Critères d'évaluation de la validité des estimateurs d'intervalle de confiance

André Gillibert

Thèse de doctorat

NNT : 2021UPASR012

université PARIS-SACLAY

Critères d'évaluation de la validité des estimateurs d'intervalle de confiance

Assessment criteria for confidence interval estimators

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°570, santé publique (EDSP)
Spécialité de doctorat : santé publique - biostatistiques
Unité de recherche : Université Paris-Saclay, UVSQ, Inserm, CESP, 94807, Villejuif, France.
Référent : Faculté de médecine

Thèse présentée et soutenue à Paris-Saclay, le 15/09/2021, par

André GILLIBERT

Composition du Jury

Jean-Christophe THALABARD Professeur, MAP5, UMR CNRS 8145, Université Paris Descartes, USPC, Paris, France	Président
Sylvie CHEVRET Professeure, Service de Biostatistiques et Information Médicale (SBIM), Hôpital Saint Louis, APHP, Université Paris, Paris, France	Rapporteur & Examinatrice
Hélène JACQMIN-GADDA Docteure, HDR, Université de Bordeaux, Inserm, Population Health U1219, ISPED, Bordeaux, France	Rapporteur & Examinatrice

Direction de la thèse

Jacques BÉNICHOU Professeur, Normandie Univ, UNIROUEN, CHU de Rouen, Unité de Biostatistique, Rouen	Directeur de thèse
Bruno FALISSARD Professeur, INSERM UMR 1178, Université Paris Sud, Maison de Solenn, Paris, France	Co-Directeur de thèse

Valorisation scientifique

Trois articles directement issus de ce travail de thèse ont été publiés :

1. Gillibert A, Bénichou J, Falissard B. Two-sided confidence interval of a binomial proportion: how to choose? arXiv:210310463 [stat] [Internet]. 17 mars 2021 [cité 22 mars 2021]; Disponible sur: http://arxiv.org/abs/2103.10463

2. Gillibert A, Bénichou J, Falissard B. The case for balanced hypothesis tests and equal-tailed confidence intervals. arXiv:210312581 [stat] [Internet]. 22 mars 2021 [cité 24 mars 2021]; Disponible sur: http://arxiv.org/abs/2103.12581

3. Gillibert A, Bénichou J, Falissard B. Best estimator for bivariate Poisson regression. arXiv:210310365 [stat] [Internet]. 17 mars 2021 [cité 22 mars 2021]; Disponible sur: http://arxiv.org/abs/2103.10365

Table des matières

1 Généralités

Cette thèse s'intéresse aux critères d'évaluation de la validité des estimateurs d'intervalles de confiance. Nous allons, dans l'introduction de ce travail, définir les concepts nécessaires à la compréhension des outils analysés et utilisés. Nous allons ensuite introduire le contexte dans lequel s'inscrit ce travail ainsi que son objectif. Puis les articles de cette thèse seront brièvement présentés et joints. Enfin, les conclusions tirées de l'ensemble de ces articles seront discutées.

1.1 Théorie utilisée

Nous nous plaçons dans le contexte de la théorie fréquentiste, qui diffère de la théorie bayésienne dans l'objet de l'aléatoire. En théorie fréquentiste, les paramètres des modèles sont considérés fixes alors que l'échantillon ainsi que les statistiques calculées sur celui-ci sont aléatoires et décrits par des distributions. En théorie bayésienne, à l'inverse, l'échantillon est considéré comme une observation figée alors que les paramètres des modèles sont décrits par des distributions reflétant l'incertitude les entourant.
La théorie fréquentiste reste, de loin, la plus largement utilisée dans les publications scientifiques médicales et c'est pourquoi nous nous restreindrons à cette théorie.

1.2 Concepts généraux de la discipline statistique

1.2.1 Construction de ce chapitre

L'ensemble des concepts présentés dans ce chapitre est bien connu, mais il fut porté une attention particulière à la cohérence des notations, l'ordre dans lequel les définitions sont introduites et enfin, à dégager des définitions en fournissant non seulement la définition mathématique, mais aussi les contre-exemples et cas limites permettant de mieux appréhender ce qui sort du champ de la définition. La plupart des contre-exemples sont l'invention originale de l'auteur de ce travail mais il est probable que d'autres auteurs aient déjà fournis des contre-exemples similaires voire identiques.

1.2.2 Espace de probabilité

1.2.2.1 Définition mathématique moderne

Un espace de probabilité est un triplet $(\Omega, \mathcal{A}, \mathbb{P})$ formé d'un ensemble Ω (univers), une tribu $\mathcal{A}$ sur Ω

(ensemble des événements, stable par passage au complémentaire, intersection et union dénombrable) et une mesure $\mathbb{P}$ (mesure de probabilité) telle que $\mathbb{P}(\Omega)=1$. Par exemple, dans un contexte dans lequel cent variables aléatoires absolument continues à valeurs dans $\mathbb{R}$ sont définies comme $X = (X_1, \dots, X_{100})$, sans faire l'hypothèse qu'elles soient indépendantes ni identiquement distribuées, on pourra définir Ω comme $\mathbb{R}^{100}$, $\mathcal{A}$ comme l'ensemble des parties Lebesgue-mesurables de Ω et $\mathbb{P}$ comme l'intégrale selon Lebesgue de la fonction de densité de probabilité sur la partie considérée. Les événements définis dans cet espace de probabilité sont alors l'appartenance de X à une partie mesurable de $\mathbb{R}^{100}$.

Comme la réunion et l'intersection de deux éléments d'une tribu appartient à la tribu, alors la réunion ou l'intersection de deux événements sont aussi des événements. Il existe l'événement vide, de probabilité $\mathbb{P}(\emptyset)=0$ et l'événement univers $\mathbb{P}(\Omega)=1$.

1.2.2.2 Définition intuitive, fréquentiste

Lors de la réalisation d'une expérience aléatoire telle que le tirage au sort de 100 sujets dans la population des résidents français réguliers, alors plusieurs événements peuvent apparaître, de manière aléatoire. Par exemple, le premier sujet peut avoir un poids compris entre 60 et 61 kg, le second sujet peut avoir une taille supérieure à 174 cm et le poids moyen des 100 sujets peut être comprise entre 65 et 75 kg. À chaque événement, on peut associer une probabilité d'occurrence reflétant la fréquence à laquelle cet événement surviendrait si l'expérience était répétée un très grand nombre de fois. Les événements peuvent se combiner par des opérateurs logiques ET, OU et NON, conduisant à de nouveaux événements tel que « 5ème sujet de 60 à 61 kg et poids moyen supérieur à 54 kg ». Dans le même espace de probabilité, on peut donc définir les probabilités associées à des variables individuelles ou des statistiques collectives.

Certaines propriétés sont garanties :

Probabilité(A ou B)=Probabilité(A)+Probabilité(B)−Probabilité(A et B)

Probabilité(non A)=1−Probabilité(A)

$0 \leq$ Probabilité(A) ≤ 1

1.2.3 Variable aléatoire

1.2.3.1 Définition mathématique moderne

Une variable aléatoire à valeurs dans un ensemble S, est une fonction de l'ensemble des événements ($\mathcal{A}$) vers S.

1.2.3.2 Définition intuitive et notations

Une variable aléatoire est une valeur que l'on peut associer au résultat d'une expérience individuelle ou collective. Une variable aléatoire représente habituellement les caractéristiques d'une entité telles que l'âge d'un patient, le nombre de lits d'un établissement de santé, ou l'aire d'une région en kilomètres carrés. Elle peut aussi représenter les caractéristiques d'un échantillon tel que l'âge moyen des sujets d'un échantillon.

On distingue la variable en elle-même de sa réalisation ; cette première est définie dans le cadre d'une expérience reproductible, alors que cette seconde est définie après que l'expérience ait été réalisée et ne comporte plus qu'une seule valeur.

Par exemple, la taille d'un sujet français tiré aléatoirement dans les listes électorales en mai 2020 est une variable aléatoire alors que 168 cm est une réalisation possible de cette variable aléatoire.

On notera les variables aléatoires par des lettres majuscules telle que X ou Θ et les réalisations des variables aléatoires par des lettres minuscules telles que x ou θ.

Nous décrirons non seulement des variables aléatoires à valeurs dans $\mathbb{R}$, mais aussi des variables aléatoires à valeurs dans $\mathbb{R}^n$ ou dans le groupe des matrices $M_{m,n}(\mathbb{R})$. Cela nous permettra notamment de décrire en une seule variable, toutes les caractéristiques de tous les sujets d'un échantillon, comme le poids, la taille et l'âge du premier au centième sujet d'un échantillon aléatoire de cent sujets ; cela est une alternative à la présentation des trois cents variables aléatoires (3 variables / sujet × 100 sujets).

1.2.4 Statistique

1.2.4.1 Définitions mathématiques

L'étymologie de statistique provient de *status*, état en latin. Son origine a d'abord été employée dans le contexte de statistiques de recensement, telles que le nombre de citoyens, la composition des foyers, la démographie. Nous distinguerons la discipline Statistique avec un grand S, d'une statistique qui est un

outil de la discipline.

De nombreuses définitions différentes d'une statistique existent et nous citerons celles qui nous paraissent cohérentes avec les définitions mathématiques des sections précédentes.

Une statistique est une variable aléatoire qui représente une propriété d'un échantillon ou d'une population. Ainsi, elle n'a pas de distinction mathématique réelle de variable aléatoire. Sa spécificité est de s'appliquer à un échantillon ou une population plutôt qu'à une entité unique telle qu'un individu. Il faut noter que même si une population est infinie on peut lui associer un espace de probabilité de dimension infinie (Ω) et ainsi, y associer des événements (dans la tribu $\mathcal{A}$) et des variables aléatoires ou des statistiques. Dans la suite de ce document, nous privilégierons cette première définition de statistique.

Une autre définition est celle d'une fonction d'un échantillon ou d'une population, souvent exprimé comme une variable aléatoire à valeurs vectorielles dans un espace de dimension n, pour un échantillon de taille n, ou dans un espace de dimension infini pour une population infinie.

Ainsi, si l'échantillon est représenté par une variable aléatoire $X = (X_1, \dots, X_n)$, la statistique T associera à n'importe quelle réalisation x de X une valeur $T(x)$. On pourra noter $T(X)$ la variable aléatoire représentant l'application de la statistique à X.

1.2.4.2 Définition intuitive

Une statistique est un résumé numérique des variables aléatoires d'un échantillon ou d'une population.

Une statistique univariée peut s'écrire comme une fonction d'une seule variable aléatoire alors qu'une statistique bivariée fait intervenir deux variables et une statistique multivariée fait intervenir au moins trois variables. Par exemple, l'âge moyen des sujets d'un échantillon est une statistique, de même que la taille minimale dans une population. La différence de moyenne d'âge entre les sujets de sexe masculin et les sujets de sexe féminin dans un échantillon est une statistique bivariée.

Une statistique peut être unidimensionnelle ($\mathbb{R}$) ou multidimensionnelle ($\mathbb{R}^n$) selon que ce résumé numérique est unique ou multiple.

Le terme de statistique, selon le contexte, peut faire référence à la variable aléatoire elle-même ou à sa réalisation.

1.2.5 Support

1.2.5.1 Définition mathématique

Le support d'une variable aléatoire X est le plus petit ensemble topologiquement fermé S tel que la probabilité que X appartienne à S soit égale à 1. Une autre formulation, c'est de dire qu'il s'agit du complémentaire de la réunion de tous les ouverts O tels que $\mathbb{P}(X \in O) = 0$, c'est-à-dire, encore l'intersection de fous les fermés F tels que $\mathbb{P}(X \in F) = 1$.

1.2.5.2 Définition intuitive

Le support d'une variable aléatoire X est l'ensemble des valeurs possibles de cette variable. Pour une variable discrète, il s'agit de l'ensemble des valeurs de probabilité non nulle alors que pour une variable absolument continue, il s'agit de l'ensemble des valeurs dont la densité de probabilité est non nulle.

1.2.6 Distribution

Une *distribution*, ou *loi de probabilité* est une fonction qui associe à tout événement $\mathrm{A} \in \mathcal{A}$ une probabilité $\mathbb{P}(\mathrm{A})$. Généralement, on parle de distribution *d'une variable* et on se restreint ainsi aux événements concernant l'appartenance de la variable à un ensemble mesurable. Pour une variable X à valeurs dans S, on pourra alors décrire la *distribution* comme une fonction de l'ensemble des parties de S dans $[0,1]$ associant à tout $s \subset S$, $\mathbb{P}(X \in s)$.

On remarquera que l'ensemble S n'est pas directement doté d'une mesure puisque la mesure de probabilité s'applique à des événements d'appartenance de X à des parties de S.

Cependant, comme X est une fonction de l'ensemble $\mathcal{A}$ des événements vers S (cf. section 1.2.3.1), l'événement d'appartenance de X à une partie s de S est définie comme la réunion des événements qui conduisent à l'appartenance de X à s:

$$\bigcup \{A \in \mathcal{A} | X(A) \in s\}$$

On peut alors définir une fonction $e: S \to \mathcal{A}$ qui à tout $s \in S$ associe cet événement d'appartenance. Même si la notion de mesurabilité ne s'applique pas directement à s, elle s'appliquera à l'image de s par cette fonction. Considérant la définition d'une distribution comme une fonction associant à s la probabilité $\mathbb{P}(X \in s)$, alors on constate que plusieurs variables distinctes, s'appliquant à un même

espace probabilisé, peuvent avoir une distribution égale bien que leur fonction e diffère.

De la distribution d'une variable, on peut souvent dériver la *fonction de répartition* (synonyme *fonction cumulative de distribution*) ainsi que, selon la nature de la variable, la *fonction de masse* ou la *fonction de densité de probabilité*.

1.2.7 Échantillon

De manière intuitive, on décrit un échantillon comme l'ensemble des observations, avec les variables aléatoires associées, obtenues lors de la réalisation d'une expérience aléatoire. Même s'il peut exister plusieurs variables aléatoires associées à chaque observation, on peut les résumer en une seule variable aléatoire sous forme de n-uplet, voire de suite ou de fonction s'il y a une infinité dénombrable ou indénombrable de variables aléatoires par observation. C'est pourquoi les caractéristiques de l'observation numéro i pourront être résumée en une variable aléatoire X_i.

Un échantillon fini réalisé peut être défini comme la suite finie des x_i.

Lorsque l'échantillon est de taille fixe n, alors un échantillon réalisé est décrit par $x = (x_1, \dots, x_n)$ et on peut définir une variable aléatoire unique décrivant l'expérience aléatoire de génération d'échantillon comme $X = (X_1, \dots, X_n)$. Il faut noter que les X_i ne sont pas forcément indépendants ni identiquement distribués.

Dans un grand nombre de problèmes de modélisation statistique, l'échantillon est considéré comme étant de taille fixe, conditionnant l'ensemble des calculs de probabilité à la taille d'échantillon.

Si l'échantillon est fini et de taille variable, avec des observations dont les réalisations x_i appartiennent à un ensemble unique S, alors on peut toujours le modéliser comme une unique variable aléatoire X prenant ses valeurs dans l'ensemble E des suites finies à valeurs dans S.

$$E = \bigcup_{n \in \mathbb{N}} S^n$$

On peut toujours analyser les propriétés d'une statistique T s'appliquant aux éléments de E.

Dans le cas des analyses intermédiaires, la variabilité de la taille d'échantillon ne peut être ignorée. Tous les tests statistiques usuels deviennent erronés s'ils ne sont pas adaptés.

1.2.8 Fonction de répartition

Pour une variable X à valeurs dans $\mathbb{R}$, on définit la fonction de répartition F : $\mathbb{R} \to \mathbb{R}$ comme la fonction qui a une valeur $x \in \mathbb{R}$ associe la probabilité que X soit inférieure ou égale à x : $F(x) = \mathbb{P}(X \leq x)$.

À partir de cette fonction on peut aisément calculer la probabilité que X soit compris entre deux valeurs quelconques x_1 et x_2 comme $\mathbb{P}(x_1 < X \leq x_2) = F(x_2) - F(x_1)$. Toute fonction de répartition pour un X à valeurs est croissante, a une limite nulle en $-\infty$ et une limite à 1 en $+\infty$.

Pour une variable $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ à valeurs dans $\mathbb{R}^n$ on définit la fonction de répartition F : $\mathbb{R}^n \to \mathbb{R}$ comme la fonction qui a une valeur $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ associe $\mathbb{P}(\forall i \in \{1, \dots, n\}, X_i \leq x_i)$, que l'on peut noter $\mathbb{P}(X \leq x)$ si on définit la relation d'ordre (non totale) par $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \leq \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ si et seulement si $\forall i \in \{0, \dots, n\}, x_i \leq y_i$.

Par soustraction de plusieurs valeurs de la fonction de répartition on peut alors calculer la probabilité que X appartienne à un pavé (produit cartésien d'intervalles bornés). En deux dimensions, $\mathbb{P}(a_1 < x_1 \leq b_1 \; et \; a_2 < x_2 < b_2) = F\left(\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right) - F\left(\begin{pmatrix} a_1 \\ b_2 \end{pmatrix}\right) - F\left(\begin{pmatrix} b_1 \\ a_2 \end{pmatrix}\right) + F\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}\right)$.

1.2.9 Variable continue

Une variable à valeurs dans $\mathbb{R}$ est dite continue si sa fonction de répartition est une fonction continue. Dans ce cas, pour tout x_0, $\mathbb{P}(X = x_0) = 0$. En effet, pour X réel, $\mathbb{P}(X = x_0) \leq \mathbb{P}(X \in [x_0, x_0 + \delta]) = F(x_0 + \delta) - F(x_0)$ pour tout $\delta \in \mathbb{R}_+$. Or, $\lim_{\delta \to 0} F(x_0 + \delta) - F(x_0) = 0$ par continuité de F. Donc $\mathbb{P}(X = x_0)$ est inférieur ou égal à cette limite, c'est-à-dire est inférieur ou égal à zéro. Comme une probabilité ne peut pas être négative, on en déduit $\mathbb{P}(X = x_0) = 0$.

La notion est aussi généralisable à $\mathbb{R}^n$, la fonction de répartition F devant alors être une fonction continue de la topologie usuelle de $\mathbb{R}^n$. La propriété $\mathbb{P}(X = x_0) = 0$ pour tout x_0 est aussi vérifiée.

1.2.10 Variable absolument continue

Une variable à valeurs dans $\mathbb{R}$ est absolument continue si sa fonction de répartition F est une fonction absolument continue. Cela est équivalent à équivaut à l'existence d'une fonction f Lebesgue-intégrable

avec pour tout a et tout x, $F(x) = F(a) + \int_a^x f(x)\mathrm{d}x$.

La fonction f est appelée *fonction de densité de probabilité* de X. Cette fonction est intéressante pour plusieurs calculs :

D'abord on peut calculer $\mathbb{P}(x \in [a,b]) = \int_a^b f(x)\mathrm{d}x$.

On peut aussi calculer l'espérance de X comme $\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x$.

Ci-dessous un exemple de fonction de répartition F et de fonction de densité de probabilité f associée à une variable absolument continue :

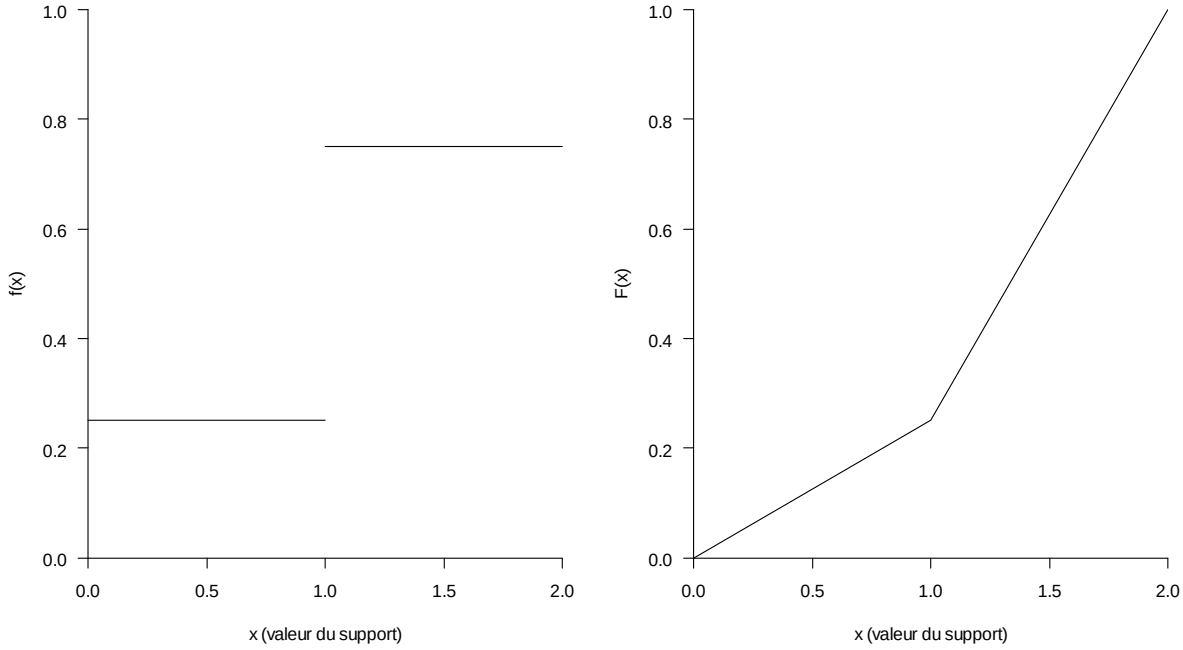


Figure 1 : exemple de fonction de répartition F et de fonction de densité de probabilité f associée à une variable absolument continue telle que F ne soit pas dérivable sur son domaine de définition complet [0,2]

La notion est généralisable à $\mathbb{R}^n$ par la fonction de densité de probabilité f telle que $F(x) = \int_{y\in\{y\leq x\}} f(y)\mathrm{d}y$ où la relation d'ordre dans $\mathbb{R}^n$ est définie par $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \leq \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ si et seulement si $\forall i \in \{0, \dots, n\}, y_i \leq x_i$.

1.2.11 Variable discrète

1.2.11.1 Définition

Une variable X est discrète si et seulement s'il existe un ensemble S dénombrable tel que $\mathbb{P}(X \in S) = 1$.

1.2.11.2 Exemple atypique

Le support n'est pas forcément un ensemble discret même si c'est le cas le plus typique. Considérons par exemple l'estimateur d'un incidence ratio, calculé comme le rapport entre deux dénombrements d'événements sur une période fixée. Afin d'éviter une division par zéro dans les cas limites, on ajoute 1 au numérateur et au dénominateur. Pour rendre le cas plus concret, considérons que les dénombrements d'événements suivent deux lois de Poisson indépendantes d'espérance respective 3 (numérateur) et 4 (dénominateur). L'ensemble des valeurs possibles de cette variable aléatoire d'incidence ratio est l'ensemble des nombres rationnels strictement positifs. Le support de l'estimateur est donc l'intervalle réel $[0, +\infty[$ selon la définition de support que nous avions donné en section 1.2.5. Avec cette définition et ce cas d'école, nous constatons qu'une variable peut être discrète mais avoir un support infini non dénombrable (à cause du support topologiquement fermé) et un ensemble de valeurs de probabilité non nulle qui, bien que dénombrable, n'est pas topologiquement discret mais bien au contraire, dense dans $\mathbb{R}^+$.

Ci-dessous une approximation de la fonction de masse ainsi que de la fonction de répartition de la distribution de l'incidence ratio décrit dans l'exemple ci-dessus :

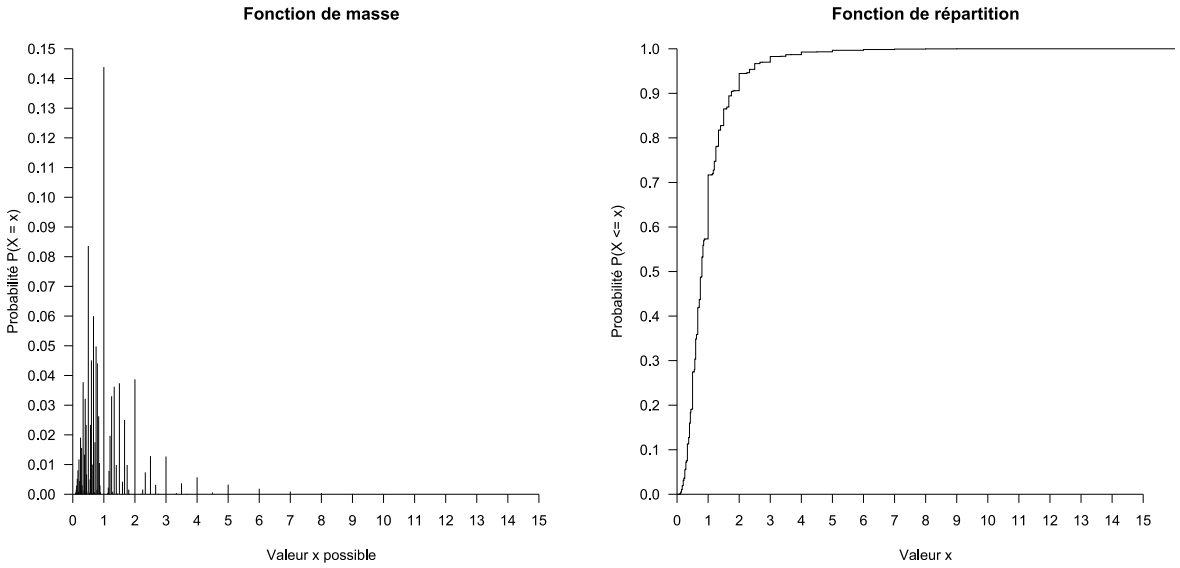


Figure 2 : fonction de masse et fonction de répartition d'une variable aléatoire discrète suivant un rapport de deux lois de Poisson d'espérances respectives 3 (numérateur) et 4 (dénominateur) après ajout de la valeur 1 au numérateur et au dénominateur.

La Figure 2 n'est pas tout à fait complète puisqu'elle ne peut décrire l'infinité des possibilités. Les valeurs de probabilité inférieures à 10^{-16} ont été supprimées de la fonction de masse conduisant à une probabilité totale des valeurs présentées sur la figure égale à $1-7.7\times10^{-16}$. La fonction de répartition a été construite afin de maintenir une erreur inférieure à 2.2×10^{-16} en tout point. Cette fonction de répartition n'est continue à gauche en aucun point rationnel strictement positif. En effet, comme $\mathbb{P}(X = x_0)$ est non nul, $F(x) \leq F(x_0) - \mathbb{P}(X = x_0)$ pour tout $x < x_0$ de telle sorte que même si $\lim_{x \to x_0, x<x_0} F(x)$ est défini, il est strictement inférieur à $F(x_0)$. La distribution n'est donc pas continue, comme toute distribution discrète, présentant forcément une discontinuité à chaque valeur de probabilité non nulle.

En pratique, on peut raisonnablement approximer un rapport de deux lois de Poisson à une distribution continue si les espérances des lois de Poisson sont suffisamment grandes et au contraire l'approximer à une variable discrète à support fini si les espérances sont suffisamment petites.

1.2.11.3 Propriétés élémentaires

La distribution d'une variable discrète peut être décrite par une *fonction de masse*, associant à toute

valeur de S sa probabilité, et zéro à toute autre valeur.

La somme finie ou infinie de ces probabilités sera égale à 1.

La famille de distributions discrètes de Poisson, par exemple, est paramétrée par une espérance λ. La fonction de masse de la loi de Poisson de paramètre λ associe à tout nombre entier la probabilité $\frac{e^{-\lambda}\lambda^k}{k!}$.

Ci-dessous cette fonction est représentée pour le paramètre λ = 1,3 :

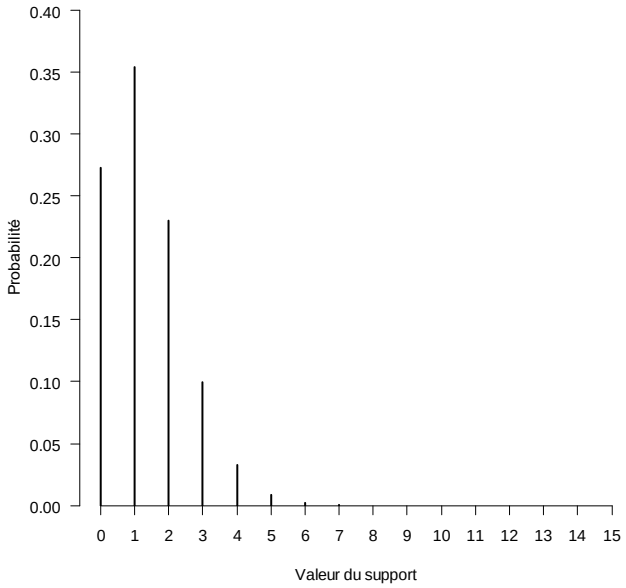


Figure 3 : représentation de la fonction de masse de la distribution de Poisson de paramètre λ=1,3 jusqu'à la valeur 15. Cette fonction est définie sur l'ensemble des entiers naturels (ℕ) tout entier mais est inférieure à 10^{-12} pour toute valeur du support supérieure ou égale à 16.

1.2.12 Distribution de mélange

Certaines distributions peuvent n'être ni discrètes ni continues, cela est notamment le cas de certaines distributions de mélange. Une variable aléatoire suit une distribution de mélange si elle peut se décrire par une succession de deux expériences : tirage au sort d'une loi de probabilité parmi plusieurs, puis réalisation d'une variable suivant la loi de probabilité sélectionnée.

On peut illustrer une distribution de mélange discret-continu par la variable aléatoire représentant la concentration en fluorodesoxyglucose marqué au fluor 18 (^{18}F-FDG), produit utilisé en tomographie par émissions de positons. Le fluor 18 est un isotope radioactif du fluor, produit dans un accélérateur de

particules. Il a une demi-vie de 110 minutes et donc, sans compter l'élimination urinaire, ne persiste jamais longtemps dans l'organisme. Considérant par exemple un patient de 70 kg absorbant 4 MBq/kg de ^{18}F-FDG, correspondant à 280 MBq soit 7.95×10^{-11} gramme ou 4.4×10^{-12} mole, ou encore 2.66×10^{12} molécules de ^{18}F. Au bout de 42 demi-vies il restera, en moyenne, moins d'une molécule dans l'organisme (en négligeant l'élimination rénale). En 4 jours soit 52 demi-vies, il est improbable qu'il reste la moindre trace. L'association particulière de ^{18}F au glucose paraît particulièrement improbable en dehors d'un usage médical. Le ^{18}F n'a pas non plus raison de se retrouver dans l'environnement en raison de sa faible demi-vie. Il paraît donc raisonnable d'imaginer que la plupart des individus ont exactement zéro molécule de ^{18}F-FDG dans leur organisme la grande majorité du temps. Pourtant certains sujets peuvent, durant une période, être exposés à des concentrations selon un continuum.

La distribution de cette variable dans une population générale se décrit donc par un pic à la valeur exacte zéro, de probabilité très élevée, suivie d'une queue de distribution continue. Sur $\mathbb{R}$, la fonction de répartition F a une discontinuité à la valeur zéro et il n'est pas possible de définir de fonction de densité de probabilité en la valeur zéro. En effet F n'est pas dérivable, avec $\lim_{x\to0,x<0}\frac{F(x)-F(0)}{x}=\lim_{x\to0,x<0}\frac{0-F(0)}{x}=+\infty$.

1.2.13 Famille de distribution

Une *famille de distribution* à valeurs dans S est une fonction associant à n'importe quel paramètre $\theta\in\Theta$ une *distribution* à valeurs dans S. Pour rappel, une *distribution* d'une variable aléatoire X à valeurs dans S est une fonction qui associe à tout sous-ensemble $T\subset S$ une mesure de la probabilité qu'une variable aléatoire X appartienne à T (cf. section 1.2.6). L'espace de paramétrisation Θ peut être n'importe quel ensemble, dont notamment des espaces vectoriels de dimension finie tels que $\mathbb{R}^n$.

Les familles de distribution représentent un concept central des statistiques paramétriques inférentielles, puisqu'on se placera généralement dans un contexte où une distribution inconnue est supposée appartenir à une famille de distribution connue, cherchant alors à estimer le paramètre θ de la distribution.

1.3 Concepts d'estimateur et estimation

1.3.1 Définition

Un estimateur est une statistique (variable aléatoire) d'un échantillon dont la caractéristique est d'approcher une statistique ou un paramètre de la population. L'estimateur est la variable aléatoire alors que l'estimation est sa réalisation.

Généralement l'expérience aléatoire consiste en le tirage au sort d'un échantillon de taille n d'observations identiquement distribuées, c'est-à-dire de même distribution inconditionnelle, et indépendantes, c'est-à-dire dont la distribution de chacune, conditionnelle à chacune autre est égale à la distribution inconditionnelle. On note $X = (X_1, \dots, X_n)$ les variables aléatoires correspondant aux n observations. On remarquera que le cas multivarié peut être simplement décrit par des X_i appartenant à $\mathbb{R}^k$. Un exemple d'estimateur, sur cet échantillon de taille n est celui de la moyenne $M = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Avec la définition que nous avons donné d'estimateur, l'estimateur de la moyenne pour un échantillon de taille n diffère de l'estimateur de la moyenne d'un échantillon de taille $n + 1$. Par extension, on peut définir un estimateur comme une fonction T associant à un échantillon X une variable aléatoire $T(X)$ dont la réalisation est l'estimation sur l'échantillon. Cela permet d'analyser les propriétés de l'estimateur en fonction de la taille d'échantillon n.

On notera que l'expérience aléatoire peut être plus complexe qu'une constitution d'un échantillon de taille n prédéfinie, notamment dans les méthodologies adaptatives (analyses intermédiaires) pour lesquelles la taille d'échantillon n est variable d'une expérience à l'autre. Nous ne nous concentrerons pas sur ces expériences adaptatives.

L'estimateur le plus simple d'une statistique définie dans une population est l'estimateur empirique. Il est obtenu en appliquant la même fonction statistique à l'échantillon qu'à la population. Si l'échantillon est exhaustif (contient toute la population) alors l'estimateur est constant et sa réalisation égale la statistique de la population. L'estimateur empirique a de bonnes propriétés pour certaines statistiques telle que la moyenne, mais est plus ou moins biaisé pour d'autres statistiques telles que le maximum.

On distingue les estimateurs ponctuels des estimateurs d'intervalle. Les premiers fournissent des

estimations à valeurs dans $\mathbb{R}$ ou $\mathbb{R}^n$ alors que les seconds ont des estimations égales à des intervalles. Ces estimateurs, par exemple, permettent la construction d'intervalles de confiance.

Plusieurs propriétés sont associées aux estimateurs ponctuels : convergence, biais, erreur, robustesse.

1.3.2 Propriétés des estimateurs ponctuels

1.3.2.1 Biais

Notons θ un paramètre de la population et $\hat{\theta}$ un estimateur ponctuel de θ sur un échantillon.

Le biais de l'estimateur $\hat{\theta}$ est, par définition, la différence entre θ et l'espérance de $\hat{\theta}$. Si $\hat{\theta}$ est absolument continue à valeurs dans $\mathbb{R}$, cette espérance est égale à $\int_{-\infty}^{+\infty} xf(x)\,\mathrm{d}x$ où f est la fonction de densité de probabilité de $\hat{\theta}$.

$$B = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

Un estimateur est dit non biaisé si son biais est nul. Un estimateur applicable à un échantillon de taille fixe est dit asymptotiquement non biaisé si la limite de son biais lorsque la taille d'échantillon tend vers l'infini est nulle. On peut aussi définir le *biais en médiane*, comme la différence entre la médiane de $\hat{\theta}$ et θ. Lorsque la distribution de $\hat{\theta}$ est asymétrique, la médiane et la moyenne peuvent différer et l'estimateur ne peut plus être à la fois non biaisé en moyenne et en médiane.

Si un estimateur est non biaisé en moyenne, d'une manière générale, il devient biaisé dès qu'une transformation monotone mais non linéaire lui est appliqué. Par exemple, l'estimateur de la variance sur un échantillon $(X_i, \ldots, X_n)$ égal à $\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$ n'est pas biaisé (en moyenne) mais sa racine carrée est un estimateur biaisé de l'écart-type. La construction d'un estimateur non biaisé (ou même de biais rapidement décroissant) de l'écart-type n'est pas chose facile car cet estimateur ne peut pas être le même pour toutes les distributions.

Dans certains cas pathologiques, ce biais est indéfini car l'espérance est indéfinie. C'est notamment le cas de l'estimateur de la médiane d'une loi de Cauchy par la moyenne de l'échantillon.

1.3.2.2 Consistance (ou convergence)

On peut d'abord définir la convergence en probabilité d'une suite de variables aléatoires $n \to X_n$ pour $n \in \mathbb{N}$.

Une telle suite converge en probabilité vers une limite l si $\forall \varepsilon > 0$, $\lim_{n \to +\infty} \mathbb{P}(\|X_n - l\| > \varepsilon) = 0$. La notion est généralisable à un espace métrique non euclidien en remplaçant $\|X_n - l\|$ par $d(X_n - l)$. La notion s'applique à une limite constante, mais se généralise aussi à la convergence vers une variable aléatoire X, en posant $\lim_{n \to +\infty} \mathbb{P}(\|X_n - X\| > \varepsilon) = 0$.

Un exemple typique de suite de variables aléatoires non convergente, c'est la suite des P-valeurs (P-values en anglais) P_n d'un test statistique sous l'hypothèse nulle, suivant de plus ou moins près une loi uniforme entre 0 et 1, où n représente la taille d'un échantillon aléatoire d'observations souvent indépendantes et identiquement distribuées dont la P-valeur est issue. Les analyses intermédiaires répétées d'observations indépendantes identiquement distribuées, sans limite de nombre ni de taille d'échantillon, conduisent alors presque sûrement, c'est-à-dire avec une probabilité égale à 1, à des P-valeurs extrêmes aussi bien arbitrairement proches de 0 que de 1 pour la plupart des statistiques usuelles. Lorsque le même échantillon est progressivement étendu, les données des P-valeurs successives sont corrélées mais cette corrélation peut être approchée de zéro en agrandissant l'échantillon de telle manière que la taille d'échantillon précédente soit infime en comparaison à la nouvelle taille et n'influence plus que de manière négligeable le calcul de la P-valeur.

On remarquera qu'une série de variables aléatoires peut converger en probabilité sans pour autant que l'espérance de $X_n - l$ soit définie. Nous pouvons, par exemple, définir les X_n comme des variables aléatoires suivant des lois de Cauchy de paramètre d'échelle $a_n = \frac{1}{n+1}$. Cette suite sera convergente en probabilité vers 0 alors qu'aucune espérance de X_n n'est définie.

On peut aussi créer une suite de variables convergente dont l'espérance est bien définie pour chaque X_n mais dont l'espérance ne converge pas. Pour ce faire considérons la suite des Y_n à valeurs dans $\mathbb{R}$, telle que pour tout $n \geq 2$, Y_n, une variable binaire telle que $\mathbb{P}(Y_n = 1) = \frac{1}{n}$, $\mathbb{P}(Y_n = 0) = \frac{n-1}{n}$ et $\mathbb{P}(Y_n = x) = 0$ pour tout autre x réel. Même si cela n'influence pas les résultats suivants, afin de définir la suite de manière précise, considérons que tous les Y_n sont indépendants. Définissons ensuite $X_n = n^2 Y_n$ de telle sorte que $\mathbb{P}(X_n = n^2) = \frac{1}{n}$ et $\mathbb{P}(X_n = 0) = \frac{n-1}{n}$. Alors $\mathbb{E}[Y_n] = \frac{1}{n}$ et $\mathbb{E}[X_n] = n^2 \mathbb{E}[Y_n] =$

n de telle sorte que $\lim_{n\to+\infty} \mathbb{E}[X_n - 0] = +\infty$ alors que $\forall\ \varepsilon > 0,\ \mathbb{P}(\|X_n - 0\| > \varepsilon) = \frac{1}{n}$ et donc, $\lim_{n\to+\infty} \mathbb{P}(\|X_n - 0\| > \varepsilon) = 0$. Ainsi, bien que X_n converge en probabilité vers 0, l'espérance de X_n ne converge pas dans $\mathbb{R}$. Si X_n était un estimateur d'une statistique nulle, on pourrait considérer que son biais d'estimation croit avec n bien qu'il soit convergent vers la statistique qu'il estime. Un autre exemple, avec une loi continue peut être fourni par $Y_n = \mathrm{Beta}(1; n)$ et $X_n = n^2 Y_n$ conduisant à une $\mathbb{E}[X_n] = \frac{n^2}{n+1}$ tendant vers $+\infty$ lorsque $n \to +\infty$ alors que X_n converge en probabilité vers 0.

Si on définit une expérience pour chaque n comme un échantillonnage d'observations $(Y_i, \dots, Y_n)$ et un estimateur $\widehat{\theta_n}$ associé à chacune de ces expériences, alors on en déduit une suite de variables aléatoires $n \to \widehat{\theta_n}$. L'estimateur est dit consistant (ou convergent) si cette suite converge en probabilité vers θ.

On remarquera qu'il est possible, et même fréquent, que quel que soit n, aussi grand soit-il, la probabilité que X_n soit très éloignée de l (supérieur à n'importe quel ε prédéfini) reste positive. Par exemple, l'estimateur M_n de la moyenne μ d'une variable aléatoire suivant une loi normale de variance non nulle vérifie la propriété malheureuse suivante : $\forall\, \Delta > 0, \forall n > 0, \mathbb{P}(|M_n - \mu| > \Delta) > 0$. En bref, quel que soit la taille d'échantillon il est possible d'avoir une valeur arbitrairement éloignée de la vraie moyenne. Heureusement, pour un Δ donné, cette probabilité est rapidement décroissante avec n.

La notion de convergence en probabilité d'un estimateur reste définissable dans une méthodologie adaptative, mais il faut préalablement définir un paramètre de l'expérience que l'on puisse faire varier, tel qu'un paramètre qui reflète directement ou indirectement la taille moyenne de l'échantillon.

Les notions de biais et de convergence sont bien distinctes.

Un estimateur peut être non biaisé mais non convergent, tel qu'un estimateur de la moyenne qui serait égal à une unique valeur sélectionnée aléatoirement dans l'échantillon. En supposant un échantillon d'observations indépendantes identiquement distribuées tirées au hasard dans une population, cet estimateur n'est pas biaisé ; cependant si la variable n'est pas constante, alors l'estimateur ne convergera pas du tout.

Un estimateur peut être biaisé mais convergent tel que l'écart-type empirique d'un échantillon d'observations suivant des lois normales indépendantes et identiquement distribuées. Comme l'exemple

fourni plus haut, il est théoriquement possible d'avoir un biais croissant avec la taille d'échantillon. Cela reste un cas purement théorique.

1.3.2.3 Convergence forte

Comme précisé dans la section précédente (1.3.2.2) une suite de variables aléatoires X_n est convergente vers une limite l si $\forall \varepsilon > 0, \lim_{n\to+\infty} \mathbb{P}(\|X_n - l\| > \varepsilon) = 0$. Elle est dite *presque sûrement convergente* si $\mathbb{P}\left(\lim_{n\to+\infty} X_n = l\right) = 1$. Il faut comprendre que $\lim_{n\to+\infty} X_n$ est une variable aléatoire dérivée de la suite infinie des X_n. Les X_n ne sont pas forcément des variables indépendantes.

Reprenons l'exemple où $Y_n \sim \mathcal{B}\left(1; \frac{1}{n}\right)$ et $X_n = n^2 Y_n$ avec des Y_n indépendants les uns des autres, tel que décrit en section 1.3.2.2. Cette suite de variables aléatoires X_n, définie pour tout $n \geq 2$, n'est pas fortement convergente. En effet, $\forall\, i > 0,\ \mathbb{P}\left[X_i < \frac{1}{2}\right] = 1 - \frac{1}{i}$ de telle sorte que $\mathbb{P}\left[\forall i \geq n, X_i < \frac{1}{2}\right] = \prod_{i=n}^{+\infty}\left(1 - \frac{1}{i}\right) = \exp\left(\sum_{i=n}^{+\infty} \log\left(1 - \frac{1}{i}\right)\right)$ car les X_i sont indépendants.

Or $\log(1 - x) < -x$ pour tout $x > 0$ réel (graphiquement, la tangente à la courbe représentatrice la fonction logarithme en 1 est au-dessus de la courbe). Ainsi, $\sum_{i=n}^{+\infty} \log\left(1 - \frac{1}{i}\right) < \sum_{i=n}^{+\infty}\left(-\frac{1}{i}\right)$. Cette dernière série tend vers $-\infty$, car il s'agit de la définition de la fonction zêta de Riemann en 1, à une constante et un signe près. Ainsi, $\mathbb{P}\left[\forall i \geq n, X_i < \frac{1}{2}\right] = \exp\left(\sum_{i=n}^{+\infty} \log\left(1 - \frac{1}{i}\right)\right)$ tend vers 0 quel que soit $n \geq 2$. On peut en déduire que, quel que soit le rang n, la suite de valeurs de X_i avec $i \geq n$ contiendra presque sûrement une valeur supérieure à ½ (ainsi que des valeurs nulles, bien sûr) de telle sorte que la suite ne pourra pas converger. Ainsi $\mathbb{P}\left(\lim_{n\to+\infty} X_n = l\right) = 0$. Bien que la suite de variables aléatoires soit convergente, elle n'est pas presque sûrement convergente.

Dans le cadre d'une suite d'estimateurs $\widehat{\theta_n}$ de θ, la suite correspondra généralement à l'application de l'estimateur à des échantillons emboîtés de taille croissante $(Y_1, \ldots, Y_n)$. Les estimateurs $\widehat{\theta_n}$ ne seront donc généralement pas indépendants et on parlera de *convergence forte* de l'estimateur si $\widehat{\theta_n}$ converge presque sûrement vers θ.

1.3.2.4 Erreur

A biais égal, on préfèrera un estimateur $\hat{\theta}$ qui fournit des résultats proches de la valeur θ. Cette proximité peut se mesurer par l'erreur quadratique moyenne $EQM(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$, la notion pouvant être généralisée à un espace Hilbertien comme $\mathbb{R}^n$ associé à sa norme usuelle par $EQM(\hat{\theta}) = \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$. Mais, il existe d'autres manières d'évaluer l'erreur, comme l'erreur absolue moyenne $\mathbb{E}[|\hat{\theta} - \theta|]$ ou l'erreur absolue médiane $Med[|\hat{\theta} - \theta|]$. On peut aussi s'abstraire du biais en étudiant la variance de l'estimateur $VAR(\hat{\theta})$, égal à l'erreur quadratique moyenne d'un estimateur modifié dont on corrigerait parfaitement le biais par l'addition d'une valeur correctrice constante.

Le choix de la mesure de l'erreur a son importance puisque la hiérarchisation de la performance de deux estimateurs peut en dépendre.

Considérons par exemple, les deux estimateurs suivants de l'espérance $1/\lambda$ d'une loi exponentielle :

1) La moyenne de l'échantillon
2) La moyenne de l'échantillon après avoir plafonné toutes les valeurs dépassant le double de la moyenne de l'échantillon ; le plafonnage des valeurs consistant à assigner au plafond toutes les valeurs le dépassant.

Pour un échantillon de taille 10, on observe empiriquement, par des simulations sous le logiciel R, que le 1er estimateur a une plus grande erreur quadratique moyenne que le second mais une moins grande erreur absolue moyenne. On remarque aussi que le second estimateur est biaisé en moyenne alors que le premier ne l'est pas.

Il existe une limite inférieure théorique à l'erreur quadratique moyenne d'un estimateur sans biais : l'inverse de l'information de Fisher $\mathcal{I}(\theta)$, qui dérive de la fonction de vraisemblance et reflète la quantité d'information sur le paramètre que l'on peut tirer de l'échantillon. Cette borne inférieure est nommée borne de Cramér-Rao. Un estimateur atteignant la borne de Cramér-Rao est dit *efficace*. Un tel estimateur n'existe pas toujours.

1.3.2.5 Robustesse

La notion de robustesse d'estimateur est liée à l'existence de valeurs atypiques dans les échantillons.

Ces valeurs atypiques ou outliers peuvent, selon les cas, être des valeurs erronées (p.ex. erreur de saisie) ou des valeurs correctes mais fortement écartées du cas typique. Ces valeurs sont susceptibles d'influencer de manière importante les statistiques telles que la moyenne, ce qui n'est pas toujours souhaitable lorsqu'on souhaite modéliser les cas typiques.

Un estimateur est dit robuste lorsqu'il est peu influencé par ces valeurs atypiques. On définit le point de rupture d'un estimateur comme la plus petite proportion d'observations dont une modification est susceptible d'engendrer un changement arbitrairement grand de l'estimateur. La moyenne sur un échantillon de taille n a un point de rupture à $1/n$, puisqu'en changeant seulement une observation on pourra affecter de manière arbitrairement grande la moyenne. Pour ajouter a à la moyenne, il suffit d'ajouter $a \times n$ à n'importe quelle observation. La médiane a un point de rupture à 50%.

Pour certaines statistiques communes, telle que la moyenne de la population, la robustesse d'un estimateur est malheureusement incompatible avec l'absence de biais voire avec la convergence pour certaines distributions. Sur un échantillon d'observations indépendantes et identiquement distribuées, la médiane est un estimateur robuste et non biaisé de la moyenne de la population lorsque la distribution est symétrique ; mais dès qu'elle est appliquée à des distributions asymétriques, c'est un estimateur asymptotiquement biaisé.

1.3.3 Propriétés des estimateurs d'intervalle de confiance

Les estimateurs d'intervalle de confiance ne sont pas évalués de la même manière que les estimateurs ponctuels, même si on peut faire une analogie entre certaines propriétés.

1.3.3.1 Couverture nominale et réelle

Un estimateur d'intervalle de confiance a une couverture nominale, souvent choisie à 95%. Cette couverture nominale, comprise entre 0 et 1 est un paramètre de la procédure d'estimation. Notons $IC_{1-\alpha}$ un estimateur d'intervalle de confiance d'un paramètre θ de la population avec une couverture nominale à $1-\alpha$. La couverture réelle de cet estimateur est $\mathbb{P}(\theta \in IC_{1-\alpha})$. Le biais de couverture est la différence entre la couverture réelle et la couverture nominale. Un estimateur d'intervalle de confiance n'est pas biaisé si la couverture réelle est égale à la couverture nominale.

Un estimateur d'intervalle est unilatéral à droite s'il est construit de telle sorte que la probabilité que θ

soit inférieur à la borne basse de l'intervalle de confiance est nulle et il est unilatéral à gauche s'il est construit de telle sorte que la probabilité que θ soit supérieur à la borne haute de l'intervalle de confiance est nulle. Pour un paramètre θ susceptible de prendre n'importe quelle valeur réelle, un estimateur intervalle de confiance unilatéral aura une borne infinie, tel que $]-\infty, U_{1-\alpha}]$ pour un intervalle de confiance unilatéral à droite et $[L_{1-\alpha}, +\infty[$ s'il est unilatéral à gauche. Un estimateur est bilatéral s'il n'est ni unilatéral à droite ni à gauche.

Notons $IC_{1-\alpha} = [L_{1-\alpha}, U_{1-\alpha}]$ un intervalle de confiance bilatéral fermé. Nous n'envisagerons pas les intervalles de confiance dont une des bornes est ouverte car cela ne présenterait d'intérêt que dans le scenario atypique ou la distribution *a priori* du paramètre θ (point de vue bayésien) serait discrète mais où l'ensemble des valeurs de probabilité non nulles de θ ne serait pas un ensemble discret, comme dans l'exemple décrit en section 1.2.11.2. Certains intervalles de confiance sont construits d'une telle manière qu'ils tendent à minimiser la différence de défaut de couverture entre les deux bornes, c'est-à-dire $|\mathbb{P}(\theta > U_{1-\alpha}) - \mathbb{P}(\theta < L_{1-\alpha})|$; ces intervalles sont dits équilibrés. On peut juger de la qualité de la couverture d'un estimateur équilibré en évaluant le biais à gauche comme $\mathbb{P}(\theta < L_{1-\alpha}) - \frac{\alpha}{2}$ et le biais à droite comme $\mathbb{P}(\theta > U_{1-\alpha}) - \frac{\alpha}{2}$. L'estimateur équilibré n'est pas biaisé si le biais à gauche comme le biais à droite sont tous deux nuls ; dans les autres cas de figure, il est plus ou moins fortement biaisé.

