
          else like=PDF('NORMAL',hi_modifie-0.5+log2(2.5),mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;


/*Méthode GSK avec excès en zéro*/

proc nlmixed data = gr2;
 parms a0=-1.0471 b0=1.2315 c0=0.6 b1=6.7608 c1=0 var1=2.1573 ;
  p0= exp(a0)/(1+exp(a0));
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
   if day_etude=0 then
     if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
     else like=(1-p0)*PDF('NORMAL',hi_modifie,mu,k);
   else
     like=PDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;


/*Méthode GSK sans excès en zéro*/

proc nlmixed data = gr2;
 parms b0=1.2315 c0=0.6 b1=6.7608 c1=0 var1=2.1573 ;
  mu = b0+b1*day_etude + eta0 ;
  k = sqrt(c0 + c1*day_etude);
  like=PDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

```
/*****Analyse de la base de données complète suivant les 4 méthodes*****/

/*Méthode développée*/

proc nlmixed data = bd_etude;
 parms a0=-0.3 a1=-0.9 b0=1.8 b1=-0.5 b2=4.9 b3=0.2 c0=2.5 c1=0.7 c2=-1.5
       c3=-0.8 var1=1.2;
  p0= exp(a0 + a1*GROUP_NB)/(1+exp(a0 + a1*GROUP_NB));
  mu = b0 + b1*GROUP_NB + b2*day_etude + b3*GROUP_NB*day_etude + eta0;
  k = sqrt(c0 + c1*GROUP_NB + c2*day_etude + c3*GROUP_NB*day_etude);
   if day_etude=0 then
     if hi_modifie=0 then like=p0+(1-p0)*CDF('NORMAL',1,mu,k);
     else
like=(1-p0)*(CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k));
   else
     if hi_modifie=0 then like=CDF('NORMAL',1,mu,k);
     else
        like=CDF('NORMAL',hi_modifie+1,mu,k)-CDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a0	-0.1752	0.3957	332	-0.44	0.6582	0.05	-0.9535	0.6031	0.016152
a1	-0.8360	1.2142	332	-0.69	0.4916	0.05	-3.2245	1.5526	0.011065
b0	1.9118	0.4766	332	4.01	<.0001	0.05	0.9742	2.8494	0.000959
b1	-0.4367	0.9538	332	-0.46	0.6474	0.05	-2.3130	1.4396	-0.007
b2	4.8466	0.4745	332	10.21	<.0001	0.05	3.9132	5.7800	-0.00302
b3	0.2354	0.9525	332	0.25	0.8050	0.05	-1.6383	2.1090	0.005712
c0	2.3605	1.0206	332	2.31	0.0213	0.05	0.3529	4.3682	-0.01232
c1	0.6709	1.9163	332	0.35	0.7265	0.05	-3.0987	4.4404	0.013928
c2	-1.4364	1.1297	332	-1.27	0.2044	0.05	-3.6586	0.7858	0.014273
c3	-0.5520	2.0076	332	-0.27	0.7835	0.05	-4.5012	3.3973	-0.01612
var1	1.2231	0.2530	332	4.83	<.0001	0.05	0.7255	1.7207	0.000139