Certains estimateurs d'intervalles bilatéraux ne recherchent pas l'équilibre mais peuvent, au contraire, tendre à minimiser l'espérance de la largeur de l'intervalle $U_{1-\alpha} - L_{1-\alpha}$ en déséquilibrant artificiellement les risques. Dans le pire des cas, l'intervalle de confiance bilatéral peut être très proche d'un intervalle unilatéral sans qu'on puisse forcément deviner de quel côté l'intervalle est unilatéral. Les intervalles bilatéraux déséquilibrés posent un certain nombre de problèmes théoriques qui seront discutés plus loin.

1.3.3.2 Largeur

L'espérance de la largeur $U_{1-\alpha} - L_{1-\alpha}$ d'un intervalle de confiance bilatéral $IC_{1-\alpha} = [L_{1-\alpha}, U_{1-\alpha}]$ reflète la stabilité de cet estimateur. Il est aussi possible d'analyser la variance de chacune des deux bornes de l'intervalle ou l'espérance de la distance entre chaque borne et le paramètre θ de la population.

Ces statistiques reflètent la précision de l'estimateur d'intervalle. Si deux estimateurs sont non biaisés on privilégiera celui qui est le plus stable (moindre largeur).

1.3.3.3 Biais asymptotique de couverture

On peut définir une suite d'estimateurs d'intervalles de confiance selon la taille d'échantillon $IC_{1-\alpha,n} = \left[L_{1-\alpha,n}, U_{1-\alpha,n}\right]$ et définir $c\left(IC_{1-\alpha,n}\right) = \mathbb{P}\left(\theta \in IC_{1-\alpha,n}\right)$ la couverture réelle pour un échantillon de taille n. On définit la couverture réelle asymptotique comme $\lim_{n\to\infty} c\left(IC_{1-\alpha,n}\right)$ et le biais asymptotique de couverture comme la différence entre la couverture nominale et la couverture réelle asymptotique. Beaucoup d'estimateurs d'intervalles sont asymptotiquement non biaisés sur des échantillons d'observations indépendantes et identiquement distribuées.

1.3.3.4 Convergence

Un intervalle de confiance bilatéral $IC_{1-\alpha} = [L_{1-\alpha}, U_{1-\alpha}]$ est convergent si $L_{1-\alpha}$ et $U_{1-\alpha}$ sont tous deux des estimateurs convergents de θ.

1.4 Modèles statistiques usuels

1.4.1 Histoire

1.4.1.1 L'erreur probable d'une moyenne

Le modèle linéaire général ainsi que les estimateurs ponctuels et d'intervalle associés représentent l'aboutissement d'analyses et de modèles de complexité et généralité croissante.

Un des premier pas est représenté par l'analyse des fluctuations d'échantillonnage par William Gosset, élève de Karl Pearson, qui publia en 1908 sous le pseudonyme de Student [1] un article détaillant l'analyse des fluctuations d'échantillonnage de $\widehat{\sigma^2}$, l'estimateur de la variance d'un échantillon d'observations indépendantes et identiquement distribuées selon une loi normale, puis l'analyse des fluctuations d'échantillonnage de $\frac{M-\mu}{\hat{\sigma}}$ où μ est la moyenne de la population et M son estimateur empirique sur l'échantillon. C'est ce travail qui servit à définir la loi T de Student suivie par $\frac{M-\mu}{\frac{1}{\sqrt{n}}\hat{\sigma}}$.

1.4.1.2 L'ANOVA

Le second pas fut le développement de l'analyse de variance (ANOVA) par Ronald Fisher, autre élève

de Karl Pearson, dont la formalisation fut réalisée dans le livre intitulé « *Statistical Methods for Research Workers* » publié en 1925 [2]. Même si le modèle d'ANOVA n'avait pas été clairement formalisé par Fisher, il peut être vu comme un ancêtre du modèle linéaire général dans lequel toutes les variables explicatives (appelés facteurs ou classes) sont qualitatives et chaque combinaison de modalités des variables est représenté par une et une seule observation. Par exemple la table 46, page 205 du livre de Fisher montre un tableau de fréquence (entre 0 et 100%) de la pluie pour (24 heures)×(12 mois)=288 observations. Chaque observation est ici, une moyenne pluviométrique sur 10 ans avec donc, en réalité environ 365.25×10 = 3652 observations qui ont été synthétisées en seulement 288. Les estimations ponctuelles des effets sont calculables comme les moyennes marginales de chaque modalité d'une variable. Les variances entre chacune des modalités d'une variable de ces moyennes sont aussi directement calculables. Il est possible d'évaluer la part de variance de Y attribuables à chacun des variables explicatives et de tester l'hypothèse selon laquelle les moyennes de Y dans toutes les modalités d'une variable explicative sont égales.

Le modèle fut ensuite généralisé aux mesures répétées, avec toujours la condition d'un design équilibré, c'est-à-dire un nombre identique d'observations dans chaque combinaison de modalités des variables explicatives, ce qui implique, entre autres, l'indépendance entre les variables explicatives, sur l'échantillon. C'est pourquoi plutôt que de parler de variables explicatives, le vocabulaire de l'ANOVA fait référence à des classes ou facteurs, qui sont soit maîtrisés par l'expérimentateur (p.ex. fertilisants sur parcelles de terrain agricole), soit naturellement à fréquence constante, comme le nombre d'heures dans une journée. Cela n'est pas compatible avec la modélisation du vivant où les variables explicatives sont aléatoires et déséquilibrées.

Le problème des ANOVA à effectifs déséquilibrés a été identifié très tôt puisqu'en 1933 Brandt proposa une première solution à ce problème [3], puis Yates améliora et généralisa la technique pour toute ANOVA à deux facteurs [4]. Il proposa trois solutions qui correspondent aux sommes de carrés de type I, II et III dans le logiciel SAS [5]. Le 1er type est une décomposition successive qui fournit des résultats dépendants de l'ordre dans lequel les facteurs sont intégrés dans le modèle, le 1er facteur n'étant pas ajusté, le 2ème étant ajusté sur le 1er et le 3ème sur les deux premiers, etc. Le 2ème et le 3ème fournissent des

résultats où chaque facteur est ajusté sur les autres (effets conditionnels aux autres), mais diffèrent dans leur comportement en présence d'interaction. Le 2ème type est adapté à la situation où la corrélation entre facteurs est naturelle, l'effet marginal du facteur est alors calculé. Le 3ème type est adapté à la situation où la corrélation entre les facteurs serait artificielle les effets du facteur d'intérêt étant alors moyennés entre les modalités d'un second facteur avec, pour pondération, la fréquence marginale de chacune des modalités du second facteur.

1.4.2 Modèle linéaire général

1.4.2.1 Définition

Le modèle linéaire général est une famille de modèles statistiques. Sur un échantillon de n observations, ces modèles permettent de décrire la distribution d'une variable à expliquer $Y = (Y_1, \dots, Y_n)$ en fonction de la distribution d'une variable $X = \begin{pmatrix} X_{1,1} & \cdots & X_{1,k} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,k} \end{pmatrix}$.

Comme précisé ci-dessus, il s'agit d'une famille de modèles, paramétrée par un coefficient $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ et une variance résiduelle σ^2. Considérant ces paramètres fixés, la variable Y s'écrit :

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

Où $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ et tous les ε_i suivent des lois normales centrées de variance σ^2 toutes indépendantes les unes des autres.

On peut aussi réécrire le modèle comme

$$\forall i \in \{0, \dots, n\}, Y_i = \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + \varepsilon_i$$

Certains auteurs ajoutent aux Y_i un coefficient β_0 constant appelé intercept, mais celui-ci peut être intégré dans le modèle décrit ci-dessus en ajoutant des variable explicatives $X_{i,0}$ constantes égales à 1 dans la population ; une simple renumérotation des indices permet de retomber sur la formulation ci-dessus. Cette formulation, outre une simplification de formule, permet la description de modèles sans intercept.

La formulation ci-dessus implique la propriété suivante sur les espérances :

$$\mathbb{E}[Y_i] = \beta_1 \mathbb{E}[X_{i,k}] + \cdots + \beta_k \mathbb{E}[X_{i,k}]$$

Les propriétés suivantes sont directement déductibles de la formulation du modèle :

1) Normalité des résidus $\varepsilon_i = Y_i - \mathbb{E}[Y_i]$

2) Homoscédasticité des résidus, c'est-à-dire, $\forall i, j\ VAR(\varepsilon_i) = VAR(\varepsilon_j) = \sigma^2$

3) Résidus indépendants et identiquement distribués

4) « Linéarité » de l'effet de l'espérance de chaque variable explicative sur l'espérance de la variable à expliquer, les valeurs des autres variables explicatives étant fixées dans l'analyse de cette relation. Au sens mathématique, il s'agit d'une relation affine plutôt que linéaire.

5) Additivité des effets : l'espérance de Y_i conditionnelle à la réalisation $x_{i,1}, \dots, x_{i,k}$ s'exprime comme une combinaison linéaire de cette réalisation.

Il existe une formulation alternative du modèle, conditionnée à X. Cela peut avoir du sens lorsque les variable $X_{i,j}$ correspondent à des conditions expérimentales contrôlées et sont donc constantes. Le conditionnement simplifie aussi le développement d'estimateurs. La formulation devient :

$$Y = x\beta + \varepsilon$$

Où $x \in \mathbb{R}^{n,k}$ est une matrice de constantes.

1.4.2.2 Variables binaires, quantitatives, qualitatives

Il existe une hypothèse de normalité sur la variable à expliquer Y mais pas sur la variables explicative X ; la question ne se pose même pas lorsqu'on conditionne sur les variables explicatives puisque x est alors constant. Les variables explicatives peuvent donc avoir n'importe quelle distribution à valeurs dans $\mathbb{R}$. Par exemple : binaires, discrètes, continues.

Les variables qualitatives n'étant généralement pas à valeurs dans $\mathbb{R}$, il faudra les recoder sous forme de variables à valeurs dans $\mathbb{R}$. Plusieurs manière de recoder existent. Par exemple, les variables qualitatives nominales à m modalités sont souvent recodées en $m - 1$ variables binaires valant 0 ou 1. Avec le contraste le plus fréquemment usité, on choisira une modalité de référence de la variable et on définira une variable binaire pour chaque autre modalité. Cette variable binaire vaudra un lorsque que la variable

qualitative est égale à la modalité concernée et zéro autrement.

1.4.2.3 Inférence statistique

Les inférences sur le modèle linéaire général sont généralement faites sur les paramètres du modèle β_i. Ainsi, on considère que l'expérience aléatoire génère des données suivant un modèle de la famille des modèles linéaires généraux mais les paramètres du modèle sont considérés comme inconnus. Toutes les propriétés des modèles linéaires généraux précédemment décrites sont donc des hypothèses sur lesquelles on repose sans forcément pouvoir les vérifier.

L'estimateur usuel de ces paramètres est l'estimateur des moindres carrés, c'est-à-dire, une fois X réalisé en x, le jeu de coefficients $\begin{pmatrix} \widehat{\beta_1} \\ \vdots \\ \widehat{\beta_n} \end{pmatrix}$ est choisi pour minimiser $\|y - x\beta\|$ avec la norme usuelle de $\mathbb{R}^n$, ce qui équivaut à minimiser $\sum_{i=1}^{n} \left(y_i - \left(\beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}\right)\right)^2$.

L'inférence statistique consiste donc à choisir des paramètres d'une famille de distribution dans un contexte où cette famille est déjà très restreinte par les hypothèses de modélisation réalisées.

1.4.2.4 Conditions de validité du modèle

Même s'il existe des outils pour vérifier les hypothèses sous-jacentes au modèle linéaire général, tels que le test de normalité de Shapiro-Wilk sur les résidus ou les tests d'interaction, ces tests sont problématiques à plusieurs sens (opinion de l'auteur) :

1) Il y a généralement une inversion entre le rôle de l'hypothèse nulle et de l'hypothèse alternative avec une recherche de non-significativité conduisant, incorrectement, à l'acceptation de l'hypothèse nulle

2) Si on a la puissance de tester une hypothèse alors on a souvent la possibilité de ne plus reposer dessus (*p.ex.* boostrap)

3) Ces hypothèses, dans des conditions réelles sont généralement fausses, même si elles peuvent être « presque » vraies

4) Les procédures consistant à orienter les choix des modèles par des résultats statistiques sont généralement fortement biaisées

J'illustrerai le problème N°3 par le problème de la normalité des résidus d'un modèle expliquant une variable biomédicale.

Si la variable à expliquer ne peut pas être négative, alors les résidus ne peuvent pas être normalement distribués car cela impliquerait une distribution de Y_i de $-\infty$ à $+\infty$. De même le modèle linéaire général est incompatible avec toute variable à expliquer ayant un maximum ou un minimum fini. Il est bien sûr incompatible avec toute variable discrète. J'ai ainsi listé presque toutes les variables à expliquer qu'il m'ait été donné de rencontrer dans ma vie de biostatisticien. Ainsi, conclure à la normalité des résidus par non-significativité du test de Shapiro-Wilk ou analyse graphique des résidus est incorrect. On peut éventuellement conclure que la déviation à la normalité n'est pas problématique, mais on ne peut pas conclure à l'absence de déviation à la normalité puisque celle-ci existe pratiquement toujours. Les autres hypothèses, telles que la linéarité et l'additivité des effets sont généralement peu plausibles dans le domaine biomédical, même si les écarts à ces hypothèses peuvent être modestes.

S'il existe pratiquement toujours des écarts aux hypothèses, comment peut-on évaluer la fiabilité des inférences réalisées ? Quelles propriétés statistiques restent valables en cas d'écart aux hypothèses ? Quels écarts aux hypothèses sont problématiques et lesquels ne le sont pas ?

Même si ce n'est pas l'objet principal de ce travail de thèse, nous allons introduire un cadre théorique permettant de fournir des éléments de réponse à ces questions.

Considérons la question : quel est le biais de l'estimation du paramètre β_i si la relation entre X_i et Y_i n'est pas linéaire ? Dans le cadre théorique du modèle linéaire général, ce paramètre n'existe pas puisque sa définition implique la linéarité de la relation. Si le paramètre n'existe pas, la notion même de biais d'estimation perd tout sens.

1.4.2.5 Cadre théorique permettant l'analyse de modèles paramétriques en cas de non-respect de leurs conditions de validité

Qu'un modèle linéaire général soit invalide ou pas, l'estimateur des moindres carrés est toujours calculable, sauf si le nombre de degrés de liberté est négatif. Si le modèle linéaire général n'est pas

valide, c'est un estimateur sans objet, au sens où il n'existe pas de paramètre de la population qu'il estime.

Considérons une suite infinie d'expériences aléatoires $n \to (Y_n, X_n)$ correspondant à la sélection d'observations indépendantes et identiquement distribuées dans une population avec $Y_n \in \mathbb{R}$ et $X_n \in \mathbb{R}^k$. On peut définir $\widehat{\beta_n} \in \mathbb{R}^k$ l'estimateur des moindres carrés du modèle linéaire général expliquant $(Y_1, \dots, Y_n)$ par $(X_1, \dots, X_n)$. Si cette suite converge en probabilité vers une valeur l (cf. section 1.3.2.2) alors nous noterons $\beta_\infty = l$ et considèrerons que β_∞ est l'objet de l'estimateur. Ainsi, nous pourrons analyser le biais $\mathbb{E}[\widehat{\beta_n}] - \beta_\infty$ ou l'erreur quadratique moyenne $\mathbb{E}\left[\left(\widehat{\beta_n} - \beta_\infty\right)^2\right]$, si ces formes sont bien définies.

1.4.3 Familles exponentielles de distributions

1.4.3.1 Définition

Les familles de distributions binomiales, de Poisson, gaussiennes (synonyme de normales), distributions gamma et distributions gaussiennes inverses sont des familles exponentielles de distribution, entre autres.

Une famille de distributions à valeurs dans S paramétrée par les θ à valeurs dans Θ forme une famille exponentielle de dimension k si sa fonction de masse (pour les lois discrètes) ou fonction de densité de probabilité (pour les lois continues) peut s'exprimer comme :

$$f(x|\theta) = h(x) \exp\left({}^t\big(\eta(\theta)\big)T(x) - A(\theta)\right)$$

Où ${}^t u$ dénote la transposition de la matrice ou du vecteur u et où les fonctions sont définies dans les espaces suivants :

$$h: S \to \mathbb{R}_+$$

$$\eta: \Theta \to \mathbb{R}^k$$

$$T: S \to \mathbb{R}^k$$

$$A: \Theta \to \mathbb{R}$$

et bien sûr la fonction $x \to f(x|\theta)$ est à valeurs dans $\mathbb{R}_+$

S et Θ sont des ensembles quelconques, sans forcément de loi de composition interne.

On peut aussi exprimer cette fonction (si h n'est jamais nul) comme :

$$f(x|\theta) = \exp\left({}^{t}\big(\eta(\theta)\big)T(x) + \log\big(h(x)\big) - A(\theta)\right)$$

Faisant alors apparaître, la somme de trois termes :

$-A(\theta)$ ne dépendant que de θ

$\log\big(h(x)\big)$ ne dépendant que de x

${}^{t}\big(\eta(\theta)\big)T(x)$ dépendant des deux

Ce dernier terme, contraint fortement la relation entre x et $\eta(\theta)$, se limitant au produit scalaire de ce paramètre canonique avec un x transformé par T.

Pour le lecteur qui ne serait pas familier avec les notations vectorielles on peut noter $\eta(\theta) = \begin{pmatrix}\eta_1(\theta)\\ \vdots \\ \eta_k(\theta)\end{pmatrix}$

et $T(x) = \begin{pmatrix}T_1(x)\\ \vdots \\ T_k(x)\end{pmatrix}$ et ${}^{t}\big(\eta(\theta)\big)T(x) = \sum_{i=1}^{k}\eta_i(x)T_i(x)$.

Le terme $A(\theta)$ est appelée constante de log-normalisation puisqu'elle permet de s'assurer que la fonction s'intègre à 1. Elle constante pour une distribution fixée (avec un θ fixé) car elle ne dépend pas de x. Cette fonction A peut se recalculer à partir de h, η et T comme le logarithme de l'intégrale (ou la série ou somme finie) de la fonction $x \to f(x|\theta)$ sur S.

En définissant $g: \Theta \to R_+^*$ par $g(\theta) = \exp\big(-A(\theta)\big)$, on peut aussi noter f comme :

$$f(x|\theta) = h(x)g(\theta)\exp\big({}^{t}\eta(\theta)T(x)\big)$$

On peut aussi reparamétrer la famille en utilisant $\tau = \eta(\theta)$ comme paramètre à la place de θ, obtenant ainsi :

$$f(x|\theta) = f(x|\tau) = h(x)\exp\big({}^{t}\tau T(x) - B(\tau)\big)$$

Où, pour un τ fixé, $B(\tau)$ est égal au logarithme de l'intégrale (typiquement intégrale de Lebesgue pour les distributions continues et série/sommes pour les distributions discrètes) de $u(x) = h(x)\exp\big({}^{t}\tau T(x)\big)$ sur S tout entier, ce qui est calculable quand bien même η ne serait pas injective. Cette intégrale de $u(x)$ étant égale, à une constante près à celle de $f(x|\theta)$ sur le même espace, elle est donc bien définie. Cette explicitation de B était nécessaire car on ne peut pas directement exprimer B à

partir de A si la fonction η n'est pas injective.

On parle de paramétrisation canonique, lorsque la fonction η est égale à la fonction identité. Cela contraint les paramètres à appartenir à $\mathbb{R}^k$. Contrairement à ce que suggère le terme canonique, cette paramétrisation n'est pas unique. Considérons par exemple un vecteur $c = \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} \in \mathbb{R}^k$ où tous les c_i sont non nuls et une paramétrisation canonique $\tau = \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_k \end{pmatrix}$ alors $\mu = \begin{pmatrix} \tau_1 c_1 \\ \vdots \\ \tau_k c_k \end{pmatrix}$ est un nouveau paramètre canonique associé à la fonction M, remplaçant la fonction T, définie par $M(x) = \begin{pmatrix} \frac{1}{c_1} T_1(x_1) \\ \vdots \\ \frac{1}{c_k} T_k(x_k) \end{pmatrix}$.

Le paramètre $\tau = \eta(\theta)$ de la forme canonique est appelé paramètre naturel. Comme précédemment décrit, le paramètre naturel n'est pas unique et selon le contexte, pourra dépendre de la paramétrisation initiale, c'est-à-dire de η et de θ.

1.4.3.2 La ou les familles exponentielles

En fixant S, Θ, h, η, T et A, on définit **une** famille exponentielle, paramétrée par les $\theta \in \Theta$. Cependant, on ne peut pas définir **la** famille exponentielle comme s'il s'agissait d'une seule famille. En effet, cette dernière serait paramétrée par S, Θ, h, η, T et A. Comme S peut être n'importe quel ensemble, alors l'ensemble des paramètres de la famille exponentielle contiendrait l'ensemble de tous les ensembles. Ce dernier est en contradiction avec la théorie des ensembles selon le paradoxe de Cantor qui est tellement simple qu'il mérite d'être redémontré ici :

Si l'ensemble de tous les ensembles E existe, alors définissons le sous-ensemble F des ensembles ne s'appartenant pas, c'est-à-dire, $F = \{x | x \notin x\}$, en appliquant le schéma d'axiomes de compréhension selon la théorie des ensembles de Zermelo-Fraenkel. La proposition suivante est-elle vraie ou fausse ?

$$F \in F$$

Si cette proposition est vraie, alors, par définition de F, nous montrons que $F \notin F$, et donc, nous arrivons à une contradiction.

Si la proposition est fausse, alors $F \notin F$ et donc, par définition de F nous pouvons dire que $F \in F$, et

nous arrivons aussi à une contradiction.

Nous montrons ainsi, que la proposition n'est ni vraie, ni fausse et donc est aussi respectivement fausse et vraie simultanément. Par équivalence de toutes les propositions fausses, toutes les propositions fausses sont vraies. Du moins si on accepte l'existence d'un ensemble de tous les ensembles, qui conduit à cette inconsistance axiomatique.

Enfin, la superfamille exponentielle contiendrait toutes les distributions bénéficiant d'une fonction g de densité de probabilité ou de masse. En effet, cette distribution peut se paramétrer par un $\theta \in \{0\}$ avec $f(x|\theta) = h(x)\exp\left({}^{t}(\eta(\theta))T(x) - A(\theta)\right)$ où $h = g,\ \eta(0) = 0,\ T(x) = 0$ pour tout $x,\ A(0) = 0,$ c'est-à-dire, $f(x|\theta) = h(x)\exp(0) = h(x)$.

1.4.3.3 Distributions de Poisson

La fonction de masse d'une distribution de Poisson (distribution discrète), définie sur l'espace $S = \mathbb{N}$ des entiers naturels s'écrit :

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Où $\lambda \in \mathbb{R}_+^*$, l'espérance de la distribution, est le paramètre de la famille de distribution.

Nous pouvons la réécrire comme :

$$f(x|\lambda) = \frac{1}{x!}\exp(\log(\lambda)x - \lambda)$$

Ainsi avec les définitions de fonction suivantes :

$$h : x \to \frac{1}{x\,!}$$

$$\eta : \lambda \to \log(\lambda)$$

$$T : x \to x$$

$$A : \lambda \to \lambda$$

On constate que la famille des lois de Poisson est une famille exponentielle de paramètre naturel $\log(\lambda)$.

1.4.3.4 Distributions binomiales

La famille des distributions binomiales est une famille de variables à valeurs dans $\mathbb{N}$ disposant de deux paramètres $n \in \mathbb{N}^*$ et $\pi \in]0;1[$ tels que la fonction de masse s'écrive :

$$f(x) = C_n^x \pi^x (1-\pi)^{n-x}$$

Où $C_n^x = \frac{n!}{(n-x)!x!}$, cette notation étant utilisée pour éviter la confusion avec la notation des vecteurs.

Si on considère la famille binomiale comme paramétrée par $\theta = (n, \pi) \in \mathbb{N}^* \times]0;1[$ alors il ne s'agit pas d'une famille exponentielle. En effet, les zéros de la fonction de masse doivent être indépendants de θ puisque $h(x)\exp\left({}^t\eta(\theta)T(x) - A(\theta)\right) \Leftrightarrow h(x) = 0$; en effet la fonction exponentielle sur $\mathbb{R}$ n'a que des valeurs strictement positives. Mais, si on considère que l'espace des valeurs de f est $\mathbb{N}$, les zéros sont dépendants de n, ce qui est incompatible avec la définition d'une famille exponentielle.

Pour définir la loi binomiale comme une famille exponentielle, il faut préalablement conditionner à n. Pour un n fixé, alors on peut définir un paramètre $\theta = \pi$ univarié tel que :

$$f(x|\theta) = C_n^x \theta^x (1-\theta)^{n-x} = C_n^x \exp(\log(\theta)\, x + \log(1-\theta)(n-x))$$

$$= C_n^x \exp(x(\log(\theta) - \log(1-\theta)) + \log(1-\theta)\, n)$$

$$= C_n^x \exp\left(x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\, n\right)$$

En définissant les paramètres suivants, nous montrons que cette famille est exponentielle :

$$h: x \to C_n^x$$

$$\eta: \theta \to \log\left(\frac{\theta}{1-\theta}\right)$$

$$T: x \to x$$

$$A: \theta \to -n \log(A - \theta)$$

Sur un espace de nombre entiers $S = \{0, \dots, n\}$. En considérant h comme nul pour tout $x > n$, on peut aussi considérer que $S = \mathbb{N}$.

Cela n'est pas sans conséquence. On pourra bénéficier des propriétés des familles exponentielles si on conditionne à n, ce qui n'est pas adapté lorsque n est variable et corrélé aux caractéristiques de l'échantillon, comme dans les expériences comportant de nombreuses analyses intermédiaires susceptibles de conduire à une interruption précoce.

1.4.3.5 Distributions gaussiennes (ou normales) univariées

On peut noter $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \mathbb{R} \times \mathbb{R}_+^*$ le paramètre bivarié de la loi normale à valeurs dans $\mathbb{R}$, où μ en

représente la moyenne et σ^2 la variance. La fonction de densité de probabilité de cette loi normale s'exprime comme :

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2) - \log(\sigma)\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \log(\sigma)\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left({}^t\begin{pmatrix}\mu/\sigma^2 \\ -1/(2\sigma^2)\end{pmatrix}\begin{pmatrix}x \\ x^2\end{pmatrix} - \frac{\mu^2}{2\sigma^2} - \log(\sigma)\right)$$

Il s'agit donc bien d'une famille exponentielle où :

$$h: x \to \frac{1}{\sqrt{2\pi}}$$

$$\eta: \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix} \to \begin{pmatrix}\mu/\sigma^2 \\ -1/(2\sigma^2)\end{pmatrix}$$

$$T: x \to \begin{pmatrix}x \\ x^2\end{pmatrix}$$

$$A: \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix} \to \frac{\mu^2}{2\sigma^2} + \log\left(\sqrt{\sigma^2}\right)$$

Un *paramètre naturel* de la loi est $\begin{pmatrix}\mu/\sigma^2 \\ -1/(2\sigma^2)\end{pmatrix}$.

Pour une variance connue, on peut aussi définir la sous-famille des lois normales de cette variance, dépendant alors d'un unique paramètre $\theta = \mu$. Chacune d'entre elles est une famille exponentielle de distributions.

1.4.4 Vraisemblance, densité de vraisemblance

D'une manière générale, la vraisemblance d'un paramètre dans une famille de distributions est une fonction du paramètre égale à la probabilité d'observer exactement l'échantillon qui a été observé pour le paramètre donné. Cette notion est adaptée aux distributions discrètes.

Si on définit une variable aléatoire *X* correspondant à un échantillon de taille fixe ou variable (cf. section 1.2.7), de variables aléatoires potentiellement corrélées et multidimensionnelles, alors nous pouvons

noter $x = (x_1, \dots, x_n)$ sa réalisation. Pour cette réalisation, nous définissons la fonction de vraisemblance L_x comme

$$L_x : \theta \to f(x|\theta)$$

Où $f : x \to f(x|\theta) = \mathbb{P}_\theta(X = x)$ est la fonction de masse de la distribution de X, paramétrée par θ.

Si les observations sont indépendantes, identiquement distribuées et l'échantillon $X = (X_1, \dots, X_n)$ est de taille fixe, alors $f(x|\theta) = \prod_{i=1}^{n} g(x_i|\theta)$ où g est la fonction de masse commune à tous les X_i.

Le logarithme de la vraisemblance est appelé log-vraisemblance. Si l'échantillon est de taille fixe n et les observations sont indépendantes et identiquement distribuées, la log-vraisemblance s'exprime alors sous forme de somme :

$$\log\big(L_x(\theta)\big) = \log(\prod_{i=1}^{n} g(x_i|\theta)) = \sum_{i=1}^{n} \log\big(g(x_i|\theta)\big).$$

Quel que soit x, cette fonction de vraisemblance est nulle pour les distributions continues. Il faut alors utiliser la notion de densité de vraisemblance pour les distributions absolument continues. Elle associe à une valeur θ la densité de probabilité en x, le paramètre θ étant fixe pour ce calcul de densité de probabilité. Par abus de langage, on parle souvent de vraisemblance et de log-vraisemblance plutôt que de densité de vraisemblance et de logarithme de densité de vraisemblance.

Une multiplication de la variable par une constante, comme lorsqu'on en change l'unité de mesure physique (p.ex. taille d'un individu en mètres → taille en centimètres) va proportionnellement diviser la densité de vraisemblance. Avec des unités d'échelle très grandes (p.ex. kilomètres pour la taille d'un être humain), il est aisément possible d'obtenir une densité de vraisemblance supérieure à 1 et donc, une log-densité de vraisemblance positive.

Il est possible de construire des hybrides entre vraisemblance et densité de vraisemblance lors de l'analyse de deux variables de nature différente, l'une étant discrète, l'autre continue. Si on note $X = (X_1, X_2)$ et que X_1 est continue à valeurs dans $\mathbb{R}^n$ et X_2 est discrète à valeurs dans $\mathbb{N}^m$ alors on peut définir $L_x : \theta \to f_1(x_1) \times f_2(x_2|X_1 = x_1)$ où f_1 représente la fonction de masse de X_1 conditionnelle à θ et $f_2(x_2|X_1 = x_1)$ représente la densité de probabilité de X_2 en x_2, conditionnelle à θ et à $X_1 = x_1$.

La vraisemblance a une propriété étonnante sur les échantillons de taille variable d'observations identiquement distribuées disposant d'une condition d'arrêt déterministe, tel qu'on en génère avec des

analyses intermédiaires : la vraisemblance est indépendante de la condition d'arrêt et peut être calculée comme si l'échantillon était de taille fixe, égale à sa réalisation n.

Considérons la fonction d'arrêt A, qui a toute réalisation x de l'échantillon associe la probabilité d'arrêt (0 ou 1 puisque la condition est déterministe). Alors la vraisemblance $L_x(\theta)$ est indépendante de A. En effet, elle s'exprime comme le produit des probabilités d'observer x_i conditionnellement aux observations x_1 à x_{i-1} et par les probabilités de non-arrêt pour $i = 1, \dots, n-1$ conditionnelles à $(x_1, \dots, x_i)$ et la probabilité d'arrêt pour $i = n$ (égale à 1). Grace au déterminisme de la condition d'arrêt selon la suite des x_i, les probas de non-arrêt et d'arrêt sont toutes égales à 1, ce qui ne change pas la vraisemblance $L_x(\theta)$.

Un simple exemple peut être illustré par le tirage aléatoire dans une distribution de Bernoulli de paramètre π, d'observations identiquement distribuées.

Considérons pour première condition d'arrêt A_1 le fait que l'échantillon ait atteint la taille 3 et pour seconde condition d'arrêt A_2 le fait que l'échantillon ait atteint la taille 3 ou qu'un tirage ayant la valeur 0 ait été réalisé.

Ci-dessous, les réalisations avec leurs probabilités associés dans les deux cas de figure :

Règle d'arrêt	A_1	
Réalisation x	Probabilité de cette réalisation $f(x\|\pi)$ ou $L_x(\pi)$	Probabilité pour $\pi = 0{,}5$
(0,0,0)	$(1-\pi)^3$	1/8
(0,0,1)	$(1-\pi)^2\pi$	1/8
(0,1,0)	$(1-\pi)^2\pi$	1/8
(0,1,1)	$(1-\pi)\pi^2$	1/8
(1,0,0)	$(1-\pi)^2\pi$	1/8
(1,0,1)	$(1-\pi)\pi^2$	1/8
(1,1,0)	$(1-\pi)\pi^2$	1/8
(1,1,1)	π^3	1/8

Règle d'arrêt	A_2		
Réalisation x	Probabilité de cette réalisation $f(x\|\pi)$ ou $L_x(\pi)$	Probabilité pour $\pi = 0{,}5$	Probabilité pour $\pi = 0{,}8$
(0)	$1-\pi$	1/2	0,200
(1,0)	$(1-\pi)\pi$	1/4	0,160
(1,1,0)	$(1-\pi)\pi^2$	1/8	0,128
(1,1,1)	π^3	1/8	0,512

Même si les réalisations possibles ne sont pas du tout les mêmes dans les deux cas de figure, on constate que la fonction de vraisemblance sous condition d'arrêt A_2 se calcule comme si la taille d'échantillon était fixe.

On peut se poser la question suivante : les observations sont-elles indépendantes ? La question de l'indépendance de X_1 et de X_2 est délicate puisque X_2 n'est pas toujours réalisée. Conditionnellement à la réalisation des deux, elles sont bien indépendantes, la première étant constante égale à 1 et la seconde suivant une variable de Bernouilli de probabilité π. Elles ne sont cependant pas identiquement distribuées conditionnellement à la réalisation des deux. 2ème point de vue : la distribution de X_2 est définie ou indéfinie selon la réalisation de X_1, donc la distribution de X_2 est dépendante de X_1, donc les variables ne sont pas indépendantes. Elles ne sont pas non plus identiquement distribuées, car la distribution de X_1 est toujours bien définie alors que celles de X_2 est inconstamment définie.

Cela a une conséquence importante : l'inférence basée sur le rapport de vraisemblance est fortement altérée en cas de condition d'arrêt. Certains bayésiens ont remarqué que le facteur bayésien permettant de s'orienter vers une ou l'autre de deux hypothèses $H_1 : \theta = \theta_1$ et $H_2 : \theta = \theta_2$, défini comme $\frac{\mathbb{P}(X|\theta=\theta_1)}{\mathbb{P}(X|\theta=\theta_2)} = \frac{L_X(\theta_1)}{L_X(\theta_2)}$ a une interprétation indépendante de la règle d'arrêt. Un seuil sur le facteur bayésien comme règle d'arrêt n'est pas possible car la probabilité d'arrêt pourrait être inférieure à 1. Comme montré par Sanborn *et al* [6] cela n'est vrai que si les hypothèses H_1 et H_2 sont simples, et complémentaires, c'est-à-dire, les distributions des X_i sont complètement définies par le paramètre θ, et que seules les hypothèses d'égalité de θ à θ_1 et à θ_2 sont possibles, c'est-à-dire, la probabilité *a priori* de $\theta \in \{\theta_1, \theta_2\}$ est égale à 1. Cela implique, entre autres, que toute valeur de θ comprise entre θ_1 et θ_2

doivent être impossibles. La plupart du temps en biostatistique, les hypothèses sont composites, avec, par exemple, une hypothèse $H_0: \theta \leq \theta_0$ et $H_1: \theta > \theta_0$. Auquel cas, en appliquant un prior sur θ, la vraisemblance de tout échantillon sous chacune des hypothèses peut être calculée par la moyenne des vraisemblances conditionnelles à tout θ pondérée par la probabilité *a priori* de ce θ. Le rapport de vraisemblance (facteur bayésien obtenu) pourra être instable si $\theta = \theta_0$; une condition d'arrêt portant sur un seuil du facteur bayésien permettra de s'arrêter presque sûrement sur l'hypothèse que l'on souhaite montrer. Un autre problème de l'approche bayésienne classique, c'est de réévaluer la distribution *a posteriori* de θ en se basant sur des rapports de vraisemblance entre tous les $\theta \in \Theta$, en ignorant à chaque fois qu'il n'y a pas que deux valeurs possibles pour θ.

Sanborn *et al* [6] montrent que même avec une hypothèse composite très simple, sur une variable de Bernoulli dont le paramètre a 25% de chances d'être égal à 0,25 et 75% de chances d'être égal à 0,75, l'inférence bayésienne est invalide. Bien que l'inférence fréquentiste soit très proche d'une inférence bayésienne avec des distributions *a priori* non informatives (*p.ex.* prior de Jeffreys), ce problème de condition d'arrêt ne fait pas débat lorsqu'il est observé sous l'angle fréquentiste. Considérons par exemple le rejet d'une hypothèse nulle $H_0: \theta = \theta_0$ par un test du rapport de vraisemblance généralisé basé sur la statistique $LLR = -2 \log\left(\frac{L_X(\hat{\theta})}{L_X(\theta_0)}\right)$. Sur un échantillon de taille fixe, la statistique LLR est généralement approximée à une loi du χ^2. En définissant, pour condition d'arrêt un seuil sur le χ^2 et donc, sur la P-valeur, on peut aisément prouver presque sûrement l'hypothèse que l'on souhaite, car cette P-valeur ne tend pas à se stabiliser avec la taille de l'échantillon (cf. section 1.3.2.2). La correction de ce biais dépend de la condition d'arrêt, la distribution (fluctuations d'échantillonnage) de LLR devant être analysée selon cette condition. Si un seuil est imposé sur le LLR pour l'arrêt, alors, évidemment la distribution ne peut prendre aucune valeur en dessous de ce seuil.

Si l'inférence bayésienne avec prior non informatif est presque équivalente à l'inférence fréquentiste et les deux inférences tendent à être équivalentes dès lors que l'échantillon contient une information bien supérieure au prior, pourquoi le problème de la condition d'arrêt (analyses intermédiaires) parfaitement connu en inférence fréquentiste serait-il absent de l'inférence bayésienne ? Cela fait encore débat en

2020, Ryan *et al* [7] concluant de manière ambiguë :

> "To demonstrate control of type I error in Bayesian adaptive designs, **adjustments to the stopping boundaries are usually required** for designs that allow for early stopping for efficacy as the number of analyses increase […] If one instead uses a strict Bayesian approach, which is currently more accepted in the design and analysis of exploratory trials, then **type I errors could be ignored** and the designs could instead focus on the posterior probabilities of treatment effects of clinically-relevant values."
>
> — Ryan *et al*, 2020 [7]
>
> (emphase ajoutée)

1.4.5 Modèle linéaire généralisé

1.4.5.1 Définition

Le modèle généralisé, comme son nom l'indique est une généralisation du modèle linéaire général. Le modèle linéaire général est le cas particulier du modèle linéaire généralisé, avec une fonction de lien identité et une loi de distribution normale.

Étant donné un échantillon de n observations décrit par une variable aléatoire à expliquer $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ et par p caractéristiques explicatives à valeurs dans $\mathbb{R}$, supposées constantes que nous représenterons sous forme d'une matrice $x = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$.

Considérons que chaque Y_i a une fonction de masse ou de densité de probabilité exprimable comme :

$$f(y|\theta,\phi) = \exp\Big(\alpha(\phi)\big(y\theta - g(\theta) + h(y)\big) + \beta(\phi,y)\Big)$$

avec $\theta \in \mathbb{R}$, $\phi \in \mathbb{R}$, $\alpha\colon \mathbb{R} \to \mathbb{R}_+^*$, $g\colon \mathbb{R} \to \mathbb{R}$, $h\colon \mathbb{R} \to \mathbb{R}$, $\beta\colon \mathbb{R}^2 \to \mathbb{R}$, selon les notations de Nelder et Wedderburn [8], de telle sorte que pour un ϕ constant, la distribution est une famille exponentielle mono-paramétrique selon θ.

Pour ϕ constant, on note $T(y) = \alpha(\phi)y$, $u(\theta) = \alpha(\phi)g(\theta)$ et $v(y) = \exp\big(\beta(\phi,y) + h(y)\big)$, la

fonction f s'exprimant alors comme :

$$f(y|\theta) = v(y)\exp\big(\theta T(y) - u(\theta)\big)$$

Ce qui correspond à une famille exponentielle mono-paramétrique dans $\mathbb{R}$, avec $\theta \in \mathbb{R}$ comme paramètre naturel. Comme précédemment décrit, ce paramètre naturel n'est pas unique (cf. section 1.4.3.1) et on pourrait aussi bien choisir $\alpha(\phi)\theta$ en redéfinissant $T(y) = y$, mais nous ne considérerons que ce paramètre naturel θ dans la suite de ce chapitre. Le paramètre ϕ, ou plutôt méta-paramètre car en dehors de la famille exponentielle analysée, est appelé paramètre d'échelle et va généralement servir à modéliser un paramètre de nuisance telle que la variance pour la loi gaussienne (qui sera analysée comme mono-paramétrique), alors que θ reflètera, à une bijection près, l'espérance de Y_i. Il est à noter que d'une manière générale, avec une paramétrisation bivariée $\begin{pmatrix}\theta\\ \phi\end{pmatrix}$, la fonction f ne correspond pas à la fonction de masse ou de densité de probabilité d'une famille exponentielle, notamment à cause de la fonction β qui n'impose aucune condition sur la manière de combiner ϕ avec y alors que la relation entre le paramètre et la variable y serait strictement contraint à une combinaison linéaire dans une famille exponentielle.

Supposant que tous les Y_i appartiennent à cette famille avec le même ϕ, et notant f une fonction bijective dérivable de $\mathbb{R}$ dans $\mathbb{R}$, appelée fonction de lien, dont le domaine de définition comprend l'ensemble des espérances possibles dans la famille de distribution, alors l'échantillon suit un modèle linéaire généralisé si tous les Y_i sont indépendants et s'il existe $\beta = \begin{pmatrix}\beta_1\\ \vdots\\ \beta_p\end{pmatrix}$ tel que :

$$\begin{pmatrix}\varphi(\mathbb{E}[Y_1])\\ \vdots\\ \varphi(\mathbb{E}[Y_n])\end{pmatrix} = x\beta$$

Par souci de simplification pour une valeur $y = \begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}$ nous noterons $\varphi(y) = \begin{pmatrix}\varphi(y_1)\\ \vdots\\ \varphi(y_n)\end{pmatrix}$. L'expression du modèle devient alors :

$$\varphi(\mathbb{E}[Y]) = x\beta$$

Cette expression implique l'additivité et la linéarité des effets des prédicteurs sur $\varphi^{-1}(\mathbb{E}[Y])$.

Généralement, la première colonne de la matrice x est égale à 1, représentant ainsi un effet présent dans toutes les observations ; il s'agit de l'ordonnée à l'origine ou intercept.

On parle de fonction de lien canonique φ, si elle fait le lien entre l'espérance d'une distribution de la famille mono-paramétrique et le paramètre canonique, c'est-à-dire $\forall\theta \in \mathbb{R}, \varphi(\mathbb{E}[f_\theta]) = \theta$ où $\mathbb{E}[f_\theta] = \int_{-\infty}^{+\infty} xf(x|\theta,\phi)dx$ pour une distribution continue ou $\mathbb{E}[f_\theta] = \sum_{x\in S} xf(x|\theta,\phi)$ où S est le support, fini ou infini dénombrable, d'une distribution discrète.

On note $\eta = \begin{pmatrix}\eta_1\\ \vdots \\ \eta_n\end{pmatrix} = \varphi^{-1}(\mathbb{E}[Y]) = x\beta$ le prédicteur linéaire de Y.

1.4.5.2 Exemple de la loi binomiale

Comme précisé en section 1.4.3.4, la famille des lois binomiales $\mathcal{B}(m;\pi)$ pour un m fixé est une famille exponentielle avec pour paramètre naturel $\theta = \log\left(\frac{\pi}{1-\pi}\right) = \text{logit}\left(\frac{\mu}{m}\right) = \text{logit}\left(\frac{1}{m}\mathbb{E}[f_\theta]\right)$. La fonction $p \to \text{logit}\left(\frac{p}{m}\right)$ est donc une fonction de lien canonique. En se conformant à la notation $f(y|\theta,\phi) = \exp\Big(\alpha(\phi)\big(y\theta - g(\theta) + h(y)\big) + \beta(\phi,y)\Big)$ de Nelder et Wedderburn [8], on a

$$\alpha(\phi) = 1$$

$$g(\theta) = m\log(1+\exp(\theta))$$

$$h(y) = 0$$

$$\beta(\phi,y) = \log\big(C_m^y\big)$$

Avec :

$$f(y|\theta,\phi) = \exp\big(y\theta - m\log(1+\exp(\theta)) + \log\big(C_m^y\big)\big)$$

Le paramètre d'échelle ϕ n'a aucune influence sur la distribution. On peut le supposer constant égal à 1. On constate que cette famille ne comprend pas la loi binomiale avec $\pi = 1$ ou $\pi = 0$, car le paramètre $\theta = \log\left(\frac{\pi}{1-\pi}\right)$ n'est défini que pour $\pi \in]0;1[$.

1.4.5.3 Exemple de la loi de Poisson

La fonction de lien canonique est $\varphi: \lambda \to \log(\lambda)$ et on peut noter $\theta = \log(\lambda)$ le paramètre naturel où λ représente l'espérance de la loi de Poisson.

$$f(y|\theta,\phi) = \frac{1}{y!}\exp(log(\lambda)y - \lambda) = \exp(\theta y - \exp(\theta) - \log(y!))$$

Et donc :

$$\alpha(\phi) = 1$$

$$g(\theta) = \exp(\theta)$$

$$h(y) = 0$$

$$\beta(\phi, y) = -\log(y!)$$

Pour la loi binomiale, les situations extrêmes avec $\pi = 1$ et $\pi = 0$ ne sont pas modélisées. Il en est de même avec la loi de Poisson qui n'est pas modélisée pour $\lambda = 0$.

1.4.5.4 Exemple de la loi Normale

Pour la loi normale $\mathcal{N}(\mu; \sigma^2)$, la fonction de lien canonique est la fonction identité. Pour le paramètre naturel $\theta = \mu$ et le paramètre d'échelle $\phi = \sigma^2$, la fonction de densité de probabilité s'exprime sous la forme suivante :

$$f(y|\theta,\phi) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right) = \exp\left(\frac{-1}{2\phi}(-2y\theta + y^2 + \theta^2) + \log\left(\frac{1}{\sqrt{2\pi\phi}}\right)\right) =$$

$$\exp\left(\frac{1}{\phi}\left(y\theta - \frac{1}{2}\theta^2 - \frac{y^2}{2}\right) + \log\left(\frac{1}{\sqrt{2\pi\phi}}\right)\right)$$

$$\alpha(\phi) = \frac{1}{\phi}$$

$$g(\theta) = \frac{1}{2}\theta^2$$

$$h(y) = -\frac{y^2}{2}$$

$$\beta(\phi, y) = \log\left(\frac{1}{\sqrt{2\pi\phi}}\right)$$

Comme précédemment précisé, $\phi = \sigma^2$ sera supposé identique pour tous les Y_i, c'est-à-dire, on supposera l'homoscédasticité.

1.4.5.5 Variables explicatives aléatoires

En section 1.4.5.1, les facteurs explicatifs (ou prédicteurs) ont été décrits sous forme d'une matrice x

supposée constante. Ce cas de figure existe lorsque les facteurs sont expérimentalement maîtrisés, mais dans le domaine médical, il est bien plus fréquent d'avoir des facteurs explicatifs aléatoires. Le modèle alternatif s'écrit :

$$Y = X\beta$$

Où X est la matrice $n \times p$ aléatoire représentant ces variables explicatives. Les méthodes d'estimation (p.ex. maximum de vraisemblance) de β du modèle linéaire généralisé avec matrice explicative fixe (cf. section 1.4.5.1) peuvent s'appliquer, en conditionnant sur la réalisation x de X. Cela peut être à l'origine de biais d'estimation s'il existe une covariance non nulle entre $\hat{\beta}$ et X. Si le modèle est exact, alors les fluctuations d'échantillonnage de $\hat{\beta}$ sont faiblement corrélés à ceux de X. Dans le cas particulier du modèle identité-gaussien, les résidus sont à l'origine des fluctuations d'échantillonnage de $\hat{\beta}$ et ces résidus sont indépendants de toute chose, y compris de X. Cela n'est pas forcément vrai s'il existe des interactions non spécifiées, puisqu'une fluctuation d'échantillonnage de X augmentant la taille d'un sous-groupe dans lequel un effet β_i est plus fort qu'un autre, ira vers une fluctuation positive de $\widehat{\beta_i}$. Par ailleurs, dans les cas limites, sur de petits échantillons, les estimateurs conditionnels à x peuvent être assez biaisés (Firth [9] et Kenne [10] ont tenté de réduire ces biais) et cette covariance entre $\hat{\beta}$ et X apparaît. La forte dépendance qui existe entre $\hat{\beta}$ et Y n'est pas un problème puisqu'elle est prise en compte par le modèle décrit en 1.4.5.1.

Le conditionnement à la taille d'échantillon n, ce dernier étant rarement bien maîtrisé dans le domaine médical, peut aussi être problématique s'il existe une dépendance entre n et la distribution de $\hat{\beta}$ conditionnelle à n. Encore une fois, sur de petits échantillons, ce problème peut apparaître. Notamment, lorsque l'aspect discret des fluctuations d'échantillonnage de $\hat{\beta}$ apparaît comme dans le scenario de l'estimation d'un simple pourcentage (modèle à intercept seul) comme décrit par Brown *et al* [11].

1.4.5.6 Estimation du maximum de vraisemblance

L'estimateur du maximum de vraisemblance permet généralement l'estimation ponctuelle du vecteur de coefficients β ; il existe d'autres estimateurs comme celui de Firth [9], de Kenne [10], les estimateurs bayésiens à *prior* informatif, les régressions pénalisées LASSO, Ridge, ElasticNet.

L'estimateur du maximum de vraisemblance reste le plus utilisé. Il s'agit du $\hat{\beta}$ maximisant la vraisemblance, c'est-à-dire tel que :

$$L(\hat{\beta}) = \max_{\beta \in \mathbb{R}^p} L(\beta)$$

Pour une réalisation y de Y, ce maximum n'est pas toujours unique. Cela arrive notamment en cas de colinéarité des facteurs explicatifs, c'est-à-dire lorsqu'une combinaison linéaire de colonnes de x est égale à une autre colonne. Par exemple, si la colonne 1 de x est la somme des colonnes 2 et 3 alors pour toute constante c, $x\beta = x\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{pmatrix} = x\begin{pmatrix} \beta_1 + c \\ \beta_2 - c \\ \beta_3 - c \\ \vdots \\ \beta_p \end{pmatrix}$. Dans ce cas, si la vraisemblance comporte un maximum, une infinité de β atteignent ce maximum de vraisemblance, puisqu'à partir de n'importe quelle solution, on peut en générer une infinité par le choix de ces constantes c. Ce problème existe dès lors qu'une combinaison linéaire de colonnes est nulle. En appliquant la méthode du pivot de Gauss à la transposée de x, on peut en mesurer le rang. Si ce rang est strictement inférieur à p, alors il existe un problème de multicolinéarité et il ne peut plus y avoir de $\hat{\beta}$ du maximum de vraisemblance unique. Cette situation se rencontre dans le cadre du modèle Âge – Période – Cohorte. Comme la période est la somme de la cohorte et de l'âge, il est impossible d'estimer simultanément les trois effets. Cela peut aussi survenir sur de petits échantillons lorsque les facteurs sont des variables aléatoires assez fortement corrélées ou de faible variance, conduisant à une colonne de variance nulle (colinéaire avec l'intercept qui est la colonne constante égale à 1) ou à deux colonnes colinéaires par hasard.

Un deuxième problème possible, c'est l'absence de maximum de vraisemblance dans $\mathbb{R}^p$ parce que la fonction de vraisemblance tend vers 1, sans jamais l'atteindre lorsqu'un β_i tend vers $+\infty$ ou $-\infty$, les autres étant fixés. Par exemple, dans une régression logistique, la présence de y_i à zéro dans certains sous-groupes peut être à l'origine de ce problème. Considérons par exemple l'estimation d'un modèle linéaire généralisé à fonction de lien logit et distribution binomiale (régression logistique) avec $m = 10$, $x = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, $y = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$. Alors :

$$x\beta = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \end{pmatrix}$$

Selon les hypothèses du modèle :

$$\mathbb{E}[Y_1] = \text{logit}^{-1}(\beta_1) = \frac{1}{1 + \exp(-\beta_1)}$$

$$\mathbb{E}[Y_2] = \text{logit}^{-1}(\beta_1 + \beta_2)$$

$$Y_1 \sim \mathcal{B}\big(m; \text{logit}^{-1}(\beta_1)\big)$$

$$Y_2 \sim \mathcal{B}\big(m; \text{logit}^{-1}(\beta_1 + \beta_2)\big)$$

Y_1 et Y_2 étant supposés indépendants, on peut décomposer la vraisemblance en L_1 et L_2 avec, selon la section 1.4.5.2 :

$$L_1(\beta) = \exp\big(y_1\beta_1 - m\log(1 + \exp(\beta_1)) + \log\big(C_m^{y_1}\big)\big)$$

$$L_2(\beta) = \exp\big(y_2(\beta_1 + \beta_2) - m\log(1 + \exp(\beta_1 + \beta_2)) + \log\big(C_m^{y_2}\big)\big)$$

et par indépendance de Y_1 et Y_2, $L(\beta) = L_1(\beta) \times L_2(\beta)$

Or, comme y_2 est nul, on peut simplifier l'expression de L_2

$$L_2(\beta) = \exp(0 - m\log(1 + \exp(\beta_1 + \beta_2)) + 0))$$

$$= \exp(-m)\,(1 + \exp(\beta_1 + \beta_2))$$

Pour un β_1 fixé, $L_1(\beta)$ est constant et strictement positif mais $\lim_{\beta_2 \to -\infty} L_2\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}\right) = 1$ avec $L_2\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}\right) <$ 1 quel que soit β_2. Ainsi, la vraisemblance a une borne supérieure mais pas de maximum.

Ce problème n'existe pas avec les estimateurs de Firth ou de Kenne.

Pour simplifier, on peut dire qu'il s'agit d'une forme de variante de la division par zéro, ou du calcul de logit(1). Ce cas de figure est insuffisamment diagnostiqué par les logiciels statistiques. C'est ce qu'on appelle le problème de « séparation complète ».

Notamment, la procédure glm.fit du logiciel R (version 4.0.2, The R Foundation for Statistical Computing, Vienne, Autriche) identifie la convergence par la stabilisation de la déviance selon l'instruction suivante :

```
(abs(dev - devold)/(0.1 + abs(dev)) < control$epsilon
```

Où dev représente la déviance non mise à l'échelle (unscaled deviance) du modèle estimé à cette

itération, devold la déviance à la précédente itération et control$epsilon est égal à la valeur 10^{-8} par défaut.

Or, lorsque β_2 est agrandi à chaque itération, la déviance se stabilise vers zéro. La petite constante 0,1 au dénominateur conduit à une perception de stabilisation par le logiciel lorsque β atteint la valeur 24,54. Le logiciel, à tort, croit que l'estimateur du modèle a convergé vers un maximum de vraisemblance alors que ce dernier n'existe pas. Cette constante 0,1 au dénominateur est pourtant nécessaire à ce que des estimations parfaites soient détectées, comme on pourrait avoir si on remplaçait y par $\binom{5}{5}$ dans l'exemple ci-dessus. Il existe une solution assez générale à l'identification automatique de ce cas de figure, soumis sous la forme du rapport de bug #17886 au bugzilla de R par l'auteur de cette thèse. Plutôt que de rechercher une divergence d'un β_i, qui est susceptible sur une échelle très variable selon la matrice x, l'idée est d'identifier la divergence d'un prédicteur linéaire $\widehat{\eta_i}$ vers $-\infty$ ou $+\infty$. Cette divergence s'identifie par le fait que la dérivée de la fonction de lien atteint des valeurs extrêmes, et notamment plutôt que de se stabiliser entre deux itérations, tend à croitre constamment en valeur absolue. Le cœur du code amélioré est présenté ci-dessous :

```
rmax <- max( abs(mu.eta.val/family$mu.eta(eta)) )
```

La valeur mu.eta.val représente la dérivée de $\hat{\eta}$ à l'itération précédente alors que family$mu.eta(eta) représente la dérivée de $\hat{\eta}$ à la dernière itération. La fonction max identifie l'observation pour laquelle ce rapport des deux dérivées est maximal en valeur absolue. Cette constante rmax tend très vite vers 1,00 lorsque le modèle converge et se stabilise généralement au-delà de 2,0 (p.ex. 2,7) lorsqu'un prédicteur linéaire tend vers $+\infty$.

Un deuxième bug est encore dû à la constante 0,1 dans l'expression de la détection de convergence :

```
(abs(dev - devold)/(0.1 + abs(dev)) < control$epsilon
```

Si la dispersion est très faible, comme on pourrait observer sur un modèle log-gaussien avec des variables Y_i comprises entre 10^{-12} et 10^{-11}, alors la déviance non mise à l'échelle est très proche de zéro même pour des valeurs prédites éloignées (de plusieurs écarts types) des valeurs prédites optimales. Le logiciel s'arrête alors avant que la convergence réelle soit atteinte à cause de cette constante 0,1.

En adaptant la constante à la dispersion, le calcul devient indépendant de la dispersion et ce bug semble

disparaître, comme décrit dans le rapport de bug #17885.

La procédure LOGISTIC du logiciel *Statistical Package for the Social Sciences* (IBM® SPSS®, version 26.0.0, 2019) génère un message d'erreur si la variable à expliquer a une variance nulle, mais dans les autres situations de séparation complète, le logiciel fournit un résultat avec des coefficients extrêmes, précisant toutefois que le nombre d'itérations a dépassé le maximum (n=20) sans atteindre la convergence. La procédure GENLOG du même logiciel, adaptée aux régressions logistiques multinomiales est incapable de détecter cette situation.

La procédure logistic du logiciel Stata® (version 13) est capable de détecter la séparation complète, refusant de fournir des estimations, mais la procédure genmod, adaptée aux modèles linéaires généralisés, y compris logit-binomiaux, ne détecte pas cette situation, même lorsque la variance de la variable à expliquer est nulle. De cette manière, la procédure genmod se comporte comme la fonction glm() du logiciel R, fournissant des estimations aberrantes tout en se vantant d'une convergence rapide, en deux itérations.

Le logiciel SAS (version 9.4 TS Level 1M2) est capable de détecter le problème de convergence avec les procédures PROC LOGISTIC (« complete separation of data points ») et PROC GENMOD (« The relative Hessian convergence criterion is greater than the limit »), fournissant un message d'avertissement, mais pas de message d'erreur. Les estimations produites sont aberrantes, comme pour les autres logiciels.

Le logiciel Julia (version 1.5.2, Septembre 2020) est incapable de détecter la séparation complète, ne fournissant aucun message d'avertissement ni d'erreur.

1.4.5.7 Déviance mise à l'échelle et non mise à l'échelle

1.4.5.7.1 Définition

Notons la vraisemblance $L_\theta(\theta) = \prod_{i=1}^{n} f(y|\theta_i, \phi)$ la vraisemblance d'un vecteur de paramètres $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$ et $L_\beta(\beta) = L_\theta\left(c\left(\varphi^{-1}(x\beta)\right)\right)$ la vraisemblance du vecteur β où c est la fonction de lien canonique, faisant le lien entre l'espérance de la loi et le paramètre de la famille exponentielle analysée.

On définit la déviance (encore dite déviance mise à l'échelle) comme la fonction qui associe à un vecteur

β la valeur $\mathrm{D}^*(\beta) = 2\left(\log(L_{sup}) - \log\left(L_\beta(\beta)\right)\right)$, où L_{sup} représente la borne supérieure de la vraisemblance théorique L_θ sur l'ensemble des paramètres $\theta = \begin{pmatrix}\theta_1 \\ \vdots \\ \theta_n\end{pmatrix}$ possibles. C'est-à-dire :

$$L_{sup} = \sup_{\theta \in \mathbb{R}^n} L_\theta(\theta)$$

Cette vraisemblance est abusivement notée L_{max} par certains auteurs, et considérée comme la vraisemblance maximale théorique, d'un modèle saturé, mais cela n'est pas toujours correct, car ce maximum n'existe pas toujours et le modèle saturé n'a pas toujours de maximum de vraisemblance. Par exemple, considérons la famille de lois binomiales mono-paramétriques $\pi \to \mathcal{B}(\pi; m)$ où m représente le nombre de répétitions et $m \times \pi$ l'espérance de la loi. La fonction de lien canonique, associant le paramètre naturel à l'espérance (cf. section 1.4.5.2) est $c: \mu \to \mathrm{logit}\left(\frac{\mu}{m}\right)$. Cette fonction est indéfinie pour $\mu = 0$ et $\mu = m$, ce qui rend impossible le choix de θ_i qui permette de construire des distributions dont l'espérance serait égale aux valeurs y_i observées dès lors qu'un des y_i est égal à 0 ou m. Ce cas de figure où $y_i \in \{0, m\}$ est pourtant systématique pour $m = 1$ (distributions de Bernoulli). Il est néanmoins possible d'approcher le cas où les $\mu_i = c^{-1}(\theta_i)$ seraient arbitrairement proches de y_i en construisant une suite de paramètres θ.

La suite qui a tout $k \in \mathbb{N}^*$ associe $o(k) = \begin{pmatrix}\mathrm{logit}\left(\frac{1}{m}(y_1 - (y_1 - m/2)/k)\right) \\ \vdots \\ \mathrm{logit}\left(\frac{1}{m}(y_n - (y_n - m/2)/k)\right)\end{pmatrix}$ a une vraisemblance $L_\theta\big(o(k)\big)$ qui tend vers L_{sup} quand $k \to +\infty$. Pour le cas où $m = 1$, cette vraisemblance tendra vers 1. D'une manière générale, on peut définir une suite $k \to o(k)$ dont la vraisemblance tende vers L_{sup}.

La déviance est alors égale à :

$$\mathrm{D}^*(\beta) = 2\left(\lim_{k \to +\infty} \log\left(L_\theta\big(o(k)\big)\right) - \log\left(L_\beta(\beta)\right)\right)$$

$$= 2\sum_{i=1}^{n}\left(\lim_{k \to +\infty}\left(\alpha(\phi)\left(y_i o_i(k) - g\big(o_i(k)\big) + h(y_i)\right) + \beta(\phi, y_i)\right)\right.$$

$$\left. - \left(\alpha(\phi)\big(y_i\theta_i - g(\theta_i) + h(y_i)\big) + \beta(\phi, y_i)\right)\right)$$

où $\theta = c\left(\varphi^{-1}(x\beta)\right)$ avec c la fonction de lien canonique.

Les termes $\beta(\phi, y_i)$, $\alpha(\phi)$ et $h(y_i)$, liés à y et au paramètre de dispersion, sont indépendants de k ou de β, ce qui permet une grande simplification :

$$\mathrm{D}^*(\beta) = 2\alpha(\phi)\sum_{i=1}^{n}\left(\lim_{k\to+\infty}\left(y_i o_i(k) - g\left(o_i(k)\right)\right) - \left(y_i\theta_i - g(\theta_i)\right)\right)$$

$$= 2\alpha(\phi)\left(z(y) - \sum_{i=1}^{n}\left(y_i\theta_i - g(\theta_i)\right)\right)$$

Où z est une fonction de y, totalement indépendante de β, d'expression similaire à $\log\left(L_{sup}\right)$ mais amputée des termes $\beta(\phi, y_i)$, $\alpha(\phi)$ et $h(y_i)$.

On définit ensuite la déviance non mise à l'échelle (unscaled deviance) comme $D(\beta) = \frac{1}{\alpha(\phi)} D^*(\beta)$ soit :

$$D(\beta) = 2\left(z(y) - \sum_{i=1}^{n}\left(y_i\theta_i - g(\theta_i)\right)\right)$$

Dont l'expression fait totalement disparaître le paramètre de nuisance ϕ. Comme $\alpha(\phi)$ est constant et strictement positif, le β minimisant D^* est aussi le β minimisant D et maximisant la vraisemblance L_β.

1.4.5.7.2 Exemple d'une famille gaussienne

Pour une famille de distributions gaussiennes, la suite constante qui à k associe $o(k) = y$ permet d'atteindre immédiatement la borne supérieure L_{sup} (qui est alors un maximum $L_{max} = L_{sup}$) et l'expression de la déviance non mise à l'échelle se simplifie :

$$D(\beta) = 2\sum_{i=1}^{n}\left(y_i^2 - g(y_i)\right) - \left(y_i\theta_i - g(\theta_i)\right)$$

où $g(\mu) = \frac{1}{2}\mu^2$ comme décrit en section 1.4.5.4.

$$D(\beta) = 2\sum_{i=1}^{n}\left(y_i^2 - \frac{1}{2}y_i^2 - y_i\theta_i + \frac{1}{2}\theta_i^2\right)$$

$$= \sum_{i=1}^{n}\theta_i^2 + y_i^2 - 2y_i\theta_i$$

$$= \sum_{i=1}^{n} (\theta_i - y_i)^2$$

Où $\theta = \varphi^{-1}(x\beta)$

L'estimateur des moindres carrés est alors équivalent à l'estimateur du maximum de vraisemblance pour toute distribution gaussienne, peu importe la fonction de lien.

1.4.5.7.3 Exemple d'une famille binomiale

Pour les familles binomiales (cf. section 1.4.5.2), $\alpha(\phi) = 1$, et donc les déviances mise et non mise à l'échelle D^* et D sont égales.

Comme décrit en section 1.4.5.7.1, la vraisemblance maximale théorique L_{sup} ne correspond pas toujours à un paramètre $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$ calculable, mais peut correspondre à la limite d'une suite o convergente vers cette borne supérieure. Cela est dû au cas où un ou plusieurs des y_i atteignent la valeur 0 ou m. Si on change la paramétrisation de la famille binomiale, remplaçant $\theta_i = c(\mu_i)$ par μ_i les espérances des distributions, puis qu'on la généralise pour inclure le cas où les μ_i atteignent 0 ou m, alors la vraisemblance maximale théorique L_{sup} devient directement calculable, sans passer par une suite convergente :

$$L_{sup} = \prod_{i=1}^{n} \left(C_m^{y_i} \left(\frac{y_i}{m}\right)^{y_i} \left(1 - \frac{y_i}{m}\right)^{m-y_i} \right)$$

Et ainsi, en notant $\mu_i = \varphi^{-1}(x\beta)$ où φ est la fonction de lien :

$$D(\beta) = D^*(\beta) = 2\left(\log\left(\frac{L_{sup}}{L_\beta(\beta)}\right)\right)$$

$$= 2\log\left(\prod_{i=1}^{n}\left(\left(\frac{y_i}{\mu_i}\right)^{y_i}\left(\frac{1 - y_i/m}{1 - \mu_i/m}\right)^{m-y_i}\right)\right)$$

$$= 2\sum_{i=1}^{n}\log\left(\left(\frac{y_i}{\mu_i}\right)^{y_i}\right) + \log\left(\left(\frac{1 - y_i/m}{1 - \mu_i/m}\right)^{m-y_i}\right)$$

Dans le cas où y_i diffère de 0 et de m, l'expression se simplifie, les exponentiations par y_i et $m - y_i$ s'exprimant comme des multiplications après le passage au log. Si y_i égale zéro, alors l'expression dans la somme devient $-m\log(1 - \mu_i/m)$ alors que si y_i égale m, l'expression devient $-m\log(\mu_i/m)$.

Finalement, la dépendance à $C_m^{y_i}$ disparaît complètement grâce au rapport entre L_β et L_{sup}.

1.4.5.7.4 Exemple d'une famille de Poisson

Pour les familles de Poisson (cf. section 1.4.5.3), $\alpha(\phi) = 1$, et donc les déviances mise et non mise à l'échelle D^* et D sont égales.

Par ailleurs, la vraisemblance $L_\theta(\theta) = \prod_{i=1}^{n} \exp(\theta y_i - \exp(\theta) - \log(y_i!))$. Comme pour la loi binomiale, il est possible de reparamétriser et étendre la distribution au cas où une ou plusieurs espérances $\mu_i = 0$ avec $L_\mu(\mu) = \prod_{i=1}^{n} \frac{1}{k!} \exp(-\mu_i)\, \mu_i^{y_i}$ de telle sorte que L_{sup} peut s'écrire comme $L_{sup} = \prod_{i=1}^{n} \frac{1}{k!} \exp(-y_i)\, {y_i}^{y_i}$.

En notant $\mu = \varphi^{-1}(x\beta)$, on en déduit :

$$D^*(\beta) = D(\beta) = 2\sum_{i=1}^{n} \log\left(\frac{\exp(-y_i)\, {y_i}^{y_i}}{\exp(-\mu_i)\, \mu_i^{y_i}}\right) = 2\sum_{i=1}^{n} \log\left(\left(\frac{y_i}{\mu_i}\right)^{y_i}\right) + \mu_i - y_i$$

Une fois encore, le calcul du terme de la somme se simplifie en $y_i \log\left(\frac{y_i}{\mu_i}\right) + \mu_i - y_i$ pour $y_i > 0$ et en $\mu_i - y_i$ pour $y_i = 0$.

1.4.5.8 Algorithme des moindres carrés itérativement repondérés

1.4.5.8.1 Notations

Contrairement à une ANOVA à facteurs équilibrés (cf. section 1.4.1.2) qui bénéficie d'une solution analytique très simple et au modèle linéaire général (cf. section 1.4.2) dont on peut trouver l'estimation des moindres carrés par inversion et produit de matrices, l'estimation du maximum de vraisemblance du vecteur β d'un modèle linéaire généralisé n'est pas directement calculable analytiquement dans la cas général. Il faut alors utiliser un algorithme itératif, se rapprochant de la solution à chaque itération.

L'algorithme des moindres carrés itérativement repondérés ou Iteratively reweighted least squares (IRWLS) est très performant, convergeant le plus souvent en moins d'une dizaine d'itérations, avec une précision d'une dizaine de chiffres numériquement significatifs. Cet algorithme est défini, ci-dessous.

Considérons un modèle linéaire généralisé avec une fonction de lien φ un vecteur de paramètres $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$, une matrice de caractéristiques explicatives $x = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$ de telle sorte que le modèle

soit basé sur un vecteur de variables à expliquer indépendantes $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ tel que chacune des distributions des Y_i soit définie par une fonction de masse ou de densité de probabilité $f(y|\theta,\phi) = \exp\Big(\alpha(\phi)\big(y\theta - g(\theta) + h(y)\big) + \beta(\phi,y)\Big)$ où $\phi \in \mathbb{R}$ est identique pour tous les Y_i et $\theta \in \mathbb{R}$ est un paramètre univarié potentiellement différent pour chaque i et en bijection avec l'espérance $\mathbb{E}[Y_i]$ tel que :

$$\varphi(\mathbb{E}[Y]) = x\beta$$

On note $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ le vecteur des réalisations de la variable réponse Y.

On cherche à estimer $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$ par $B = \begin{pmatrix} B_1 \\ \vdots \\ B_n \end{pmatrix}$.

On définit *l'espace de réponse* comme l'ensemble des valeurs possibles pour $\mathbb{E}[Y_i]$, tel que l'intervalle $]0;1[$ si les Y_i sont des variables de Bernoulli. On définit ensuite *l'espace linéaire*, comme l'ensemble des valeurs possibles de $\varphi(\mathbb{E}[Y])$; généralement égal à $\mathbb{R}$ tout entier.

1.4.5.8.2 Principe de l'algorithme

Partant d'une estimation (fortement écartée de β ou non) de coefficients de régression B, l'algorithme calcule $u = \varphi^{-1}(xB)$ les prédictions du modèle linéaire généralisé, dans l'espace de réponse, basé sur ces coefficients de régression. Ensuite, l'algorithme calcule $z = \varphi(u) + (y-u)\varphi'(u)$, des prédictions volontairement décalées, dans l'espace linéaire, dans l'idée de régresser linéairement (viser z dans l'espace linéaire) afin de corriger le biais d'estimation dans l'espace réponse.

On estime une régression linéaire des moindres carrés, pondérée par des $w_i = \frac{1}{V(u_i)\times(\varphi'(u_i))^2}$ expliquant z à partir de x afin d'obtenir un nouveau vecteur de coefficients B.

On réitère jusqu'à convergence, c'est-à-dire, stabilité de B ou stabilité de la déviance.

1.4.5.8.3 Algorithme détaillé

La 1ère étape consiste à une initialisation de la réponse $u = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$.

Pour $i = 1, \dots, n$. La meilleure initialisation dépend de la distribution des variables réponse :

On peut proposer $u_i = \frac{(y_i+0,5)}{m+1}$ pour la distribution binomiale $\mathcal{B}(m;\pi_i)$.

On peut proposer $u_i = y_i$ pour les distributions Gamma, gaussienne ou gaussienne inverse.

Le logiciel R (version 4.0.2) utilise $u_i = y_i + 0{,}1$ pour la distribution de Poisson.

La valeur $u_i = (y_i + \bar{y})/2$ convient pour la plupart des distributions.

Ensuite, on calcule $z = \varphi(u) + (y - u)\varphi'(u)$.

Puis, on calcule la matrice $n \times n$ diagonale présentant sur la diagonale les poids w_1 à w_n.

$$w_i = \frac{1}{V(u_i) \times \left(\varphi'(u_i)\right)^2}$$

Où $V(\mu) = \alpha(\phi) \times \mathrm{var}(Y)$ représente la variance qu'aurait une variable aléatoire d'espérance μ appartenant à la même famille de lois après élimination de la dépendance au paramètre de dispersion ϕ. Par exemple, pour la loi normale $\phi = \sigma^2$ représentera la variance de la loi, $\alpha(\phi) = \frac{1}{\phi}$, $\mathrm{var}(\mu) = \phi$ et $V(\mu) = 1$ pour tout μ. Pour la loi binomiale $V(\mu) = \mathrm{var}(\mu) = \frac{(m-\mu)\times\mu}{m}$ et $\alpha(\phi) = 1$.

$\mathrm{var}(u_i) \times \left(f'(u_i)\right)^2$ représente donc, par la méthode delta, une estimation de la variance du prédicteur linéaire $\widehat{\eta_i} = \varphi(u_i)$ pour l'observation i. On se basera sur un modèle linéaire général pondéré par les w_i, et donc par les inverses des variances estimées des prédicteurs linéaires. On peut noter que la multiplication par $\alpha(\phi)$ n'a aucune influence sur l'estimation du modèle linéaire général car c'est une constante, indépendante de i. Cet usage de la fonction V plutôt que var permet de s'abstraire de la connaissance de la dispersion ϕ pour le calcul des poids ; cette dernière pourra être estimée à la fin de l'algorithme seulement.

Ensuite, on estime les coefficients B d'un modèle linéaire général pondéré dont la réponse est représentée par z, les covariables par x et la matrice de poids $w = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix}$.

$$B = ({}^t xwx)^{-1}({}^t xwz)$$

Où ${}^t x$ représente la transposée de la matrice x.

Puis, on calcule les valeurs prédites par les coefficients actuels afin d'obtenir une nouvelle estimation du prédicteur u

$$u = \varphi^{-1}(xB)$$

Puis on réitère l'algorithme avec le nouveau u jusqu'à convergence.

Note : $({}^{t}xwx)^{-1}$ représente une estimation de la matrice de variance-covariance non mise à l'échelle (unscaled covariance matrix). Pour la matrice de variance-covariance mise à l'échelle, il faudra multiplier tous les termes de la matrice par $\frac{1}{\alpha(\hat{\phi})}$. Ce paramètre de dispersion $\frac{1}{\alpha(\hat{\phi})}$ peut être estimé comme :

$$\frac{1}{\alpha(\hat{\phi})} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - u_i)^2}{V(u_i)}$$

Où $u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \varphi^{-1}(xB)$ est l'espérance prédite pour Y a la dernière itération de l'algorithme.

En effet les $(y_i - u_i)^2$ représentent des estimations de la variance des Y_i alors que $V(u_i) = \alpha(\hat{\phi}) \times \text{var}(u_i)$ où les $\text{var}(u_i)$ sont d'autres estimations de la variance des Y_i. Le rapport entre les deux est alors une estimation de $\alpha(\hat{\phi})$. La division par $n-p$ plutôt que par n permet de rectifier le surajustement dû au fait que le paramètre B est estimé sur l'échantillon, tendant à une réduction des espérances des $(Y_i - U_i)^2$ par rapport aux espérances des $(Y_i - \mu_i)^2$. En bref, on divise par le nombre de degrés de liberté.

Cas particulier :

Si la distribution est gaussienne et la fonction de lien est l'identité, alors, lors de la 1ère itération, $z_i = \varphi(u_i) + (y_i - u_i) \times \varphi'(u_i) = u_i + (y - u_i) = y_i$ sera totalement indépendant du choix du u initial. Le vecteur B sera égal aux coefficients du modèle linéaire général estimé par les moindres carrés dès la 1ère itération et restera parfaitement stable à toute itération ultérieure.

Ainsi, on vérifie que le modèle linéaire généralisé identité-gaussien est équivalente à la régression linéaire des moindres carrés. Cela est en cohérence avec le fait que l'estimateur des moindres carrés est équivalent à celui du maximum du vraisemblance pour les familles de distributions gaussiennes mono-paramétriques (cf. section 1.4.5.7.2).

1.4.5.8.4 Exemple d'implémentation minimaliste

L'algorithme ci-dessous implémente la méthode des moindres carrés itérés sous le logiciel R, en se basant sur un minimum de fonctions et faisant un minimum de vérifications.

Afin d'alléger au maximum l'algorithme, plutôt que de partir d'un vecteur u, il part d'un vecteur B, supprimant alors la dépendance à l'appel de la fonction de lien. Par ailleurs, plutôt que de reposer sur la déviance pour identifier la convergence, l'algorithme réalise un nombre d'itérations fixe qui devrait généralement suffire à obtenir une très bonne estimation :

```
# fonction d'estimation d'un modèle linéaire généralisé
# paramètre y : vecteur de longueur n représentant les réponses yi
# paramètre x : matrice des facteurs explicatifs de dimension n×p
# paramètre fam : famille de distribution et fonction de lien
# paramètre fam$family : nom commun de la distribution (character)
# paramètre fam$linkinv : réciproque de la fonction de lien
# paramètre fam$mu.eta : fonction dérivée de la réciproque de la fonction de lien
# paramètre fam$variance : fonction V de variance non mise à l'échelle
# paramètre start : vecteur de paramètres B initial (de longueur p)
# paramètre niterations : nombre d'itérations
# valeur de renvoi : liste contenant
#   $coef = le paramètre B de maximum de vraisemblance
#   $vcov = la matrice de variance-covariance de l'estimateur de ce paramètre B
#   $dispersion = une estimation du paramètre de dispersion
glmx=function(y, x, fam, start=rep(0, ncol(x)), niterations=20) {
      # vérifier la présence des éléments clés de la famille et fonction de lien
      req = c("family", "linkinv", "mu.eta", "variance")
      if (!all(req %in% names(fam))) {
            stop("missing items in fam")
      }
      # vecteur de paramètres initiaux du modèle
      B = start
      for(i in 1:niterations) { # nombre fixe d'itérations
            # calcul du prédicteur linéaire
            eta = x %*% B
            # et du prédicteur sur l'espace des réponses
            mu = fam$linkinv(eta)
            # calcul du z permettant d'itérer la régression linéaire
            z = eta + (y - mu)/fam$mu.eta(eta)
            # calcul des poids
            w = diag(as.vector(fam$mu.eta(eta)^2/fam$variance(mu)))
            # matrice de variance-covariance non mise à l'échelle
            vc = solve(t(x) %*% w %*% x)
            # régression linéaire des moindres carrés
```

```
            B = vc %*% t(x) %*% w %*% z
      }
      # à ce stade, le vecteur B a été estimé
      # nous allons estimer la dispersion 1/alpha(phi)
      if (fam$family %in% c("poisson", "binomial")) {
            dispersion=1
      } else {
            dispersion = sum((y - mu)^2/fam$variance(mu))/(nrow(x)-ncol(x))
      }
      # renvoyons les paramètres estimés
# ainsi que la matrice de variance-covariance mise à l'échelle
      list(coef=as.vector(B), vcov=dispersion * vc, dispersion=dispersion)
}

# famille et fonction de lien correspondant au modèle logit-binomial
logistic=list(
      family  = "binomial",
      linkinv = function(eta) {1/(1+exp(-eta))},
      mu.eta  = function(eta) {exp(-eta)/(1+exp(-eta))^2},
      variance= function(mu) {mu*(1-mu)}
)
# famille et fonction de lien correspondant au modèle identité-gaussien
gaussian=list(
      family  = "gaussian",
      linkinv = identity,
      mu.eta  = function(eta) {rep(1, length(eta))},
      variance= function(mu) { rep(1, length(mu))}
)

# exemple 1
glmx(y=c(0,1,1,1,0), x=cbind(1, c(0,0,1,1,1)), logistic)
# exemple 2
glmx(y=c(0,1,1,1,0), x=cbind(1, c(0,0,1,1,1)), gaussian)
```

Plutôt que de reposer sur la dérivée φ' de la fonction de lien, cet algorithme repose sur la dérivée $(\varphi^{-1})'$ de la réciproque de la fonction de lien, afin de se conformer aux conventions de la fonction `make.link()` du logiciel R. L'algorithme en est à peine modifié. En effet pour tout μ, $\varphi'(\mu) = \frac{1}{(\varphi^{-1})'(\varphi(\mu))}$.

1.4.5.9 Fonctionnalité additionnelle : offset

Il est possible, en ne changeant presque rien à la théorie et aux calculs, d'ajouter un vecteur offset $o =$

$\begin{pmatrix} o_1 \\ \vdots \\ o_n \end{pmatrix}$ constant à un modèle linéaire généralisé. En reprenant les notations de la section 1.4.5.1, l'expression du modèle devient :

$$\varphi(\mathbb{E}[Y_i]) = o + x\beta$$

Le vecteur offset o, comme la matrice de modèle x, doit être connu avant estimation de β. Il s'agit de variables observées ou de facteurs contrôlés.

On remarquera que les modèles linéaires généralisés avec et sans offsets sont mathématiquement équivalents, la distinction n'ayant de sens que pour l'estimation.

En effet, tout modèle sans offset peut être transformé en un modèle avec pour offset $o = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ et tout modèle avec offset peut être transformé en un modèle sans offset en ajoutant une colonne à x contenant les offsets o_i et en ajoutant un élément correspondant à β, dont la valeur est égale à 1.

La différence concerne uniquement l'estimation. Dans un modèle sans offset, les procédures d'estimation que nous avons vues considèrent que tous les β_i sont inconnus. Si certains β_i sont connus et que l'on cherche à estimer les autres, alors on peut additionner les produits des β_i par les colonnes correspondantes de x sous forme d'un unique offset o. Les colonnes de x correspondant aux β_i connus sont ensuite supprimées, conduisant à un modèle avec offset dont le vecteur β ne contient plus que des paramètres inconnus. Cette fonctionnalité est utile aux tests d'hypothèses sur un paramètre. On estime un premier modèle en fixant certains paramètres sous forme d'un offset, que l'on compare, par un test du rapport de vraisemblance, à un second modèle sur une matrice de modèle plus large, où ces paramètres sont librement estimés par le maximum de vraisemblance.

La présence d'un offset change le lien entre β et le prédicteur linéaire η puisque $\eta = o + x\beta$. L'algorithme des moindres carrés itérativement repondérés décrit en section 1.4.5.8.3 est légèrement modifié :

L'initialisation de u est inchangée. Le calcul du z, des poids w_i et de la régression linéaire $B = ({}^t xwx)^{-1}({}^t xwz)$ sont aussi inchangés.

La dernière étape $u = \varphi^{-1}(xB)$ est modifiée. Elle devient $u = \varphi^{-1}(o + xB)$. Le calcul de la déviance

non mise à l'échelle (cf. section 1.4.5.7) est aussi légèrement modifié, la première étape du calcul n'étant plus $\eta = x\beta$ mais $\eta = o + x\beta$.

1.4.5.10 Fonctionnalité additionnelle : poids

Comme pour l'offset, la fonctionnalité de poids ne change pas la nature du modèle mais juste la méthode d'estimation. On suppose toujours que $\varphi(\mathbb{E}[Y_i]) = o + x\beta$, mais on associe un poids plus fort à certaines observations que d'autres dans le calcul du maximum de vraisemblance. Cela se fait à partir d'un vecteur $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ où les $v_i \in \mathbb{R}_+^*$. Le calcul de la déviance non mise à l'échelle (élément à minimiser) est modifié par le poids :

$$D(\beta) = \sum_{i=1}^{n} v_i D_i(\beta)$$

Où $D_i(\beta)$ représente la déviance non mise à l'échelle partielle, pour l'observation numéro i.

Si les poids sont tous égaux à 1, alors l'estimation est inchangée par rapport à un modèle sans poids. Autrement, une observation dont le poids est plus grand que les autres, influencera de manière plus importante l'estimation du vecteur β. On remarquera, que pour une distribution gaussienne, on retombe sur la somme des carrés pondérée.

On peut intégrer ces poids dans le calcul de la méthode des moindres carrés itérativement repondérés (cf. section 1.4.5.8.3), en multipliant les poids calculés pour l'étape de la régression linéaire par le vecteur v :

$$w_i = \frac{v_i}{V(u_i) \times \left(\varphi'(u_i)\right)^2}$$

1.4.5.11 Fonctionnalité additionnelle : loi binomiale de dénominateur variable

Selon notre formulation du modèle linéaire généralisé pour la loi binomiale (cf. section 1.4.5.2), le paramètre de dénominateur (nombre de tentatives de la loi binomiale) m doit être identique pour toutes les observations. La fonction de lien, elle-même, est dépendante de ce m, car elle doit prendre des valeurs comprises entre 0 et m.

On peut relâcher cette hypothèse en définissant une loi binomiale « remise à l'échelle » égale à une loi

binomiale $\mathcal{B}(m;\pi)$ divisée par m, de telle sorte que toutes les valeurs observables sont comprises entre 0 et 1, bornes incluses. On peut ensuite définir une fonction de lien indépendante de m telle que les fonctions $p \to \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ où $p \to \text{probit}(p) = z_p$ où z_p représente le quantile p de la loi normale centrée réduite.

En définissant un vecteur de dénominateurs $k = \begin{pmatrix} k_1 \\ \vdots \\ k_n \end{pmatrix}$ dont toutes les valeurs doivent être explicitées avant estimation de β, et un vecteur de réalisation des réponses $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ où tous les $y_i \in [0;1]$, on peut définir la vraisemblance comme :

$$L_\beta(\beta) = \prod_{i=1}^{n} \left(\mathbb{P}(\mathcal{B}(k_i; k_i\mu_i) = k_i y_i)\right)^{v_i k_i}$$

Où $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \varphi^{-1}(o + x\beta)$, o est le vecteur d'offsets (cf. section 1.4.5.9) et les v_i représentent les poids (cf. section 1.4.5.10).

On peut aussi écrire la déviance non mise à l'échelle (cf. section 1.4.5.7.3) :

$$D(\beta) = 2\sum_{i=1}^{n} v_i k_i \left(\log\left(\left(\frac{y_i}{\mu_i}\right)^{k_i y_i}\right) + \log\left(\left(\frac{1-y_i}{1-\mu_i}\right)^{m-k_i y_i}\right)\right)$$

Où $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \varphi^{-1}(o + x\beta)$ avec $\mu_i \in]0;1[$ pour tout $i \in \{0, \dots, n\}$.

C'est cela qui permet de parler de modèle logit-binomial alors, qu'au sens strict, toute loi binomiale avec dénominateur $m \geq 2$ est susceptible d'avoir une espérance en dehors du domaine de définition de la fonction logit.

2 Contexte

Les estimateurs statistiques sont généralement biaisés car reposant sur des approximations asymptotiques ou des hypothèses erronées. Par exemple, les intervalles de confiance de Student pour l'estimation d'une moyenne reposent sur une hypothèse de normalité [1] qui est fausse pour la plupart des variables cliniques ou biologiques. En effet, toute loi normale peut prendre toute valeur dans l'intervalle $]-\infty, +\infty[$ alors que la plupart des variables cliniques ou biologiques ne peuvent pas prendre de valeur négative. Néanmoins, le théorème central limite assure la validité asymptotique de la méthode de Student ; puisque la normalité passera du statut d'hypothèse à celui d'approximation sur des échantillons finis de grande taille. L'estimateur de Wald des intervalles de confiance des coefficients d'une régression logistique reposera aussi sur une approximation normale asymptotiquement correcte, d'un estimateur discret (cf. section 1.2.11.2).

Les échantillons étant toujours finis, il est alors utile d'évaluer empiriquement, l'ampleur des biais induits par les approximations et de déterminer des conditions dans lesquelles ils sont tolérables. Pour ce faire, des critères d'évaluation de ces biais doivent être déterminés. Pour un estimateur ponctuel $\hat{\theta}$ d'un paramètre unique $\theta \in \mathbb{R}$, on se basera généralement sur le biais $B = \mathbb{E}[\hat{\theta}] - \theta$ correspondant à l'écart entre l'espérance de l'estimateur et le véritable paramètre (cf. section 1.3.2.1). Pour un estimateur $IC_{1-\alpha}$ d'intervalle de confiance de θ, de couverture nominale $1-\alpha$, le biais de couverture, égal à $1-\alpha-\mathbb{P}(\theta \in IC_{1-\alpha})$ sera généralement utilisé. Il s'agit de la différence entre la couverture réelle $\mathbb{P}(\theta \in IC_{1-\alpha})$ et la couverture nominale $1-\alpha$, paramètre de l'algorithme choisi généralement comme égal à 95%.

2.1 Intervalles de confiance équilibrés ou étroits

Un intervalle de confiance bilatéral ayant une borne basse ($L_{1-\alpha}$) et une borne haute ($U_{1-\alpha}$), deux situations où $\theta \notin IC_{1-\alpha}$ se présentent : $L_{1-\alpha} > \theta$ et $U_{1-\alpha} < \theta$. Ces deux situations correspondent, respectivement, à une surestimation et une sous-estimation de θ par l'intervalle de confiance. Si la couverture réelle est $1-\alpha'$ alors la non-couverture α' peut se décomposer en la somme des risques de

surestimation, $\mathbb{P}(L_{1-\alpha} > \theta)$, et de sous-estimation, $\mathbb{P}(U_{1-\alpha} < \theta)$. Alors que la couverture réelle α' peut être égale ou très proche de la couverture nominale α, il peut exister un très grand déséquilibre entre les risques de surestimation et de sous-estimation. Par exemple, un risque de surestimation de 4.9% associé à un risque de sous-estimation de 0,1% conduit à une couverture réelle de 95%.

Un tel déséquilibre peut être volontairement recherché afin de réduire la largeur de l'intervalle de confiance. Par exemple, Zieliński décrit une variante de l'intervalle de confiance d'un pourcentage de Clopper-Pearson [12] ne se basant plus sur les quantiles $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de lois beta mais sur les quantiles γ et $1-\alpha+\gamma$, conduisant virtuellement à des risques nominaux de surestimation et sous-estimation respectivement égaux à γ et $\alpha-\gamma$. Le γ minimisant la largeur $U_{1-\alpha}-L_{1-\alpha}$ de l'intervalle de confiance sera choisi, conduisant à un déséquilibre des risques de surestimation et sous-estimation dès que le γ optimal diffère de $\frac{\alpha}{2}$, mais conservant la somme de ces deux risques $\gamma+\alpha-\gamma=\alpha$. La stratégie de Zieliński se rapproche de celle décrite par Woodward *et al* [13] proposant d'utiliser des quantiles déséquilibrés pour tout intervalle de confiance pouvant s'exprimer à partir d'une fonction de répartition. Zieliński n'est pas le seul à avoir volontairement déséquilibré les estimateurs d'intervalles de confiance de proportions dans l'objectif d'en diminuer la largeur. Des intervalles de confiance déséquilibrés entre leurs risques mais de largeur minimale ont été recherchés pour les proportions par Crow, en 1956, [14] qui formalise bien le fait que les risques de sous-estimation et surestimation peuvent dépasser $\frac{\alpha}{2}$, ainsi que Blyth et Still en 1983 [15] puis Casella en 1986 [16] qui généralisa le concept sous la forme de modification d'un jeu d'intervalles existant, afin d'en minimiser la largeur. Dans un autre contexte, Jackson et Bowden proposent de déséquilibrer les intervalles de confiance de la variance inter-étude τ^2, statistique d'hétérogénéité utilisée dans les méta-analyses à effets aléatoires [17]. Selon Jackson et Bowden, la minimisation de la largeur d'intervalle de confiance est parfois obtenue en visant un risque de surestimation à 0% et un risque de sous-estimation à 5%, mais, sans fournir d'argument précis, ils découragent fortement l'usage d'un tel déséquilibre, suggérant plutôt de viser 1% de surestimation et 4% de sous-estimation. Ils notent néanmoins que le déséquilibre 0% vs 5% correspond à un intervalle de confiance unilatéral.

Cette minimisation de la largeur des intervalles est retrouvée dans le cadre des intervalles de crédibilité bayésiens, qui peuvent s'utiliser comme des intervalles de confiance fréquentistes avec une distribution *a priori* non informative. Cela concerne les intervalles de crédibilité de densité de probabilité *a posteriori* maximale, ou Highest Probability Density (HPD) regions [18]. Dans le cas où les distributions *a posteriori* sont unimodales, l'intervalle HPD est l'intervalle dont les deux extrémités sont de même densité de probabilité et pour lequel l'intégrale de la distribution *a posteriori* entre les bornes est égale à $1-\alpha$. Un tel intervalle aura des queues de probabilité déséquilibrées si la distribution a posteriori n'est pas symétrique ; l'intégrale de la densité de probabilité entre $-\infty$ et la borne basse de l'intervalle de crédibilité pourra différer de $\frac{\alpha}{2}$ ainsi que l'intégrale de densité de probabilité entre la borne haute et $+\infty$ alors que la somme des deux égalera α.

On peut ainsi opposer, la recherche d'un intervalle de confiance le plus étroit (shortest confidence interval ou SCI) et la recherche d'un équilibre entre les risques de surestimation et sous-estimation (equal-tailed confidence intervals ou ETCI). Le SCI et l'ETCI peuvent coïncider si les fluctuations d'échantillonnage de la statistique estimée ou d'une statistique pivot sont symétriques, mais généralement, ils diffèrent [19]. Par ailleurs, le SCI n'existe pas toujours, mais il reste possible de rechercher de s'approcher de cet objectif d'optimalité sans toutefois l'atteindre.

Alors que certains recherchent les intervalles les plus étroits, d'autres proposent des intervalles à queues équilibrées [20] ou des intervalles unilatéraux [21–24] dans des situations de fluctuations d'échantillonnages asymétriques. L'intersection de deux intervalles unilatéraux de confiance permet de construire un intervalle bilatéral à queues équilibrés. Inversement, l'assignation d'une borne d'un intervalle bilatéral équilibré à $+\infty$ ou $-\infty$ engendre un intervalle unilatéral dont le risque nominal est contrôlé. L'évaluation des propriétés statistiques des intervalles unilatéraux ou des intervalles bilatéraux équilibrés est alors un problème très proche, notamment en ce qui concerne l'évaluation des biais de couverture.

Les SCI et ETCI s'opposent donc, aussi bien dans leur construction que dans leur évaluation, puisque l'équilibre des défauts de couverture n'aura aucune importance pour l'évaluation des SCI alors qu'elle sera cruciale pour les ETCI. La largeur de l'intervalle de confiance reste un paramètre à minimiser aussi

bien pour les SCI que les ETCI, mais les SCI ayant une contrainte en moins, ils seront généralement plus étroits.

Se pose alors la question : quand doit-on utiliser les intervalles les SCI, et quand doit-on utiliser les ETCI ? Des éléments théoriques aident à comprendre les enjeux.

Les intervalles SCI, dans le pire des cas, sont des intervalles unilatéraux dont la borne unilatérale n'est pas spécifiée, car dépendante de l'échantillon. Cela a pour conséquence une grande difficulté à interpréter isolément chacune des bornes de l'intervalle de confiance. Par exemple, si on est tenté d'affirmer qu'un paramètre est supérieur à une valeur inférieure ou égale à la borne basse de son intervalle de confiance le risque statistique réellement pris est imprécisément identifié ; c'est-à-dire que l'usage de l'intervalle comme test d'hypothèse va conduire à des tests peu puissants ou ayant une inflation du risque alpha. En effet, selon que le déséquilibre de l'intervalle soit en faveur d'un rétrécissement ou élargissement à gauche, le risque alpha pourra être doublé par rapport à celui que l'on aurait avec un intervalle équilibré (5% plutôt que 2,5%) ou au contraire, fortement diminué mais associé à une perte de puissance statistique. Dans le cas extrême, l'intervalle sera unilatéral à droite avec une inférence extrêmement conservatrice sur la borne inférieure.

Le déséquilibre des intervalles est acceptable à partir du moment ou sous-estimation et surestimation sont considérés comme des risques équivalents, la somme des deux risques devant alors se rapprocher du niveau nominal α, sans que chacun des risques aient à être contrôlés séparément. Cela suppose donc, que l'on tolère une surestimation plus fréquente à condition de sous-estimer plus rarement et vice versa. Plaçons-nous dans le contexte de l'évaluation, par intervalle de confiance, de l'effet d'une intervention dans un essai clinique sur un critère de jugement clinique, un effet positif correspondant à un bénéfice. Peut-on raisonnablement accepter d'augmenter le risque de surestimation de l'effet du traitement, sur l'argument qu'en contrepartie, l'effet du traitement sera rarement sous-estimé. Les deux mouvements vont dans le sens favorable à l'hypothèse du bénéfice du traitement ! Ce mouvement peut d'ailleurs être formalisé par le déplacement des deux bornes d'un intervalle de confiance équilibré vers le haut afin de construire l'intervalle déséquilibré. En dehors du contexte d'un essai clinique, dès lors que le signe et l'ampleur d'un effet n'est pas neutre, ce problème se pose. Que l'on s'intéresse aux effets secondaires

d'une intervention, pour lesquels le risque de sous-estimation est particulièrement préoccupant, ou que l'on s'intéresse à la nocivité d'une exposition environnementale pour laquelle la sous-estimation est à éviter, les risques de sous-estimation et surestimation ne peuvent être échangés librement. Il n'est pas rare que les deux bornes de l'intervalle soient toutes deux d'intérêt. Par exemple, afin de décider de la politique à adopter face à une exposition environnementale nocive, la borne basse permettra de savoir quelle nocivité minimale est « garantie » et induira directement ou indirectement une prise de décision sur les mesures minimales à adopter (poursuite de la recherche scientifique, politique visant à diminuer l'exposition, etc.) alors que la borne haute pourra éventuellement réfréner des politiques trop fortes, si elle s'avère suffisamment basse. Néanmoins, il est important de pouvoir juger de chacune des deux bornes isolément. Se baser sur un intervalle de confiance presque unilatéral mais dont la direction n'est pas connue conduira alors, selon les situations, à une réaction insuffisante ou excessive. Ainsi, il semblerait que la plupart du temps, l'interprétation d'un intervalle bilatéral soit celle de l'intersection de deux intervalles unilatéraux. Un intervalle bilatéral déséquilibré serait tolérable si les conséquences d'une sous-estimation et d'une surestimation étaient les mêmes. Or, dans le cadre d'une exposition potentiellement nocive, une surestimation conduirait à un excès de réaction alors qu'une sous-estimation conduirait à un défaut de réaction. Il est préférable de contrôler chacun des deux risques séparément.

Face à ces éléments, il semble plus pertinent, dans la recherche biomédicale, de rechercher des intervalles de confiance équilibrés, et donc, d'analyser séparément les biais des risques nominaux $\frac{\alpha}{2}$ de surestimation et sous-estimation.

Un autre problématique, peut-être soulevée concernant les tentatives de réduction de largeur d'intervalles de confiance : elles ne sont pas invariantes par transformation monotone. Par exemple, un log-odds ratio dans une régression logistique, sur un échantillon de taille suffisante, suivra des fluctuations d'échantillonnage symétriques, proches d'une loi normale, conduisant à un intervalle de confiance SCI proche d'un intervalle ETCI. Si on transforme la statistique par la fonction exponentielle afin de calculer l'intervalle de confiance d'un odds ratio, alors la dérivée de l'exponentielle étant forcément plus élevée au point de la borne haute qu'au point de la borne basse, un intervalle plus court sera obtenu en décalant les deux bornes de l'intervalle de confiance de l'odds ratio vers le bas. Ce

mouvement permettra de descendre plus fortement la borne haute que la borne basse, conduisant à un intervalle plus étroit. Par exemple, un intervalle de confiance bilatéral à 95% de Wald du log-odds ratio [−log(2) ; +log(2)] deviendra l'intervalle [0,50 ; 2,00] de l'odds ratio avec risques équilibrés. Après raccourcissement de l'intervalle selon la méthode de Woodward *et al* [13], l'intervalle de l'odds ratio deviendra [0,42 ; 1,84] avec des risques nominaux de surestimation et sous-estimation respectivement égaux à 0,75% et 4,25%. Si on s'intéresse au log-odds ratio, la méthode de Woodward *et al* ne modifie pas l'intervalle de Wald original. Cette problématique s'applique aussi au biais de l'estimateur ponctuel. Un estimateur ponctuel non biaisé, en espérance, du log-odds ratio conduira à un estimateur ponctuel biaisé de l'odds ratio. Le bais médian, quant à lui, est invariant aux transformations monotones.

2.2 Tests d'hypothèse bilatéraux et unilatéraux

Le problème qui suit a été introduit et discuté par Kaiser [25], mais il semble encore être peu pris en compte de nos jours. Pour contextualiser le problème, les tests d'hypothèses fréquentistes reposent sur le rejet d'une hypothèse nulle H0, par le fait que l'observation réalisée est peu compatible avec cette hypothèse nulle, du fait d'une faible P-valeur=$\mathbb{P}(observation|H0)$. Le rejet de l'hypothèse nulle permet de conclure à sa négation, qui est appelée hypothèse alternative H1. L'échec à rejeter l'hypothèse nulle, ne permet ni de rejeter, ni d'accepter H0, ni H1. L'inférence n'est valide que dans le cas où H1 est la négation de H0. Les formulations telles que H0 : $\mu_1=\mu_2$ et H1 : $\mu_1>\mu_2$, où μ_1 et μ_2 représentent des moyennes dans deux populations, sont donc incorrectes dans le cas général [26], sauf si $\mu_1<\mu_2$ a été préalablement prouvée fausse. Afin de formuler correctement H0 et H1, il est préférable de commencer par la formulation de H1, hypothèse de recherche dont on souhaite prouver la véracité [26], pour ensuite formuler H0 comme sa négation. Si, par exemple, on souhaite montrer qu'une mesure biologique est reproductible, ne présentant que rarement une erreur de mesure majeure, susceptible d'entraîner une modification de la prise en charge du patient, alors on voudra prouver que ces écarts majeurs (deux mesures successives écartées de plus de 40%, par exemple) surviennent rarement (*p.ex.* dans moins de 3% des cas). L'hypothèse alternative sera alors H1 : $\pi < 0,03$ où π représente la proportion de paires de mesures avec discordance majeure. L'hypothèse nulle, par négation, sera définie comme H0 : $\pi \geq 0,03$.

Avec cette formulation de l'hypothèse nulle, on constate que les fluctuations d'échantillonnage et la P-valeur ne sont pas aisément calculables car elles diffèrent selon la valeur de π, qui n'est pas parfaitement définie. Dans ce cas, on va définir la P-valeur dans le pire des cas, c'est-à-dire, la valeur maximale des P-valeurs conditionnelles à π pour tout $\pi \geq 0{,}03$. Généralement, le test d'hypothèse sera construit d'une telle manière que le pire des cas sera le scenario où $\pi = 0{,}03$. C'est pourquoi, on peut être tenté de définir H0 : $\pi = 0{,}03$, mais cela obscurcirait le fait que la statistique du test doit être construite de telle manière à ce que ce scenario soit le pire des cas. Par ailleurs, la construction de tests d'hypothèses dans lesquels H0 et H1 ne sont pas la négation l'une de l'autre peut conduire à des conclusions fallacieuses telles que l'on peut en observer avec le test de Gehan-Breslow-Wilcoxon dont l'hypothèse nulle est la superposition parfaite des courbes de survie et dont l'hypothèse alternative a pu être interprétée comme la supériorité de la médiane de survie du groupe dont la médiane observée est la plus grande [27].Cette erreur d'interprétation peut même toucher les créateurs d'une procédure, se dupant eux-mêmes dans l'article original, comme dans l'exemple N°2 de Fleming *et al* [28] dans lequel les auteurs concluent à une meilleure survie avec traitement par radiothérapie + 5-fluorouracile qu'avec le traitement contrôle alors que les courbes se croisent et qu'il y a une tendance à une survie moins bonne à long terme dans le groupe radiothérapie + 5-fluorouracile.