```
/*Méthode Nauta, point milieu*/

proc nlmixed data = bd_etude;
 parms a0=-0.3 a1=-0.9 b0=1.8 b1=-0.5 b2=4.9 b3=0.2 c0=2.5 c1=0.7 c2=-1.5
       c3=-0.8 var1=1.2;
  p0= exp(a0 + a1*GROUP_NB)/(1+exp(a0 + a1*GROUP_NB));
  mu = b0 + b1*GROUP_NB + b2*day_etude + b3*GROUP_NB*day_etude + eta0;
  k = sqrt(c0 + c1*GROUP_NB + c2*day_etude + c3*GROUP_NB*day_etude);
   if day_etude=0 then
     if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
     else like=(1-p0)*PDF('NORMAL',hi_modifie+log2(1.5),mu,k);
   else
     if hi_modifie=0 then like=PDF('NORMAL',0,mu,k);
     else like=PDF('NORMAL',hi_modifie+log2(1.5),mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a0	0.4114	0.1683	332	2.44	0.0151	0.05	0.08024	0.7425	-0.00001
a1	-0.2229	0.2373	332	-0.94	0.3484	0.05	-0.6897	0.2440	-6.01E-6
b0	2.7010	0.1909	332	14.15	<.0001	0.05	2.3254	3.0766	-0.00014
b1	0.02673	0.2613	332	0.10	0.9186	0.05	-0.4873	0.5408	-0.00004
b2	4.1441	0.1998	332	20.74	<.0001	0.05	3.7511	4.5371	-0.00014
b3	-0.2296	0.2775	332	-0.83	0.4086	0.05	-0.7754	0.3162	-0.00004
c0	1.3389	0.3492	332	3.83	0.0002	0.05	0.6521	2.0258	0.000234
c1	-0.07641	0.4665	332	-0.16	0.8700	0.05	-0.9940	0.8412	0.000094
c2	0.03054	0.4299	332	0.07	0.9434	0.05	-0.8151	0.8762	0.000126
c3	0.3675	0.6009	332	0.61	0.5412	0.05	-0.8145	1.5495	0.000056
var1	0.7931	0.1929	332	4.11	<.0001	0.05	0.4136	1.1725	0.000134

```
/*Méthode GSK avec excès en zéro*/

proc nlmixed data = bd_etude;
 parms a0=-0.3 a1=-0.9 b0=1.8 b1=-0.5 b2=4.9 b3=0.2 c0=2.5 c1=0.7 c2=-1.5
       c3=-0.8 var1=1.2;
  p0= exp(a0 + a1*GROUP_NB)/(1+exp(a0 + a1*GROUP_NB));
  mu = b0 + b1*GROUP_NB + b2*day_etude + b3*GROUP_NB*day_etude + eta0;
  k = sqrt(c0 + c1*GROUP_NB + c2*day_etude + c3*GROUP_NB*day_etude);
   if day_etude=0 then
     if hi_modifie=0 then like=p0+(1-p0)*PDF('NORMAL',0,mu,k);
     else like=(1-p0)*PDF('NORMAL',hi_modifie,mu,k);
   else like=PDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a0	0.3197	0.1761	332	1.81	0.0704	0.05	-0.02680	0.6662	-0.0001
a1	-0.2349	0.2502	332	-0.94	0.3485	0.05	-0.7270	0.2573	0.000748
b0	2.0480	0.1890	332	10.83	<.0001	0.05	1.6761	2.4198	-0.00407
b1	0.02706	0.2600	332	0.10	0.9172	0.05	-0.4844	0.5385	-0.00003
b2	4.2119	0.1979	332	21.28	<.0001	0.05	3.8226	4.6013	-0.00408
b3	-0.2297	0.2760	332	-0.83	0.4058	0.05	-0.7727	0.3132	0.000846
c0	1.3351	0.3354	332	3.98	<.0001	0.05	0.6754	1.9949	-0.00169
c1	-0.07579	0.4458	332	-0.17	0.8651	0.05	-0.9528	0.8012	0.00023
c2	0.02944	0.4126	332	0.07	0.9432	0.05	-0.7822	0.8411	0.000718
c3	0.3524	0.5794	332	0.61	0.5435	0.05	-0.7874	1.4922	0.001164
var1	0.7903	0.1839	332	4.30	<.0001	0.05	0.4285	1.1520	-0.00225

```
/*Méthode GSK sans excès en zéro*/

proc nlmixed data = bd_etude;
 parms b0=1 b1=-0.4 b2=5.7 b3=0 c0=1.6 c1=0.18 c2=-0.58 c3=0 var1=1.2;
  mu = b0 + b1*GROUP_NB + b2*day_etude + b3*GROUP_NB*day_etude + eta0;
  k = sqrt(c0 + c1*GROUP_NB + c2*day_etude + c3*GROUP_NB*day_etude);
  like=PDF('NORMAL',hi_modifie,mu,k);
  ll=log(like);
 model hi_modifie ~ general(ll);
 random eta0~ normal(0, var1) subject=PID;
run;
```

NOTE: GCONV convergence criterion satisfied.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
b0	0.8802	0.1108	332	7.95	<.0001	0.05	0.6624	1.0981	0.002168
b1	0.1710	0.1581	332	1.08	0.2802	0.05	-0.1400	0.4819	0.001737
b2	5.3771	0.1351	332	39.81	<.0001	0.05	5.1114	5.6428	0.000244
b3	-0.3711	0.1939	332	-1.91	0.0565	0.05	-0.7525	0.01038	0.000519
c0	1.4200	0.2127	332	6.68	<.0001	0.05	1.0016	1.8385	0.000051
c1	0.06276	0.2976	332	0.21	0.8331	0.05	-0.5227	0.6483	0.000095
c2	0.1814	0.3225	332	0.56	0.5742	0.05	-0.4531	0.8159	-0.00025
c3	0.06661	0.4604	332	0.14	0.8851	0.05	-0.8391	0.9723	-0.00022
var1	0.6286	0.1253	332	5.02	<.0001	0.05	0.3821	0.8751	-0.00043

Annexe 11 : Références.

- **Jos Nauta,**
"Statistics in Clinical Vaccine Trials", Springer, October 25, 2010

- **Kaifeng Lu,**
"On efficiency of Constrained Longitudinal Data Analysis versus Longitudinal Analysis of Covariance", Biometrics 66, 891-896, September 2010

- **Peter L. Bonate, Crystal Sung, Karen Welch and Susan Richard,**
"Conditional modeling of Antibody titers using a zero-inflated random effects model : Application to Fabrazyme", Journal of Pharmacokinetics and Pharmacodynamics, 2009, 36:443-459.

- Rodolphe Thiébaut, Hélène Jacqmin-Gadda,
"Mixed models for longitudinal left-censored repeated measures", INSERM E0338, ISPED, Université Victor Segalen Bordeaux 2.

- V.Shankar, J.Milton and F.Mannering,
"Modeling accident frequencies as zero-altered probability processes : An empirical inquiry", Accid. Anal. And Prev., Vol. 29, N° 6, pp. 829-837, 1997.

- Ane Nodtvedt, Ian Dohoo, Javier Sanchez, Gary Conboy, Luc DesCôteaux, Greg Keefe, Ken Leslie and John Campbell,
"The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows", The Canadian Journal of Veterinary Research, 66:249-257, 2002.