Kaiser [25] précise qu'un test bilatéral formulé avec des hypothèses H0 : $\mu_1=\mu_2$ et H1 : $\mu_1\neq\mu_2$, permet, si son résultat est significatif, de conclure à l'existence d'une différence entre μ_1 et μ_2 mais ne permet pas de conclure que $\mu_1>\mu_2$ ou $\mu_1<\mu_2$. Pour faire une telle inférence, un test à trois hypothèses pourrait être réalisé (H1 : $\mu_1<\mu_2$, H2 : $\mu_1=\mu_2$ et H3 : $\mu_1>\mu_2$). Un tel test est dit *bilatéral directionnel*. En plus des erreurs α et β, une erreur γ est introduite, correspondant à une conclusion à l'existence d'une différence dans le sens opposé de la différence réelle ($\mu_1<\mu_2$ alors que $\mu_1>\mu_2$ ou vice versa). Pour l'inférence bilatérale non directionnelle, cette erreur γ n'existe pas car on ne doit pas conclure au sens de la différence lorsque l'on rejette H0. Dans la situation où $\mu_1>\mu_2$, mais où on observe $m_1<m_2$ et que l'on rejette $\mu_1=\mu_2$ dans un test bilatéral non directionnel, alors la conclusion $\mu_1\neq\mu_2$ est correcte. Dans la pratique, en recherche biomédicale, les tests bilatéraux sont presque toujours interprétés directionnellement, selon la direction observée : $\mu_1 < \mu_2$ si $m_1 < m_2$ et $\mu_1 > \mu_2$ si $m_1 > m_2$. Il apparaît

alors que la théorie doit s'adapter à cette pratique, afin de développer des outils statistiques fiables pour ces usages. Les conclusions $\mu_1>\mu_2$ et $\mu_1<\mu_2$ n'étant pas équivalentes, les risques doivent être contrôlés séparément. Imaginons un essai clinique randomisé dans lequel on compare le traitement chirurgical au traitement médical de la hernie discale sur la douleur à 1 an. En l'absence de préjugé sur la supériorité d'un traitement par rapport à l'autre, une test bilatéral directionnel paraît raisonnable. Supposons que les deux traitements soient équivalents. Un test statistique bilatéral non directionnel qui contrôle le risque alpha de l'hypothèse H0 : $\mu_1=\mu_2$ à 5% pourra avoir un risque de 4,9% de conclure en faveur du traitement chirurgical et 0,1% en faveur du traitement médical. On peut alors considérer que par rapport à un test bilatéral directionnel équilibré qui aurait des risques respectifs de 2,5% et 2,5%, il existe à la fois une inflation, par un facteur deux, du risque de conclure en faveur du traitement chirurgical et une perte de puissance pour conclure en faveur du traitement médical si celui-ci est meilleur. Dans le pire des cas, le test bilatéral s'approcherait d'un test unilatéral dans une direction non contrôlée. Un cadre théorique bilatéral directionnel, distinguant les risques allant dans chacune des deux directions, permettra de s'assurer de leur contrôle distinct. Dans un cadre théorique plus classique, il est possible de réaliser deux tests unilatéraux allant dans des sens opposés ($\mu_1>\mu_2$ puis $\mu_1<\mu_2$), chacun des deux ayant un risque contrôlé [29], l'inférence étant alors équivalente à celle de tests bilatéraux directionnels à risques équilibrés.

Les tests unilatéraux ont été critiqués pour leur incapacité à identifier une différence allant dans le sens opposé à celle qui était attendue [30] ; critique non applicable aux tests bilatéraux directionnels. Le point de vue de Kaiser a été aussi critiqué comme futile. Le risque unilatéral est, au pire, doublé dans une approche non directionnelle et le risque de troisième espère serait très faible, anecdotique [30]. Cependant, ce doublement du risque unilatéral est associé à une perte de puissance dans la direction opposée, le seuil de significativité unilatéral se rapprochant fortement de zéro. Le risque de troisième espèce, quant à lui, peut frôler les 5% lorsque la différence réelle est très faible. Par ailleurs, une théorie en inadéquation avec l'usage pratique, peut conduire à une perte de ressources humaines sur des problèmes de faible pertinence, tels que les nombreux développements statistiques ayant attrait au rejet de l'hypothèse de superposition parfaite de deux courbes de survie mais n'aidant pas à identifier laquelle

est « supérieure » à l'autre [31–33].

Il a aussi été remarqué que l'hypothèse nulle d'égalité stricte ($\mu_1=\mu_2$) est rarement plausible [34] ; rejeter cette hypothèse n'est pas informatif car la probabilité *a priori* qu'elle soit fausse est souvent de 100%. L'important est de pouvoir identifier la direction et l'amplitude de la différence, ce qui est plus aisément réalisé en s'aidant d'intervalles de confiance. Les tests d'hypothèses par P-valeurs ont été critiqués, notamment dans un communiqué de l'American Statistical Association (The ASA Statement on *p*-values) [35] comme n'étant pas une mesure de taille d'effet et donc pouvant induire en erreur par la mise en évidence d'un effet futile sur un grand échantillon. Cela est un artefact de la tendance à poser une hypothèse nulle d'égalité stricte. L'existence d'un effet cliniquement important [36] peut être prouvé par un test d'hypothèse unilatéral avec une hypothèse alternative H1 : effet > $\Delta_{important}$ et une hypothèse nulle H0 : effet ≤ $\Delta_{important}$. Ce seuil peut aussi être combiné à un cadre bilatéral directionnel (aussi appelé trilatéral) [37]. Cette approche souffre malheureusement d'une limite majeure freinant son adoption : la différence minimale cliniquement importante ($\Delta_{important}$) est subjective, ou n'est pas reproductible par les méthodes objectives qui la définissent [38]. La présentation d'un intervalle de confiance bilatéral permet au lecteur de l'article ou du rapport d'analyse de comparer la borne basse (ou haute) de l'intervalle de confiance à la valeur seuil $\Delta_{important}$ qu'il aura lui-même défini ; il n'est plus tributaire du choix des auteurs. Cela suppose néanmoins que cette inférence unilatérale soit valide ; reposant sur l'équilibre des risques des intervalles bilatéraux (cf. section 2.1).

2.3 Couverture moyenne et minimale

Certains estimateurs d'intervalles, se vantant d'être exacts, tels que l'estimateur de Clopper-Pearson [39] d'une distribution binomiale ont été critiqués pour leur strict conservatisme [40]. Cela est explicable par le fait qu'ils garantissent une couverture supérieure ou égale à la couverture nominale pour toute valeur théorique de la proportion π à estimer. Pour une couverture nominale donnée (*p.ex.* 95%), il existe des discontinuités de la fonction associant la couverture réelle à la proportion théorique π, comme décrit par Brown *et al* [11]. Ces discontinuités surviennent à chacune des bornes des intervalles de confiance déterminés pour les numérateurs, discrets, de 0 à N où N est la taille d'échantillon. Ce problème existe

pour tout estimateur reposant sur des événements discrets, comme les estimateurs d'intervalles de confiance de proportions binomiales, mais aussi, les régressions logistiques, régressions de Poisson et modèles de Cox. Cela est dû à la nature discrète des fluctuations d'échantillonnage sur de petits échantillons. La seule solution qui ait été trouvée à ce problème qui puisse faire disparaître les discontinuités de la couverture est l'ajout d'aléatoire (typiquement issue d'un ordinateur) dans la procédure de calcul de l'intervalle de confiance [41] ; cela détruit la reproductibilité des analyses. Si on souhaite construire des intervalles de confiance déterministes, un choix doit être fait quant à la couverture : doit-on garantir que la couverture réelle soit supérieure ou égale à la couverture nominale pour toute valeur théorique des paramètres d'intérêt (π pour les proportions binomiales), ou doit-on rapprocher autant que possible la couverture réelle moyenne de la couverture nominale ?

Que l'on s'intéresse à une proportion binomiale [11] une régression logistique [42] ou tout autre modèle linéaire généralisé ou extension, les estimateurs sont généralement évalués conditionnellement à la taille d'échantillon et à la matrice d'exposition. Pourtant, ces paramètres sont rarement contrôlés en recherche biomédicale. Quand bien même des calculs de nombre de sujets nécessaires sont réalisés dans beaucoup d'études, la taille exacte d'échantillon diffèrera notablement d'une expérience à l'autre portant sur le même sujet. C'est pourquoi il est très rare que deux études incluses dans une même méta-analyse aient exactement la même taille d'échantillon. En supposant que la couverture d'un intervalle de confiance suive une fonction en dents de scie selon la taille d'échantillon, oscillant autour de la couverture nominale, alors les excès de couverture pourraient compenser les défauts de couverture pour tendre vers une proportion des intervalles de confiance contenant la valeur réelle très proche du niveau de confiance nominal. Même lorsqu'une seule étude existe sur un sujet, à partir du moment où il existe une incertitude sur la valeur réelle du paramètre que l'on cherche à estimer, même pour une taille d'échantillon fixée, la couverture réelle est imprévisible en raison des oscillations de couverture dépendantes de la valeur de ce paramètre. On peut alors faire une interprétation mixte bayésienne et fréquentiste de la couverture : une probabilité de 50% d'avoir une couverture à 97% associée à une probabilité de 50% d'avoir une couverture à 93% est équivalent à une couverture à 95%. Par analogie, une expérience constituée d'un jet d'une pièce sur pile (50%) ou face (50%), suivie d'un tirage, selon le résultat à pile ou face, d'une

valeur y suivant une loi de Bernoulli de paramètre π=0,97 ou π=0,93 conduit à une distribution de Y parfaitement égale à une distribution de Bernoulli de paramètre π=0,95. Enfin, les variables d'exposition des modèles multivariés (modèles linéaires généralisés à effets fixes ou mixtes) sont souvent elles-mêmes aléatoires sauf dans des situations très expérimentales, rarement rencontrées en recherche clinique ou en épidémiologie. Tout ceci suggère que le conservatisme conditionnel strict, c'est-à-dire, une couverture réelle garantie comme supérieure ou égale à la couverture nominale quels que soient la taille d'échantillon, les paramètres théoriques et les covariables, n'est pas souhaitable.

La couverture moyenne d'un estimateur d'intervalle de confiance d'un paramètre θ sera calculée à partir d'une distribution *a priori* η de θ, comme l'intégrale du produit de la densité de η par la couverture conditionnelle à θ sur l'espace Ω des valeurs de θ. Cette couverture moyenne reflète la couverture réelle dans la théorie mixte bayésienne-fréquentiste décrite ci-dessus. Cette procédure est décrite par Wang [43]. Malheureusement, les couvertures moyennes de proportions binomiales données par Wang ainsi qu'Agresti et Coull [40] sont pour des lois *a priori* uniformes sur l'intervalle $]0,1[$ ou des lois beta de forte variance. Dans ces couvertures moyennes, un défaut de couverture d'un intervalle de confiance binomial pour une proportion théorique π=0,20 pourra être compensée par un excès de couverture pour la proportion théorique π=0,80. Dans une situation où l'on a un connaissance assez précise du paramètre à estimer, telle que le taux de décès des patients atteints de cancer non à petites cellules non métastatiques à 30 jours d'une lobectomie curative, qui est entre 1% et 3% en France [44], alors un défaut de couverture systématique qui porterait sur cet intervalle engendrerait un biais dans l'estimation. Afin d'évaluer les propriétés statistiques des estimateurs il est donc nécessaire de fournir des distributions η *a priori* de bien moindre variance. Nous parlerons de *couvertures moyennes locales*, pour décrire les couvertures moyennes sur des distributions *a priori* de faible variance, par opposition à des *couvertures moyennes globales* qui seraient obtenues sur des distributions *a priori* non informatives ou faiblement informatives.

Une dernière considération permet de réaliser l'intérêt des *couvertures moyennes locales*. Les paramètres théoriques estimés dans les différentes études portant sur une même question de recherche diffèrent en raison de différences dans les populations, les mesures ainsi que d'autres différences

méthodologiques. C'est pourquoi, des modèles à effets aléatoires sont utilisés dans les méta-analyses, reposant, dans la théorie fréquentiste, sur l'existence d'une variance inter-étude du paramètre théorique estimé. Si cette variance est prise en compte, alors la réalisation d'une étude peut se décrire comme un processus stochastique en deux étapes : à la première étape, le paramètre théorique θ est réalisé à partir d'une variable aléatoire Θ. À la seconde étape, une estimation $\hat{\theta}$ de θ est réalisée ainsi qu'un intervalle de confiance $IC_{1-\alpha}$ de θ. Cette procédure a pour conséquence que la couverture réelle de $IC_{1-\alpha}$, c'est-à-dire, la probabilité qu'il contienne le θ, ne peut plus être calculé conditionnellement à une valeur fixée de θ mais doit être moyennée sur l'ensemble de la distribution de Θ. D'un point de vue calculatoire, cette couverture peut se calculer presque de la même manière que la *couverture moyenne locale* avec distribution *a priori* bayésienne, mais la théorie sous-jacente est purement fréquentiste.

2.4 Synthèse des considérations théoriques et définition d'un objectif

Selon les considérations de la section 2.1, les intervalles de confiance devraient être généralement équilibrés plutôt que de largeur minimale. Selon les considérations de la section 2.2, les tests d'hypothèses devraient être conçus pour avoir une inférence unilatérale ou bilatérale directionnelle valide, c'est-à-dire, avec une maîtrise séparée des risques de conclure à la supériorité et à l'infériorité. Selon les considérations de la section 2.3, le contrôle des *couvertures moyennes locales* des intervalles de confiance est préférable au contrôle strict des couvertures conditionnelles.

Ces considérations, guidées par l'usage pratique qui est fait des outils statistiques, devraient être prises en compte dans l'évaluation des propriétés des estimateurs statistiques. Mêmes si elles ne sont pas nouvelles, ces considérations ne semblent pas avoir été combinées pour analyser les estimateurs les plus communs, tels que les trois estimateurs classiques des modèles linéaires généralisés : Wald, le score de Rao et le rapport de vraisemblance.

C'est pourquoi, l'objectif de cette thèse était d'établir des critères d'évaluation des estimateurs prenant en compte l'ensemble de ces considération et de réévaluer les estimateurs usuels des modèles les usuels sur ces critères.

3 Méthodes

Dans cette section, nous définirons les critères d'évaluation des estimateurs d'intervalles de confiance qui nous aideront à réévaluer la performance des estimateurs usuels des modèles usuels.

3.1 Risques unilatéraux conditionnels et moyens locaux

Étant donné un paramètre θ de la population et un estimateur d'intervalle de confiance bilatéral $IC_{1-\alpha}$ dont les valeurs sont des intervalles fermés bornés de $\mathbb{R}$, définissons les bornes basse et haute de l'intervalle de confiance $L_{1-\alpha}$ et $U_{1-\alpha}$, tel que $IC_{1-\alpha} = [L_{1-\alpha}; U_{1-\alpha}]$. La *couverture nominale* ou *niveau de confiance nominal* sont définis comme égaux à $1-\alpha$. Les risques unilatéraux nominaux α_U et α_L sont égaux tous deux à $\frac{\alpha}{2}$ pour un intervalle de confiance à risques équilibrés et sont interprétables comme les risques désirés de surestimation (α_L) et de sous-estimation (α_U) de θ par l'intervalle de confiance $IC_{1-\alpha}$. Les *risques unilatéraux réels* $\alpha'_L = \mathbb{P}(L_{1-\alpha} > \theta)$ et $\alpha'_U = \mathbb{P}(U_{1-\alpha} < \theta)$ correspondent aux risques de surestimation et sous-estimation et peuvent différer des risques nominaux si les intervalles de confiance sont biaisés. Lorsqu'on parlera de *risques unilatéraux*, sans préciser s'il s'agit des risques réels ou nominaux, il sera fait référence aux risques réels.

Les rapports entre risques unilatéraux réels et nominaux α_L/α'_L et α_U/α'_U décrivent les biais unilatéraux de couverture. Le seuil distinguant un biais acceptable d'un biais inacceptable est subjectif. Pourtant, c'est ces seuil qui permettraient de définir, de manière pratique, les conditions de validité. Du point de vue de l'auteur de ce travail, des rapports α_L/α'_L et α_U/α'_U dépassant 2 ou inférieurs à 1/2 peuvent être considérés comme nettement biaisés. En dessous de 1,50 ou 1/1,50=0,667, le biais est tolérable. Ces seuils sont arbitraires, de la même manière que le choix usuel du niveau de confiance 95% est arbitraire.

La notion de *risques unilatéraux moyens locaux* α''_U et α''_L est plus difficile à définir dans un cadre complètement général. Le concept est de supprimer le conditionnement des analyses à la taille d'échantillon, aux covariables voire aux paramètres théoriques d'un modèle statistique, qui seront alors supposés aléatoires. De nouvelles variables aléatoires indépendantes les unes des autres seront définies, telles que la taille d'échantillon N dont la réalisation sera n, où le paramètre à estimer Θ dont la

réalisation sera θ. Ces variables aléatoires suivront des lois de distribution de faible variance, où la notion de *faible variance* n'est pas mathématique mais est au sens commun du terme ; donc hautement subjective. Par exemple, on pourra considérer que la taille d'échantillon varie ± 10% selon une loi *ad hoc*, telle que la partie entière d'une loi log-normale. Dans certaines situations, telles que le comptage d'événements incidents, on pourra s'aider de la loi de Poisson, réduisant alors l'arbitraire du choix. Cette faible variance est largement inférieure à la variance que l'on observe dans les différentes expériences conduisant à une méta-analyse sur un sujet précis. C'est l'usage d'une faible variance qui permettra de parler de *moyenne locale*. L'expérience aléatoire sera ensuite définie comme un processus à deux étapes. La première étape sera la réalisation des variables aléatoires sur des paramètres que l'on considère comme habituellement comme fixes ou sur lesquels on conditionne : taille d'échantillon N, paramètre Θ, matrice d'exposition X. La seconde étape sera la génération aléatoire de l'expérience conditionnelle aux réalisations de la première étape. Enfin, on pourra définir les *risques unilatéraux moyens locaux* comme $\alpha_L'' = \mathbb{P}(L_{1-\alpha} > \Theta)$ et $\alpha_U'' = \mathbb{P}(U_{1-\alpha} < \Theta)$ où les probabilités $\mathbb{P}$ se définissent dans le contexte de l'expérience à deux étapes sus-décrite. Dans ce même cadre théorique, il est possible de redéfinir toutes les statistiques habituelles : largeur moyenne d'intervalle de confiance, biais moyen ou médian d'estimateur ponctuel, *etc*.

Par opposition, aux *risques unilatéraux moyens locaux* α_L'' et α_U'', on parlera de *risques unilatéraux conditionnels* α_L' et α_U' lorsque l'expérience aléatoire est en une seule étape, avec les paramètres habituels supposés fixes. Les *risques unilatéraux conditionnels* peuvent être aisément analysés comme des fonctions des principaux paramètres de l'expérience, telle que la taille d'échantillon où les valeurs du paramètre θ. Les risques *unilatéraux moyens locaux* seront analysés comme des fonctions des espérances des paramètres $\mathbb{E}[N]$, $\mathbb{E}[\Theta]$ après avoir défini les distributions des variables aléatoires N et Θ dans des familles mono-paramétriques, directement paramétrées par leur espérance ou indirectement paramétrables par leur espérance.

D'un point de vue pratique, on peut calculer les *risques unilatéraux moyens locaux* comme des moyennes des *risques unilatéraux conditionnels* pondérée par la distribution multivariée définie à la première étape de l'expérience aléatoire, telle que la distribution de la variable aléatoire (N, Θ).

3.2 Demi-largeur des estimateurs d'intervalles de confiance

À couverture égale, on privilégie généralement les estimateurs d'intervalles de confiance les plus étroits, car fournissant une meilleure précision statistique avec la même quantité de données.

En sections 2.1 et 2.2 nous avons montré l'intérêt de considérer les risques unilatéraux que nous avons définis en section 3.1 (Risques unilatéraux conditionnels et moyens locaux). En appliquant les mêmes considérations aux largeurs des intervalles de confiance, il paraît peu pertinent de considérer la largeur totale de l'intervalle de confiance. Lorsqu'on s'intéresse à la borne inférieure de l'intervalle de confiance, comme dans un essai clinique de supériorité, l'usage d'un estimateur d'intervalle de confiance ayant en moyenne les deux bornes d'intervalle de confiance plus bas qu'un autre, bien que pouvant éventuellement rétrécir l'intervalle global, handicaperait toute tentative de s'assurer d'un effet minimal du traitement. Si on interprète l'intervalle de confiance comme un outil permettant de faire des tests d'hypothèse pour des seuils d'effets désirés, par comparaison directe de la borne de l'intervalle au seuil, alors une baisse des deux bornes conduirait à une perte de puissance. C'est pourquoi nous définirons séparément les demi-largeurs d'intervalle de confiance à gauche et à droite, comme $w_L = \mathbb{E}\left[\hat{\theta} - L_{1-\alpha}\right]$ et $w_U = \mathbb{E}\left[U_{1-\alpha} - \hat{\theta}\right]$ où $\hat{\theta}$ est un estimateur ponctuel de la statistique d'intérêt et $L_{1-\alpha}$ et $U_{1-\alpha}$ représentent respectivement les borne basse et haute de l'intervalle de confiance.

Par la même approche qu'en section 3.1, il est possible de définir les *demi-largeurs moyennes locales* w_L'' et w_U'' dans une expérience à deux étapes, en levant le conditionnement à la taille d'échantillon, à la matrice des covariables et en supposant éventuellement que le paramètre θ est aléatoire plutôt que fixe. Cette notion de demi-largeur est dépendante du choix d'un estimateur ponctuel $\hat{\theta}$. Dans les analyses que nous ferons, l'estimateur fréquentiste le plus usuel sera employé, c'est-à-dire, l'estimateur du maximum de vraisemblance.

3.3 Performances statistiques des tests d'hypothèses

Dans ce travail les performances statistiques (risque alpha, puissance) de certains tests d'hypothèses ne seront qu'indirectement analysées. Les trois tests d'hypothèses les plus fréquemment usités sont : Wald, le score de Rao et le test du rapport de vraisemblance généralisé aussi connu sous le nom de test du

rapport de vraisemblance [45]. Par inversion de test, des intervalles de confiance peuvent être construits à partir des tests d'hypothèses réalisant des inférences sur un paramètre $\theta \in \mathbb{R}$. Cette approche permet la construction des intervalles de confiance de Wald, Rao et du rapport de vraisemblance généralisé, mais s'applique à d'autres tests [46]. Il est aussi possible de construire un test d'hypothèse unilatéral ou bilatéral directionnel à partir d'un estimateur d'intervalle de confiance d'une statistique $\theta \in \mathbb{R}$, par comparaison directe de θ aux bornes de l'intervalle de confiance. Ainsi, la description des performances statistiques (risques moyens locaux unilatéraux et demi-largeurs) des estimateurs d'intervalle informera indirectement sur les propriétés des tests d'hypothèses associés, dans leur interprétation unilatérale ou bilatérale directionnelle au sens défini par Kaiser [25] et discuté en section 2.2.

Certains test d'hypothèses, tels que le test du χ^2 de Pearson d'indépendance entre deux variables qualitatives à trois modalités ne sont pas directement inversibles en intervalles de confiance d'une statistique d'interprétation intuitive ; ils n'offrent pas non plus d'interprétation directionnelle au sens défini par Kaiser [25] et discuté en section 2.2. Les propriétés statistiques de ces tests ne seront, ni directement, ni indirectement analysées dans notre travail.

3.4 Calculs numériques

Il est habituel de faire appel à des simulations de Monte Carlo pour évaluer les propriétés statistiques des estimateurs. Une approche différente a été appliquée dans ce travail, en raison de la simplicité des modèles analysés. Cette méthode est équivalente aux résultats que l'on obtiendrait asymptotiquement avec une infinité de simulations. Elle repose sur l'usage des distributions binomiales et de Poisson exactes, avec un algorithme parcourant la totalité ou la presque totalité de l'espace discret des fluctuations d'échantillonnage et cumulant les statistiques conditionnelles aux réalisations des lois de Poisson et binomiale, avec une pondération par la probabilité de la réalisation. Le support de la loi de Poisson étant infini, il n'est pas algorithmiquement possible de parcourir toutes les réalisations imaginables, mais en négligeant les queues de distribution représentant une probabilité cumulée de moins d'un pour un milliard, on obtient des résultats presque exacts.

4 Publications

4.1 The case for balanced hypothesis tests and equal-tailed confidence intervals

4.1.1 Présentation de l'article

Même si, dans l'ordre chronologique, l'article intitulé « *The case for balanced hypothesis tests and equal-tailed confidence intervals* » [47] (Annexe 1) n'est pas le premier de la thèse, il introduit le rationnel justifiant le choix des critères d'évaluation présentés en section 3. Il expose une partie des considérations présentées en section 2, appliquées au domaine de la survie, dans lequel l'écart existant entre la théorie non directionnelle et l'interprétation directionnelle qui en est généralement faite a des conséquences particulièrement graves. En résumé, alors que l'usage d'un test bilatéral non directionnel sur des comparaisons de moyennes conduira au pire à un doublement du risque alpha ou une perte de puissance tel que décrit en section 2.2, de nombreux tests développés pour comparer des courbes de survie pourront conduire à un risque de troisième espèce approchant les 50% lorsque leur interprétation est directionnelle. Pour rappel, le risque de troisième espèce γ, est la probabilité de conclure à l'existence d'un effet significatif dans le sens opposé à l'effet réel. Certes, ce risque ne surviendra que dans un contexte d'erreur humaine d'interprétation du résultat du test, puisque le test permet juste de rejeter l'hypothèse de superposition parfaite des courbes de survie, n'aidant pas à savoir laquelle aurait la meilleure médiane ou le meilleur taux de survie à 5 ans, ou la meilleure moyenne. Mais une revue épidémiologique de la littérature montre que les erreurs d'interprétation sont fréquentes avec ces tests, y compris par les auteurs des tests eux-mêmes dans les exemples fournis avec l'article original. Par ailleurs, la même erreur d'interprétation théorique, aux conséquences bien moins graves est faite à chaque fois qu'une conclusion directionnelle est faite à partir d'un test de Student ou d'un test du χ^2 décrits comme bilatéraux, comme retrouvé dans une grande partie des articles du domaine biomédical en 2021.

4.1.2 Regard historique

Un éclairage peut être apporté à l'article présenté en Annexe 1 par une narration des développements successifs des tests présentés dans l'article. Il s'agit de tests apparentés au test du log-rank et à l'estimateur de Kaplan-Meier, ayant été analysés ou développés dans la situation de non-proportionnalité des risques.

4.1.2.1 Prérequis : notion de proportionnalité des risques

L'hypothèse des risques proportionnels, décrite dans le modèle de Cox, consiste en une constance du Hazard Ratio à tous les temps t. Plus précisément, il est fait l'hypothèse que le rapport entre les fonctions de Hazard correspondant aux deux courbes est constant. Cette fonction de Hazard, appelée aussi fonction de risques instantanés, est interprétable comme la limite, quand delta tend vers zéro, de la probabilité de survenue de l'événement entre le temps t et t+delta, conditionnelle à une absence d'événement jusqu'au temps t, c'est-à-dire, $h(t) = \lim_{\delta \to 0} \mathbb{P}(T \in [t; t+\delta] | T \geq t)$ où T est la variable aléatoire représentant le délai avant événement. Les écarts à l'hypothèse des risques proportionnels sont particulièrement problématiques lorsqu'il existe une inversion des risques entre deux périodes, puisque cela signifie qu'un des deux groupes, a un risque d'événement (*p.ex.* décès) moindre, dans une période, mais un risque plus élevé, dans une autre période. Il n'est alors pas évident d'identifier le groupe dont la survie sans événement est la meilleure. Notamment, les conclusions peuvent différer selon que l'on s'intéresse à la moyenne de survie, la médiane de survie, ou le taux de survie à un temps prédéfini. Il existe néanmoins une situation qui devrait, à mon sens, faire consensus : lorsque la fonction de survie S(t) est constamment supérieure, pour tout temps t, dans un groupe que dans l'autre, alors il paraît raisonnable de considérer que la survie est meilleure dans ce groupe ; quand bien même la fonction de Hazard pourrait lui être défavorable sur certaines zones temporelles, elles ne suffiraient pas à rattraper complètement l'avantage pris sur les zones favorables. Mais dans le pire des cas, les courbes de survie se croisent complètement, de telle sorte, par exemple, que la survie à court terme est meilleure dans un groupe alors que la survie à long terme est meilleure dans l'autre groupe.

4.1.2.2 Tests développés pour des risques proportionnels (1965-1972)

Dans la famille des tests de comparaison de courbes de survie basés sur les rangs, le test de Gehan [48]

et du log-rank [49] ont d'abord été proposés en 1965 par Gehan et en 1966 par Mantel. Le test du log-rank, bien connu, est équivalent au test du score d'un modèle de Cox [50], et ne pose pas de problème lorsque l'hypothèse des risques proportionnels respectée ; on notera néanmoins que la question du poids donnée aux événements précoces et tardifs est discuté par Mantel [49], qui précise que le test du log-rank donne un fort poids aux événements précoces. Le test de Gehan [48] est interprétable comme un test du log-rank dans lequel les observations seraient pondérées par le nombre de sujets à risque au moment de chacun des événements ; ainsi, il donne un poids encore plus important aux événements précoces que tardifs que le test du log-rank. En situation de risques proportionnels, les inférences sont équivalentes, même si Peto en 1972 montre que le test du log-rank est plus puissant lorsqu'il n'y a pas d'ex-aequo sur les censures et que le rythme de censure est homogène dans les deux groupes [51].

4.1.2.3 Développements en cas de non-proportionnalité des risques (1975-1991)

Les tests de Gehan et du log-rank tentent de rejeter l'hypothèse nulle de superposition parfaite des courbes, sous laquelle l'hypothèse des risques proportionnels est forcément respectée. En cas de non-respect de l'hypothèse des risques proportionnels, le rejet de cette hypothèse nulle conduit à une conclusion correcte, puisque cette hypothèse nulle est forcément fausse dans cette situation, quand bien même aucune interprétation directionnelle ne peut être directement faite. C'est pourquoi les développements théoriques semblent s'être concentrés sur la puissance statistique, sans se préoccuper de la direction de la conclusion statistique. Pour commencer, Lee *et al*, en 1975, analysèrent la puissance statistique de divers tests, notamment le log-rank et le Gehan, en cas de non proportionnalité des risques [52] ; ils conclurent que le test de Gehan est le plus puissant dans cette situation, mais en réalité, cette conclusion repose sur les spécificités de leur scenario de référence. Tarone et Ware, en 1977, a contrario, notèrent qu'aucun des deux tests n'est uniformément plus puissant que l'autre, et que la puissance de l'alternative ; ils proposèrent un nouveau test, intermédiaire entre le log-rank et le Gehan, avec un poids égal à la racine carrée du nombre de sujets à risque. Tarone et Ware en discutant extensivement du fait que la fonction de poids optimale pour rejeter l'hypothèse nulle de superposition parfaite des courbes, est fortement dépendante de l'alternative, révèlent un premier problème. En effet, si l'hypothèse alternative n'est pas la négation de l'hypothèse nulle, comment le rejet de l'hypothèse nulle permet-elle

l'acceptation de l'hypothèse alternative ? Nous discuterons plus en détail de ce problème en section 4.1.3.2 en page 87.

En 1991, Fleming & Harrington généralisèrent les divers tests des rangs grâce à une famille de fonctions de poids $G^{\rho,\lambda}(t) = \left(\hat{S}(t)\right)^{\rho}\left(1-\hat{S}(t)\right)^{\lambda}$ paramétrable par deux coefficients $\rho \in [0;+\infty[$ et $\lambda \in [0;+\infty[$ [53]. Cette famille de poids, permet de fournir un poids fort aux événements précoces, tardifs ou tout intermédiaire entre les deux extrêmes, et permet même de focaliser les poids sur les événements intermédiaires. Les deux paramètres ρ et λ peuvent alors être choisis afin de maximiser la puissance selon la forme des courbes anticipées. L'inférence étant conditionnelle aux paramètres ρ et λ, il n'est pas statistiquement valide, de choisir ρ et λ *a posteriori* ; malheureusement, il n'est pas forcément facile d'anticiper les écarts à l'hypothèse des risques proportionnels.

4.1.2.4 Kaplan-Meier pondéré (1989)

En parallèle du développement des statistiques dérivées du log-rank, interprétables comme des moyennes pondérées des hazard ratio aux différents temps, des comparaisons basées sur les différences des fonctions de survie S(t) furent développés, en s'aidant de l'estimateur de Kaplan-Meier. En 1989, Pepe et Fleming décrivirent un test de comparaison de courbes de survie de Kaplan-Meier, pondérée par une fonction de poids W, c'est-à-dire $WKM = \int_0^T W(t)\left(\widehat{S_1}(t) - \widehat{S_2}(t)\right)\mathrm{d}t$. Pour la fonction de poids $W(t) = 1$, il s'agit de la différence d'espérances de vie restreintes au temps T [54]. Même si les auteurs se concentrent sur les calculs de puissance de rejet de l'hypothèse de superposition parfaite des courbes, on peut remarquer que cette famille de test n'est pas susceptible de conduire à des erreurs de troisième espèce lorsqu'une courbe domine complètement l'autre avec une inversion des risques proportionnels insuffisante pour conduire à un croisement des courbes de Kaplan-Meier. En effet, la supériorité d'une courbe de survie persiste pour toute valeur de t, peu importe la fonction de poids W, la statistique WKM garde toujours le même signe. Pepe et Fleming précisent que la fonction de poids doit être choisie *a priori*, et donnent un exemple de fonction de poids assez générale, basée sur le taux de censure, donnant un poids particulièrement faible aux observations tardives, ce qui se justifie par une volonté de puissance statistique, car ces observations tardives étant peu nombreuses, leurs fluctuations d'échantillonnage sont

fortes.

4.1.2.5 Maximum du Kaplan-Meier pondéré (2001)

Shen et Cai publièrent, en 2001 un article décrivant un test [55] basé sur une combinaison des idées de Fleming et Harrington [53], pour le choix de la famille de fonction de poids $G^{\rho,\gamma}$, de Pepe et Fleming [54] pour la statistique de Kaplan-Meier pondérée et d'une idée originale consistant à définir pour statistique de test, la valeur maximale de la différence de Kaplan-Meier pondérée, sur l'ensemble de l'espace de paramètres (ρ, γ). En d'autres termes, les paramètres de la famille de fonctions de poids sont choisis *a posteriori* afin de maximiser la différence entre les groupes. L'objectif de cette statistique est de fournir une puissance maximale lorsque les courbes se croisent. L'estimation des fluctuations d'échantillonnage de cette statistique maximale est réalisée sous l'hypothèse nulle de superposition parfaite des courbes. La statistique du test étant un maximum de différences relatives (positives ou négatives) pondérées entre les courbes, elle permet, d'après les auteurs, une inférence unilatérale, selon que l'on s'intéresse à la différence entre le groupe A et B, ou à la différence entre le groupe B et A. L'exemple fourni sur une base de données réelle d'un essai clinique randomisé sur le dépistage du cancer du sein, fait une inférence unilatérale.

Même si ce test, de Shen et Cai [55], est particulièrement élaboré et repose sur des années de théorisation et de développement, il conduit à une inférence unilatérale erronée et révèle les problèmes qui s'étaient accumulés durant des années. Ce test statistique peut conduire à un nouveau type d'erreur statistique : conclure simultanément à la supériorité et à l'infériorité, toutes deux significatives, de chacune des deux courbes de survie par rapport à l'autre. Ce risque d'erreur approche 100% sur de grands échantillons dès que les courbes se croisent, c'est-à-dire, spécifiquement dans le scenario pour lequel ce test a été conçu. Comment est-ce possible ? D'abord, il faut comprendre que sous l'hypothèse nulle de superposition parfaite des courbes, les deux courbes observées seront toujours relativement proches l'une de l'autre, à tous temps, surtout sur un grand échantillon. Quelle que soit la fonction de poids choisie, la statistique du test restera faible, de telle sorte que la valeur maximale restera raisonnablement petite. Intéressons-nous maintenant à un scenario de grand échantillon, avec une incertitude statistique négligeable, et deux courbes A et B qui se croisent. La courbe A correspondrait à une survie précoce meilleure et une survie

tardive moins bonne que la courbe B. Si on s'intéresse à la différence A-B dans l'idée de montrer la supériorité de A par rapport à B (statistique unilatérale), alors le paramétrage de $G^{\rho,\lambda}$avec un ρ très élevé et un λ nul, conduirait à une statistique de Kaplan-Meier pondérée (WKM) très largement positive. Le maximum des WKM sur l'espace (ρ, λ) serait donc très positif, et largement supérieur à toute fluctuation d'échantillonnage sous l'hypothèse de superposition des courbes, permettant alors de conclure à la supériorité significative de la survie de A par rapport à B. De même, si on s'intéresse à B-A, le paramétrage $G^{\rho,\lambda}$ avec un λ très élevée et un ρ nul, conduirait à un WKM très largement positif, permettant de conclure à la supériorité significative de la survie de B par rapport à A. Les auteurs notèrent néanmoins que des bornes supérieures doivent être prédéfinies sur les paramètres ρ et γ pour s'assurer que les poids soient lisses (*to assure smoothness for the weights*). Comme les auteurs le suggérèrent par la phrase « *In addition, for large ρ and λ, the weight function would take values near zero for all values of t except at the peak* », ces bornes supérieures sont nécessaires afin que le maximum existe ; autrement, les valeurs ρ et γ tendraient toutes deux vers l'infini, avec un ratio calculé pour mettre l'intégralité du poids sur le point de différence positive et maximale entre les courbes de Kaplan-Meier.

4.1.3 Considérations contextuelles au test de Shen et Cai

Considérant que l'inférence du test de Shen et Cai est en inadéquation avec l'interprétation unilatérale qui en est proposée par les auteurs, il est intéressant de proposer des réflexions méthodologiques permettant de prévenir la reproduction du même schéma.

4.1.3.1 Importance de définir la supériorité ou l'infériorité

Avant de conclure à une survie meilleure dans un groupe que dans un autre, il paraît nécessaire de définir ce qu'on entend par une survie *meilleure* pour des courbes de survie se croisant. Faut-il donner l'avantage à la survie précoce ou à la survie tardive ? C'est en laissant ce choix à un algorithme que le test de Shen et Cai arrive à la conclusion simultanée de supériorité de A à B et de B à A, grâce au non-dit en italique : « A est supérieur à B *parce qu'elle a une meilleure survie précoce* » et « B est supérieur à A *parce qu'elle a une meilleure survie tardive* ». Avec des tests tel que celui de Fleming et Harrington, le choix est librement donné au statisticien, pourvu qu'il fasse le choix *a priori*, en aveugle des données. On peut être tenté d'adapter le choix des poids selon la forme attendue de la relation, par un pari, plus

ou moins risqué, selon l'information disponible *a priori* sur les courbes de survie. Cette approche pourtant souffre du même problème que le test de Shen et Cai. En effet, si on anticipe, avec justesse, que le groupe A aura une meilleure survie précoce que le groupe B, et que l'on souhaite montrer la supériorité de A à B, alors on choisira une fonction de poids favorisant la survie précoce, alors que si on souhaite montrer la supériorité de B à A dans le même scenario, on choisira une fonction de poids favorisant la survie tardive. Deux statisticiens différents, partant des mêmes hypothèses et s'aidant des mêmes données, pourraient ainsi arriver à des conclusions diamétralement opposées ; chacun concluant ce qu'il souhaite. C'est pourquoi le choix de favoriser la survie précoce ou tardive ne devrait pas se faire sur la différence attendue, mais sur ce qui doit être considéré comme cliniquement préférable, indépendamment de toute connaissance de la réalité des courbes.

Il n'existe actuellement pas de consensus chez les biostatisticiens, mais les médico-économistes semblent avoir une piste intéressante. Considérant qu'une année de vie a la même valeur chez tous les individus, la courbe correspondant à l'espérance de vie la plus longue doit être considérée comme la meilleure. Il est aussi possible de pondérer par la qualité de vie (QALY). Cette approche souffre d'une limitation technique : soit, le suivi doit être prolongé jusqu'au décès de 100% de la cohorte, ce qui est rarement réalisable dans des délais raisonnables, soit des modèles paramétriques (Weibull) doivent être utilisés pour extrapoler, avec une fiabilité très douteuse, soit les courbes doivent être extrapolées avec les données d'une cohorte historique externe à l'étude, avec un certain nombre d'hypothèses plus ou moins fragiles. Cependant, à défaut de répondre de manière précise à une question pertinente, on pourrait être tenté de répondre de manière très précise à une question distincte. Cela pourrait conduire à une erreur de 3ème espèce.

Ce problème de définition de la supériorité n'a pas été discuté par Shen et Cai ni par les articles le précédant sur la même lignée de développements statistiques [51,56,53,54], à l'exception de Mantel, qui avait débuté une réflexion, mais avait suggéré qu'il était insoluble :

> « *What is required is a value or utility function defining the value of living to a certain age or for a certain period of time beyond therapy. Given the value function one can determine the average value for exact survival patterns. […] An economist might objectively determine the*

utility of surviving from any one age to any other age, perhaps doing so separately for each sex and race combination. With differing normative assumptions, the theologian and the philosopher would provide quite different value functions. »

— Nathan Mantel, 1966 [49].

Pourtant, en se désintéressant de ce problème, il existe un risque de conclure à la supériorité d'une courbe de survie à une autre quand bien même l'économiste, le théologien et le philosophe seraient tous en accord sur son infériorité.

4.1.3.2 Importance de nommer précisément les termes

Pour pouvoir accepter une hypothèse alternative par rejet de l'hypothèse nulle, celle-ci doit être formulée comme la négation de celle-là. Le rejet de l'hypothèse nulle de superposition parfaite des courbes permet de conclure à l'existence d'une différence entre les courbes, en au moins un de leurs points, mais ne permet pas de conclure sur la direction de la différence.

Une définition précise des termes statistiques permettrait d'éviter des formulations inappropriées des hypothèses. Le terme d'hypothèse alternative semble être polysème, puisqu'il est utilisé en des sens bien distincts. Afin de pouvoir discuter de ce problème, nous allons tenter de fournir un ensemble cohérent de termes qui serviront de base à la réflexion. Pour l'estimation d'une différence de moyennes, nous allons distinguer plusieurs paramètres et hypothèses : la différence de moyennes *réelle* $\mu_1 - \mu_2$, la différence de moyennes *observée* $m_1 - m_2$, l'hypothèse *alternative*, c'est-à-dire, l'hypothèse que l'on souhaite prouver par un test d'hypothèse, telle que $\mu_1 - \mu_2 > \Delta$, et enfin, l'hypothèse *nulle*, formulée comme la négation de l'alternative, respectivement $\mu_1 - \mu_2 \leq \Delta$. Le concept se généralise aussi aux comparaisons de courbes de survie : il existe une *réalité*, définie sur la population entière, par deux courbes parfaitement définies, point par point ; cette réalité n'est connue que dans les simulations statistiques et le développement d'outils statistiques. Il existe une *observation* de deux courbes, sur un échantillon, ainsi qu'une statistique de test obtenue à partir de ces deux courbes. Il existe une hypothèse *alternative* portant sur l'ensemble des courbes réelles possibles. Cette hypothèse peut être formulée comme l'appartenance des paires de courbes à un ensemble particulier, tel que l'ensemble des paires de courbes pour lesquelles l'espérance (aire sous la courbe) de la première est supérieure à l'espérance de

la seconde. L'hypothèse alternative, est une assertion logique, qui ne peut être que vraie ou fausse. Enfin l'hypothèse nulle est la négation logique de l'hypothèse alternative. Les hypothèses nulle et alternative sont des choix du statisticien, fondés sur les hypothèses de recherche, alors que la distribution réelle n'est en rien dépendante de lui.

Le terme d'*alternative* utilisé par Mantel [49], Peto [51], Tarone et Ware [56], Pepe et Fleming [54], Shen et Cai [55] est utilisée au singulier, pour faire référence à la paire de courbes *réelle*, ou au pluriel, pour faire référence à une famille de paires de courbes dans lesquelles la paire de courbes *réelle* se situe : dans tous les cas, cette définition porte sur les courbes *réelles* plutôt que sur l'hypothèse que le statisticien souhaite tester. Cette ambiguïté sur la définition de l'hypothèse alternative, existe depuis les premières publications de Neyman et Pearson sur les tests du rapport de vraisemblance, dès 1933 [57]. En effet, Neyman et Pearson commencent par définir des tests d'hypothèses simples, dans lesquelles la distribution statistique sur laquelle le test porte est entièrement définie sous l'hypothèse nulle. Pour commencer, Neyman et Pearson présentent une hypothèse nulle comme alternative toutes deux simples. Cela correspondrait à une situation où le paramètre θ d'une distribution ne peut prendre que deux valeurs (θ_0 sous H_0 et θ_1 sous H_1) sans qu'il soit possible que θ prenne une autre valeur. Ensuite, Neyman et Pearson considèrent une hypothèse nulle simple avec de multiples alternatives H_1, H_2, …, prises dans une famille d'alternatives et précisent :

> « *As will appear below when discussing illustrative examples, in certain cases the family of best critical regions is not the same for each of the admissible alternatives* H_1 , H_2, *... ; while in other cases a single common family exists for the whole set of alternatives. In the latter event the basis of the test is remarkably simple. If we reject* H_0 *when the sample point,* Σ, *falls into* ω_0, *the chance of rejecting it when it is true is* ε, *and the risk involved can be controlled by choosing from the family of best critical regions to which* ω_0 *belongs, a region for which* ε *is as small as we please.* »
>
> — Neyman & Pearson, 1933 [57]

Pour résumer leur propos, la plupart du temps, il n'est pas nécessaire de connaître la distribution *réelle*

(ici dénommée alternative) pour pouvoir rejeter une hypothèse nulle. Dans ce cadre théorique, il n'est pas non plus précisé que l'ensemble des alternatives doit correspondre à la négation de l'hypothèse nulle. Il n'y a pas non plus de distinction entre distribution *réelle* et hypothèse *alternative*.

Dans un chapitre sur les hypothèses composites, Neyman et Pearson [57] introduisent les hypothèses nulles composites portant sur l'appartenance d'une distribution à une famille de distributions ; ainsi, certains paramètres de la distribution ne sont pas spécifiés dans ces hypothèses composites. Malheureusement, ils ne traitent que le cas où la statistique du test est indépendante des paramètres non spécifiés. Ces hypothèses composites sont adaptées aux situations où il existe un paramètre de nuisance (p.ex. la variance dans un test de comparaison de moyennes) mais ne peuvent pas correspondre à une spécification partielle de paramètres inconnus telle que $H_0: \mu_0 \leq \mu_1$. Il faut admettre qu'il n'est plus possible de contrôler parfaitement le risque de première espèce dans cette dernière situation ; on peut juste garantir que cette erreur soit inférieure ou égale à un seuil prédéfini, qui sera généralement obtenu dans le scenario où μ_0 serait égal à μ_1.

4.1.3.3 Utilité d'un cadre théorique permettant une inférence directionnelle

Le cas de la comparaison des courbes de survie répond à l'argument des défenseurs de l'inférence directionnelle sur des tests non conçus comme directionnels [58], qualifiant les propos de Kaiser [25] de *reductio ad absurdum*, notamment lorsqu'il écrivait *"we cannot logically make a directional statistical decision or statement when the null hypothesis is rejected on the basis of the direction of the difference in the observed sample means"*. L'exemple de l'analyse de survie permet de comprendre qu'une inadéquation entre la théorie et l'usage peut avoir des conséquences réelles sur la validité de l'inférence statistique. L'approche des tests trilatéraux proposée par Goeman *et al* [37], paraît intéressante, car répondant à la fois au problème de la futilité des différences et au problème de la directionnalité des conclusions :

$$H_0: -\Delta \leq \theta \leq +\Delta \; (equivalence)$$

$$H_+: \theta > \Delta$$

$$H_-: \theta < -\Delta$$

Elle introduit néanmoins le problème du choix, subjectif, du Δ.

4.2 Two-sided confidence interval of a binomial proportion: how to choose?

L'article intitulé « *Two-sided confidence interval of a binomial proportion: how to choose?* » [59] (Annexe 2) revisite un des problèmes statistiques les plus basiques, l'estimation d'une proportion binomiale, avec les critères d'évaluation de performances des estimateurs décrits en section 3. Une revue de la littérature sur le sujet est faite, avec une tentative de lister l'ensemble des estimateurs d'intervalles qui ont fait l'objet de publication, puis les estimateurs sont tous soumis aux évaluations de performances et des conclusions sont faites sur le ou les estimateurs présentant les meilleures propriétés théoriques et pratiques.

4.2.1 Justification des choix statistiques et méthodologiques

Un total de 55 estimateurs d'intervalles de confiance de proportions binomiales a été réimplémenté dans le logiciel R avant d'être analysés. Des modifications minimes ont été faites à ces estimateurs en supposant qu'elles seraient faites *a posteriori* par les statisticiens : notamment toute borne de confiance négative était mise à zéro, alors que toute borne supérieure à un était plafonnée à un. Pour les estimateurs indéfinis pour les proportions observées à 0 et 1, tel que l'estimateur de Wald dans une régression logistique à intercept seul, l'estimateur de Clopper-Pearson a été utilisé pour substituer les valeurs indéfinies ; en effet l'intervalle de Clopper-Pearson est l'intervalle « exact » le plus classique, implémenté dans la plupart des logiciels statistiques.

Les risques unilatéraux moyens locaux et les demi-largeurs moyennes locales ont été analysés pour un risque bilatéral nominal de 5% (10% en analyse de sensibilité). Il n'a pas été nécessaire de recourir à des simulations de Monte Carlo puisque des calculs pouvaient être faits aisément à partir des distributions exactes. Par exemple, pour estimer le risque unilatéral gauche de l'intervalle de confiance de Clopper-Pearson conditionnel à une distribution binomiale de dénominateur $n = 32$ et de proportion théorique $\pi = 0{,}089$, les intervalles de confiance ont été calculés pour tous les numérateurs de 0 à 32, ensuite, tous les intervalles de confiance dont la borne basse est supérieure à $\pi = 0{,}089$ ont été identifiés. Dans cet exemple, ces intervalles de confiance correspondent à tous les numérateurs inférieurs ou égaux à 6, car pour 6/32, l'intervalle de confiance est [0,072 ; 0,364] alors que pour 7/32, l'intervalle est [0,093 ; 0,400]. Ainsi, il existe une erreur de surestimation de la proportion réelle $\pi = 0{,}089$ pour

tout numérateur observé supérieur ou égal à 7. Ensuite, la fonction de masse de la distribution binomiale $\mathcal{B}(32; 0{,}089)$ a servi à calculer la somme de toutes les probabilités correspondant aux numérateurs de 7 à 32, soit 0,02037. En négligeant les erreurs d'arrondi des nombres à virgule flottante 64 bits (IEEE-754), les calculs sont exacts. Les risques moyens locaux ont été calculés en approximant la distribution continue du paramètre supposé aléatoire (proportion théorique π) a une distribution discrète à 512 valeurs régulièrement espacées, permettant alors de calculer le risque unilatéral moyen local comme la moyenne des risques unilatéraux conditionnels à chacune de ces 512 valeurs de π, pondérée par leurs probabilités. Un affinement de la distribution par une distribution discrète plus fine (p.ex. 1024 valeurs plutôt que 512) n'avait qu'un impact imperceptible sur les résultats.

Une analyse graphique des risques unilatéraux moyens locaux et des demi-largeurs moyennes locales devait être faite, mais quelques problèmes ont nécessité une adaptation de la représentation, tout particulièrement pour les risques unilatéraux moyens locaux. D'abord, la largeur d'un intervalle de confiance d'une proportion binomiale est très dépendante du numérateur. Par exemple, pour l'estimateur d'intervalle de confiance de Clopper-Pearson la demi-largeur à droite pour 0/2048 est 12,2 fois plus petite que pour 1024/2048. Cela rendrait difficile la lecture d'une figure qui présenterait ces demi-largeurs pour tout l'intervalle des numérateurs de 0 à 2048, car des ordonnées très différentes seraient présentées. Ce problème a été résolu en présentant le rapport entre les demi-largeurs locales moyennes de chaque estimateur d'intervalle de confiance, avec celles d'un estimateur d'intervalle de confiance de référence. L'estimateur de Clopper-Pearson mid-P a été choisi comme référence, car il contrôle très bien les risques moyens locaux unilatéraux, évitant que des irrégularités de l'intervalle de référence impactent la présentation graphique. Par ailleurs, pour un grand dénominateur, comme 2048, la majorité des estimateurs d'intervalles ne diffèrent notablement que pour des numérateurs proches de 0 ou de 2048, puisqu'au milieu, le théorème central limite rend l'approximation normale valide et toutes les méthodes sont presque équivalentes. Il paraissait nécessaire de focaliser l'attention sur les numérateurs bas, tel que l'intervalle des numérateurs [0 ; 32]. C'est pourquoi, l'espérance du numérateur $\lambda = n \times \pi$ a été présentée en abscisse, avec une échelle allant de 0 à 32. Cette présentation permet aussi de présenter le cas limite de la loi de Poisson (dénominateur $n \to +\infty$ avec λ constant) sur une même figure. Il n'a pas

été jugé nécessaire de présenter les numérateurs très proches du dénominateur (p.ex. intervalle [2016 ; 2048]) car les estimateurs étaient presque tous équivariants, c'est-à-dire, ayant des comportements transposables que l'on s'intéresse à estimer la proportion d'événements ou la proportion de non-événements.

4.2.2 Interprétation et généralisation des résultats

4.2.2.1 Convergence vers la loi normale

Comme la généralisation de la problématique et des résultats ne transparaît pas forcément dans l'article, une mise en perspective est faite ci-dessous. D'abord, la convergence d'une proportion binomiale vers une loi normale, pour les proportions éloignées de 50% est particulièrement lente et l'approximation de la variance observée à la variance réelles pose d'autant plus problème qu'il existe une grande covariance entre l'estimateur de la variance et l'estimateur ponctuel de la proportion ; c'est donc un scénario assez défavorable, dans lequel de nombreux estimateurs statistiques sont fortement biaisés. L'estimation d'une proportion est un cas limite de l'estimation d'une différence entre deux proportions où une des deux proportions serait estimée sur un échantillon de taille infinie. L'asymétrie des fluctuations d'échantillonnage et la covariance entre la moyenne et l'estimateur de la variance s'estompent dans le scenario de l'estimation d'une différence de deux proportions sur des échantillons de taille égale à condition que les proportions ne soient pas très différentes. Le cas limite de l'estimation d'une seule proportion est particulièrement défavorable. Le cas limite d'une proportion binomiale π sur un échantillon de taille n où n tendrait vers l'infini mais $n\pi$ serait constant est décrit par la loi de Poisson. Dans cet article, nous montrons que le cas limite de la loi de Poisson est le plus fortement biaisé.

4.2.2.2 Fluctuations d'échantillonnage discrètes

Sur un autre aspect, l'estimation d'une proportion binomiale est un cas limite pathologique d'une comparaison de deux proportions où une des deux proportions serait calculée sur un échantillon de taille infinie. L'aspect discret des fluctuations d'échantillonnage est particulièrement marqué, du moins dans une analyse conditionnelle à la taille d'échantillon n et la proportion π. Ce problème qui a fait couler beaucoup d'encre [11] est beaucoup moins marqué sur les statistiques relatives telles que les rapports d'incidences (cf. section 1.2.11.2) car le cardinal du support de l'estimateur (nombre de valeurs possibles

prises par les estimations) est bien plus grand dans ces statistiques que pour une proportion binomiale unique. Ce problème de fluctuations discrètes est encore atténué dans les modèles multivariés. Les résultats de l'article mettent en évidence l'intérêt des intervalles de confiance construits par inversion d'un test basée sur une mid-P-valeur [60]. D'une manière assez générale, la P-valeur est définie comme $p = \mathbb{P}(S \geq s|H_0)$ où S est la variable aléatoire reflétant la statistique d'un test et s est sa réalisation sur l'échantillon. La P-valeur est donc la probabilité, sous l'hypothèse nulle, d'observer une statistique au moins aussi extrême que celle qui a été observée sur l'échantillon. Doit-on utiliser une inégalité stricte ou large pour la comparaison de S à s ? Lorsque les fluctuations d'échantillonnage de S sont continues, la question n'est pas pertinente car la probabilité d'égalité parfaite est nulle et les deux approches sont équivalentes. Pour une fluctuation discrète de S, l'approche de l'inégalité large tend à produire des statistiques conservatives alors que l'inégalité stricte produit des statistiques trop libérales. La mid-P-valeur est une approche intermédiaire. Elle est égale à $\frac{1}{2}\mathbb{P}(S = s|H_0) + \mathbb{P}(S > s|H_0)$. Cette mid-P-valeur, est aussi égale à la moyenne des deux P-valeurs avec inégalité stricte ($\mathbb{P}(S > s|H_0)$) et large ($\mathbb{P}(S \geq s|H_0)$). Cette solution intermédiaire, appliquée à l'estimateur des proportions binomiales, conduit à un intervalle de confiance qui n'est ni conservatif ni libéral lorsqu'il est analysé avec les *risques unilatéraux moyens locaux* (cf. section 3.1). La mid-P-valeur est généralisable à des cas bien plus complexes, tel que l'estimation des odds ratio dans des régressions logistiques basées sur la loi binomiale exacte, telle que décrites par Hirji *et al* [42].

4.2.2.3 Relation entre demi-largeur et risque unilatéral

L'analyse des demi-largeurs moyennes locales a montré un miroir parfait des risques moyens locaux unilatéraux : un point avec une demi-largeur moyenne plus étroite était systématiquement associée à un risque unilatéral plus grand alors qu'une demi-largeur moyenne plus large était associée à un risque unilatéral plus petit. Cela est explicable par la grande simplicité de la distribution binomiale, qui, conditionnellement au dénominateur n, toujours bien connu, est mono-paramétrique. Le seul paramètre qui soit considéré comme un paramètre de nuisance par certains estimateurs, est la variance, notamment pour l'estimateur de Wald qui approxime la loi binomiale mono-paramétrique à une loi bi-paramétrique normale. En conséquence, les estimateurs sont tous stables. Cela diffère de scenarii plus complexes, tels

que l'estimation de l'effet d'un traitement avec ajustement sur de nombreuses covariables fortement déséquilibrées entre les groupes. Dans ces scenarii plus complexes, il est possible que deux estimateurs contrôlent tout aussi bien les risques unilatéraux, mais que l'un soit bien plus stable que l'autre, avec une largeur nettement plus étroite.

Les demi-largeurs conditionnelles s'avèrent être une fonction continue de la proportion π, contrairement aux couvertures conditionnelles. Les discontinuités de la fonction de couverture conditionnelle aux bornes des intervalles de confiance, sont explicables par le fait qu'une modification infime de π, sans changer notablement la distribution binomiale, pourra brutalement changer le statut d'un intervalle de confiance, qui se mettra à contenir de justesse, ou au contraire, arrêtera de contenir de justesse, la valeur π. Ces discontinuités ne sont pas du tout retrouvées sur les demi-largeurs conditionnelles car celles-ci ne dépendent pas du fait que les intervalles de confiance contiennent ou pas π. Ainsi, les moyennes locales ne sont pas nécessaires à lisser les demi-largeurs, car celles-ci sont déjà lisses. Les demi-largeur moyennes locales sont alors presque égales aux demi-largeurs conditionnelles.

4.2.2.4 Comparaison des estimateurs classiques

Trois grands estimateurs généraux d'intervalles de confiance sont aussi décrits : Wald, score de Rao et l'inversion d'un test de rapport de vraisemblance généralisé avec, dans les trois cas, une approximation à une loi normale ou une loi du χ^2. L'estimateur de Wald, déjà connu pour être médiocre, s'avère être encore pire lorsque les risques unilatéraux sont analysés. L'estimateur du score de Rao, qui conduit à l'intervalle de Wilson [61] s'avère avoir des propriétés bien moins bonnes que décrites par Agresti et Coull [40] lorsque les risques unilatéraux sont analysés. L'inversion du test de rapport de vraisemblance généralisé a d'excellentes propriétés statistiques sauf dans les cas les plus extrêmes où l'espérance du nombre d'événements $n\pi$ est inférieure à deux. Ces résultats permettent de privilégier l'estimateur du maximum de vraisemblance

4.3 Best estimator for bivariate Poisson regression

Ce troisième article [62] (Annexe 3) se rapporte au scenario de l'estimation du rapport de deux variables suivant des lois de Poisson. Si on considère que l'estimation de l'espérance λ d'une loi de Poisson est

le cas limite de l'estimation d'une proportion binomiale dans laquelle les biais sont maximaux, comme vus dans l'article précédent, en raison d'une covariance maximale entre l'espérance et la variance ainsi que d'un coefficient d'asymétrie (skewness) maximal, alors il est pertinent de s'intéresser à la loi de Poisson pour le problème bivarié, décrit comme l'estimation du rapport de deux lois de Poissons indépendantes. Cette loi de Poisson est typiquement utilisée pour décrire des rapports de taux d'incidences.

Lorsqu'on conditionne sur le nombre total d'événements (Poisson) ou sur la taille d'échantillon (binomial), Le problème de Poisson bivarié s'avère proche du problème binomial univarié. Si on considère deux variables $Y_1 \sim \mathcal{P}(\lambda_1)$ et $Y_2 \sim \mathcal{P}(\lambda_1)$ suivant deux lois de Poisson indépendantes alors la distribution de Y_1 conditionnelle à $Y_1 + Y_2 = y_1 + y_2$ est une distribution binomiale $\mathcal{B}\left(\frac{\lambda_1}{y_1+y_2}; y_1 + y_2\right)$.

Cette équivalence est décrite en Annexe 4 (section 11). Il est aussi constaté que les inférences faites par le score de Rao et le rapport de vraisemblance réalisées sur un modèle linéaire généralisé binomial univarié (intercept seul) est équivalent à l'inférence du même estimateur réalisée sur un modèle linéaire généralisé Poisson bivarié (intercept + covariable binaire).

Le problème diffère néanmoins légèrement dans une approche d'évaluation inconditionnelle des risques. Même en supposant λ_1 et λ_2 constants, le nombre d'événements $Y_1 + Y_2$ est variable, suivant une loi de Poisson d'espérance $\lambda_1 + \lambda_2$. Lorsque les risques moyens locaux sont analysés, notamment dans l'hypothèse qu'une taille d'échantillon binomiale est elle-même une variable aléatoire N, alors les problèmes se rapprochent de nouveau.

L'analyse qui est faite dans l'article confirme la supériorité du rapport de vraisemblance au score de Rao ou à la méthode de Wald. La robustesse du rapport de vraisemblance sur de petits échantillons résout un problème de fond concernant les conditions de validité des estimateurs : si celles-ci sont évaluées *a posteriori* et s'avèrent violées, que peut-on faire ? Ne fournir aucun résultat engendrerait un biais de présentation sélective des résultats (selective reporting bias), qui n'est pas forcément dû au hasard, et pourrait parfois être le fruit d'effets particulièrement forts sur l'échantillon. Cela biaise alors l'estimateur. Des estimateurs dont la validité sont très larges, n'étant invalides que lorsqu'on se rapproche de la division par zéro systématique, permettent de se retrouver très rarement dans cette

situation désagréable.

Les méthodes de Wald et du score présentent des défauts de couverture majeurs pour un niveau de confiance à 0,999 et à $1\text{-}10^{-6}$, indirectement interprétables comme l'inflation du risque alpha d'un test d'hypothèse. Cela doit être pris en considération dans l'interprétation d'un petit p, notamment en cas de correction de multiplicité des tests pouvant descendre le niveau de significativité largement en dessous du seuil 5% conventionnel. La méthode du rapport de vraisemblance reste très robuste dans ce contexte extrême.

5 Discussion

5.1 Intérêt des moyennes locales

L'approche des *risques moyens locaux* revêt un intérêt tout particulier lorsque les fluctuations d'échantillonnage sont discrètes, conduisant à un lissage, réfrénant alors l'envie de reposer sur des estimateurs strictement conservatifs, et favorisant les approches telles que les mid-P-valeurs. Cette approche a-t-elle un intérêt au-delà des cas des fluctuations discrètes ? La réponse n'est pas évidente, mais renverser la question apporte une nouvelle perspective. Pourquoi conditionnerait-on toujours les analyses des propriétés statistiques à des paramètres qui sont manifestement variables ? N'y a-t-il pas un risque d'oublier une source de variance, voire de covariance dans certaines situations ? La méthode de conditionnement aux paramètres de nuisance est très largement répandue et, n'engendre pas forcément de biais important lorsque la covariance entre le paramètre de nuisance et la statistique estimée est faible ou nulle. Ce problème est parfois sous-estimé. Par exemple, l'offset dans une régression de Poisson sert souvent à prendre en compte une durée de suivi, reflétant ainsi le risque d'un ou plusieurs événements incidents. Pour un calcul de ratio standardisé de mortalité, l'offset d'une régression de Poisson peut aussi servir à exprimer le risque théorique de décès pour chacun des sujets analysés. Dans ces deux contextes, la covariance entre le risque théorique et le risque observé peut être fort, voire très fort. La régression de Poisson étant conditionnelle à l'offset, elle ignore cette covariance. Dans une étude estimant le ratio standardisé de mortalité du Pemphigus en France [63] dont les statistiques ont été réalisées par l'auteur de cette thèse, l'intervalle de confiance à 95% obtenu par boostrap, prenant en compte cette covariance était (1,46 à 1,93) alors que celui obtenu à partir de la régression de Poisson était (1,30 à 2,13), soit une erreur type 1,77 fois plus grande sur l'échelle logarithmique, correspondant encore à une variance 3,1 fois plus grande. En bref, ignorer la covariance positive entre le risque observé et le risque théorique était équivalent à sous-estimer d'un facteur 3,1 la taille d'échantillon. L'intervalle de confiance conditionnel à l'offset avait un niveau de confiance environ égal à 0,9995, soit un risque de non-couverture cent fois inférieur au risque nominal.

5.2 Intérêt des analyses unilatérales

Cette question a suffisamment été discutée dans ce travail. L'analyse des propriétés unilatérales, aussi bien pour les intervalles de confiance que pour les tests d'hypothèses, permet de rapprocher la théorie de la réalité des besoins et des usages qui sont déjà faits. Un écart entre la théorie et la pratique conduit à une perte d'énergie dans des outils statistiques peu utiles, notamment les nombreux intervalles ayant pour objectif d'être les plus étroits possibles par déséquilibre volontaire, mais aussi parfois à des erreurs d'interprétation plus graves comme pour les comparaisons de courbes de survie.

5.3 Limites du travail

Dans l'article concernant les proportions binomiales, les risques unilatéraux moyens locaux ont été analysés en faisant varier, soit la proportion π théorique, soit la taille d'échantillon N, mais pas les deux en mêmes temps. Étant donné que le principal effet de cette variabilité des paramètres est de lisser les risques de sous-estimation et surestimation, atténuant, voire faisant disparaître les fluctuations des risques, cette variance bidimensionnelle tendrait à lisser encore plus les fluctuations d'échantillonnage. Le choix d'analyser séparément les deux cas de variabilité, de N et de π, avait été fait en raison de la lourdeur calculatoire du scenario de double variance. En effet, afin d'avoir des courbes parfaites, les calculs ne se sont pas basés sur des simulations ; ils ont été faits à partir des distributions binomiales ou de Poisson exactes.

La notion de demi-largeur décrite en section 3.2 souffre de deux limitations : elle est dépendante d'un estimateur ponctuel et elle n'est pas invariable aux transformations monotones. Si un estimateur ponctuel biaisé est utilisé, on surestimera ou sous-estimera la demi-largeur par rapport à ce qui serait obtenu avec un estimateur non biaisé. Néanmoins, ce biais sera identique entre deux estimateurs d'intervalle de telle sorte qu'on pourra toujours les comparer sur leur demi-largeur à condition de les baser sur le même estimateur ponctuel. Ce problème pourrait être complètement résolu en s'intéressant à la différence entre la borne de l'intervalle de confiance et la valeur théorique du paramètre θ plutôt que son estimation ponctuelle. Néanmoins, cela ne changerait la demi-largeur que d'une constante et fournirait une interprétation légèrement moins intuitive. La deuxième limite mentionnée est la fragilité

aux transformations monotones. Cette question se pose notamment dans le cadre des régressions de Poisson et des régressions logistiques pour lesquels on peut s'intéresser aux paramètres du modèle sur l'échelle du prédicteur linéaire (log-odds ratios et log-incidence ratios) ou après transformation exponentielle (odds ratios et incidence ratios). Avec de telles transformations, il est théoriquement possible que le classement de deux estimateurs d'intervalles de confiance soit renversé. Ce problème pourrait être résolu par l'usage de la médiane de la différence entre le paramètre théorique θ et la borne de l'intervalle de confiance.

Ces problèmes sont probablement futiles en comparaison de la très forte covariance entre le risque alpha unilatéral moyen local et la demi-largeur. D'une manière générale, un intervalle libéral sera plus étroit et vice versa. À risque alpha égal, il arrive parfois qu'un intervalle soit plus stable et plus étroit qu'un autre, mais cette situation se rencontre dans des modèles complexes. Cela est parfois observé avec la pondération par l'inverse du score de propension qui est moins efficace que la pondération par le complément à un du score de propension lorsque des scores extrêmes conduisent à des poids extrêmes d'un petit nombre de sujets, entraînant des fluctuations d'échantillonnage démesurées [64]. Lorsque les risques alpha unilatéraux ou bilatéraux réels diffèrent, il faudrait trouver un moyen de calculer des largeurs ou demi-largeurs ajustées sur ces risques. Malheureusement, après moult tentatives sur les intervalles de confiance de proportions binomiales, aucune solution satisfaisante n'a été trouvée. Une simple modification du niveau de confiance nominal afin d'égaliser les niveaux de confiance réels n'est pas possible car le biais de couverture n'est pas global mais local. Une rectification du niveau de confiance nominal afin que le niveau de confiance réel soit respecté pour une proportion suivant une loi uniforme sur l'intervalle $]0;1[$ n'a souvent aucun effet car la couverture nominale moyenne générale peut être égale à la couverture réelle alors que les problèmes de biais sont locaux. Des procédures de modification du jeu d'intervalles, à appliquer avant la comparaison des demi-largeurs, dans l'objectif de rééquilibrer les risques unilatéraux locaux a été envisagée, mais toutes les procédures envisagées altéraient complètement les propriétés du jeu d'intervalles qui n'avaient alors plus aucun rapport avec le jeu originel. Les comparaisons des demi-largeurs ainsi « ajustées » ne dépendaient plus que de la procédure de modification des intervalles. Les tentatives d'altérer localement le jeu d'intervalles par

modification du niveau de confiance nominal, séparément en chaque point π, conduisait à des interprétations aberrantes car c'est un jeu d'intervalle complet que l'on doit analyser : pour chaque proportion π, il est très facile de trouver un jeu d'intervalle optimal, complètement et étroitement centré autour de π, mais c'est dans la globalité que le problème se pose. Le problème de la comparaison de l'efficacité de deux estimateurs différemment biaisés reste ouvert. En l'absence de solution à ce problème, l'absence de biais important est un prérequis à la comparaison de la précision de plusieurs estimateurs.

Ce travail s'est limité aux situations les plus simples. Notamment, les fréquents scenario de modèles linéaires généralisés avec variables d'ajustements n'ont pas été analysés. Certains problèmes supplémentaires apparaissent dans ces situations, notamment en cas de corrélations fortes entre les covariables.

5.4 Tests d'hypothèses

Comme décrit dans les méthodes, les performances des tests d'hypothèses n'ont pas été analysés dans ce travail. Néanmoins une relation peut être établie entre les performances statistiques des estimateurs d'intervalle et celles des tests d'hypothèses que l'on peut établir par comparaison directe d'une borne d'un intervalle de confiance à la valeur comparée. Le risque alpha d'un test d'hypothèse unilatéral sera superposable aux risques α'_L et α'_U définis en section 3.1. Une demi-largeur moyenne plus étroite, telle que définie en section 3.2, sera généralement associée à une puissance plus élevée du test unilatéral associable à la borne considérée de l'intervalle de confiance. Une demi-largeur moyenne plus étroite, telle que définie en section 3.2, sera généralement associée à une puissance plus élevée du test unilatéral associable à la borne considérée de l'intervalle de confiance. Néanmoins, la corrélation entre la puissance et la demi-largeur n'est pas parfaite, et il est possible d'imaginer un estimateur d'intervalle qui aurait une demi-largeur moyenne plus étroite qu'un autre mais conduirait à une inférence de moindre puissance lorsqu'il serait utilisé comme test d'hypothèse unilatéral. Cela est explicable par le fait que les distributions des demi-largeurs d'intervalles ne peuvent être résumées à leur espérance ; la puissance devrait se calculer à partir de la fonction de répartition de la distribution de demi-largeur.

Les niveaux de confiance 90%, 95% et 99% sont fréquemment usités pour les intervalles de confiance,

mais les autres niveaux de confiance sont rarement retrouvés dans la littérature. Les tests d'hypothèses sont généralement présentés avec une P-valeur plutôt que de manière binaire (significatif ou pas). L'exactitude de cette P-valeur, même lorsqu'elle est faible (p.ex. 3 pour 10 000) est souhaitable d'autant plus que des corrections de multiplicité des tests peuvent leur être appliquées. Cela a pu être indirectement évalué par des niveaux de confiance extrêmes, tels que 99,9 % ou 99,9999 %. C'est pour cela que l'article intitulé « Best estimator for bivariate Poisson regression » [62] comprend une analyse à de tels niveaux de confiance, mettant en évidence une certaine robustesse de l'estimateur du rapport de vraisemblance, mais des biais inacceptables (multiplications des risques alpha par un facteur > 100) des estimateurs de Wald et du score de Rao. Un petit p à 10^{-6} selon le test du score peut alors survenir plus d'une fois sur 10 000 sous l'hypothèse nulle si les deux groupes sont fortement déséquilibrés en taille et que les effectifs sont faibles dans le plus petit groupe, tel que 4 événements sur 10 patients-années dans un groupe contre 100 événements sur 250 patients-années dans l'autre groupe.

Même si l'analyse des tests d'hypothèses est indirecte, les résultats comparant les estimateurs les plus communs sont suffisamment contrastés pour en permettre l'interprétation. Dans les scenarii univariés et bivariés analysés, les demi-largeurs d'intervalle de confiance sont indissociables des risques unilatéraux, tels que décrit dans la section des limites (5.3), ce qui conduit à de fortes limites d'interprétation des demi-largeurs ; le même problème affecterait toute tentative de calculer les puissances des tests d'hypothèses.

6 Conclusion

Les considérations qui ont été à l'origine de ce travail ne sont pas nouvelles mais semblent être encore rarement prises en compte. Ce travail se destine, premièrement, aux statisticiens qui voudraient développer ou évaluer de nouvelles ou anciennes méthodes statistiques, leur suggérant de prendre en considération l'interprétation directionnelle et les invitant à réfléchir aux sources de variance et de covariance afin de ne pas toujours conditionner l'évaluation des statistiques sur les paramètres de nuisance ou la taille d'échantillon. Il est notamment possible d'adopter une approche différente (conditionnelle ou inconditionnelle) lors de la construction de l'outil et lors de son évaluation.

Deuxièmement, ce travail se destine aux statisticiens soucieux d'appliquer des procédures adaptées à leur problème et souhaitant les interpréter correctement. Avant d'utiliser un test d'hypothèse, il faut retenir l'importance de bien identifier l'hypothèse nulle qu'il rejette et se méfier des formulations dans lesquels l'hypothèse alternative n'est pas la négation de l'hypothèse nulle.

Troisièmement, ce travail se destine aux enseignants qui souhaitent transmettre une théorie en adéquation avec les usages qui en sont faits. Dans l'enseignement des tests d'hypothèses, débuter par la formulation de l'hypothèse alternative avant de reformuler sa négation, permettrait de pousser à une réflexion quant aux objectifs de la recherche, aux résultats attendus ainsi qu'aux résultats souhaités plutôt que de formuler une hypothèse nulle d'indépendance de toutes les variables en jeu. On remarquera qu'il arrive parfois que les résultats attendus sont supérieurs aux résultats souhaités. Par exemple, on peut attendre une supériorité d'une intervention mais se satisfaire d'une non-infériorité en raison d'autres avantages de l'intervention.

Même si les explorations réalisées ne concernent que des situations statistiques très simples, elles mettent en évidence un avantage très net de l'estimateur du rapport de vraisemblance sur celui de Wald et du score de Rao.

Il serait souhaitable de pousser les investigations plus loin, en réévaluant les propriétés des estimateurs dans les modèles linéaires généralisés à plusieurs covariables, mais aussi pour les modèles de survie, notamment le modèle de Cox et l'estimateur de Kaplan-Meier associé à la variance de Greenwood.

7 Références

1. Student. The Probable Error of a Mean. Biometrika. 1908;6(1):1-25.

2. Fisher R. VII. Intraclass correlations and the analysis of variance. In: Statistical methods for research workers. Oliver&Boyd. Edinburgh; 1925. p. 176-210.

3. Brandt AE. The analysis of variance in a 2×s table with disproportionate frequencies. J Am Stat Assoc. 1933;28(182):164-73.

4. Yates F. The analysis of multiple classifications with unequal numbers in the different classes. J Am Stat Assoc. 1934;29(185):51-66.

5. Herr DG. On the history of ANOVA in unbalanced, factorial designs: The first 30 years. Am Stat. 1986;40(4):265-70.

6. Sanborn AN, Hills TT. The frequentist implications of optional stopping on Bayesian hypothesis tests. Psychon Bull Rev. 2014;21(2):283-300.

7. Ryan EG, Brock K, Gates S, Slade D. Do we need to adjust for interim analyses in a Bayesian adaptive trial design? BMC Med Res Methodol. 2020;20(1):150.

8. Nelder JA, Wedderburn RWM. Generalized Linear Models. J R Stat Soc Ser Gen. 1972;135(3):370-84.

9. Firth D. Bias Reduction of Maximum Likelihood Estimates. Biometrika. 1993;80(1):27-38.

10. Kenne Pagui EC, Salvan A, Sartori N. Median bias reduction of maximum likelihood estimates. Biometrika. 2017;104(4):923-38.

11. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. Stat Sci. 2001;16(2):101-17.

12. Zieliński W. The Shortest Clopper–Pearson Confidence Interval for Binomial Probability. Commun Stat - Simul Comput. 2009;39(1):188-93.

13. Arthur Woodward J, Liu W-C, Bonett DG. Shortest two-tailed confidence intervals. Appl Math Comput. 1997;84(1):65-76.

14. Crow EL. Confidence Intervals for a Proportion. Biometrika. 1956;43(3/4):423-35.

15. Blyth CR, Still HA. Binomial Confidence Intervals. J Am Stat Assoc. 1983;78(381):108-16.

16. Casella G. Refining binomial confidence intervals. Can J Stat. 1986;14(2):113-29.

17. Jackson D, Bowden J. Confidence intervals for the between-study variance in random-effects meta-analysis using generalised heterogeneity statistics: should we use unequal tails? BMC Med Res Methodol. 2016;16(1):118.

18. Turkkan N, Pham-Gia T. Computation of the highest posterior density interval in bayesian analysis. J Stat Comput Simul. 1993;44(3-4):243-50.

19. Ferentinos KK, Karakostas KX. More on Shortest and Equal Tails Confidence Intervals. Commun Stat - Theory Methods. 2006;35(5):821-9.

20. Laud PJ. Equal-tailed confidence intervals for comparison of rates. Pharm Stat. 2017;16(5):334-48.

21. Zhou X-H, Gao S. One-Sided Confidence Intervals for Means of Positively Skewed Distributions. Am Stat. 2000;54(2):100-4.

22. Angus JE. Bootstrap One-Sided Confidence Intervals for the Log-Normal Mean. J R Stat Soc Ser Stat. 1994;43(3):395-401.

23. Tony Cai T. One-sided confidence intervals in discrete distributions. J Stat Plan Inference. 2005;131(1):63-88.

24. HALL P. Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters. Biometrika. 1982;69(3):647-52.

25. Kaiser HF. Directional statistical decisions. Psychol Rev. 1960;67(3):160-7.

26. Cho H-C, Abe S. Is two-tailed testing for directional research hypotheses tests

legitimate? J Bus Res. 2013;66(9):1261-6.

27. Kinoshita M, Hashimoto N, Izumoto S, Okita Y, Kagawa N, Maruno M, et al. Immunohistological profiling by B-cell differentiation status of primary central nervous system lymphoma treated by high-dose methotrexate chemotherapy. J Neurooncol. 2010;99(1):95-101.

28. Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Modified Kolmogorov-Smirnov Test Procedures with Application to Arbitrarily Right-Censored Data. Biometrics. 1980;36(4):607-25.

29. Freedman LS. An analysis of the controversy over classical one-sided tests. Clin Trials. 2008;5(6):635-40.

30. Lombardi CM, Hurlbert SH. Misprescription and misuse of one-tailed tests. Austral Ecol. 2009;34(4):447-68.

31. Lin X, Wang H. A New Testing Approach for Comparing the Overall Homogeneity of Survival Curves. Biom J. 2004;46(5):489-96.

32. Yang S, Prentice R. Improved Logrank-Type Tests for Survival Data Using Adaptive Weights. Biometrics. 2010;66(1):30-8.

33. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. PLOS ONE. 2015;10(1):e0116774.

34. Bonett DG, Wright TA. Comments and recommendations regarding the hypothesis testing controversy. J Organ Behav. 2007;28(6):647-59.

35. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70(2):129-33.

36. Bain Barbara A., Dollaghan Christine A. The Notion of Clinically Significant Change. Lang Speech Hear Serv Sch. 1991;22(4):264-70.

37. Goeman JJ, Solari A, Stijnen T. Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. Stat Med. 2010;29(20):2117-25.

38. Woaye-Hune P, Hardouin J-B, Lehur P-A, Meurette G, Vanier A. Practical issues encountered while determining Minimal Clinically Important Difference in Patient-Reported Outcomes. Health Qual Life Outcomes. 2020;18(1):156.

39. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404-13.

40. Agresti A, Coull BA. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. Am Stat. 1998;52(2):119-26.

41. Stevens WL. Fiducial Limits of the Parameter of a Discontinuous Distribution. Biometrika. 1950;37(1/2):117-29.

42. Hirji KF, Mehta CR, Patel NR. Computing Distributions for Exact Logistic Regression. J Am Stat Assoc. 1987;82(400):1110-7.

43. Wang H. Exact average coverage probabilities and confidence coefficients of confidence intervals for discrete distributions. Stat Comput. 2009;19(2):139-48.

44. Die Loucou J, Pagès P-B, Falcoz P-E, Thomas P-A, Rivera C, Brouchet L, et al. Validation and update of the thoracic surgery scoring system (Thoracoscore) risk model. Eur J Cardiothorac Surg. 2020;58(2):350-6.

45. Buse A. The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. Am Stat. 1982;36(3a):153-7.

46. Fay MP. Confidence intervals that match Fisher's exact or Blaker's exact tests. Biostatistics. 2010;11(2):373-4.

47. Gillibert A, Bénichou J, Falissard B. The case for balanced hypothesis tests and equal-tailed confidence intervals. ArXiv210312581 Stat [Internet]. 2021 [cité 24 mars 2021]; Disponible sur: http://arxiv.org/abs/2103.12581

48. Gehan EA. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika. 1965;52(1/2):203.

49. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966;50(3):163-70.

50. Cox DR. Regression Models and Life-Tables. J R Stat Soc Ser B Methodol. 1972;34(2):187-202.

51. Peto R. Rank tests of maximal power against Lehmann-type alternatives. Biometrika. 1972;59(2):472-5.

52. Lee ET, Desu MM, Gehan EA. A Monte Carlo study of the power of some two-sample tests. Biometrika. 1975;62(2):425-32.

53. Fleming TR, Harrington DP. 7.5 Some Versatile Test Procedures. In: Counting Processes and Survival Analysis. John Wiley & Sons; 2011. p. 277-84.

54. Pepe MS, Fleming TR. Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data. Biometrics. 1989;45(2):497.

55. Shen Y, Cai J. Maximum of the Weighted Kaplan-Meier Tests with Application to Cancer Prevention and Screening Trials. Biometrics. 2001;57(3):837-43.

56. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. Biometrika. 1977;64(1):156-60.

57. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond Ser Contain Pap Math Phys Character. 1933;231(694-706):289-337.

58. Bakan D. The test of significance in psychological research. Psychol Bull. 1966;66(6):423.

59. Gillibert A, Bénichou J, Falissard B. Two-sided confidence interval of a binomial proportion: how to choose? ArXiv210310463 Stat [Internet]. 2021 [cité 22 mars 2021]; Disponible sur: http://arxiv.org/abs/2103.10463

60. Berry G, Armitage P. Mid-p Confidence Intervals: A Brief Review. J R Stat Soc Ser

Stat. 1995;44(4):417-23.

61. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209-12.

62. Gillibert A, Bénichou J, Falissard B. Best estimator for bivariate Poisson regression. ArXiv210310365 Stat [Internet]. 2021 [cité 22 mars 2021]; Disponible sur: http://arxiv.org/abs/2103.10365

63. Jelti L, Cordel N, Gillibert A, Lacour J-P, Uthurriague C, Doutre M-S, et al. Incidence and Mortality of Pemphigus in France. J Invest Dermatol. 2019;139(2):469-73.

64. Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights. Am J Epidemiol. 2019;188(1):250-7.

8 Annexe 1 : article N°1

The case for balanced hypothesis tests and equal-tailed confidence intervals

André GILLIBERT[ab]*†, Jacques BÉNICHOU[bc] and Bruno FALISSARD[a]

[a] INSERM UMR 1178, Université Paris Sud, Maison de Solenn, Paris, France.

[b] Department of Biostatistics and Clinical Research, CHU Rouen, Rouen, F 76031, France

[c] Inserm U 1181, Normandie University, Rouen, France

* Correspondence to: André GILLIBERT, Department of Biostatistics and Clinical Research, CHU Rouen, Rouen, F 76031, France

†E-mail: andre.gillibert@chu-rouen.fr

8.1 Abstract

Introduction: there is an ongoing debate about directional inference of two-sided hypothesis tests for which some authors argue that rejecting $\theta = \theta_0$ does not allow to conclude that $\theta > \theta_0$ or $\theta < \theta_0$ but only that $\theta \neq \theta_0$, while others argue that this is a minor error without practical consequence.

Discussion: new elements are brought to the debate. It is shown that the directional interpretation of some non-directional hypothesis tests about Receiver Operating Characteristic (ROC) and survival curves may lead to inflated type III error rates with a probability of concluding that a difference exists in the opposite side of the actual difference that can reach 50% in the worst case. Some of the issues of directional tests also apply to two-sided confidence intervals (CIs). It is shown that equal-tailed CIs should be preferred to shortest CIs. New assessment criteria of two-sided CIs and hypothesis tests are proposed to provide a reliable directional interpretation: partial left-sided and right-sided α error rates for hypothesis tests, probabilities of overestimation and underestimation α_L and α_U and interval half-widths for two-sided CIs.

Conclusion: two-sided CIs and two-sided tests are interpreted directionally. This implies that directional interpretation be taken in account in the development and evaluation of confidence intervals and tests.

8.2 Introduction

Hypothesis tests may be performed on a statistic θ, such as the effect of a treatment or an exposition. For a prespecified θ_0, a two-sided hypothesis test is typically based on the rejection of the null hypothesis that $\theta = \theta_0$ and allows the conclusion of the alternative hypothesis that $\theta \neq \theta_0$.

Uncertainty about an estimation of the statistic θ may also be expressed as confidence intervals (CIs). The most used CIs are two-sided with a confidence level set at $1 - \alpha$, typically 95%. Two-sided CIs procedures are designed to guarantee that the probability $\Pr(\theta \in CI)$ that the CI contain the statistic θ be as close as possible as the confidence level $1 - \alpha$.

Although hypothesis tests and CIs have different definitions, they are related. For a given hypothesis test procedure, the set of θ_0 values that are not rejected by the hypothesis test is a CI. For a given CI, the hypothesis that $\theta = \theta_0$ can be rejected when the CI does not contain θ_0.

There is an ongoing debate about whether two-sided hypothesis tests should be used or replaced by one-sided, directional or three-sided hypothesis tests [1–4]. The literature about directional inference on CIs is scarce but similar theoretical considerations apply.

First, we will summarize the arguments for and against the use of two-sided tests. Second, we will see that proponents of both positions agree on how statistics should be interpreted but they disagree on the severity of the consequences of the use of two-sided tests that were not designed for directional inference. Third, we will show that the consequences of improper interpretation of two-sided inference can be more severe that had been previously reported. Fourth, we will show that these considerations should be taken in account when designing and assessing new and old statistical tools. Fifth, we will draw a parallel between hypothesis tests and CIs to transfer hypothesis tests considerations to CIs.

8.3 Arguments for and against two-sided hypothesis tests

Kaiser [5] argues that rejecting the hypothesis $\theta = \theta_0$ should lead to the conclusion that $\theta \neq \theta_0$ but not that $\theta > \theta_0$ or that $\theta < \theta_0$ according to the sign of the observed difference [5]. He proposes a directional theoretical framework to permit directional conclusions, with three hypotheses:

H1 : $\theta < \theta_0$

H2 : $\theta = \theta_0$

H3 : $\theta > \theta_0$

This allows Kaiser to define the error γ (type III error) of concluding about the existence of an effect in the opposite direction of the actual effect. In the classical two-sided framework, this error does not exist, because rejecting $\theta = \theta_0$ with $\hat{\theta} < \theta_0$ (observed value less than the test target) while $\theta > \theta_0$ (real value greater than the test target) is a correct conclusion rather than an error.

John E. Overall argues that in regulated fields, such as drugs, the risk of error is taken only on one side [6]. Showing the superiority of the treatment to the placebo while it is equal or inferior may lead to the marketing of an inefficacious or harmful treatment (right-sided type I error). Showing that the treatment is significantly inferior to the placebo leads to the same conclusion than failing to show its superiority: in both cases, the treatment will not be marketed. Therefore, this risk should add to the type II error rate

rather than the type I error rate.

Lombardi and Hurlbert [2] argue that one-sided tests are unable to detect effects opposite to what was expected. They argue that performing a second one-sided test in the opposite direction is possible but doubles the α error rate. Indeed, a balanced two-sided test with $\alpha = 0.05$ is equivalent to two one-sided tests with $\alpha = 0.025$. Lombardi and Hurlbert argue that Kaiser directional two-sided theory (H1, H2, H3) is practically equivalent to the classical two-sided theory (H0: $\theta = \theta_0$ vs H1: $\theta \neq \theta_0$) with directional conclusions based on whether $\hat{\theta} > \theta_0$ or $\hat{\theta} < \theta_0$ and that γ (type III) errors are negligible.

Ludbrook [7] and Koch [8] suggest choosing the test on the *a priori* scientific hypothesis. The test alternative hypothesis is the scientific hypothesis performed before collecting data of the study and should be one-sided when a hypothesis is done in a specific direction. When no specific hypothesis is done, Ludbrook suggests that the test should be two-sided.

Ruxton and Neuhäuser [1] support two-sided tests but assume that their interpretation will be directional. They notice that the type I error may not split equally in two $\alpha/2$ errors but do not see that as a problem.

Freedman [9] argues that, even if two one-sided tests are performed (one in each direction), the actual risk taken is not doubled compared to a single one-sided test. This can be understood with a simple example. If a comparison of two population means, μ_X and μ_Y must be performed with a single observation in each sample group, the one-sided test concluding that $\mu_X > \mu_Y$ if $m_X > m_Y$ has a one-sided type I error rate at 50%. It means that, if $\mu_X \leq \mu_Y$ (H0), in the worst case ($\mu_X = \mu_Y$) one out of two experiments will conclude that $\mu_X > \mu_Y$. A non-directional two-sided test concluding that $\mu_X \neq \mu_Y$ whenever $m_X \neq m_Y$ has a type I error rate equal to 100% if the distribution is continuous. A directional test concluding that $\mu_X > \mu_Y$ when $m_X > m_Y$ and that $\mu_X < \mu_Y$ when $m_X < m_Y$ has a type III error rate equal to 50% if one assumes that $\mu_X \neq \mu_Y$. Indeed, in the worst-case scenario where $\mu_X \approx \mu_Y$ but $\mu_X \neq \mu_Y$, the conclusion will be correct half of the time and wrong half of the time. Tukey argues that only the sign of the difference $\mu_X - \mu_Y$ is unknown and that the perfect equality $\mu_X = \mu_Y$ is never plausible [10]. The case where the difference $|\mu_X - \mu_Y|$ is negligible (equivalence) is another debate [4].

8.4 Summary of arguments

There is general agreement that conclusions should be directional, either with pre-specification of the direction (one-sided test), or with an *a posteriori* specification of the direction of the effect, according to the observed effect. The questions are about how to formulate the null and alternative hypotheses, whether to perform a classical two-sided test, one or two one-sided tests or more elaborate tests, and what type I error rate should be taken in each direction. As Freedman [9] showed, there is no need of summing the two one-sided error rates when performing a two-sided test, since the risk is only taken in the direction of the conclusion. Each error rate should be controlled separately, typically set at 2.5% each.

Until now, the debate focused on how tests should be used and interpreted, but not how they should be conceived and assessed. It was also assumed that the use of unbalanced two-sided test may, at worst, double the one-sided type I error rate, from $\alpha/2$ to α. We will show that interpreting some two-sided hypothesis tests directionally can raise the one-sided α and even γ error rates to 50%. We will show that some statistical tools have been developed to gain statistical power or precision while, at the same time, making directional inference impossible or highly biased.

8.5 Directional interpretations with risk of error up to 50%

8.5.1 Venkatraman's test

Venkatraman's test [11] compares receiver operating characteristic (ROC) curves. The null hypothesis of this test is that the two ROC curves are perfectly superimposed, which implies that the two areas under curves (AUC) are equal. It is based on a statistic summing absolute values of differences between the two ROC curves. When the two curves cross, it may have a high power to detect that they are not superimposed but gives no indication on which one has the greater AUC. Figure 1 shows crossing ROC curves from a simulated dataset, based on shifted log-normal distributions, where AUCs are close to each other. The dataset contains 200 observations, 100 with positive diagnostic and 100 with negative diagnostic. The real AUCs are 0.763, for the solid curve and 0.759 for the dashed curve, but the observed difference of AUCs is opposite to the real difference. Venkatraman's test is significant at the 5%

threshold (p=0.03). A naïve interpretation of Venkatraman's test as a directional two-sided test comparing AUCs leads to a type III error, *i.e.* concluding that a difference exists in the opposite of the real direction. With 1000 Monte Carlo simulations the type III error rate (γ risk) was estimated at 44%, while the statistical power was estimated at 50% and type II error rate (β risk) at 6%. Since Venkatraman's test compares ROC curves rather than AUCs, it could be argued that this is a misinterpretation of Venkatraman's test. That is true, but how frequent is this misinterpretation?

A review of the 133 documents citing Venkatraman's article found by Google® Scholar on March 18, 2021, show that Venkatraman's test is mostly interpreted as a directional test comparing AUCs of ROC curves. After exclusion of duplicates (n=8), methodological articles (n=29), documents not published in scientific journals (n=31) and studies not actually using Venkatraman's test on real-life data (n=19), 46 articles were analyzed. A total of 32 (69.2%) articles improperly interpreted at least one significant Venkatraman's test as evidence of a higher actual AUC for the measurement having a higher observed AUC, 7 (15.2%) interpreted non-significant differences as evidence of the equivalence of AUCs (n=4) or ROC curves (n=3) and 7 (15.2%) described AUCs (n=5) or ROC curves (n=2) as non-significantly different without further precision. In one study [12] the probability of a type III error was very high (~50%). Salcedo *et al* assessed the diagnostic AUC for 63 patients *vs* 622, with a caregiver symptom severity scale (CSS) and a caregiver symptom count scale (CSC). The CSS and CSC ROC curves crossed (see Figure 1 of Salcedo *et al* [12]). The CSS and CSC AUCs were respectively estimated at 0.79 (95% CI: 0.75 to 0.83) and 0.78 (95% CI: 0.74 to 0.83). Due to ROC curves crossing, the Venkatraman's test was significant (p=0.001) and authors concluded that the CSS outperformed the CSC. Due to the construction of CSC and CSS, a high covariance between the two scales is expected, but the observed AUC difference is so small (0.01) that it is probably smaller than the random error. It is possible that the CSC AUC is actually slightly higher than the CSS even though the opposite has been observed on the sample, leading to a type III error.

In order to identify the planned usage of Venkatraman's test, the example provided by Venkatraman in his original publication was analyzed [11]. The difference of ROC curves being non-significant in the example, Venkatraman concluded that "*Both the tests do not reject the hypothesis that the ROC curves*

are equal, suggesting that the effect of volume is the same for both dose levels." Therefore, this test is interpreted as an equivalence test by failed rejection of the null hypothesis, which is known to be an inappropriate interpretation [13].

8.5.2 Survival curves crossing

When Cox's proportional hazards assumption is violated, log-rank tests and Cox models are not recommended. Li *et al* reviewed 20 tests of comparison of survival curves having no proportional hazard assumptions [14], published in 14 articles; several tests with different tuning parameters could be analyzed for the same articles. All these tests were designed to reject the null hypothesis that the curves are perfectly superimposed. Fourteen (70%) tests published in ten articles gave no indication as to which curve is the "best" [15–24]. Indeed, the absolute values of differences or squared differences were summed or a maximum of differences between curves or hazard functions was calculated, making it impossible to identify the direction and position of the difference. Six (30%) tests published in four articles gave different weights to early and late events [25–28]; all were based on the log-rank or Kaplan-Meier with modified weights. Weights of Gehan-Breslow and Tarone-Ware are higher for early events than for late events. Although it may help to reject the null hypothesis of superimposed curves, it tends to favor the survival curve having a poor long-term outcome. Figure 2 shows survival curves crossing. Gehan-Breslow and Tarone-Ware when interpreted as directional tests from their Z statistic, lead to the conclusion that the curve with poor long-term survival (blue solid curve) has better overall survival since early events are much more weighted than late events. The log-rank test gives less weight to early events and concludes that the curve with good long-term survival (dashed red curve) has better survival. Therefore, tests designed to compare curves that cross, are either unable to provide a direction of the difference or tend to favor the curve with poor life expectancy. This is probably because most authors only assessed the statistical power to reject the null hypothesis of perfect superimposition of curves without directional interpretation in mind.

We performed a review of the literature that used the Gehan-Breslow test, with the keyword "Gehan[TW]" on PubMed, including all articles published from January 2000 up to March, 19th 2021.

A total of 202 references were found. After exclusion of methodological articles (n=24), articles that did not apply any Gehan-Breslow test on any real-life dataset (n=36), and one article for which the full text could not be retrieved, the interpretation of the test was analyzed in 141 articles. A total of 68 of 141 (48.2%) articles interpreted the test as a comparison of an unrelated survival statistic, such as survival rate at a specific time (n=26), survival rate at time of last follow-up (n=6), median of survival (n=23), mean of survival (n=9) or hazard ratio (n=4). This was identified by sentences such as "*There was a median 48-month longer survival in patients with carboplatin HSR receiving carboplatin desensitization when compared to patients without carboplatin HSR (p = 0.0094)*". A total of 25 of 141 (17.7%) articles had a graphical interpretation of the difference suggested by sentences such as "*Kaplan-Meier survival curves showed significantly worse patient survival (figure 1) (p =0.004) and graft survival (figure 2) (p 0.02) for low T*". A total of 10 of 141 (7.1%) articles had a directional interpretation of the comparison without any specification of the statistic compared, with non-specific terms such as "*better survival*" or "*worse survival*" while 26 of 141 (18.4%) articles had a non-specific non-directional interpretation of the statistic such as "*When comparing the survival rates related to the alloy used, the Gehan-Wilcoxon test showed no significant differences*" and 8 of 141 (5.7%) provided no interpretation or an unclear interpretation, often due to the use of several different P-values (log-rank, Gehan-Breslow, etc.) or presentation of different statistics (median survival time, crude survival rate, etc.) in the same sentence. Finally, 4 of 141 (2.8%) articles interpreted the test as a comparison of early survival with sentences such as "*statistically significant delay of TRAMP-PSA tumor growth at early time points (Gehan-Breslow test, p = 0.002)*".

Articles interpreting the Gehan-Breslow test as a comparison of an unrelated survival statistic (*e.g.* median survival time) may, in the worst case have a type III error rate (γ risk) close to 50%. Indeed, if medians are almost equal in two groups, the observed difference of medians would be randomly positive or negative in half of the experiments, although the Gehan-Breslow test may be almost always significant on a large sample.

Authors may not be taken responsible of misinterpretations of the tests they developed. However, some of them fooled themselves. For instance, Fleming *et al* [24] developed a modified Kolmogorov-Smirnov

test for comparison of survival curves, that do not provide any indication on the direction of the difference. They provided an example from a real-life clinical trial in bile duct cancer and wanted to "*test whether patients treated with RöRx+5-FU would survive longer than control patients*" with their own procedure. Their test was able to show that curves were not superimposed but provided a directional interpretation of their test. Shen and Cai also provided a directional interpretation of a non-directional test in the example illustrating their article [20]. The absence of proper directional interpretation is less obvious for the test of Shen and Cai than for other tests because it can be used as a "one-sided" test trying to show evidence that a curve is better than the other, by choosing an optimal weight function, either favoring early or late events, to prove the hypothesis in that specific direction. Unfortunately, sampling fluctuations are computed under the null hypothesis of superimposed curves so that in cases of blatantly crossing curves, the test could show simultaneously the one-sided superiority of the first curve to the second (*e.g.* by choosing heavy early weights) and the superiority of the second curve to the first (*e.g.* by choosing heavy late weights). Indeed, according to the direction indicated by the statistician, the weights are automatically chosen by the procedure to prove what he wants.

From the 10 articles about tests with no possible directional interpretation cited in the systematic review of Li *et al* [14], 8 (80%) gave examples about trials in human or animal [16–21,23,24] for which a directional interpretation is expected, 1 (10%) [22] gave an example about a prognostic study for which a directional interpretation is expected and 1 (10%) [15] gave no example. From these 10 articles, only the two articles mentioned in the previous paragraph [20,24] gave a directional interpretation of their test in the example, all other concluded only that the curves significantly differed.

8.6 Consequences of inadequacy between theory and practice

The theory is often based on a non-directional null hypothesis while in practice, tests are interpreted directionally. In the next paragraphs, the consequences of this inadequacy between theory and practice are analyzed.

8.6.1 Development of tests without proper directional interpretation

Proponents of two-sided tests agree that the interpretation of two-sided tests is directional. This is

confirmed by the literature review about articles using Venkatraman's and Gehan-Breslow's tests. This directional interpretation is not taken in account in many articles developing new statistical methods, especially for the problem of comparison of survival curves that cross. Due to an inappropriate null hypothesis, authors may ignore the fact that the main problem of crossing survival curves is not to prove that curves differ but to define a relevant criterion to identify which one is the "best". Authors developing new methods can fool themselves and fool others, as the literature reviews in the previous sections has shown.

8.6.2 Improper assessment of tests properties

Outside of survival or ROC curves analyses, even with simple one-dimensional statistics, ignoring the directional interpretation of two-sided tests may lead to an inappropriate assessment of the properties of the statistical test. A two-sided hypothesis test is designed to control the α risk of concluding $\theta \neq \theta_0$ when $\theta = \theta_0$, but may not control separately the one-sided risks of concluding that $\theta > \theta_0$ and that $\theta < \theta_0$ when $\theta = \theta_0$. In the worst case, the two-sided test may be equivalent to a one-sided test in an unknown direction, with all the α risk on one side. This would lead to a doubling of the α risk to prove one alternative hypothesis (*e.g.* $\theta > \theta_0$) and major loss of power to prove the other alternative hypothesis (respectively $\theta < \theta_0$). For instance, a test used to assess the superiority of a treatment to another may either have a very low statistical power to show the superiority or a doubled α risk, because of the unpredictable imbalance of the two-sided test. This could be especially unfair when there is no reason to favor one treatment compared to the other, such as a comparison of the medical to the surgical treatment of lumbar disc herniation. An imbalanced test could unfairly favor one treatment. Therefore, since two-sided tests are used directionally, they should be assessed with that use in mind, with a balance of the α risk, separated in two partial left-sided and right-sided risks controlled at $\alpha/2$. A two-sided test should be equivalent to two one-sided tests in opposite directions. As Freedman pointed out [9], the actual risk taken when performing two one-sided tests, is the one-sided risk, hence $\alpha/2$ when performing a balanced two-sided hypothesis test. As many hypothesis tests are quite balanced (*e.g.* Wald's tests of coefficients in generalized linear models with large sample sizes), most analyses described in the literature as two-sided with directional conclusions, actually take an $\alpha/2$ (usually 2.5%) risk.

Unfortunately, the two-sided theoretical framework encourages authors of new tests to assess the global two-sided α risk without control of its balance between the left and right side of the alternative hypothesis.

8.7 Considerations for confidence intervals

A two-sided CI estimator, at confidence level $1-\alpha$, must be built so that the frequentist probability that the CI contains the actual parameter is $1-\alpha$ and the probability that the CI does not contain the actual parameter is α. This α is the sum of the overestimation risk (α_L), the probability that the CI is above θ, and the underestimation risk (α_U), the probability that the CI is below θ. Two-sided CI estimators may be equal tailed, so that the underestimation and overestimation risks are equal or almost equal (*i.e.* $\alpha_U \approx \alpha_L \approx \alpha/2$). Unequal-tailed estimators may control the overall coverage but may have very different underestimation (*e.g.* 0.005) and overestimation (*e.g.* 0.045) risks. Unequal tailed CIs may be intentionally built to shorten the CI, such as Zieliński's interval [29]. As Zieliński show, a shortest CI may be built from an equal-tailed CI by moving both CI bounds upwards or downwards, increasing one risk (overestimation or underestimation) and decreasing the other (respectively underestimation or overestimation). Equal-tailed CI may be interpreted as the intersection of two one-sided CI while unequal-tailed CI bounds should not be interpreted separately since the actual risk associated to each bound is not well controlled.

In a phase III drug randomized controlled trial, the treatment effect may be estimated by CI. When filling a new drug application, the treatment effect may be claimed to be equal or greater than the lower bound of the confidence interval. The actual error rate is the one-sided error associated to the lower bound (α_L). For an equal tailed CI, this risk is $\alpha/2$, usually 2.5%. For an unequal-tailed CI, the actual error rate is unknown, between 0 and α, and so, may be up to 5% for a 95% CI. Similarly, claiming that the frequency of adverse effects is less than the upper boundary of a two-sided CI has an error rate that is not well controlled by unequal-tailed CIs. Overestimation and underestimation are not equivalent, and one error rate cannot be traded for the other. Moving down the lower bound of the CI (decreasing the probability of overestimation) should not allow moving down the upper bound (increasing the probability of

underestimation) since both movements tend to favor the hypothesis of low frequency of adverse effects. Directional interpretations are done in all clinical and epidemiological contexts: to claim that an epidemiological exposure is dangerous because its effect is higher than the lower bound of the 95% CI or to claim that the sensitivity of a screening test is above the lower bound of the 95% CI. Therefore, CIs should be developed and assessed with these directional interpretations in mind. Equal-tailed CIs should be preferred over shortest CIs that may, in the worst case, be one-sided CIs in unpredictable directions. The probabilities of overestimation (α_L) and underestimation (α_U) should be both assessed when assessing CIs, rather than their sum $\alpha_L + \alpha_U = \alpha$. The argument for the use of shortest CIs is that they have better statistical precision, but this is not necessarily true when interpreting CIs directionally. When assessing the efficacy of a treatment, the expectancy of the lower bound should be as high as possible while the risk of overestimation is kept controlled to $\alpha/2$. That would guarantee the maximal statistical power and chance of concluding that the efficacy is greater than any predefined threshold. The expectancy of the lower bound can be indirectly assessed by the CI half-width, equal to the expectancy of the difference between the point estimate and the lower bound of the CI. A shortest CI, constructed by moving both bounds of an equal-tailed CI in the same direction, such as Zieliński's interval [29] or the highest density probability Bayesian credible intervals with a non-informative prior [30], may increase a half-width (*e.g.* Left half-width) while decreasing the other (respectively right half-width) at the same time, in an unpredictable way. This may sometimes reduce the statistical power when the CI is interpreted as a hypothesis test. Therefore, rather than assessing the total width of CIs, we suggest that the two half-widths of CIs, with their associated α_L and α_U risks be assessed when analyzing the statistical performances of CIs. This is equivalent to analyzing a two-sided CI as the intersection of two one-sided CIs.

8.8 Conclusion

Two-sided hypothesis tests and two-sided CIs are mostly used with a directional interpretation, although the theoretical framework on which they are based does not allow this interpretation. This gap between theory and practice has no practical consequences for many common tests, such as the likelihood ratio

test for generalized linear models, but can sometimes lead to severely biased inference, such as for tests designed to compare survival or ROC curves that cross.

Entirely changing the biomedical research practice is not easy, especially because the most common tools have no major bias. However, the development of new statistical tools and assessment of existing tools should be performed with the directional interpretation in mind. We propose to assess separately, the partial left-sided and right-sided type I error rates for hypothesis tests, to guarantee their balance. We propose to assess separately the underestimation and overestimation risk of CIs as well as left and right half-width of the CIs. When developing new tests that cannot have a directional interpretation, we suggest to keep in mind the risk that it may be misinterpreted.

8.9 Supplementary Material

Detailed information about articles included in the literature reviews of the Venkatraman and the Gehan-Breslow tests is available on the Open Science Framework at https://osf.io/zds8b/

8.10 References

1. Ruxton GD, Neuhäuser M. When should we use one-tailed hypothesis testing? Methods in Ecology and Evolution. 2010;1(2):114–7.
2. Lombardi CM, Hurlbert SH. Misprescription and misuse of one-tailed tests. Austral Ecology. 2009;34(4):447–68.
3. Pradhan V, Evans JC, Banerjee T. Binomial confidence intervals for testing non-inferiority or superiority: a practitioner's dilemma. Stat Methods Med Res. 2016 Aug;25(4):1707–17.
4. Goeman JJ, Solari A, Stijnen T. Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. Statistics in Medicine. 2010;29(20):2117–25.
5. Kaiser HF. Directional statistical decisions. Psychological Review. 1960;67(3):160–7.
6. Overall JE. Tests of one-sided versus two-sided hypotheses in placebo-controlled clinical trials. Neuropsychopharmacology. 1990 Aug;3(4):233–5.

7. Ludbrook J. Should we use one-sided or two-sided P values in tests of significance? Clin Exp Pharmacol Physiol. 2013 Jun;40(6):357–61.
8. Koch GG. One-sided and two-sided tests and ρ values. Journal of biopharmaceutical statistics. 1991;1(1):161–70.
9. Freedman LS. An analysis of the controversy over classical one-sided tests. Clin Trials. 2008;5(6):635–40.
10. Tukey JW. The Philosophy of Multiple Comparisons. Statist Sci. 1991 Feb;6(1):100–16.
11. Venkatraman ES. A permutation test to compare receiver operating characteristic curves. Biometrics. 2000;56(4):1134–8.
12. Salcedo S, Chen Y-L, Youngstrom EA, Fristad MA, Gadow KD, Horwitz SM, et al. Diagnostic efficiency of the Child and Adolescent Symptom Inventory (CASI-4R) Depression Subscale for identifying youth mood disorders. Journal of Clinical Child & Adolescent Psychology. 2018;47(5):832–46.
13. Harris AH, Fernandes-Taylor S, Giori N. “Not statistically different” does not necessarily mean “the same”: The important but underappreciated distinction between difference and equivalence studies. JBJS. 2012;94(5):e29.
14. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. PLOS ONE. 2015 Jan 23;10(1):e0116774.
15. Gill RD. 5.4 Rényi-type tests. In: Censoring and stochastic integrals. Amsterdam: Mathematisch Centrum; 1980. p. 135–8.
16. Schumacher M. Two-Sample Tests of Cramér--von Mises- and Kolmogorov--Smirnov-Type for Randomly Censored Data. International Statistical Review / Revue Internationale de Statistique. 1984 Dec;52(3):263.
17. Kraus D. Adaptive Neyman’s smooth tests of homogeneity of two samples of survival data. Journal of Statistical Planning and Inference. 2009 Oct;139(10):3559–69.

18. Lin X, Xu Q. A new method for the comparison of survival distributions. Pharmaceut Statist. 2010 Jan;9(1):67–76.

19. Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. J Royal Statistical Soc B. 2007 Nov 2;0(0):071103032514002-???

20. Shen Y, Cai J. Maximum of the Weighted Kaplan-Meier Tests with Application to Cancer Prevention and Screening Trials. Biometrics. 2001 Sep;57(3):837–43.

21. Koziol J. A Two Sample CRAMÉR-VON MISES Test for Randomly Censored Data. Biom J. 1978;20(6):603–8.

22. Lee S-H. On the versatility of the combination of the weighted log-rank statistics. Computational Statistics & Data Analysis. 2007 Aug;51(12):6557–64.

23. Lin X, Wang H. A New Testing Approach for Comparing the Overall Homogeneity of Survival Curves. Biometrical Journal. 2004;46(5):489–96.

24. Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Modified Kolmogorov-Smirnov Test Procedures with Application to Arbitrarily Right-Censored Data. Biometrics. 1980 Dec;36(4):607.

25. Tarone RE, Ware J. On Distribution-Free Tests for Equality of Survival Distributions. Biometrika. 1977;64(1):156–60.

26. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics. 1989 Jun;45(2):497–507.

27. Fleming TR, Harrington DP. 7.5 Some Versatile Test Procedures. In: Counting Processes and Survival Analysis. John Wiley & Sons; 2011. p. 277–84.

28. Gehan EA. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika. 1965 Jun;52(1/2):203.

29. Zieliński W. The Shortest Clopper–Pearson Confidence Interval for Binomial Probability. Communications in Statistics - Simulation and Computation. 2009 Dec 8;39(1):188–93.

30. Turkkan N, Pham-Gia T. Computation of the highest posterior density interval in bayesian analysis. Journal of Statistical Computation and Simulation. 1993 Jan 1;44(3–4):243–50.

8.11 Figures

Figure 1: Receiver Operating Curves (ROC) from simulated datasets for which a naïve interpretation of Venkatraman's test leads to a type III error.

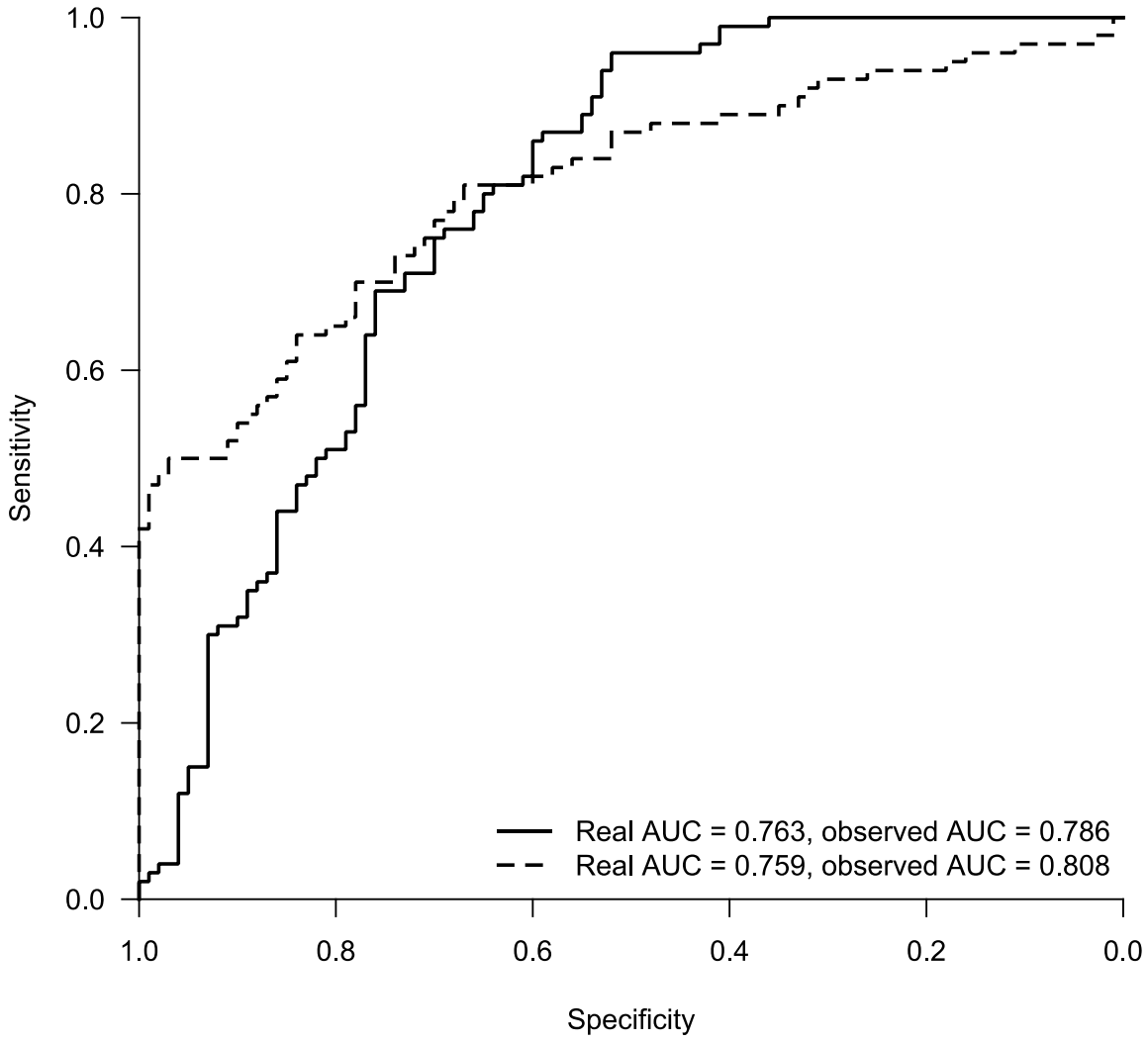
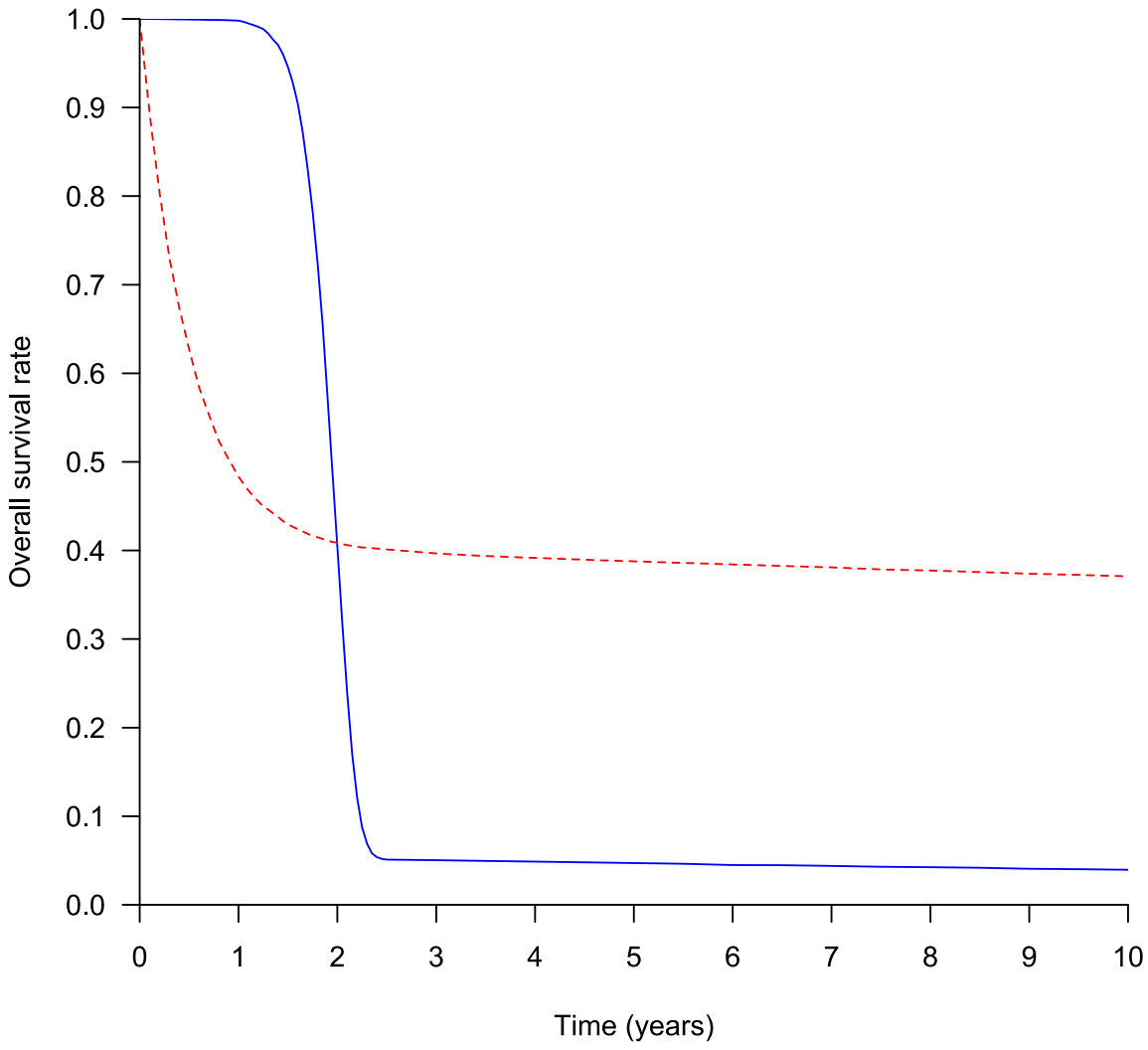


Figure 2: simulated survival curves crossing for which Gehan's test finds that the survival is better for subjects with poor long-term survival (blue solid line) than for subjects with poor short-term survival (red dashed line).



9 Annexe 2 : article N°2

Two-sided confidence interval of a binomial proportion: how to choose?

André GILLIBERT[ab]*†, Jacques BÉNICHOU[bc] and Bruno FALISSARD[a]

[a] INSERM UMR 1178, Université Paris Sud, Maison de Solenn, Paris, France.

[b] Department of Biostatistics and Clinical Research, Rouen University Hospital, Rouen, France

[c] Inserm U 1219, Normandie University, Rouen, France

* Correspondence to: André GILLIBERT, Department of Biostatistics and Clinical Research, Rouen University Hospital, Rouen, France

†E-mail: andre.gillibert@chu-rouen.fr

Abstract

Introduction: estimation of confidence intervals (CIs) of binomial proportions has been reviewed more than once but the directional interpretation, distinguishing the overestimation from the underestimation, was neglected while the sample size and theoretical proportion variances from experiment to experiment have not been formally taken in account. Herein, we define and apply new evaluation criteria, then give recommendations for the practical use of these CIs.

Materials & methods: Google® Scholar was used for bibliographic research. Evaluation criteria were (i) one-sided conditional errors, (ii) one-sided local average errors assuming a random theoretical proportion and (iii) expected half-widths of CIs.

Results: Wald's CI did not control any of the risks, even when the expected number of successes reached 32. The likelihood ratio CI had a better balance than the logistic Wald CI. The Clopper-Pearson mid-P CI controlled well one-sided local average errors whereas the simple Clopper-Pearson CI was strictly conservative on both one-sided conditional errors. The percentile and basic bootstrap CIs had the same bias order as Wald's CI whereas the studentized CIs and BC_a, modified for discrete bootstrap distributions, were less biased but not as efficient as the parametric methods. The half-widths of CIs mirrored local average errors.

Conclusion: we recommend using the Clopper-Pearson mid-P CI for the estimation of a proportion except for observed-theoretical proportion comparison under controlled experimental conditions in which the Clopper-Pearson CI may be better.

KEYWORDS: binomial confidence interval, coverage bias, equal-tailed confidence intervals, local average errors, Clopper-Pearson mid-P confidence interval.

9.1 Introduction

Estimating the confidence interval (CI) of a proportion is one of the most basic statistical problems and an everyday task for many statisticians. Most of statistical software provides two estimators, an "approximate" and an "exact" estimator. The "approximate" CI estimator is usually Wald's CI, namely

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad (1)$$

where $\hat{p}$ is the observed proportion and n is the sample size. Wald's estimator may be the best known, but has been criticized for its coverage bias deemed unacceptable [1,2]. The "exact" CI estimator is usually the Clopper-Pearson CI [3] (*e.g.*, in R, SAS, Stata). The word "exact" is misleading since no deterministic estimator has complete control over the coverage due to the binomial distribution discreteness.

The unsolvable problem of exact coverage led to the development of many estimators: Newcombe reviewed seven estimators in 1998, then Pires reviewed as many as 20 different estimators in 2008 [4] and the development of new methods was still active as of 2018 [5]. Agresti and Coull [1] argue that the Clopper-Pearson estimator is too conservative, with actual coverage always greater or equal to the nominal coverage (strict conservatism) and suggest that some approximate estimators (e.g. Wilson and Agresti-Coull) may have better control over the average coverage than an exact estimator.

Some standard criteria for assessing the validity of CI estimators have been widely used in systematic reviews, such as minimal coverage control, average coverage control and interval width [1,2,4,6–8]. In our opinion, two important issues have not been sufficiently addressed.

The first issue is the balance of one-sided errors: equal-tailed CIs or unequal-tailed CIs. This is needed for a directional interpretation of CIs where overestimation and underestimation are distinguished [9]. For instance a 95% unequal-tailed CI estimator, applied to the rate of adverse events of a treatment, has an unpredictable probability between 0 and 5% of underestimating the proportion of adverse events, another unpredictable probability between 5% and 0% of overestimating the proportion with a sum of the two probabilities equal or close to 5%. In order to prove the safety of the treatment, the actual proportion could be claimed to be less than the upper boundary of the CI of the rate of adverse events; doing so, the actual risk taken is unpredictable, between 0% and 5%. An equal-tailed 95% CI estimator

controls the probability of overestimating the actual proportion (equal or close to 2.5%) and the probability of underestimating it (equal or close to 2.5%). The actual risk taken when claiming that the rate of adverse events is less than the upper boundary would be controlled (close to 2.5%). In our opinion, directional interpretation is important and equal-tailed CIs should be preferred.

The second issue is the variability of the actual proportion and sample size from one experiment to another. Although most CI estimators are conditional to the sample size, this sample size is not actually constant from one experiment to another. Most meta-analyses show different sample sizes for all included research studies. Moreover, they show some heterogeneity of actual proportions or effects. Consequently, the actual CI coverage biases should be judged in a realistic setting where experiment replication is performed with a variable sample size, a variable proportion or both. This variability should smooth coverage oscillations described by Brown, Cai and DasGupta, attributed to discreteness of the binomial distribution [2].

The objective of this article is to systematically review binomial proportion and Poisson CI estimators and assess them with evaluation criteria taking in account the two issues mentioned above: directional interpretation and variability of the actual proportion or sample size.

9.2 Methods

9.2.1 Systematic review

A bibliographic search with the keywords “binomial”, “confidence” and “interval” was conducted on the Google® Scholar database in July 2017 (updated in October 2018), looking for articles defining CIs of a binomial proportion. Articles were sorted by relevance according to Google’s algorithm and the first 400 results were screened on their title then their summary. The references of systematic review articles were used to identify original references. The CI estimators considered as redundant will be mentioned but not presented. Estimators were considered redundant when they were very accurate approximations of other estimators and were graphically indistinguishable on error and width figures. Deterministic behavior and simple interpretation was considered a requirement. Consequently, randomized or fuzzy CIs have not been implemented. Even though some randomness is included in bootstrap CIs, they have been considered as deterministic since stability is achieved with a high number

of bootstrap samples. The exact asymptotic solution to bootstrapping has been found, so bootstrap CIs presented in Appendix 1 are perfectly deterministic.

9.2.2 Evaluation criteria

9.2.2.1 One-sided conditional errors

Conditional errors are defined as coverage errors conditional to a constant sample size and constant actual proportion p. The experiment is a draw from a random binomial distribution $\mathcal{B}(n;p)$. The conditional lower bound error is noted α'_L and is defined as the actual probability that the CI is below the actual proportion p. Similarly, we denote α'_U, the actual probability that the CI is above the actual proportion p.

9.2.2.2 One-sided local average errors

We define a random two-steps random binomial experiment. The first step is to draw an actual proportion p from a random variable P, taking in account the variability of the actual proportion from experiment to experiment. The second step is to draw an actual binomial variable x from a sample of size n following a binomial distribution $\mathcal{B}(n;p)$. Then, a CI is computed from x and n. The local average coverage is defined as the probability that the CI contains the realization p of P.

The random variable P will be modeled as a logit-normal variable such that the typical odds ratio of the actual proportion between two experiments is OR_S. That is, $\log(P) \sim \mathcal{N}(\mu;\sigma^2)$ and $\exp(\sigma) = OR_S$. For the primary analysis, OR_S will be set at 1.20 (arbitrary choice). Sensitivity analyses with $OR_S = 1.10$ and $OR_S = 1.05$ will be performed. We will denote $p_0 = E[P]$ the expected proportion and $\lambda = np_0$ the expected number of events. The α''_L risk is the actual probability that the CI is below the realization of P while α''_U is the actual probability that the CI is above the realization of P.

A sensitivity analysis will be performed with a random sample size N and a constant proportion p rather than a random proportion P and a constant sample size n. It is hypothesized that both models will equally smooth error curves.

The Poisson distribution will be seen as the asymptotic case of the binomial distribution when the sample size tends towards positive infinity.

9.2.2.3 Local average half-widths

For a number of successes x and a number of trials n, the lower and upper half-widths of the CI $[L_{1-\alpha}(x,n)\,;U_{1-\alpha}(x,n)]$ are defined as $w_L = \frac{x}{n} - L_{1-\alpha}(x,n)$ and $w_U = U_{1-\alpha}(x,n) - \frac{x}{n}$, the distances between the point estimate and the CI bounds. The local average half-widths are defined as the expected half-widths in experiments with a random theoretical proportion P, as described in the previous section, and are denoted $w_L'' = E[W_L]$ and $w_U'' = E[W_U]$. Expected half-widths will be analyzed as functions of λ, the expected number of successes.

9.2.2.4 Relative local average half-widths

Local average half-widths are highly dependent on the number of successes. This makes CI widths difficult to compare on graphical figures. In order to make comparisons easier, relative local average half-widths have been defined as the ratio between the local average half-width of the CI estimator and the local average half-width of a reference CI estimator for the same actual proportion. The Clopper-Pearson mid-P estimator has been selected as reference for its good statistical properties.

9.2.2.5 Other evaluation criteria

The following few desirable properties of CIs have been assessed (shown in Appendix 3):

1) Consistency of CI inference and p-values of hypothesis tests [10];
2) Generalizability to bivariate and multivariate models;
3) Theoretical simplicity and the existence of an analytical solution. This motivated Agresti and Coull when they defined their CI [1,11];
4) Equivariance: The consistency of the CI of successes with the CI of failures [12–14];
5) Strict monotonicity of CI bounds along x, n and α [15];
6) Deterministic procedure. Randomized CIs, based on a computer-generated random number, produce different CIs for the exact same data set. Conditional risks are smoothed by randomization, but the practical use of these CI requires a rigorous analysis difficult to apply in practice [16].

9.2.3 Graphical representations

Most interpretations will be graphical. Coverage bias will be expressed by local average errors. In order to help identify acceptable and unacceptable bias, reference lines will be drawn on graphical figures at thresholds $0.025 \times 1.50 = 0.0375$ and $0.025/1.50 = 0.016667$ for nominal one-sided risks at 0.025. These thresholds seemed relevant to article authors' but are arbitrary.

9.2.4 Validation

All analyzed CI have been implemented by the first author of this article with the R statistical software except the Blyth-Still-Casella CI, the C++ implementation by Keith Winstein (commit 850c75a35f816aa22fd6050453f1b7df2c5773c6) [17]. It has been verified to be consistent with StatXact on a few examples, in order to avoid software bugs. When an implementation was available in an R package (Hmisc, binom, exactci, DescTools, PropCIs), consistency has been verified.

In order to avoid transcription errors in formulas shown in tables of this article, the algorithms have been implemented again, from these formulas, two month later, by the same author. Once the second implementation was completed, an automatic consistency check was performed. A few transcription errors were fixed thanks to this second check. This double implementation has not been performed for the most complex CIs (Wang [18], Schilling-Doi [19] and Blyth-Still-Casella [20]) but consistency with author's original script had been verified. A third implementation has been performed by the same author, before manuscript finalization, months later, for the nine CIs presented in Table 1.

9.3 Results

9.3.1 Flow chart of CI estimators

From the systematic review 64 CI estimators were found. Four additional custom CI estimators were built by fixing problems in existing estimators (Appendix 1, two Pan 2002 modifications, a BC_a bootstrap modification and a modification of the likelihood ratio interval), leading to a total of 68 CI estimators. Thirteen estimators were excluded. Approximations of Molenaar [21], Pratt [22], Blyth's equation C improving the Molenaar approximation [22] and the Chen CI [23] were all accurate approximations of the Clopper-Pearson CI [3] and so were considered redundant. The estimator of Zhou [24] and the average

randomized inference model (ARIM) estimator of Lu [5] were redundant with Bartlett's Arc-Sine estimator while the inference model (IM) estimator of Lu [5] was redundant with the Clopper-Pearson mid-P estimator. The estimators of Crow was redundant with Blaker's [25]. Probit and logit transformations for Wald's CI were redundant. Estimators of Stevens [16], Geyer [26] and Zieliński [27] were randomized or fuzzy. The Wang 2017 [28] estimator is deterministic but relies on ranks of successes in addition to the sufficient statistics (number of successes and number of trials); it has not been analyzed. Eventually, 55 CI estimators have been extensively analyzed. Nine are presented in this article (see Table 1) and the 46 others (including boostrap CIs and closed-form skewness corrected CIs) are presented in Appendix 1.

9.3.2 Definition of CI estimators

The lower bounds $L_{1-\alpha}(x,n)$ of estimators are described in Table 1. Upper bounds of these nine CI estimators can be computed by equivariance [12], *i.e.* $U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)$. In all cases, bounds outside the $[0,1]$ interval were set to the nearest valid proportion, 0 or 1. For example, when Wald's CI had a negative lower boundary, it was set at zero. These nine CI estimators have been selected because they are widely used (Wald, Clopper-Pearson), have been recommended (modified Wilson, modified equal-tailed Jeffreys [2]), are special cases of widely used general estimators or models (Likelihood ratio, Wald logit), have good statistical properties (Bartlett Arc-Sine, Clopper-Pearson mid-P) or illustrate a specific problem (Blaker).

Wilson's CI can be obtained by inversion of chi-square tests without transformations. It has an analytical solution (root of a 2nd degree equation). The modified Wilson CI, based on a Poisson approximation for small values of k, was described by Brown, Cai and DasGupta, (see page 112) [2] but these authors did not specify the threshold x^* for $n > 100$. Since they did not analyze the behavior of the CI for $n > 100$, they did not make a recommendation (personal communication of the first author). The already sufficient convergence to the Poisson distribution made us retain the threshold $x^* = 3$ for $n > 100$.

Bartlett's Arc-Sine CI is a normal-approximation CI with a variance-stabilizing transformation. Variance stabilization has been improved from the standard Arc-Sine CI transformation, adding 0.5 success and 0.5 failures.

The Wald logit CI is the normal-approximation asymptotic CI that statistical software typically computes for intercept-only logistic regressions. It is indefinite for $x = 0$ and $x = n$. We supplement its definition by the Clopper-Pearson CI for $x = 0$ and $x = n$.

The likelihood ratio CI is defined by inverting a chi-square test on the deviance function. This CI is well defined even for $x = 0$ and $x = n$, but in order to compare its performances with those of the Wald logit CI, we applied the same Clopper-Pearson substitution for $x = 0$ and $x = n$. The unmodified likelihood ratio CI is presented in, Figures A.8 and A.10 of Appendix 1.

Jeffreys CI is an equal-tailed Bayesian credible CI with the non-informative Jeffreys prior $Beta\left(\frac{1}{2};\frac{1}{2}\right)$. Brown *et al* [2] proposed a slight modification to improve its frequentist properties. The lower bound of the CI is set at zero when the number of successes (x) is zero or one and the upper bound is set to the same exact upper Poisson boundary as that of Wilson's modified CI when the number of successes is zero.

In table 1, we defined three exact binomial CIs that can be constructed by test inversion: (i) Clopper-Pearson, (ii) Clopper-Pearson mid-P and (iii) Blaker. The first is constructed from a one-sided exact binomial test with non-strict inequality: $\Pr(X \geq x)$. The second is similar to the first but uses a "half strict" inequality: $\Pr(X \geq x) + \frac{1}{2}\Pr(X = x)$. Blaker's CI is based on a two-sided test. Blaker's test ***P***-value, for $x > 0.50$ is equal to $\Pr(X \geq x) + \Pr(X \leq y)$ where y is the largest number such as $\Pr(X \leq y) \leq \Pr(Y \leq y)$. For $x < 0.50$, the CI can be defined by equivariance.

9.3.3 Results common to all CI estimators

Local average errors (Figure 1) are smoothed in comparison to conditional errors (Figure 2) having large amplitude oscillations. For a constant expected number of successes λ, coverage bias is higher in absolute value for larger values of n. The results for $n = 2048$ are very close to those for asymptotic Poisson CIs (see Figure 2 and Figures A.4 and A.5).

Local average interval half-widths mirror local average errors (Figure 3). Where a CI is shorter than another, it has a higher local average error and where it is larger, it has lower local average error.

9.3.4 Specific CI results

Wald's CI has a high two-sided bias, even for an expected number of successes equal to 32, and is a very unbalanced unequal-tailed CI (see Figures 1-2). The right local average error α'_U tends to 1 when the expected true proportion p_0 tends to zero because Wald's CI width is zero when the number of successes x is null.

Modified Wilson's, modified Wald logit and Blaker's CIs have lower absolute bias but are not equal-tailed either. The biases of these three CIs estimators have an opposite sign to Wald's CI estimator bias, while the modified likelihood ratio CI has a small bias in the same direction as Wald's CI. The modified Wald logit CI has a lower bound local average error spike (α''_L) equal to 0.097 for $n = 2\,048$ and an expected number of successes $np_0 = \lambda = 0.11$.

Blaker's and Clopper-Pearson CIs are both conservative, slightly less so for Blaker's CI (Figures 1-2). For proportions close to zero, Blaker's CI conditional right α'_U error (Figure 2) is very close to Clopper-Pearson α'_U error but Blaker's α'_L errors can get much higher with one-sided conditional error oscillations up to 0.05 while Clopper-Pearson's CI one-sided conditional error oscillations never exceed 0.025.

Local average errors (Figure 1) with Bartlett's Arc-Sine CI, modified equal-tailed Jeffreys and Clopper-Pearson mid-P CIs are close to each other. The modified equal-tailed Jeffreys CI, for proportions close to zero, has a larger right local average half-width than Bartlett's Arc-Sine and Clopper-Pearson mid-P CIs (Figure 3) leading to a more conservative upper boundary for an expected number of successes close to 4 (Figure 1). Bartlett's Arc-Sine CI local average errors get closer to nominal than Anscombe and Freeman-Tukey CIs (see Figure A.7).

The modified likelihood ratio CI has a mild one-sided local average bias, lower than the Wald, modified Wald logit and modified Wilson CIs (Figure 1). The unmodified approximate likelihood ratio CI has a high one-sided error spike for an expected number of successes close to 2.3 (Figure A.8).

Percentile and basic bootstrap CIs (Figure A.8) have high local average biases. The basic bootstrap CI has higher biases than Wald's CI while the percentile bootstrap CI is slightly less biased than Wald's. Unmodified BC_a bootstrap is highly biased and is not equivariant; a modification of equation 3.2 of Efron [29] taking in account the discreteness of the bootstrap distribution, provides equivariance and

reduces the bias (Appendix 1). Modified BC_a bootstrap, smoothed BC_a bootstrap and studentized BC_a bootstrap CIs are all conservative in terms of local average error. Due to division by zero errors, the studentized bootstrap CI cannot be computed for $\min(x, n-x) \leq 4$ and has been replaced by the Clopper-Pearson CI.

Not all two-sided unequal-tailed CIs have the same imbalance of errors. Extreme examples are shown in Appendix 1. The Pan 2002 Wald t CI (Table A.1 and Figure A.7) for sample size $n = 2048$ and expected number of successes $np_0 = 8.01$ has $\alpha_L'' = 0.005$ and $\alpha_U'' = 0.045$ while the Rubin logit CI (Table A.3 and Figure A.8) for the same experiment has $\alpha_L'' = 0.045$ and $\alpha_U'' = 0.003$.

Other evaluation criteria are shown in Appendix 3. All the nine CIs shown in Table 1 are equivariant and deterministic. Generalizability to bivariate and multivariate models applies to the Wald, Wald logit, likelihood ratio, Clopper-Pearson and Clopper-Pearson mid-P CIs. Strict monotonicity of CI bounds along x, n and α is guaranteed for all the nine CIs except Blaker's CI, the modified equal-tailed Jeffreys CIs and Wald's CI. No analytical solution is known for the likelihood ratio, Blaker and Clopper-Pearson mid-P CIs while the six other CIs have one.

9.3.5 Continuity correction

Continuity corrections make CIs more conservative on both sides (Appendix 1, Figures A.7-9). Regarding Wilson's and Wald's CIs make them less liberal on one side (respectively left and right side) and more conservative on the other (respectively right and left side).

9.3.6 Sensitivity analyses

For a typical odds ratio between the actual proportions of two experiments $OR_S = 1.20$, graphically, local average error oscillations can be seen for an expected number of successes below 2 (Figure 1). For an actual proportion with less random fluctuation, $OR_S = 1.05$, large amplitude oscillations are graphically visible when the expected number of successes is below 8 (Appendix 1, Figure A.2), with a maximum local average one-sided error equal to 0.0381 for the Clopper-Pearson mid-P CI.

The 90% CI absolute coverage biases are larger than 95% CI biases (as seen from comparing Figure 1 with Figure A.3) but relative biases, i.e., ratios of actual to nominal errors, are smaller.

The random sample average risks α''' of a constant proportion p with random sample size N are close

to local average α'' errors for random proportion P with a constant sample size n (see Figure 1 and A.6).

9.3.7 Validity conditions of Wald's CI

Validity conditions of Wald's CI are assessed in Appendix 1 (Tables A.12-14). The simple condition $\min(x, n-x) > 40$ is enough to control the one-sided local average error for a 95% CI albeit not perfectly as the actual one-sided error may be up to 1.5 times higher (that is 0.0375) than the nominal error (0.025).

9.4 Discussion

9.4.1 Summary of main findings

Wald's CI has much higher biases than other CI estimators and is unequal-tailed. The modified Wilson CI and modified Wald logit CIs have lower local average biases but are unequal-tailed too. Wald's logit CI has higher biases than the likelihood ratio CI.

Bartlett's Arc-Sine, Clopper-Pearson mid-P and modified Jeffreys equal-tailed CIs are close to each other in terms of the local average error control and expected CI half-widths. They properly control the local average errors and are equal-tailed.

9.4.2 Recommendations

Statistical software (e.g. R, SAS, SPSS) provides two CI estimators for generalized linear models coefficients: likelihood ratio CI and Wald's CI. We showed lower coverage bias for the likelihood ratio CI than for Wald's CI. This suggests that likelihood ratio CIs may be better for logistic regressions. This is consistent with Agresti's [30] recommendation suggesting the use of the likelihood ratio CI when Wald's CI is discordant with it.

In our opinion, Bartlett's Arc-Sine, Clopper-Pearson mid-P and modified Jeffreys equal-tailed CIs have the best statistical properties and are practically equivalent. The modified Jeffreys CI is based on *ad hoc* modifications [2] and has a lower bound that is not strictly monotone with the number of successes since the lower bound is equal to zero for a number of successes equal to zero or one. Therefore, we cannot recommend the equal-tailed Jeffreys CI. The Clopper-Pearson's mid-P CI has been generalized to logistic regressions by Hirji [31], making it theoretically more attractive than Bartlett's Arc-Sine CI.

Therefore, we recommend using the Clopper-Pearson's mid-P CI in almost all scenarios.

When comparing an observed proportion to a theoretical proportion under highly controlled experimental conditions or in a strongly regulated domain such as clinical trials, we recommend the Clopper-Pearson CI. Indeed, the theoretical proportion p is not variable anymore, as it is defined in a protocol, and the sample size may be quite controlled; oscillations are not smoothed anymore and the Clopper-Pearson CI (strictly conservative) may be safer than the Clopper-Pearson mid-P CI (control local average errors but not conditional errors).

9.4.3 Originality of this work

Newcombe distinguished mesial (the CI bound nearest to 0.50) and distal (the CI bound nearest to 0 or 1) one-sided errors [7] but averaged them over the whole $]0\,,1[$ interval assuming a uniformly distributed random proportion. Agresti and Coull [1] analyzed a random proportion following a beta distribution with 0.10 expectancy, which is similar to the local average error for an expected true proportion $p_0 = 0.10$. However, they did not analyze other theoretical random proportions or one-sided errors. To our knowledge, the influence of the sample size for a fixed expected number of successes has not been graphically presented in any systematic review of binomial proportion CIs. These new evaluation criteria, including one-sided errors of two-sided CI, make the originality of this work.

9.4.4 Validity conditions of Wald's CI

Wald's CI has low coverage even for quite high number of success and failures, leading some authors to recommend against teaching this CI in elementary courses, favoring Agresti-Coull or Wilson's CI [1,2,11]. Recent textbooks such as Fritz and Berger in 2015 [32] may mention that Wald's CI is biased, citing Agresti and Coull, but then give the old validity condition $np, n(1-p) \geq 5$. The condition $\min(x, n-x) > 40$ controls the one-sided local average errors for a 95% CI. It may be taught in place of $\min(x, n-x) \geq 5$ *or* 10.

9.4.5 Poisson distribution

As the binomial distribution is asymptotically equivalent to the Poisson distribution when the sample size is infinite and the expected number of successes is held constant, results for binomial CIs for a large sample size ($n = 2048$) can be extrapolated to the Poisson CIs. Convergence to the Poisson distribution

has been shown in A.4 and A.5.

9.4.6 Bootstrap

According to Carpenter [33], under usual conditions, the theoretical convergence rate of bootstrap 'normal', studentized and percentile is of the order $O\left(\frac{1}{\sqrt{n}}\right)$ whereas the theoretical convergence rate of the studentized and BC_a bootstrap is of the order $O\left(\frac{1}{n}\right)$. The 'normal' bootstrap is equivalent to Wald's CI. The asymptotical convergence rate of these CIs is reflected in the very high biases of the percentile and basic bootstrap CIs and much lower biases of the modified BC_a and studentized bootstrap CIs. No bootstrap CI controls the nominal error as well as the Clopper-Pearson mid-P CI.

9.4.7 Implementations

We recommend using the Clopper-Pearson mid-P CI. Its computer implementation is available in the exactci package for the R statistical software and in SAS version 9.4 through the MIDP option of the EXACT statement of PROC FREQ. Macros or programs for SPSS, Stata, SAS (for older version), Python, Minitab, MYSTAT/SYSTAT, Microsoft Excel, LibreOffice and nine other software and platforms are given in Appendix 2. They are distributed under the free software Creative Commons CC0 license terms. Online implementation of the procedure is available [34,35]. The R source code used to calculate confidence intervals, figures and tables of this document is distributed under the CC0 license in Appendix 4.

9.5 Practical example

Dellas *et al* [36] published a prognostic study in a population of hemodynamically stable patients with acute symptomatic pulmonary embolism. The objective was to predict complicated courses within 30 days (death, catecholamine administration, mechanical ventilation and resuscitation) from clinical parameters, Heart-type Fatty Acid Binding Protein (H-FABP) and multidetector computed tomography (MDCT). The simplified Pulmonary Embolism Severity Index (sPESI) was used to assess the clinical risk.

Complication risks were reported by frequency and 95% CI of proportions. Several frequencies were low: one adverse outcome for 225 patients was reported in the group defined by H-FABP ≤ 6 ng/mL

and sPESI = 0. Two adverse outcomes for 46 patients were reported in the group defined by H-FABP > 6 ng/mL and sPESI = 0. The reported CIs were 1/225 (0.4%, 95% CI: 0 to 1.3%) and 2/46 (4.3%, 95% CI: 0 to 10.9%). The CI estimator was not reported in the methods section. Different estimators yield quite different results, as shown in Table 2.

Dellas *et al* most probably used the percentile bootstrap estimator, as could be verified from the many low frequency proportions in the article with numerators ranging from 1 to 6. For the proportion 1 / 225, the upper boundary of the proportion ranged from 0.9% to 3.1%, depending on the estimator and the lower boundary could be negative, zero or positive.

It's not easy to estimate the actual confidence level Della *et al* had, since the actual proportion is unknown. It can be assumed, from the Clopper-Pearson mid-P CI that the actual proportion is less than 2.18% (upper boundary of the 95% CI). We assume that the authors would have noticed that the percentile bootstrap yields a [0 ; 0] CI when the observed proportion is zero and would have used the Clopper-Pearson 95% CI (classical exact CI) in place. In that conservative scenario, the actual local average lower bound and upper bound risks would have been respectively 0.93% and 7.4%. Therefore, the CI risks are unbalanced and far from the 2.5% nominal risk. The classical two-sided conditional coverage for this CI, for this theoretical proportion (2.18%) is equal to 94.6%, with actual risks of 1.1% of overestimation (α'_L risk) and 4.3% of underestimation (α'_U risk). This would appear to be a small bias. This example illustrates that the classical assessment of estimator bias by two-sided conditional coverage may provide results quite different from one-sided local average errors. In our opinion, the one-sided local average errors are more relevant as they take in account the variability of the number of subjects and/or actual proportion from one experience to another and allow interpretation of each CI boundary separately. This also shows that while bootstrap methods are asymptotically unbiased, on small or unbalanced samples they may be unreliable. This also illustrates the need of specifying the actual statistical method used in scientific articles.

9.6 Conclusion

The binomial proportion CI problem may seem trivial, but some aspects of the problem may be missed, such as tails equality or local average error control.

In this light, we assessed 55 CI estimators and give the following recommendation: use the Clopper-Pearson mid-P CI in all scenarios but the comparison of an observed proportion to a theoretical proportion in a strongly controlled or heavily regulated experimental environment, such as a clinical trial; in that case, use the Clopper-Pearson CI.

9.7 References

1. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* 1998; **52**:119–126.
2. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical science* 2001; **16**:101–117.
3. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**:404-413.
4. Pires AM, Amado C. Interval estimators for a binomial proportion: Comparison of twenty methods. *REVSTAT–Statistical Journal* 2008; **6**:165–197.
5. Lu H, Jin H, Wang Z, Chen C, Lu Y. Prior-free probabilistic interval estimation for binomial proportion. *TEST* June 2018.
6. Vollset SE. Confidence intervals for a binomial proportion. *Statistics in Medicine* 1993; **12**:809-824.
7. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine* 1998; **17**:857–872.
8. Sakakibara I, Haramo E, Muto A, Miyajima I, Kawasaki Y. Comparison of five exact confidence intervals for the binomial proportion. *American Journal of Biostatistics* 2014; **4**:11.
9. Kaiser HF. Directional statistical decisions. *Psychological Review* 1960; **67**:160.

10. Vos PW, Hudson S. Evaluation Criteria for Discrete Confidence Intervals. *The American Statistician* 2005; **59**:137-142.
11. Agresti A, Coull BA. Comment on "Interval estimation for a binomial proportion" by Brown, Cai and DasGupta (2001). *Statistical science* 2001; **16**:117–120.
12. Blyth CR, Still HA. Binomial Confidence Intervals. *Journal of the American Statistical Association* 1983; **78**:108-116.
13. Santner TJ. Comment on "Interval estimation for a binomial proportion" by Brown, Cai and DasGupta (2001). *Statistical science* 2001; **16**:126-128.
14. Schilling MF, Doi JA. A coverage probability approach to finding an optimal binomial confidence procedure. *The American Statistician* 2014; **68**:133–145.
15. Vos PW, Hudson S. Problems with Binomial Two-Sided Tests and the Associated Confidence Intervals. *Australian & New Zealand Journal of Statistics* 2008; **50**:81-89.
16. Stevens WL. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* 1950; **37**:117-129.
17. Winstein K. *Efficient Routines for Biostatistics.*; 2018.
18. Wang W. An iterative construction of confidence intervals for a proportion. *Statistica Sinica* 2014:1389–1410.
19. Schilling MF, Doi JA. A coverage probability approach to finding an optimal binomial confidence procedure. *The American Statistician* 2014; **68**:133–145.
20. Casella G. Refining binomial confidence intervals. *Canadian Journal of Statistics* 1986; **14**:113-129.
21. Molenaar W. Simple Approximations to the Poisson, Binomial, and Hypergeometric Distributions. *Biometrics* 1973; **29**:403–407.

22. Blyth CR. Approximate Binomial Confidence Limits. *Journal of the American Statistical Association* 1986; **81**:843-855.

23. Chen X, Zhou K, Aravena JL. Explicit Formula for Constructing Binomial Confidence Interval with Guaranteed Coverage Probability. *Communications in Statistics - Theory and Methods* 2008; **37**:1173-1180.

24. Zhou XH, Li CM, Yang Z. Improving interval estimation of binomial proportions. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 2008; **366**:2405-2418.

25. Crow EL. Confidence intervals for a proportion. *Biometrika* 1956; **43**:423-435.

26. Geyer CJ, Meeden GD. Fuzzy and randomized confidence intervals and P-values. *Quality Control and Applied Statistics* 2006; **51**:649.

27. Zieliński W. The Shortest Clopper-Pearson Randomized Confidence Interval. *REVSTAT–Statistical Journal* 2017; **15**:141–153.

28. Wang W. A "Paradox" in Confidence Interval Construction Using Sufficient Statistics. *The American Statistician* March 2017:1-6.

29. Efron B. Better bootstrap confidence intervals. *Journal of the American statistical Association* 1987; **82**:171–185.

30. Agresti A. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons; 2015.

31. Hirji KF, Mehta CR, Patel NR. Computing Distributions for Exact Logistic Regression. *Journal of the American Statistical Association* 1987; **82**:1110.

32. Fritz M, Berger PD. *Improving the User Experience through Practical Data Analytics: Gain Meaningful Insight and Increase Your Bottom Line*. Morgan Kaufmann; 2015.

33. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 2000; **19**:1141-1164.

34. Gillibert A. Clopper-Pearson mid-P interval calculator. URL http://andre.gillibert.fr/stats/cpmidp.html (accessed October 24, 2017).

35. OpenEpi - Confidence intervals for a proportion. URL http://www.openepi.com/Proportion/Proportion.htm (accessed October 24, 2017).

36. Dellas C, Lobo JL, Rivas A, Ballaz A, Portillo AK, Nieto R, del Rey JM, Zamorano JL, Lankeit M, Jiménez D. Risk stratification of acute pulmonary embolism based on clinical parameters, H-FABP and multidetector CT. *International Journal of Cardiology* 2018; **265**:223-228.

37. Bartlett MS. The Square Root Transformation in Analysis of Variance. *Supplement to the Journal of the Royal Statistical Society* 1936; **3**:68-78.

38. Anscombe FJ. The transformation of poisson, binomial and negative-binomial data. *Biometrika* 1948; **35**:246-254.

39. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 2000; **28**:783–798.

40. Lancaster HO. The combination of probabilities arising from data in discrete distributions. *Biometrika* 1949; **36**:370-382.

41. Berry G, Armitage P. Mid-P confidence intervals: a brief review. *The Statistician* 1995; **44**:417–423.

9.8 Tables and figures

Table 1: definition of lower bounds of the confidence intervals, upper bounds being defined by equivariance $\boldsymbol{U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)}$ *according to sample size* $\boldsymbol{n}$ *and number of sucesses* $\boldsymbol{x}$

Name	**Lower bound** $\boldsymbol{L_{1-\alpha}(x,n)}$
Wald[a]	$\max\left(0, \frac{x}{n} - \kappa\sqrt{\frac{x(n-x)}{n^3}}\right)$
[2] Modified Wilson[ab]	$\begin{cases} \frac{1}{2n}\chi^2_{\alpha,2x}\ if\ 1 \le x \le x^* \\ \frac{x + \frac{\kappa^2}{2} - \kappa\sqrt{\frac{x(n-x)}{n} + \frac{\kappa^2}{4}}}{n+\kappa^2}\ otherwise \end{cases}$ where $x^* = 2$ for $n \le 50$ and $x^* = 3$ for $n > 50$
[37,38] Bartlett Arc-sine[a]	$\sin^2\left(\max\left(0, \operatorname{asin}\left(\sqrt{\frac{x+\frac{1}{2}}{n+1}}\right) - \frac{\kappa}{2\sqrt{n+\frac{1}{2}}}\right)\right)$
[2] Modified Wald logit[ac]	$\begin{cases} \operatorname{logitinv}\left(\log\left(\frac{x}{n-x}\right) - \kappa\sqrt{\frac{n}{x(n-x)}}\right)\ if\ 0 < x < n \\ \sqrt[n]{\alpha/2}\ if\ x = n \\ 0\ if\ x = 0 \end{cases}$
[6]Modified likelihood ratio[a]	$\begin{cases} \inf\left\{q \middle\| \log\left(\left(\frac{x}{nq}\right)^x \left(\frac{n-x}{n(1-q)}\right)^{n-x}\right) \le \frac{1}{2}\kappa^2\right\}\ if\ 0 < x < n \\ \sqrt[n]{\alpha/2}\ if\ x = n \\ 0\ if\ x = 0 \end{cases}$
[2]Modified equal-tailed Jeffreys[d]	$\begin{cases} \beta iCDF(\alpha/2; x+1/2, n-x+1/2)\ if\ 2 \le x < n \\ \sqrt[n]{\alpha/2}\ if\ x = n \\ 0\ if\ x \le 1 \end{cases}$
[39]Blaker	$\inf\{q \mid \text{bpval}(q,x,n) > \alpha\}$ where

	$\text{bpval}(p,x,n)$ $= \begin{cases} \Pr(X \le x \text{ or } X \ge \inf\{x' \mid \Pr(Y \ge x') \le \Pr(Y \le x)\}) \text{ if } p \ge \frac{x}{n} \\ \text{bpval}(1-p, n-x, n) \text{ if } p < \frac{x}{n} \end{cases}$ Where $X \sim \mathcal{B}(n,p)$ and $Y \sim \mathcal{B}(n,p)$
[2,3]Clopper-Pearson[d]	$\beta iCDF\left(\frac{\alpha}{2}; x, n-x+1\right)$
[40,41]Clopper-Pearson mid-P	$\inf\left\{q \mid \Pr(X \ge x) - \frac{1}{2}\Pr(X = x) > \frac{\alpha}{2} \text{ where } X \sim \mathcal{B}(n,q)\right\}$

[a]We denote by $\kappa = z_{1-\alpha/2}$ the quantile $1-\alpha/2$ of the normal distribution $\mathcal{N}(0,1)$

[b]$\chi^2_{q,df}$ is the q quantile of the χ^2 distribution with df degrees of freedom

[c]The reciprocal of the logistic transformation is defined by $\text{logitinv}(t) = \frac{\exp(t)}{1+\exp(t)}$

[d] $\beta iCDF(q;\alpha,\beta)$ is the q th quantile of the beta distribution whose shape parameters are α and β

Table 2: different CI estimators applied to the proportions defined in the article of Dellas et al.

	Observed proportion	
	1 / 225 (0.4%)	2 / 46 (4.3%)
Estimator		
Percentile bootstrap	0 to 1.3%	0 to 10.9%
Basic bootstrap	−0.4% to 0.9%	−2.2% to 8.7%
Wald CI	−0.4% to 1.3%	−1.5% to 10.2%
Clopper-Pearson	0.01% to 2.4%	0.5% to 14.8%
Clopper-Pearson mid-P	0.02% to 2.2%	0.7% to 13.6%
Wilson	0.08% to 2.5%	1.2% to 14.5%
Wald logit	0.06% to 3.1%	1.1% to 15.8%
Likelihood ratio	0.03% to 1.9%	0.7% to 12.8%

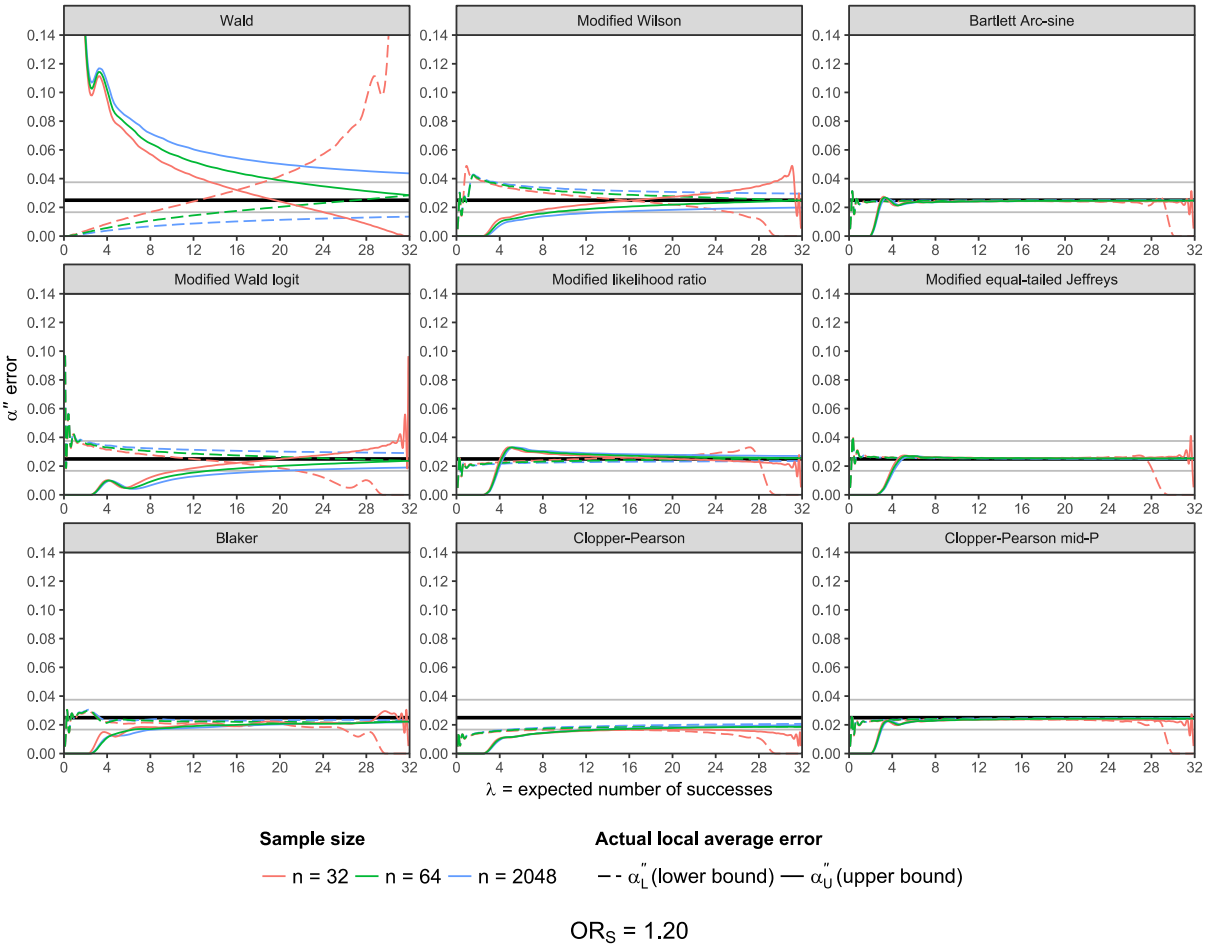


Figure 1: one-sided local average errors of nine 95% confidence interval estimators according to different sample sizes (red for $\boldsymbol{n = 32}$*, green for* $\boldsymbol{n = 64}$ *and blue for* $\boldsymbol{n = 2048}$*), with random actual* $\boldsymbol{P}$ *proportion following a logit-normal distribution with typical odds ratio of the actual proportion between two experiments equal to* $\boldsymbol{OR_S = 1.20}$*. The abscissa is the expected number of successes* $\boldsymbol{np_0}$ *and the ordinate is the probability that the lower bound of the confidence interval is greater than the true proportion* $\boldsymbol{p}$ *(left local average error: dashed lines) or the probability that the upper bound of the confidence interval is lower than the true proportion* $\boldsymbol{p}$ *(right local average error: solid line).*

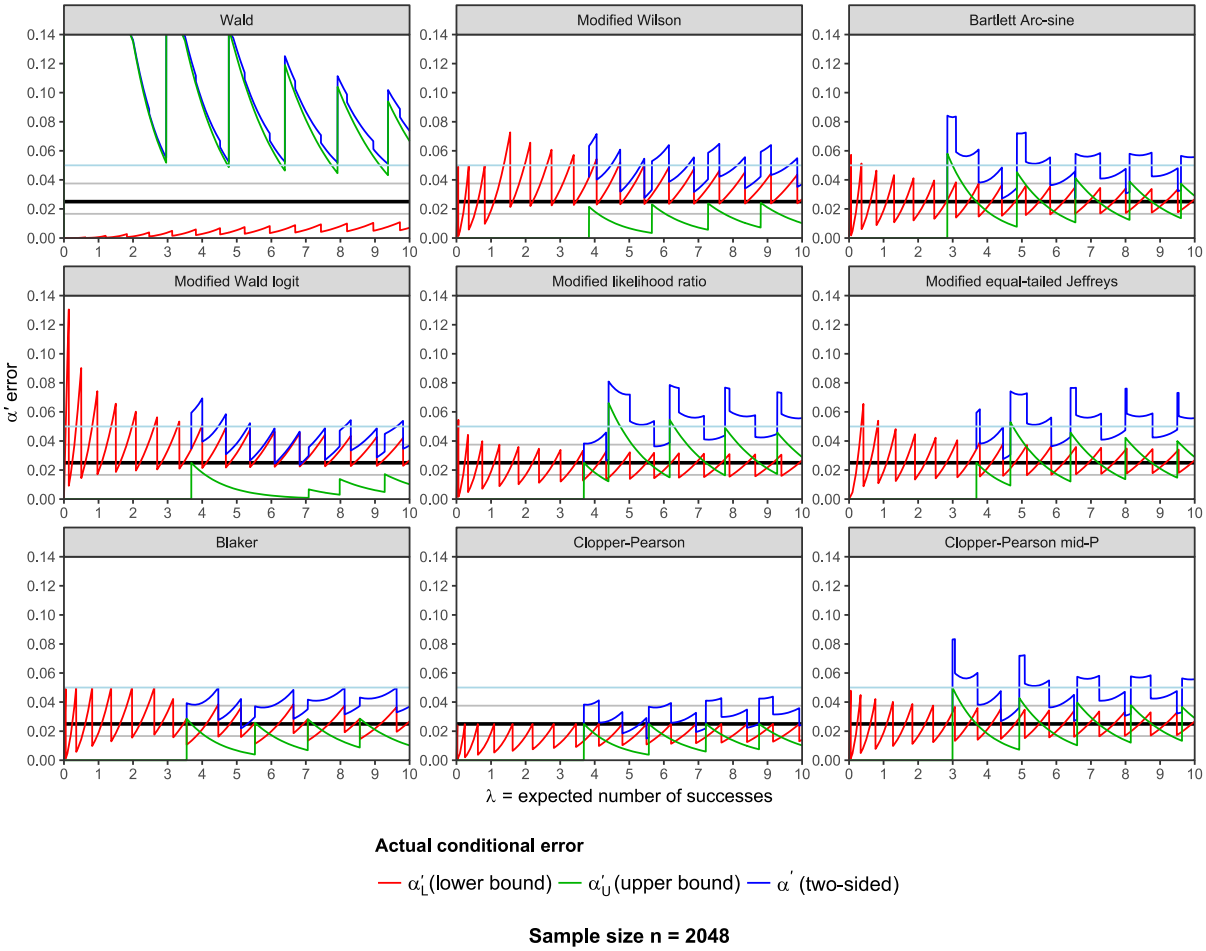


Figure 2: one-sided and two-sided conditional errors of nine 95% confidence interval estimators for a sample of size $\boldsymbol{n = 2048}$ *and a constant theoretical* $\boldsymbol{p}$ *proportion. The abscissa is the expected number of successes* $\boldsymbol{np}$ *and the ordinate is the risk that the lower bound of the confidence interval is greater than the true proportion* $\boldsymbol{p}$ *(left conditional error: red), the risk that the upper bound of the confidence interval is lower than the true proportion* $\boldsymbol{p}$ *(right conditional error: green) or the risk that the confidence interval does not contain the true proportion* $\boldsymbol{p}$ *(two-sided risk: blue).*

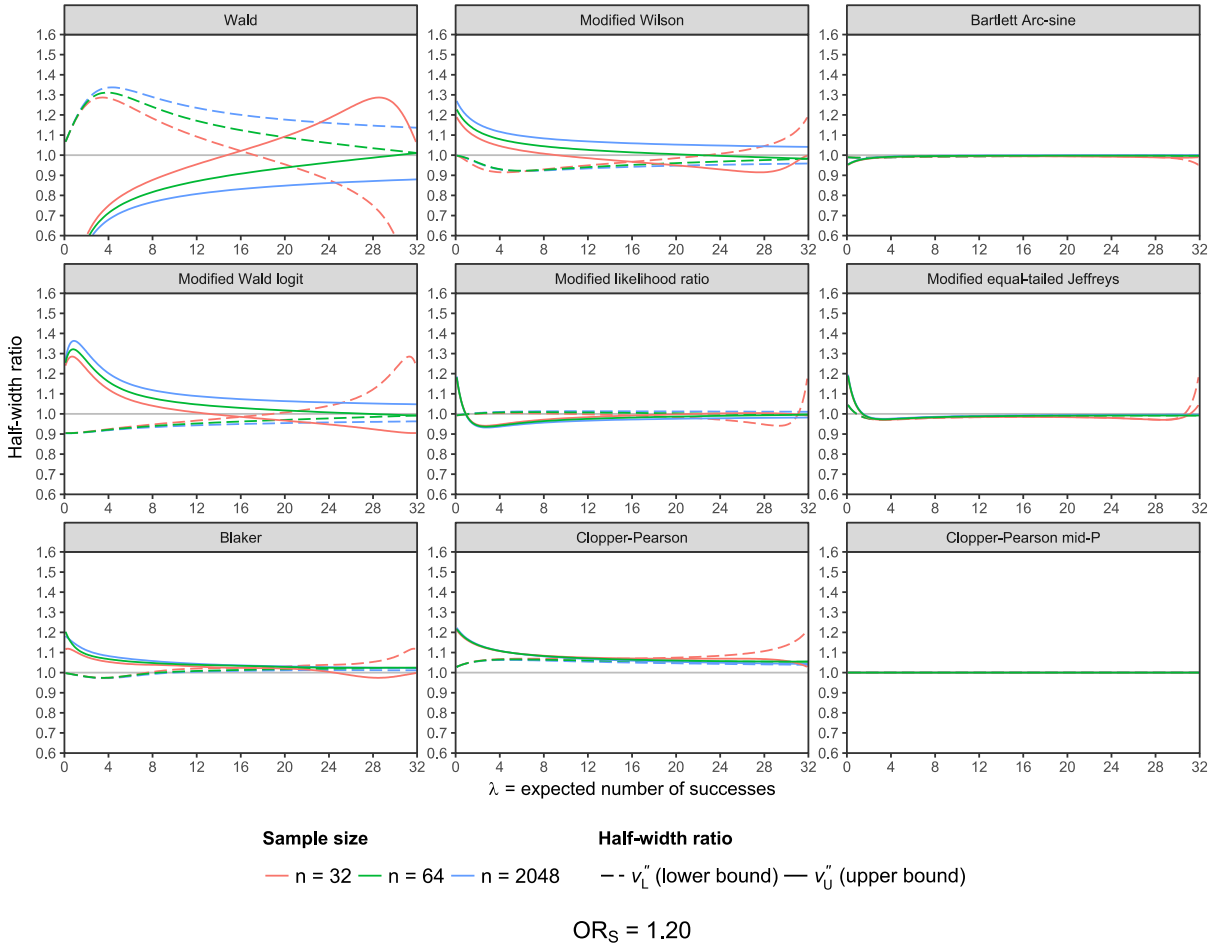


Figure 3: relative local average half-widths of nine 95% confidence interval estimators for a sample of size $\boldsymbol{n = 2048}$ *with a random actual* $\boldsymbol{P}$ *proportion following a logit-normal distribution with a typical odds ratio of the actual proportion between two experiments equal to* $\boldsymbol{OR_S = 1.20}$*. The relative half-width is the local average half-width of one of the nine intervals divided by the local average half-width of the Clopper-Pearson mid-P interval for the same* $\boldsymbol{x}$*,* $\boldsymbol{n}$ *and* $\boldsymbol{p_0}$ *(expected* $\boldsymbol{P}$*) parameters. The abscissa is the expected number of successes* $\boldsymbol{np}$ *and the ordinate is the left relative local average half-width (dashed lines) or the right relative local average half-width (solid lines).*

9.9 Appendices

All appendices are available on the Open Science Framework (https://osf.io/gqyem/)

10Annexe 3 : article N°3

Best estimator for bivariate Poisson regression

André GILLIBERT[ab]*†, Jacques BÉNICHOU[bc] and Bruno FALISSARD[a]

[a] INSERM UMR 1178, Université Paris Sud, Maison de Solenn, Paris, France.

[b] Department of Biostatistics and Clinical Research, CHU Rouen, Rouen, F 76031, France

[c] Inserm U 1181, Normandie University, Rouen, France

* Correspondence to: André GILLIBERT, Department of Biostatistics and Clinical Research, CHU Rouen, Rouen, F 76031, France

†E-mail: andre.gillibert@chu-rouen.fr

Abstract

INTRODUCTION: Wald's, the likelihood ratio (LR) and Rao's score tests and their corresponding confidence intervals (CIs), are the three most common estimators of parameters of Generalized Linear Models. On finite samples, these estimators are biased. The objective of this work is to analyze the coverage errors of the CI estimators in small samples for the log-Poisson model (i.e. estimation of incidence rate ratio) with innovative evaluation criteria, taking in account the overestimation/underestimation unbalance of coverage errors and the variable inclusion rate and follow-up in epidemiological studies.

METHODS: Exact calculations equivalent to Monte Carlo simulations with an infinite number of simulations have been used. Underestimation errors (due to the upper bound of the CI) and overestimation coverage errors (due to the lower bound of the CI) have been split. The level of confidence has been analyzed from 0.95 to $1-10^{-6}$, allowing the interpretation of P-values below 10^{-6} for hypothesis tests.

RESULTS: The LR bias was small (actual coverage errors less than 1.5 times the nominal errors) when the expected number of events in both groups was above 1, even when unbalanced (e.g. 10 events in one group *vs* 1 in the other). For 95% CI, Wald's and the Score estimators showed high bias even when the number of events was large (≥20 in both groups) when groups were unbalanced. For small P-values ($<10^{-6}$), the LR kept acceptable bias while Wald's and the score P-values had severely inflated errors (×100).

CONCLUSION: The LR test and LR CI should be used.

10.1 Introduction

Generalized Linear Models (GLMs) are a family of statistical models, including the widely used logistic regressions and Poisson regressions. Several hypothesis tests and estimators of CIs (CIs), exist for parameters of these models. The best known tests are Rao's score test (also known as the score test or Lagrange multiplier Test), Wald's test and the generalized likelihood ratio test (GLRT) [1]. These three tests are equivalent when there is no nuisance parameter [2], *i.e.* a parameter such as the base incidence rate (*e.g.* incidence rate of group 1) for estimating an incidence rate ratio: it is not the parameter that we wish to estimate, but our estimation of the incidence rate ratio depends on it, so that uncertainty of this nuisance parameters generates an error or a bias in the estimation of the parameter that we wish to estimate. Since nuisance parameters stabilize when the sample size grows, the three tests are asymptotically equivalent for GLMs. The statistics of Rao's score test [3] and of the GLRT (after transformation) [4] have both an asymptotic chi-square distribution. Wald's statistic, for GLMs, is asymptotically normally distributed [5]. By test inversion, CIs estimators can be built from these three hypothesis tests.

> Note that the generalized likelihood ratio test (GLRT) is often abbreviated to likelihood ratio test (LRT), but is actually not equivalent. The GLRT statistic is the ratio of the likelihood of the observed data under the null hypothesis and the maximum likelihood (ML) estimate over the complete space of the tested parameter while the LRT defined in Neyman-Pearson's lemma is based on the ratio of likelihood for two fully specified parameters θ_0 and θ_1 [6]. The GLRT can be used to reject a hypothesis when all values in a parameter space are possible while the LRT relies on a parameter space restricted to two values, is the latter being uncommon; it can help choose between two hypotheses when no other hypothesis is possible. In the rest of this document, the term likelihood ratio (LR) is considered synonymous to the generalized likelihood ratio since the GLRT is, by far, the most used method and is usually called LR.

Statistical software do not rely on the same default estimator (e.g. likelihood ratio CI for R, Wald's CI for SAS, SPSS, Stata) and may even be inconsistent: for instance, by default R (version 3.6.2) uses

Wald's P-values and likelihood ratio CIs for GLMs.

On finite samples, some estimators and tests may behave better than others. The standard estimators (Score, Wald, LR) are supplemented by more anecdotal estimators such as the penalized likelihood estimators (Firth [7], Kenne [8]) and Hirji's exact estimator [9]. Since these estimators have been designed to reduce bias, we will analyze their behavior as well.

A quick review of articles (articles reviewed by one researcher, with their supplementary material if available) published between September and November 2018 in the British Medical Journal, The Lancet, The Journal of the American Medical Association, The New England Journal of Medicine, and the Annals of Internal Medicine, showed that, out of 203 research articles, 198 were non-meta-analysis articles of which 58 used a logistic regression (fixed effects, conditional, mixed effects or GEE) for primary, secondary or post hoc analysis, of which 46 tested hypothesis or provided CIs of odds ratio (12 logistic regressions were used for multiple imputation, propensity score matching or were an intermediate step for a more complex calculation). Out of the 46 original research articles using a logistic regression with a hypothesis test or CI on an odds ratio, 9 (20%) gave some information about the estimator or test used and 6 (13%) [10–15] gave explicitly an estimator or test name while the other 3 (7%) [16–18] gave indirectly the estimator, by sharing the script or software package used. Of the 6 articles with explicit estimator or test sharing, one of them [11] added the specification of the test after retraction and replacement of the article due to errors in statistical analyses. Therefore, most articles do not specify the estimator used.

The Poisson distribution is asymptotically equivalent to the Binomial distribution if the proportion tends towards zero. The Gaussian approximation of the binomial distribution is best when the actual proportion is close to 50%, leading to small bias in logistic regressions. For low proportions and high proportions, the binomial distribution is highly skewed and the Gaussian approximation is poor. For a logistic regression, the two extreme scenarii are equivalent since a high proportion can be transformed to a low proportion by inverting the outcome (analyzing non-events rather than events). Therefore, we concentrate our analysis in the scenario of low proportion. A logit-binomial and log-Poisson model are equivalent when the denominator tends towards infinity. Therefore, we considered that the log-Poisson

model is the worst case scenario (most biased for Wald's CI) of the logistic regression. That is why we analyzed the Poisson regression only.

The objective of this work was to assess the coverage bias of bivariate Poisson regression CI estimators with innovative evaluation criteria in order to define their validity conditions and provide useful advice on which estimator to use and when to use it.

10.2 Rationale and Methods

10.2.1 Scenario analyzed

The assessment of the ratio of two Poisson distribution has been analyzed. We note $Y_1 \sim P(\lambda_1)$ and $Y_2 \sim P(\lambda_2)$ two independent variables following Poisson distributions and r_1 and r_2 the offset (e.g. number of subjects-years of follow-up). The ratio $\rho = \frac{\lambda_2/r_2}{\lambda_1/r_1}$ can be interpreted as an incidence rate ratio. The estimation of this IRR is equivalent to the estimation of $\tau = \frac{\lambda_2}{\lambda_1}$ after a change of variable. Indeed, comparing ρ to a value ρ_0 in a GLM with offset is equivalent to comparing τ to a value $\rho_0 \times \frac{r_2}{r_1}$ in a GLM without offset. That is why we only analyzed the scenario without offset. The model can be written:

$$\log(E[Y_1]) = \beta_0$$

$$\log(E[Y_2]) = \beta_0 + \beta_1$$

With $J = \exp(\beta_1)$ the IRR that we wish to estimate.

The plane of all possible combinations of (λ_1, λ_2) will be analyzed from $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ to $\lambda_1 = 104$ and $\lambda_2 = 104$ by increments of 0.05.

10.2.2 Evaluation criteria: one-sided unconditional risks

The coverage bias of two-sided 95% CIs is usually assessed by the coverage error defined as the actual coverage minus the nominal coverage. For the realization of a CI, a coverage fault can be an overestimation (lower boundary of the CI above the actual value) or an underestimation (upper boundary of the CI below the actual value). The overall coverage error is computed by adding the probability of overestimation to that of underestimation; doing so, they are somehow assumed to be equivalent.

Underestimating an incidence rate ratio in 4.9% of cases and overestimating it in 0.1% of cases will be seen as an unbiased CI estimator by this statistic, since the overall coverage will be 100% − 4.9% − 0.1%=95%, but will falsely reassure when answering to the research question "may this exposition raise the incidence of that disease?". Therefore, we think that, in most cases a two-sided CI should behave as the intersection of two one-sided CIs with the same risk of non-coverage. That is why we will analyze separately the non-coverage due to the lower and upper boundaries of the CI estimators. A balanced CI estimators is wished.

We define $\alpha_L = \alpha_U = \alpha$ the nominal probabilities that a two-sided CI be strictly above or strictly below the actual statistic $\rho = \lambda_2/\lambda_1$. This definition seeks for balanced CIs. For a two-sided 95% CI, $\alpha = 0.025$.

Assuming that λ_1 and λ_2 are fixed, we respectively define the ***conditional risks*** as α'_L and α'_U as the actual probabilities that the CI estimator is strictly above or strictly below the actual statistic I. That is, if $Y_1 \sim P(\lambda_1)$ and $Y_2 \sim P(\lambda_2)$ are two random variables following Poisson distributions, and $C(y_1, y_2) = [L(y_1, y_2); U(y_1, y_2)]$ is a CI estimator with a lower boundary L and a upper boundary U, computed from the observed number of events y_1 and y_2, then $\alpha'_L = \Pr\left(L(Y_1, Y_2) > \frac{\lambda_2}{\lambda_1}\right)$ and $\alpha'_U = \Pr\left(U(Y_1, Y_2) < \frac{\lambda_2}{\lambda_1}\right)$.

In studies estimating incidence rate ratios, λ_1 and λ_2 depend on the duration of follow-up and inclusion rate. The inclusion rate is rarely controlled and can be considered random and the follow-up is not always perfectly controlled. Therefore, we can consider that r_1 and r_2 are random variables. Although the incidence rate ratios $I_1 = \frac{\lambda_1}{r_1}$ and $I_2 = \frac{\lambda_2}{r_2}$ may be constant, the randomness of r_1 and r_2 leads to random $\lambda_1 = I_1 \times r_1$ and $\lambda_2 = I_2 \times r_2$. We note Λ_1 and Λ_2 the random variables representing the expected number of events, assuming they follow log-normal distributions with a geometrical standard deviation equal to 1.10 and expectancies respectively equal to λ_1 and λ_2. The random distribution of Λ_1 and Λ_2 were chosen arbitrarily. In sensitivity analyzes, the geometrical standard deviation 1.10 was changed. We define random variables Y_1'' and Y_2'' built from a two-steps procedure: realizing Λ_1 and Λ_2 as l_1 and l_2, then, drawing y_1'' and y_2'' from Poisson distributions with expectancy l_1 and l_2. In this two-steps

experiment, l_1 and l_2 represents the expected number of events in both groups after the total duration of follow-up (r_1 and r_2) are realized and before the actual number of event occur. The values y_1'' and y_2'' represent the observed number of events in both groups at the end of the experiment.

We define the ***unconditional risks*** α_L'' and α_U'' as $\alpha_L'' = \Pr\left(L(Y_1'', Y_2'') > \frac{\Lambda_2}{\Lambda_1}\right)$ the actual probability that the confidence interval be strictly above the actual incidence rate ratio and $\alpha_U'' = \Pr\left(U(Y_1'', Y_2'') < \frac{\Lambda_2}{\Lambda_1}\right)$ the probability that the confidence interval be strictly below the actual incidence rate ratio.

The nominal coverage will be 95% for the primary analysis, i.e. $\alpha = \alpha_L = \alpha_U = 0.025$. There is no consensus limit of what is an "acceptable" coverage bias. The limit has been arbitrarily set to $\frac{\alpha}{1.5} = 0.01667$ and $\alpha \times 1.5 = 0.0375$. Less strict (1.25) and more strict (2 and 10) multiplicative thresholds are shown in figures.

10.2.3 Evaluation criteria: interval relative half-width

As the coverage errors are separately assessed for both boundaries of the CI, the width of the CI is better split in two halves and assessed separately: the lower and upper half-widths are equal to the distance between ML point estimate (even for Firth's and Kenne's CI) and, respectively, the lower and upper boundaries of the CI.

The half-width is highly dependent on the number of events observed. For instance, for $y_1 = 1$ and $y_2 = 10$ the upper half-width of the ML LR 95% CI and Hirji's mid-P 95% CI are respectively equal to 173.4 and 210.1 (ratio of the two half-widths = 0.83), but for $y_1 = 100$ and $y_2 = 100$ the half-widths are respectively equal to 0.3200 and 0.3206 (ratio = 0.9982). In these two scenarii, the half-widths are different by order of magnitudes (~200 *vs* ~0.3) but the ratio of half widths are much closer (0.83 *vs* 0.9982). Therefore, a figure showing crude half-widths for a large range of values would be illegible but a figure showing ratios of half-widths of a CI to a reference CI is much more legible. Therefore, we displayed ratios of half-widths of estimators relative to each other.

10.2.4 Estimators

The following CI estimators of bivariate log-Poisson regressions have been analyzed:

1) The CI constructed by inversion of Rao's score test (score CI)

2) Wald's asymptotic normal CI (Wald's CI)
3) The profile likelihood ratio CI based on ML estimates (ML LR CI)
4) Wald's CI in a GLM fitted by Firth's penalized likelihood ratio estimator [7] (Wald-Firth CI)
5) Non-Penalized LR CI in a GLM where the point estimates are estimated by Firth's penalized LR estimator (LR1-Firth CI)
6) Penalized LR CI in a GLM where the point estimate is estimated by Firth's penalized LR estimator (LR2-Firth CI)
7) Wald's CI [8] in a GLM fitted by Kenne's penalized LR estimator (Wald-Kenne CI)
8) Hirji's exact estimator of the GLM CI [9] (Hirji's CI)
9) Hirji's exact estimator of the GLM CI [9] with mid-P value modification [19] (Hirji mid-P CI)

All these CI estimators can be constructed by inversion of hypothesis tests. Consequently, the results shown for CIs can be extrapolated to corresponding hypothesis tests.

When the number of events y_1 or y_2 is equal to zero, some estimators do not converge (Rao's score, Wald's CI, the ML LR CI). Penalized LR estimators (Firth, Kenne) and exact estimators do converge unless both y_i are zero. For consistency, it has been assumed that whenever one or the other y_i is zero, no CI will be computed. Therefore, all results of this work are conditional to the fact that both y_i are strictly positive.

10.2.5 Computation method of CI

10.2.5.1 The Wald, ML LR and score CI

As these estimators are conditional to $y_1 + y_2$, they have been shown to be equivalent to the estimation of a binomial proportion in a univariate (intercept-only) logistic regression after a change of variable. Applying the function $p \to \frac{p}{1-p}$ to both boundaries of the CI of the binomial distribution yields the bivariate Poisson regression CI. This gave simple analytical solutions to Wald's and the score CI. The ML LR CI is harder to compute and required numerical inversion of the family of function $r(x) = a \times \log(x) + b \times \log(x+1) + c$ where a, b and c are constant parameters of the function. After a change of variable $y = \log(x)$, the Newton-Raphson algorithm converged in three iterations to more than 10 decimal places (data not shown).

10.2.5.2 Firth's LR CI

The LR CI has been computed by inverting a hypothesis test. The Newton-Raphson algorithm with numerical derivation was used. The hypothesis test has been performed by fitting the model without constraining the slope (alternative hypothesis, 2 degrees of freedom) and with a fixed slope (null

hypothesis, 1 degree of freedom). The fitting process of both models was performed by Firth's penalized likelihood. Then, either the log-likelihood (LL) or the penalized log-likelihood (PLL) of both fitted models was computed, respectively for LR1-Firth and LR2-Firth CIs. Twice the difference of LL or PLL, was approximated to a chi square distribution at one degree of freedom under the null hypothesis.

10.2.5.3 Hirji's exact CI

Hirji described his estimator for the logit-binomial regression [9] but it can be generalized to the Poisson distribution. Indeed, a Poisson distribution is equivalent to a binomial distribution with a sample size (parameter n) tending towards infinity. Like the Wald, ML LR and score CI, Hirji's exact CI is equivalent to the binomial exact CI. It can be based on the Clopper-Pearson CI with or without mid-P modification.

10.2.6 Computation method of non-coverage probabilities

10.2.6.1 Conditional risks

Coverage errors, defined as one-sided risks, as seen in section 10.2.2, are dependent on λ_1 and λ_2 the expected number of events (*e.g.* crude incidence of events) but not on r_1 and r_2, the number of patients-years of follow-up. Therefore, all our calculations are only based on λ_1 and λ_2, and the set of possible values for λ_1 and λ_2 is explored up to $\lambda_1 = 104$ and $\lambda_2 = 104$ by increment of 0.05, leading to a large matrix of coverage errors, graphically represented as a colored surface.

Rather than based on Monte Carlo sampling, coverage errors have been computed by almost exact probability calculation. This made it possible to quickly estimate coverage errors for a confidence level equal to $1 - 10^{-6}$. The algorithm for the lower bound risk α_L' conditional to Poisson parameters λ_1 and λ_2 is described below:

1) Input of the algorithm : λ_1 and λ_2, the expected number of events, and α_L the nominal one-sided coverage error
2) Compute quantiles 10^{-9} and $1 - 10^{-9}$ for both $P(\lambda_1)$ and $P(\lambda_2)$ distributions
3) Sets the overall sum to zero
4) For all y_1 and y_2 (target observation) between the respective quantiles of Poisson distributions (matrix of all possible pairs of observations (y_1, y_2))

a. Compute the CI for y_1 and y_2

b. If the CI is entirely above λ_2/λ_1, adds the probability that $Y_1 = y_1$ and $Y_2 = y_2$ (product of both probabilities) to the overall sum, with $Y_1 \sim P(\lambda_1)$ and $Y_2 \sim P(\lambda_2)$

At the end of the loop, the overall sum is the probability that the CI is entirely above the actual ratio λ_2/λ_1. The same algorithm applies to α'_U but the upper boundary is compared to λ_2/λ_1. A small adjustment of the matrix of probabilities has been applied to make sure that the sum of all probabilities of (y_1, y_2) pairs is equal to 1, rather than $(1-10^{-9})^2$ as would be expected from the quantiles of the Poisson distribution described at step 1 of the algorithm. This algorithm is equivalent to Monte Carlo sampling with an infinite number of simulations, with an error less than 10^{-9}.

Conditional risks have been computed for all λ_1 and λ_2, from 0 to 40 and from 100 to 104 by increment of 0.05. This defines the size of figures pixels.

10.2.6.2 Unconditional risks

The log-normal distributions of Λ_1 and Λ_2 define a bivariate distribution. The bivariate distribution, assuming Λ_1 and Λ_2 are independent, was approximated to a bivariate discrete distribution with a 20×20 grid (40×40 for the sensitivity analysis with geometrical standard deviation = 1.20). The support of this discrete distribution was calculated to align on the grid (precision 0.05) that had been used for the computation of conditional risks. Then, the probabilities of each (l_1, l_2) pair was multiplied by the conditional α'_L or α'_U risk, then summed up to estimate the unconditional α''_L and α''_U risks.

10.3 Sensitivity analyzes

For unconditional risks, the geometrical standard deviation has been set at 1.10 in the primary analysis and changed to 1.05 and 1.20 in sensitivity analyzes.

10.4 Secondary analyzes

The two-sided confidence levels has been set at 0.95 in the primary analysis and changed to 0.80, 0.90, 0.999 and $1-10^{-6}$ in secondary analyzes.

10.5 Results

10.5.1 Unconditional coverage

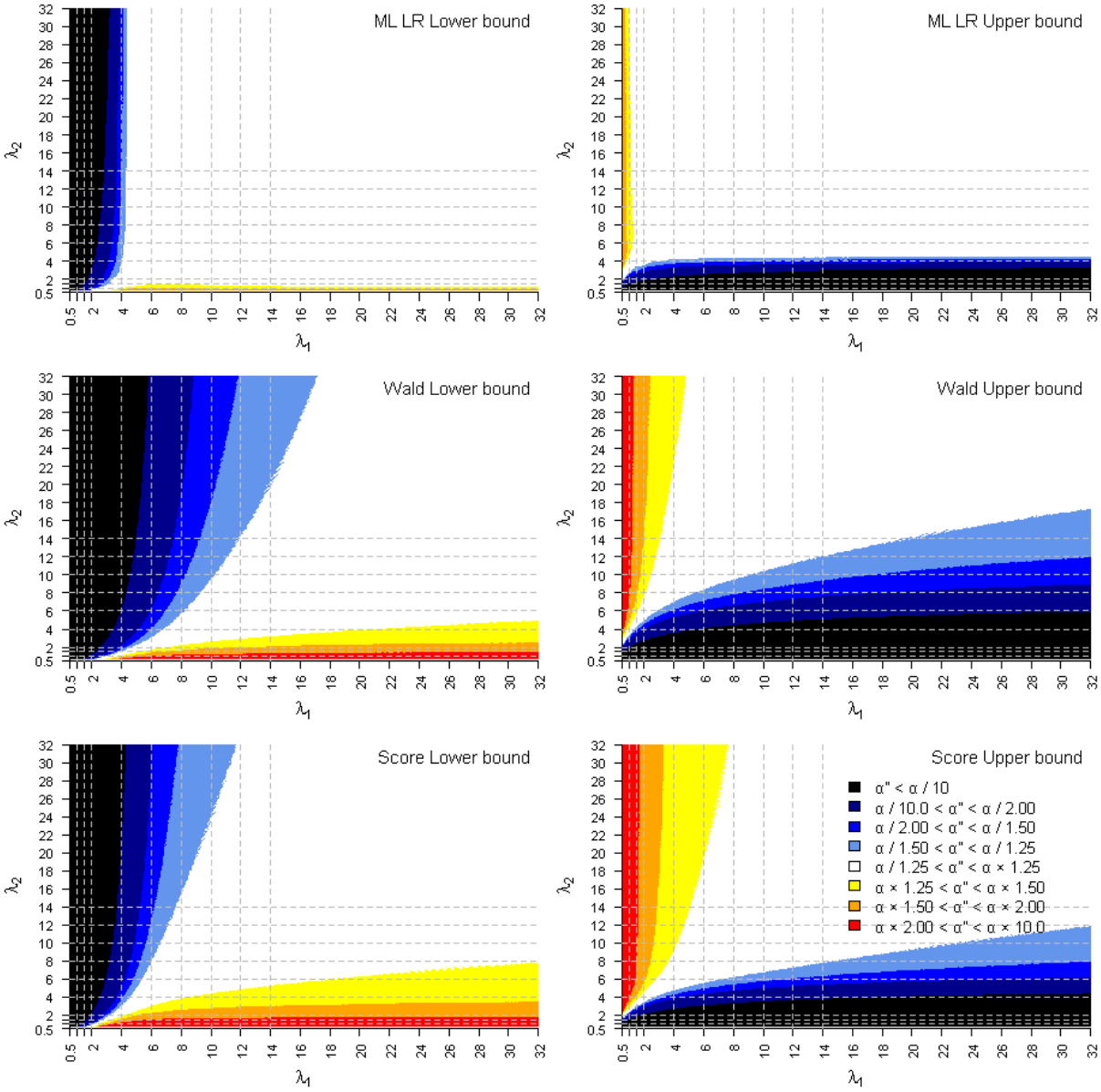


Figure 4 : unconditional α_L'' and α_U'' risks for the 95% two-sided likelihood ratio confidence interval (ML LR), 95% two-sided Wald confidence interval and 95% two-sided Score confidence interval of a ratio Λ_2/Λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the expectancies of the Poisson distribution parameters assuming these parameters follow a log-normal distribution having a geometrical standard deviation equal to 1.1. The white, yellow and light blue zones show the zones of

tolerated risk deflation/inflation by a factor less than 1.5.

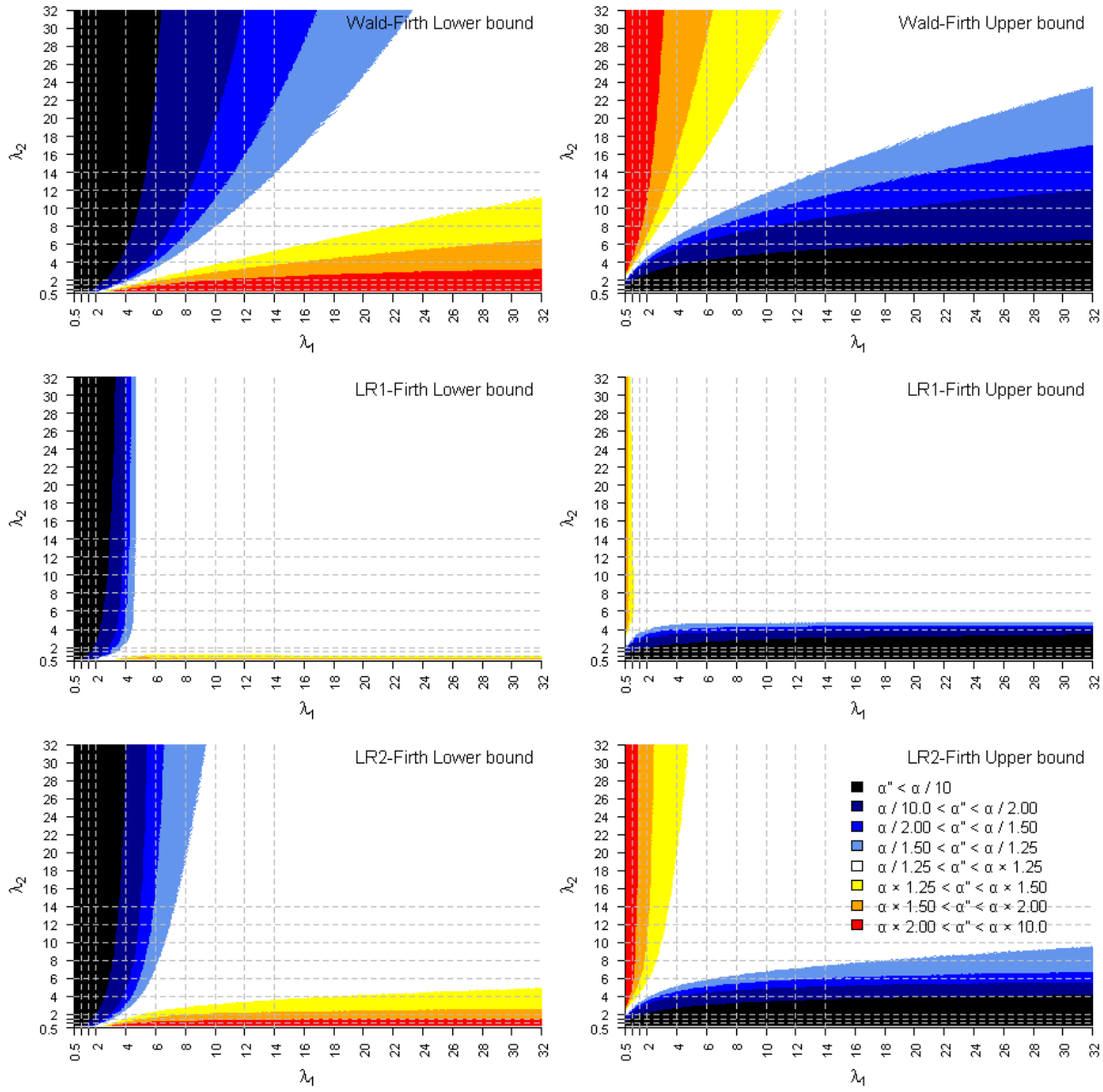


Figure 5 : unconditional α_L'' and α_U'' risks for the 95% two-sided Wald-Firth, LR1-Firth and LR2-Firth confidence intervals of a ratio Λ_2/Λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the expectancies of the Poisson distribution parameters assuming these parameters follow a log-normal distribution having a geometrical standard deviation equal to 1.1. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

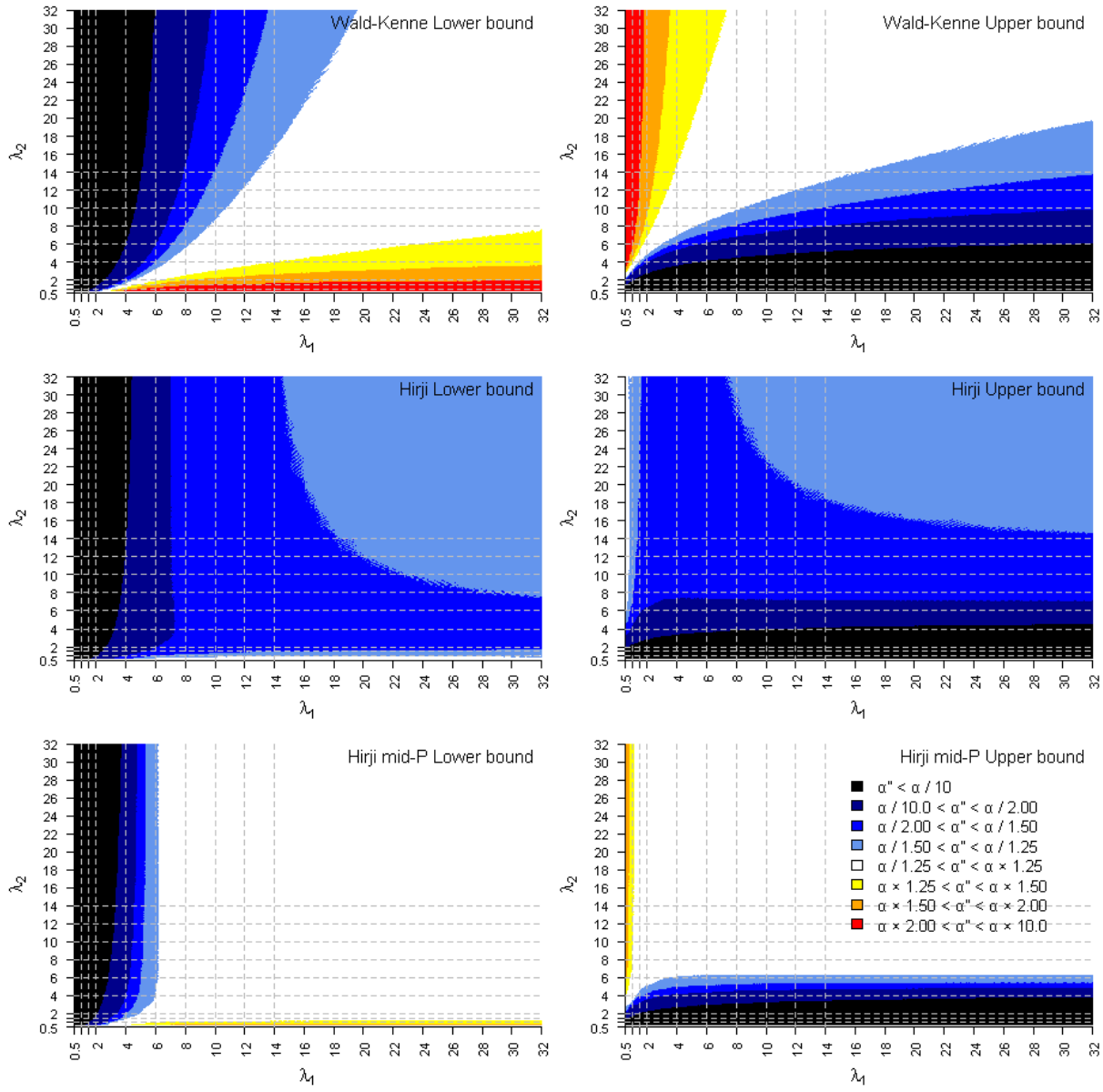


Figure 6 : unconditional α_L'' and α_U'' risks for the 95% two-sided Wald-Kenne, Hirji and Hirji mid-P confidence intervals of a ratio Λ_2/Λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the expectancies of the Poisson distribution parameters assuming these parameters follow a log-normal distribution having a geometrical standard deviation equal to 1.1. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

Figure 4 shows that the ML LR CI estimator is much less biased than the Wald's and Score CI estimators, for both boundaries. The main bias of the ML LR CI estimator is due to an unavoidable over-coverage of the upper boundary when λ_2 (numerator) is close to zero and λ_2/λ_1 is small. Indeed, if the ratio λ_2/λ_1

is small enough to be smaller than the upper boundary of the CI for an observed denominator equal to 1, then, the upper boundary will never be below λ_2/λ_1 and the risk associated to the boundary will be zero. Similarly, a very small λ_1 and high λ_2/λ_1 ratio leads to an over-coverage of the lower boundary. There is some under-coverage for expected number of events (λ_1 or λ_2) smaller than 1. The score CI is not uniformly better than Wald's CI: it has less over-coverage but has more under-coverage.

Although Firth's estimator may be less biased than the ML estimator for a point estimate, Figure 5 shows that the associated Firth-Wald CI is more biased than Wald's CI associated to the ML estimator (Figure 4). The LR1-Firth CI is very slightly more conservative (more over-coverage, less under-coverage) than the ML LR CI (Figure 4). Approximation of the differences of penalized deviances to a chi-square distribution seems to lead to biased CI as seen in the LR2-Firth interval.

Figure 6 shows that the Kenne-Wald CI estimator is less biased (less over-coverage, less under-coverage) that the Firth-Wald CI estimator (Figure 5), but is still more biased than the standard Wald CI estimator (Figure 4). Hirji's CI, without mid-P modification, is strictly conservative, with much over-coverage and no under-coverage. The mid-P modification provides properties close to that of the ML LR CI estimator (Figure 4), but slightly more conservative.

10.5.2 Conditional coverage

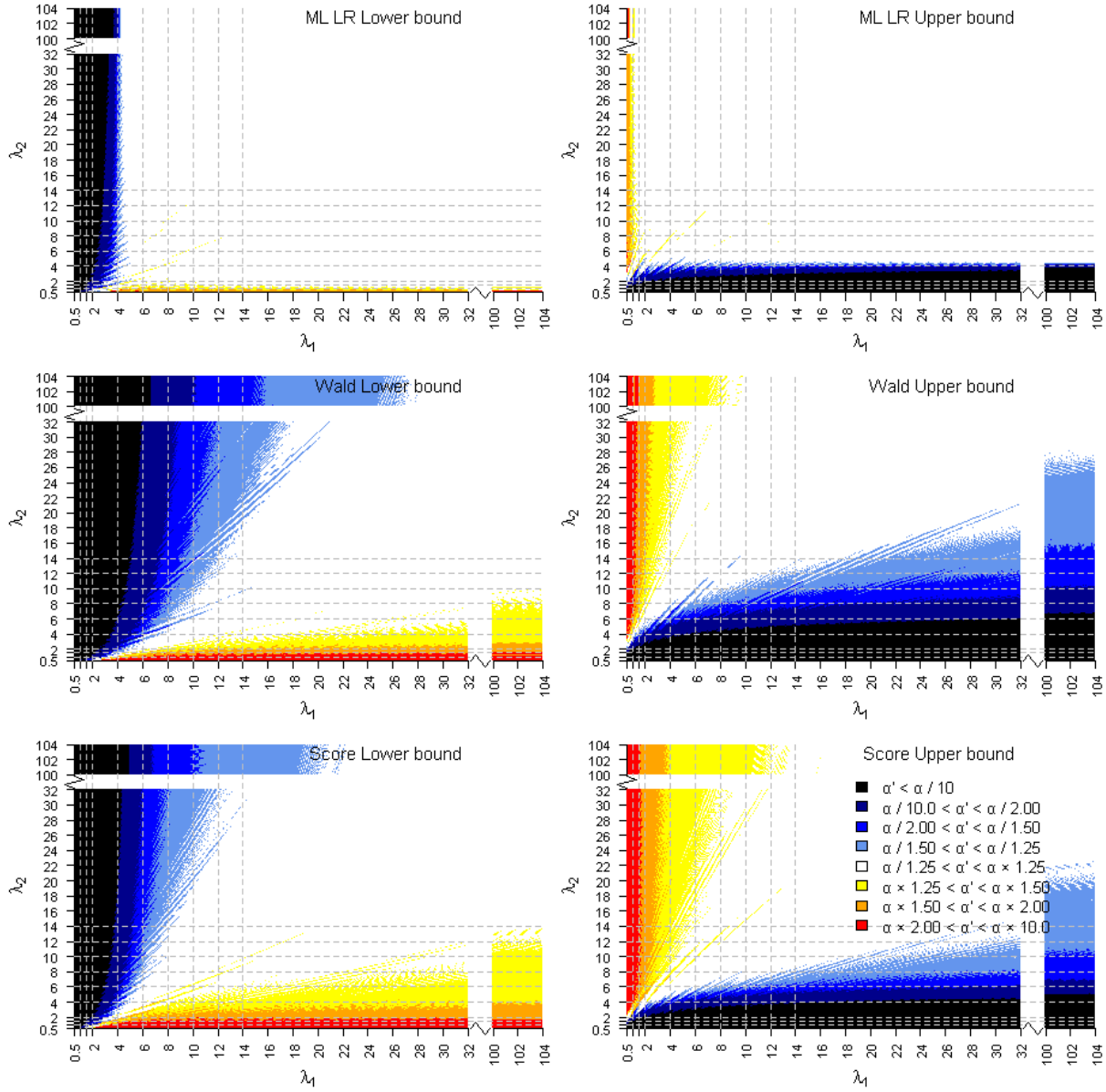


Figure 7: conditional α'_L and α'_U risks for the 95% two-sided likelihood ratio confidence interval (LR), 95% two-sided Wald confidence interval and 95% two-sided Score confidence interval of a ratio λ_2/λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the parameters of the Poisson distributions. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

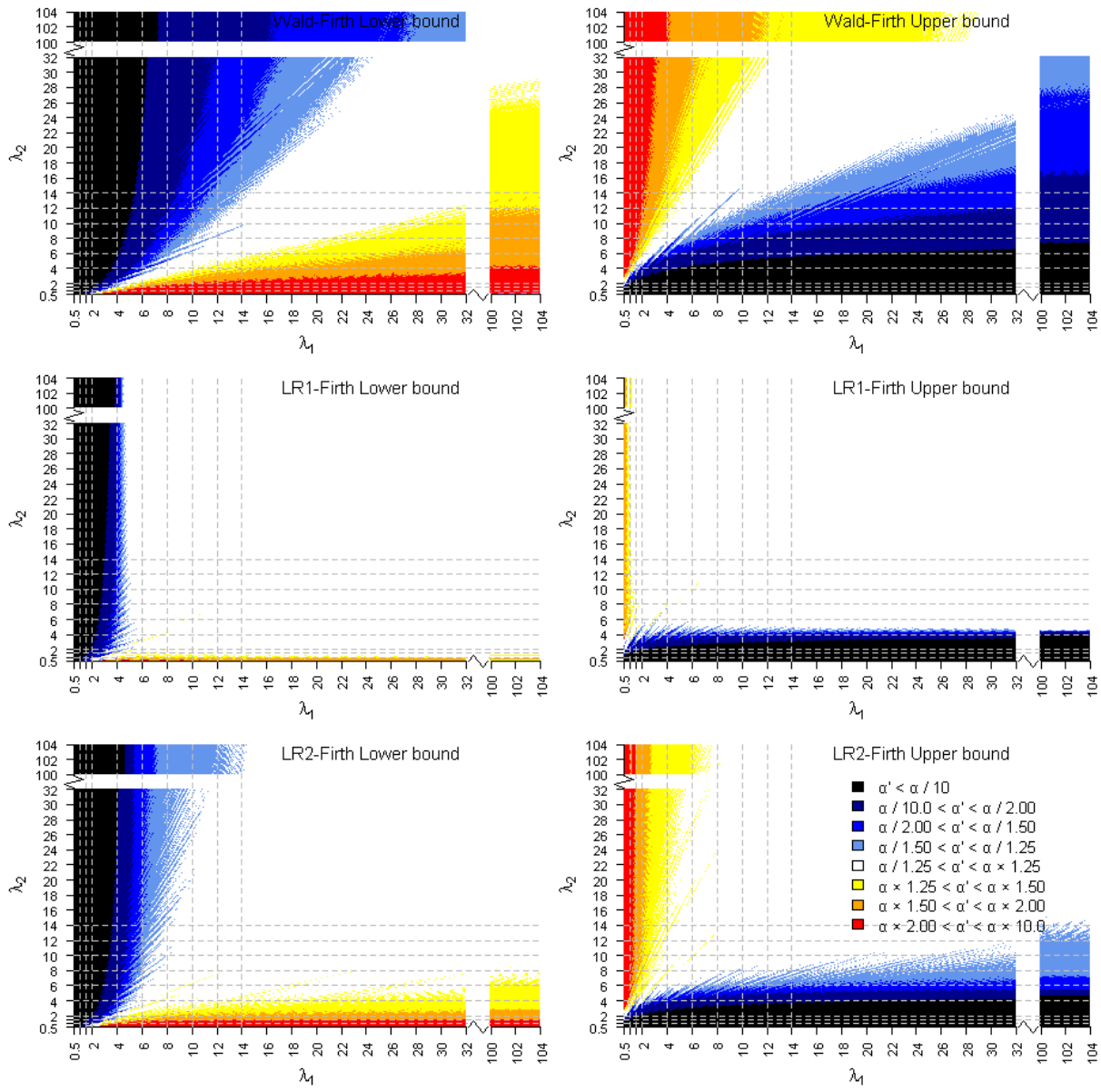


Figure 8: conditional α'_L and α'_U risks for the 95% two-sided Wald-Firth, LR1-Firth and LR2-Firth confidence intervals of a ratio λ_2/λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the parameters of the Poisson distributions. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

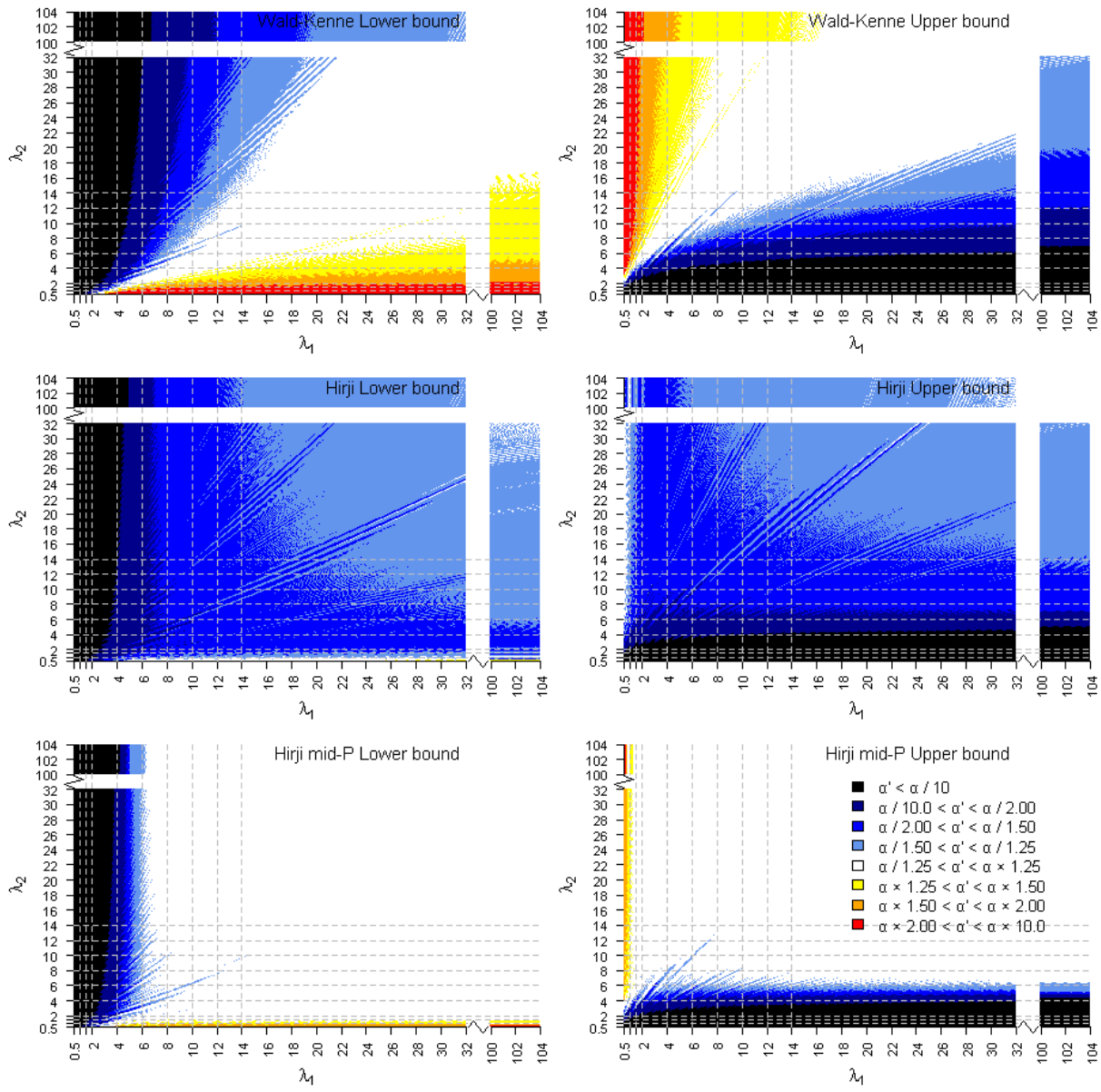


Figure 9: conditional α'_L and α'_U risks for the 95% two-sided Wald-Kenne, Hirji and Hirji mid-P confidence intervals of a ratio λ_2/λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the parameters of the Poisson distributions. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

Figures 4 to 6 show that the plane of unconditional risks is not perfectly smooth as had been shown for the estimation of a single proportion by Brown, Cai and DasGupta [19], but otherwise confirm what was shown in figures 1 to 3. Oscillations are moderate because even if λ_1 and λ_2 are assumed to be constant, $y_1 + y_2$ is variable.

10.5.3 Interval half-widths

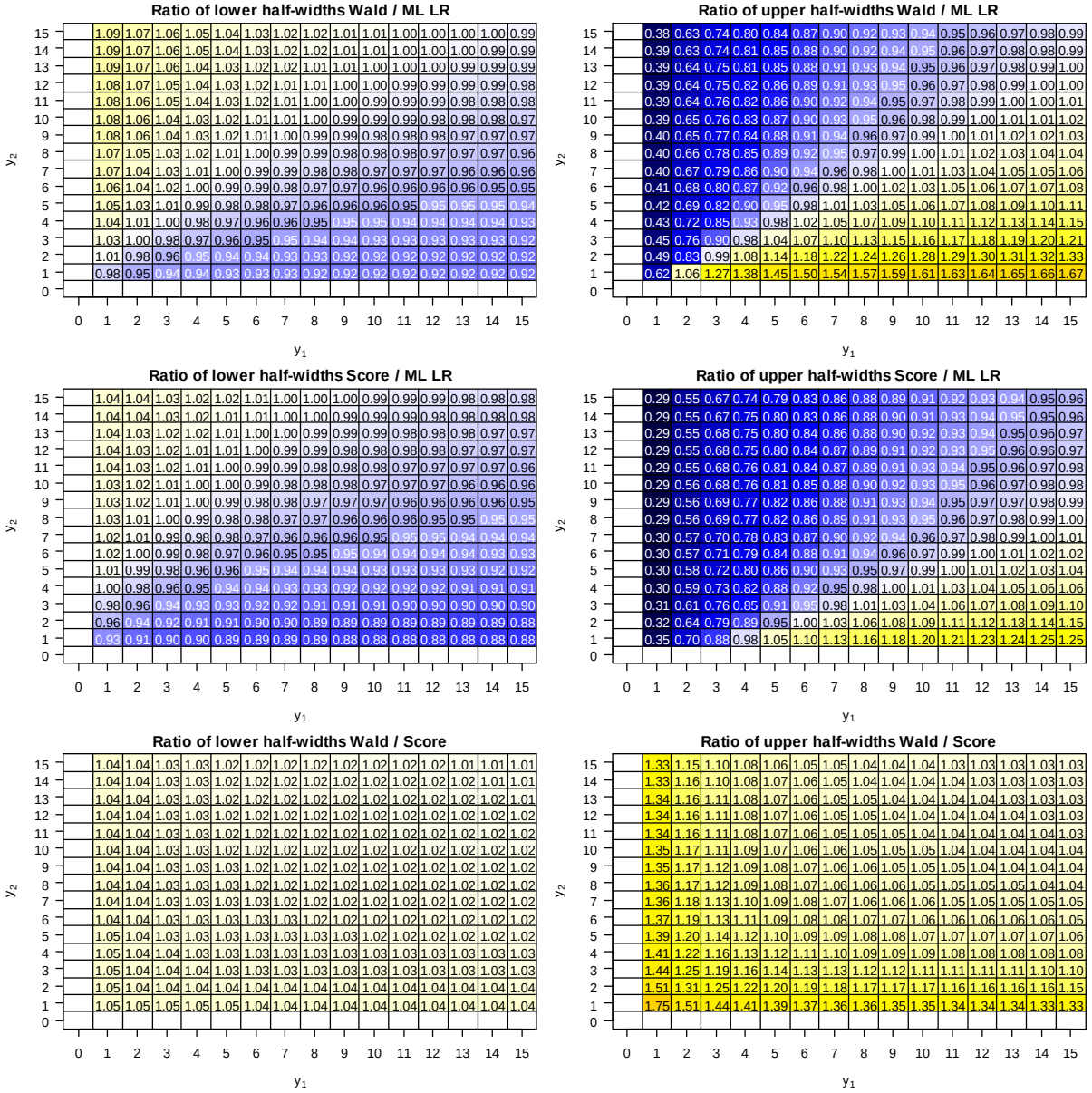


Figure 10: Ratio of half-widths (ML point estimate minus CI boundary) of Wald's, Score and ML LR 95% CI respectively to each other for y_1 (x axis) and y_2 (y axis), the observed number of events, from 1 to 15.

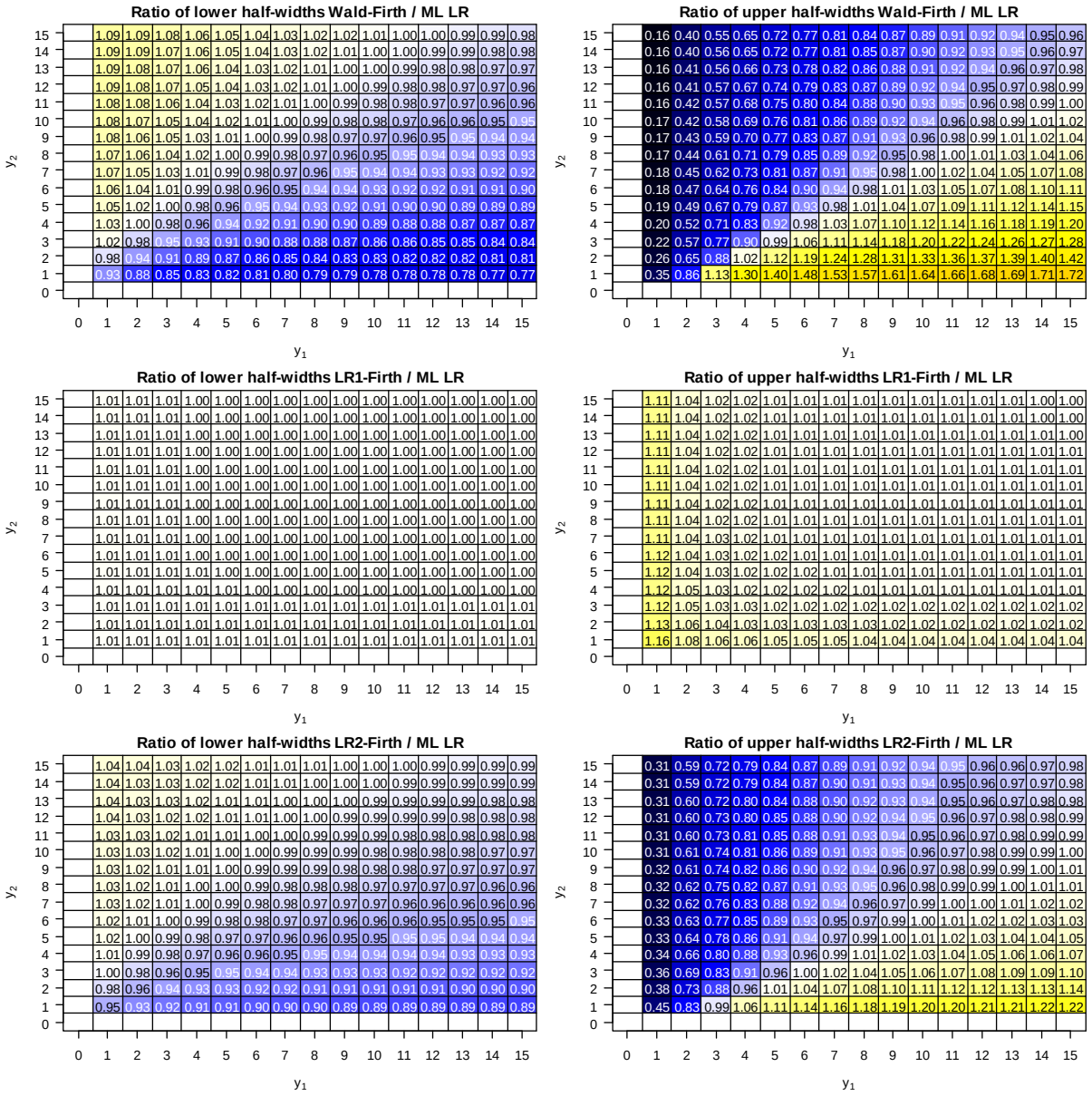


Figure 11: Ratio of half-widths (ML point estimate minus CI boundary) of the Wald-Firth, LR1-Firth and LR2-Firth 95% CI compared to the ML LR CI for y_1 (x axis) and y_2 (y axis), the observed number of events, from 1 to 15.

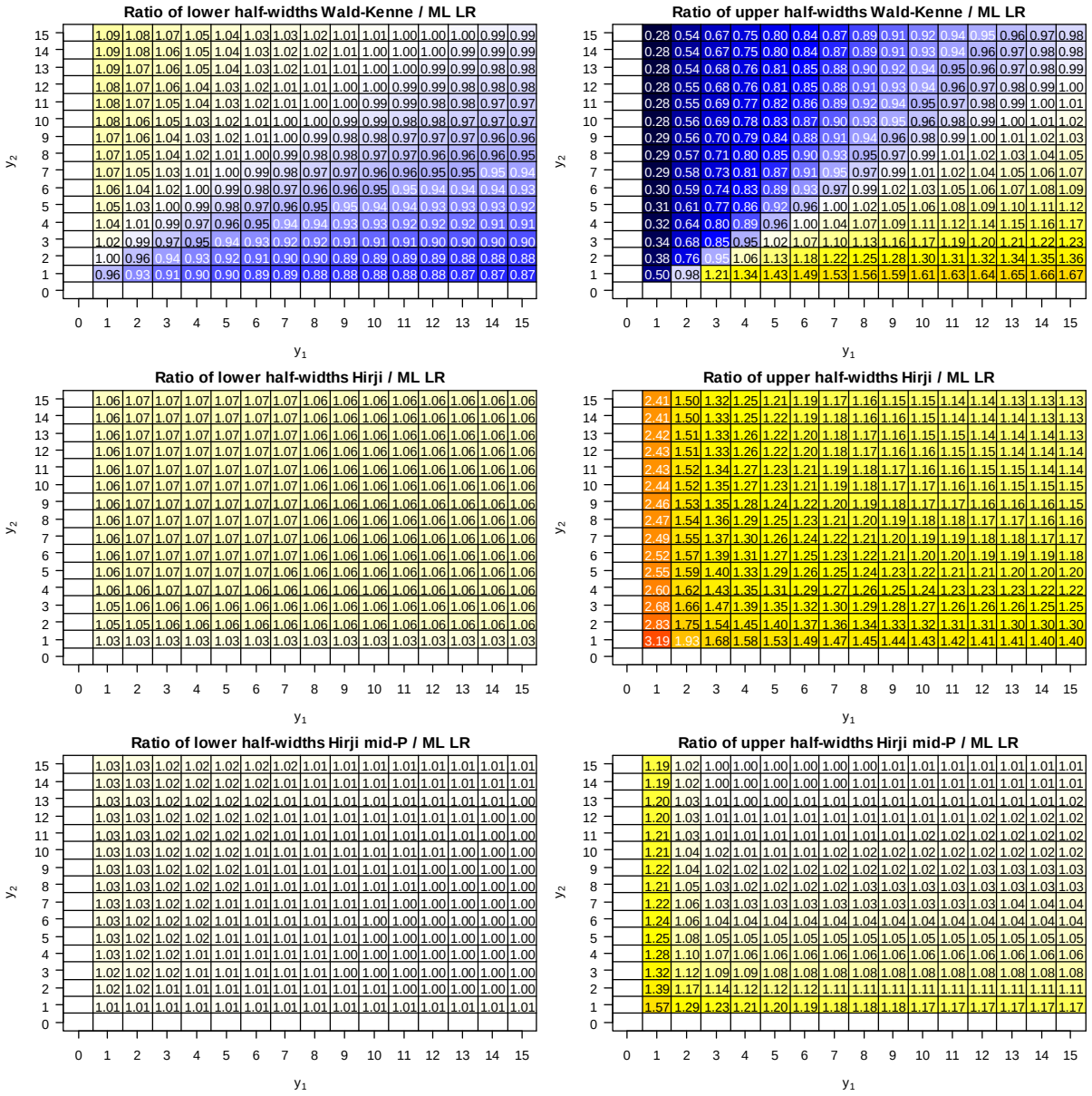


Figure 12: Ratio of half-widths (ML point estimate minus CI boundary) of the Wald-Kenne, Hirji and Hirji mid-P 95% CI compared to the ML LR CI for y_1 (x axis) and y_2 (y axis), the observed number of events, from 1 to 15.

Figures 7 to 10 show the half-widths of CI estimators relatively to the ML LR CI. Where a CI estimator is larger, it tends to over-cover and where it is shorter, it tends to under-cover.

10.5.4 Sensitivity analyzes: change in random distribution of $\boldsymbol{\Lambda_1}$ and $\boldsymbol{\Lambda_2}$ for unconditional risks

The log-normal distributions for Λ_1 and Λ_2 had a geometrical standard deviation equal to 1.10 in the primary analysis. Setting this geometrical standard deviation to 1.20 (respectively 1.05) changed, on

average, the absolute coverage error of 1.4×10^{-4} (resp 7.5×10^{-5}) on the whole set of points of Figures 1 to 3, with a 99th percentile equal to 1.0×10^{-3} (resp 6.2×10^{-4}) and maximum equal to 7.7×10^{-3} (resp 1.0%). The visual difference on the figures was negligible (not shown).

10.5.5 Secondary analyzes: change in confidence level

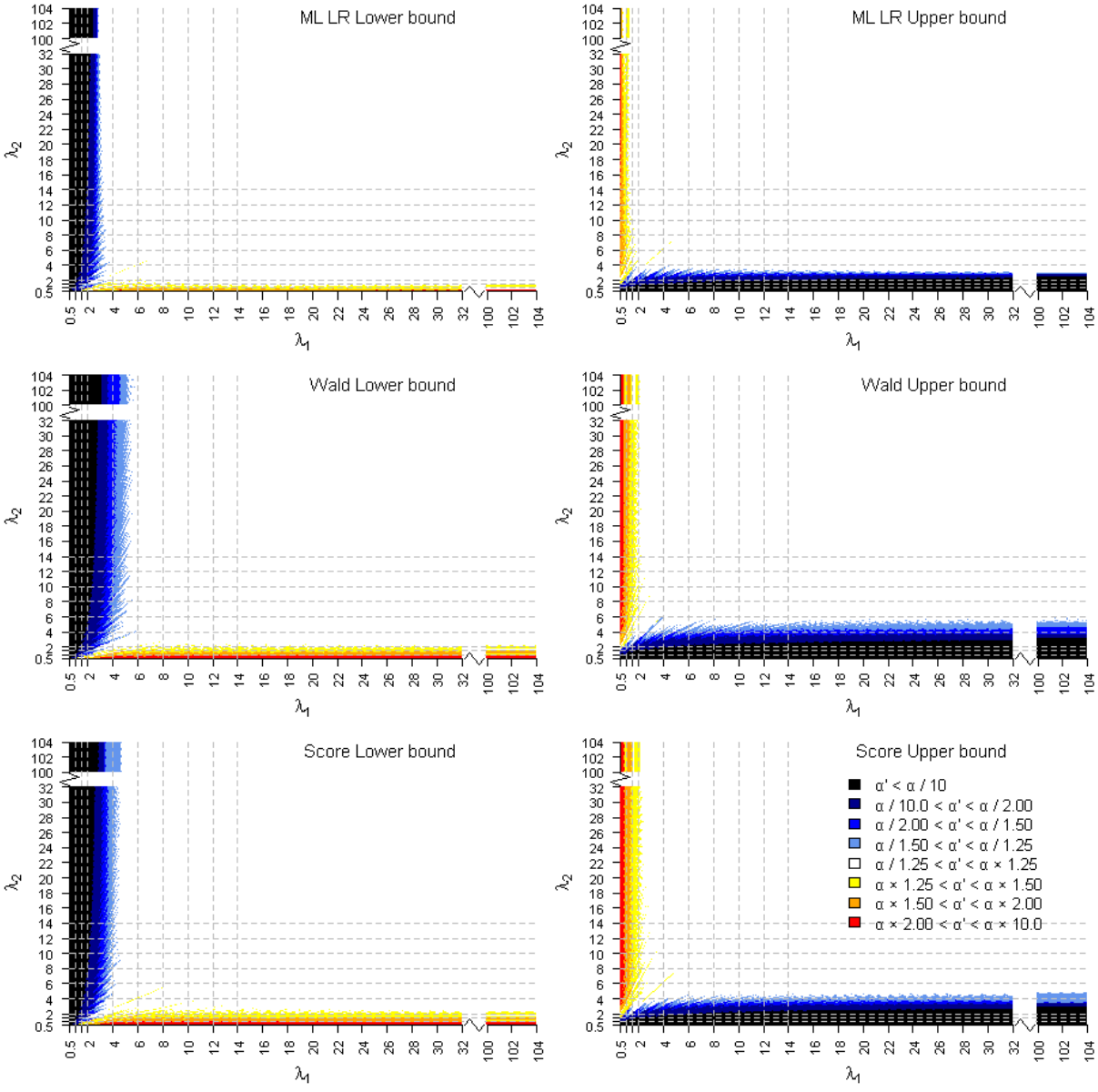


Figure 13: conditional α'_L and α'_U risks for the 80% two-sided likelihood ratio confidence interval (LR), 80% two-sided Wald confidence interval and 80% two-sided Score confidence interval of a ratio λ_2/λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the parameters of the Poisson distributions. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation

by a factor less than 1.5.

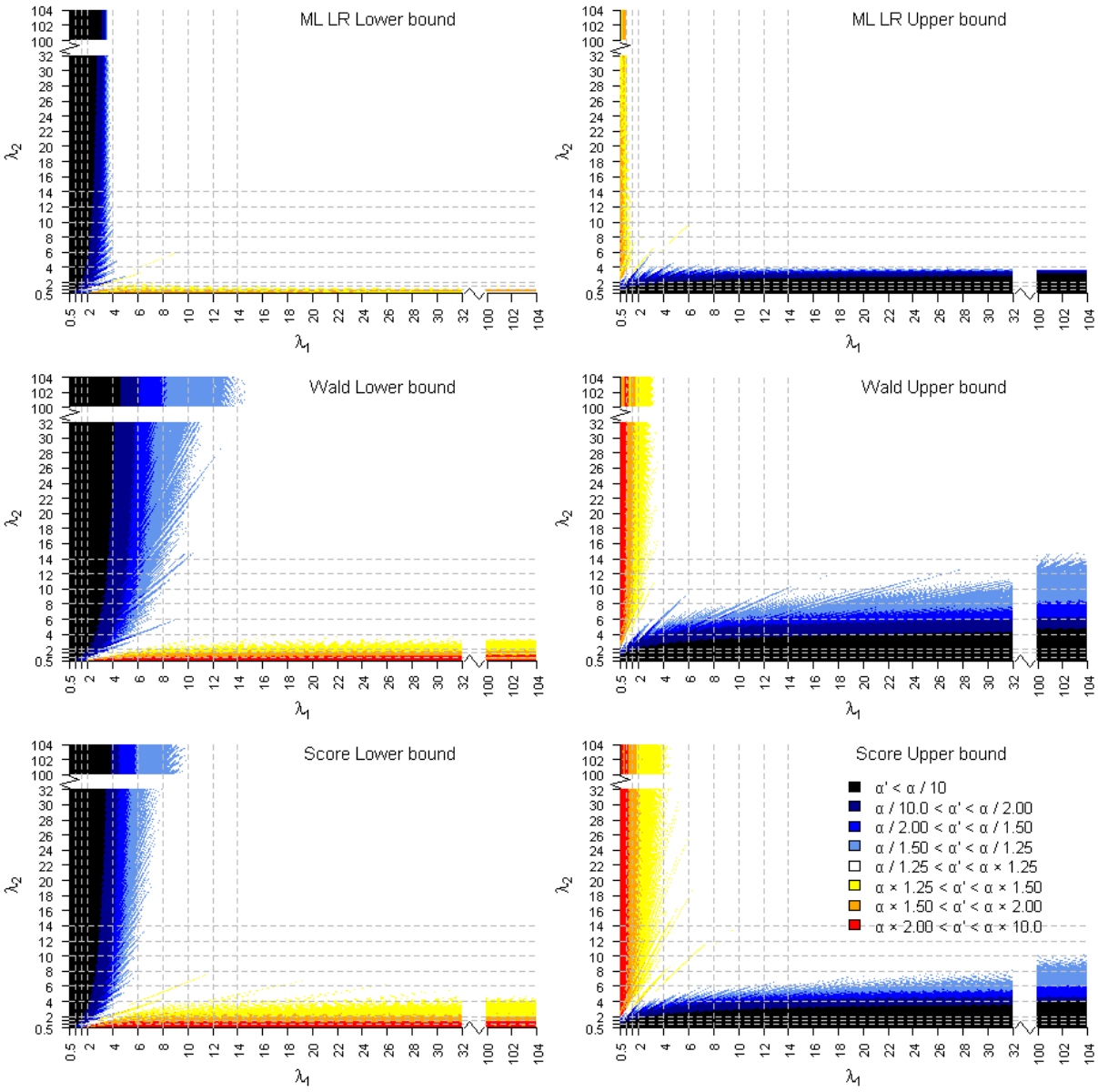


Figure 14: conditional α'_L and α'_U risks for the 90% two-sided likelihood ratio confidence interval (LR), 90% two-sided Wald confidence interval and 90% two-sided Score confidence interval of a ratio λ_2/λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the parameters of the Poisson distributions. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

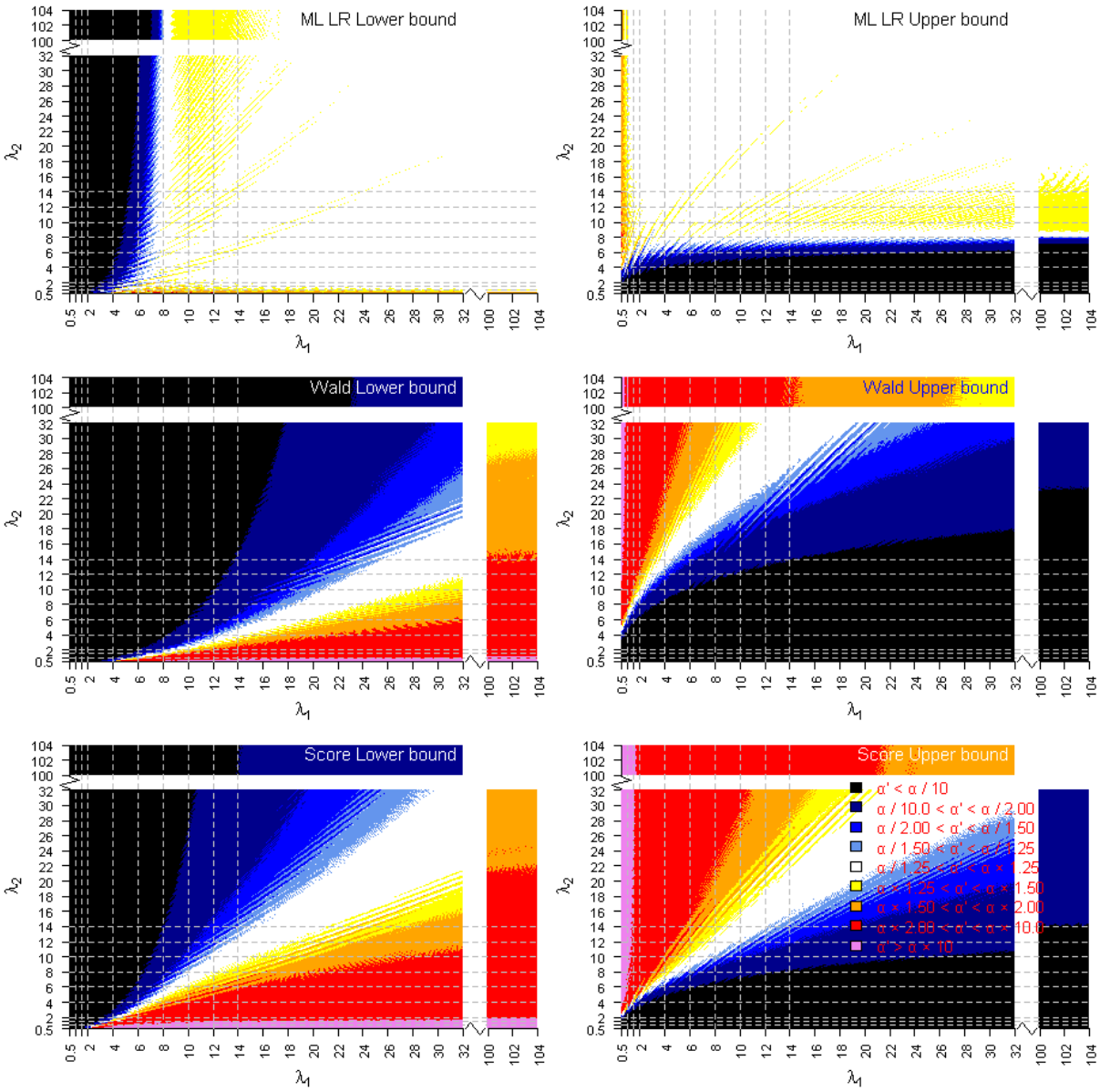


Figure 15: conditional α'_L and α'_U risks for the 99.9% two-sided likelihood ratio confidence interval (LR), 99.9% two-sided Wald confidence interval and 99.9% two-sided Score confidence interval of a ratio λ_2/λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the parameters of the Poisson distributions. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

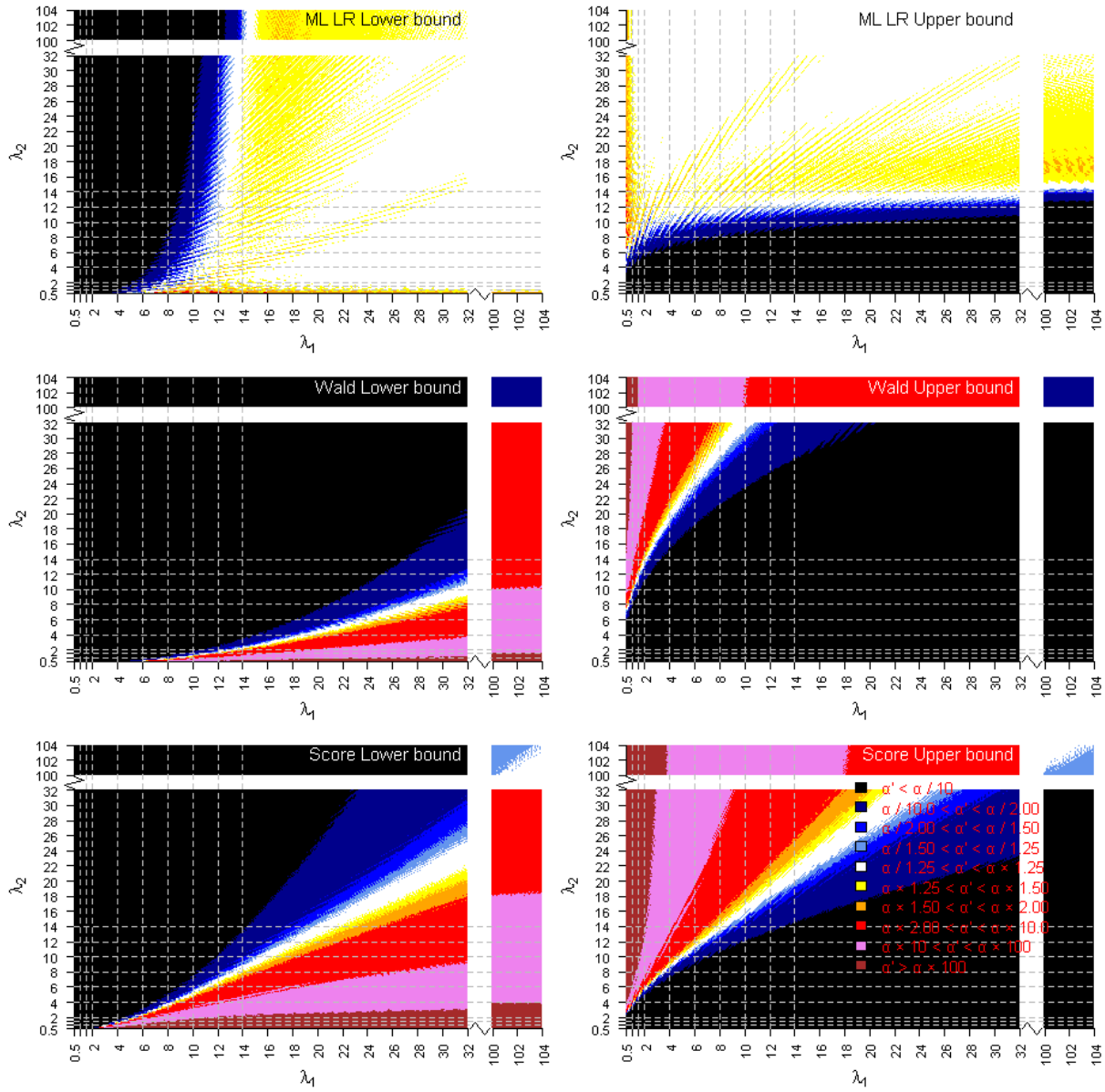


Figure 16: conditional α'_L and α'_U risks for the 99.9999% two-sided likelihood ratio confidence interval (LR), 99.9999% two-sided Wald confidence interval and 99.9999% two-sided Score confidence interval of a ratio λ_2/λ_1 of two Poisson variables, according to λ_1 (x axis) and λ_2 (y axis) the parameters of the Poisson distributions. The white, yellow and light blue zones show the zones of tolerated risk deflation/inflation by a factor less than 1.5.

Secondary analyzes show that coverage errors are worse for confidence levels close to 1. The ML LR CI behaves much better than Wald' and the Score CI for high confidence levels. Its main flaw is the unavoidable over-coverage when one or the other λ_i is small. For $2 \times \alpha = 10^{-6}$ (cf. Figure 16), the ML

LR CI has some light to moderate under-coverage even for quite large λ_1 and λ_2 values (above 20).

When both λ_i are large but that λ_2/λ_1 is far from 1, Wald's and the Score CI are biased. For instance, with $\lambda_1 = 100$ and $\lambda_2 = 400$ and $2 \times \alpha = 10^{-6}$, the conditional risk $\alpha'_L = \alpha \times 0.09$ and $\alpha'_U = \alpha \times 2.34$ for Wald's CI, $\alpha'_L = \alpha \times 0.18$ and $\alpha'_U = \alpha \times 2.91$ for the Score CI while $\alpha'_L = \alpha \times 1.08$ and $\alpha'_U = \alpha \times 0.96$ for the ML LR CI.

With $\lambda_1 = 100$ and $\lambda_2 = 400$ and $2 \times \alpha = 10^{-3}$, the conditional risk $\alpha'_L = \alpha \times 0.57$ and $\alpha'_U = \alpha \times 1.34$ for Wald's CI, $\alpha'_L = \alpha \times 0.63$ and $\alpha'_U = \alpha \times 1.41$ for the Score CI while $\alpha'_L = \alpha \times 1.05$ and $\alpha'_U = \alpha \times 0.97$ for the ML LR CI.

10.6 Discussion

10.6.1 Unconditional and conditional risks

Some "exact" CI estimators, such as Hirji's estimator [9], are designed to be strictly conservative. The actual coverage is always greater or equal to the nominal coverage. This estimator, like all other estimators described in this work, is conditional to the total number of events $y_1 + y_2$. For some values of $y_1 + y_2$ the actual coverage is close to 95% while it is closer to 97.5% for other values of $y_1 + y_2$. Hirji's estimator for the bivariate Poisson problem is equivalent to the Clopper-Pearson CI for a binomial distribution and the oscillations alongside $y_1 + y_2$ found by Brown Cai and DasGupta [19] apply to Hirji's estimator too. Since actually, $Y_1 + Y_2$ is random, oscillations are averaged and the actual coverage is always strictly greater than 95%. When analyzing the α'_L and α'_U risks with a random $Y_1 + Y_2$, Hirji's CI is too conservative (see Figure 9) leading to a wide interval (see Figure 12). As discussed in the rationale of the Materials & Methods section, not only Y_1 and Y_2 are random, but λ_1 and λ_2 should be considered random too, leading to even smoother unconditional coverage errors α''_L and α''_U. The actual variance of the distributions of Λ_1 and Λ_2 is unknown but does not matter much as long as it is continuous, as shown in the sensitivity analysis.

Averaging risks is not a new idea in the assessment of CI estimator bias. For instance, Agresti and Coull had shown that "exact" CI estimators tended to be too conservative when the average risk was analyzed [20]. Nevertheless, they had averaged the risk over the whole space of theoretical values. They

concluded that the score CI (Wilson's score CI) behaved well, while actually, undercoverage for some parameters may compensate over-coverage for other distant values, as we observed on Figure 4. Our unconditional coverage analysis performs local averaging only which is more relevant to a particular setting where parameters λ_1 and λ_2 are moderately variable.

10.6.2 What is the best CI estimator?

The score and Wald's CI showed coverage bias much higher than the ML LR CI or Hirji's estimator with mid-P. Penalized likelihood estimators (Firth's and Kenne's estimators) perform poorly with Wald's confidence boundaries. One may notice that the Wald-Firth estimator is equivalent to Wald's CI in a maximum likelihood model after adding 0.5 events in both groups (analysis not shown). The main advantage of penalized likelihood estimators over the ML estimator is the bias reduction of the point estimate in small samples and availability of an estimation when the number of events is zero in one group. The LR1-Firth CI behaves well but cannot be computed when zero events are observed in one group (complete separation). Indeed, the likelihood profile is monotone. The LR2-Firth CI can be computed in case of complete separation but has poor coverage properties. It is equivalent to a ML LR CI after adding 0.5 events in both groups (analysis not shown). Hirji's estimator is too conservative while Hirji's mid-P estimator is quite good but a bit wider and more conservative than the ML LR CI. The ML LR CI being theoretically simple, widely available and having very good coverage properties, we recommend its use in almost all scenarios.

10.6.3 When is the ML LR CI valid?

In the simple bivariate scenario, the ML LR CI always has tolerable coverage bias, for low or high confidence levels, as soon as the expected number of events is above 1 in both groups. When an experiment is performed, nobody knows the expected number of events λ_1 and λ_2. One can safely assume that both $\lambda_1 > 1$ and $\lambda_2 > 1$ if both $y_1 \geq 3$ and $y_2 \geq 3$. The highest risk of having both $y_i \geq 3$ while one of the λ_i is actually below or equal to 1 occurs when one of the λ_i is high but the other is equal to 1. In that case, the risk of having both $y_i \geq 3$ is 8% according to the quantiles of the Poisson distribution $P(1)$. When one or both y_i are between 1 and 3, we suggest to provide the ML LR CI if the estimation was planned, anyway. Not doing so would bias the literature towards overestimating the

smaller λ_i because the CI would not be provided when it is underestimated. We recommend that, in the first place, the estimation of λ_2/λ_1 be not attempted at all if there are prior information that suggests that either λ_1 or λ_2 is below 1. Indeed, the statistical precision will be unacceptable well before the statistical estimator is invalid. When $y_1 = 0$ or $y_2 = 0$ we assumed that the CI would not be provided so that all results of our analysis are conditional to both $y_i \geq 1$. Therefore, this selective reporting bias is taken in account in our results.

10.6.4 Unusual scenario where one $\boldsymbol{\lambda_i}$ is large but not the other

Sometimes, one λ_i may be expected to be large while the other is expected to be close to zero although the offsets (e.g. number of patients-years at risk) are not very different. This may happen when the outcome is expected to (almost) completely disappears in a group. In that case the estimation of the absolute risk difference in a linear model may be more relevant than the risk ratio and we suggest not to attempt estimation of the risk ratio in the first place. The second scenario occurs when the two groups have very different offsets (e.g. number of patients-years at risk) although the actual ratio assessed $\frac{\lambda_2/r_2}{\lambda_1/r_1}$ may not be very far from 1. In that case one of the λ_i may be close to zero and the estimation of the ratio almost impossible due to poor statistical precision. In that case, we advise to change the design of the study (e.g. cohort -> case-control). In all other cases, both λ_i are expected to be large enough to warrant good coverage to the ML LR CI.

10.6.5 Hypothesis tests

Since all CI estimators analyzed can be obtained by inversion of hypothesis tests, our results apply to hypothesis tests as well. As expected, approximations are best for large P-values, and bias may be enormous in a scenario where the significance level is very close to zero (see Figure 15 and Figure 16). This may happen in a multiple-testing scenario with thousands of tests. In that case, the Wald and score tests are very biased, even when the number of events is quite large ($\lambda_1 = 100$ and $\lambda_2 = 400$). Due to different offsets in both groups, the ratio actually tested may be equal to 1 but the offset (e.g. number of patients-years at risk) r_2 may be four times larger than r_1 so the $\lambda_2 = 4 \times \lambda_1$ too. Even outside of the multiple testing scenario, Wald's and the Score CI may provide a wrong impression of confidence when

the P-value is very small.

10.6.6 Software implementation

By default R statistical software uses the ML LR CI for generalized linear models with the function confint but the function summary on GLMs uses Wald's method for hypothesis tests. The package glmglrt, created by the author of this article and available on the Comprehensive R archive Network (CRAN), adds a column “LR P value” to the summary.glm function without changing other columns, for backwards compatibility. This function is compatible with most features of GLM (weights, offsets, dropped coefficients). When convergence in the null hypothesis is impossible (quite frequent for Intercept of log-binomial models), it provides "NA" in place of the P-value.

SAS (Statistical Analysis System Institute, Inc., Cary, NC) software provides options "LRCI" and "type3" to the MODEL statement in the GENMOD procedure, providing respectively ML LR CI and LR tests. Unfortunately, if a categorical variable is included in a model, it's automatically recoded as several binary variables but the "type3" statement tests the hypothesis that all coefficients of the variable are equal to zero at once. Manual recoding of categorical variables to binary variables is required.

Stata (StataCorp LP, TX, USA) provides the command `pllf` [21] to compute profile-likelihood CIs in various models and LR hypothesis tests can be performed by the `lrtest` command; the latter requires fitting two models and saving estimates with the `estimates store` command.

SPSS does not provide any easy way of computing ML LR CI or performing LR tests.

10.6.7 Limits of this work

The case where one y_i is zero (complete separation) has been excluded from all analyzes. Therefore, our results apply to the case where the statistician do not provide any CI in that case. These results do not apply to a statistician who prefer to provide Hirji's exact CI or a penalized likelihood CI when one of the y_i is zero. As discussed above, the study should be designed to avoid these cases in the first place.

The logistic regression has not been analyzed. The Poisson regression has been considered as the limit case of the logistic regression, but actually, the logistic regression may behave quite differently when proportions are close to 50%. Since the bias of Wald's CI is mainly due to the skewness of the ML estimator (data not shown), the case where both proportions are close to 50% may be optimal, while the

case where one proportion is close to 0 or 100% and the other is on the opposite side of the scale, should be the worst.

The effect of additional covariates (multivariate regression) has not been analyzed. The problems found in the simple bivariate scenario are expected to be found in multivariate regressions, but additional problems, such as inflation of effects due to overfitting are expected with the ML LR CI. The overfitting problem should be slightly weaker with penalized likelihood CIs. One may note that penalized likelihood estimation is not the same as penalized regression (LASSO, Ridge Regression, Elastic-Net). The former avoids point estimate bias but should have a very small effect on over-fitting.

Half-widths of CI estimators can hardly be compared when their coverage bias is not identical. Computing half-widths after correction of the coverage bias by inflation/deflation of the confidence level was attempted. A global correction would not make sense as it could not correct local coverage errors. A local correction made more sense but would transform so much the CI estimator that all CI estimators would become almost identical, with half-width differences that would be due to the method of correction rather than to real differences. Fortunately, for complex multivariate statistics, some estimators may have much better performances (lower variance and same bias) than others, but for statistics as simple as a rate ratio, that does not seem to be the case.

10.7 Conclusion

From theoretical considerations and results of this work, we advise not to attempt assessment of an incidence ratio when the frequency in one group is expected to be less than 1 on average. Otherwise we advise the LR hypothesis test be performed and the ML LR CI be computed for the rate ratio estimated in Poisson regressions with log link as long as there is no group where the number of events is zero (complete separation). In case of complete separation, one may not publish the rate ratio CI (strategy assessed in this work), or Firth's estimator may be considered (not assessed in this work).

10.8 References

1. Buse A. The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. Am Stat. 1982;36(3):153–7.
2. De Maio A, Kay SM, Farina A. On the invariance, coincidence, and statistical equivalence of the GLRT, Rao test, and Wald test. IEEE Trans Signal Process. 2009;58(4):1967–1979.
3. Rao CR. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Math Proc Camb Philos Soc. 1948 Jan;44(1):50–7.
4. Wilks SS. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. Ann Math Stat. 1938 Mar;9(1):60–2.
5. Hauck WW, Donner A. Wald's Test as Applied to Hypotheses in Logit Analysis. J Am Stat Assoc. 1977;72(360):851–3.
6. Neyman J, Pearson ES, Pearson K. IX. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond Ser Contain Pap Math Phys Character. 1933 Feb 16;231(694–706):289–337.
7. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993 Mar 1;80(1):27–38.
8. Kenne Pagui EC, Salvan A, Sartori N. Median bias reduction of maximum likelihood estimates. Biometrika. 2017 Dec 1;104(4):923–38.
9. Hirji KF, Mehta CR, Patel NR. Computing Distributions for Exact Logistic Regression. J Am Stat Assoc. 1987;82(400):1110–7.
10. Olgin JE, Pletcher MJ, Vittinghoff E, Wranicz J, Malik R, Morin DP, et al. Wearable Cardioverter–Defibrillator after Myocardial Infarction. N Engl J Med [Internet]. 2018 Sep 26 [cited 2019 Oct 23]; Available from: https://www.nejm.org/doi/10.1056/NEJMoa1800781
11. Dyrbye LN, Burke SE, Hardeman RR, Herrin J, Wittlin NM, Yeazel M, et al. Association of Clinical Specialty With Symptoms of Burnout and Career Choice Regret Among US Resident Physicians. JAMA. 2018 Sep 18;320(11):1114–30.

12. Azoulay E, Lemiale V, Mokart D, Nseir S, Argaud L, Pène F, et al. Effect of High-Flow Nasal Oxygen vs Standard Oxygen on 28-Day Mortality in Immunocompromised Patients With Acute Respiratory Failure: The HIGH Randomized Clinical Trial. JAMA. 2018 Nov 27;320(20):2099–107.
13. Alexander S, Fisher BT, Gaur AH, Dvorak CC, Luna DV, Dang H, et al. Effect of Levofloxacin Prophylaxis on Bacteremia in Children With Acute Leukemia or Undergoing Hematopoietic Stem Cell Transplantation: A Randomized Clinical Trial. JAMA. 2018 Sep 11;320(10):995–1004.
14. Vollenhoven RF van, Hahn BH, Tsokos GC, Wagner CL, Lipsky P, Touma Z, et al. Efficacy and safety of ustekinumab, an IL-12 and IL-23 inhibitor, in patients with active systemic lupus erythematosus: results of a multicentre, double-blind, phase 2, randomised, controlled study. The Lancet. 2018 Oct 13;392(10155):1330–9.
15. Zhang WR, Craven TE, Malhotra R, Cheung AK, Chonchol M, Drawz P, et al. Kidney Damage Biomarkers and Incident Chronic Kidney Disease During Blood Pressure Reduction: A Case–Control Study. Ann Intern Med. 2018 Nov 6;169(9):610.
16. Lowe WL, Scholtens DM, Lowe LP, Kuang A, Nodzenski M, Talbot O, et al. Association of Gestational Diabetes With Maternal Disorders of Glucose Metabolism and Childhood Adiposity. JAMA. 2018 Sep 11;320(10):1005–16.
17. Goldacre B, DeVito NJ, Heneghan C, Irving F, Bacon S, Fleminger J, et al. Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource. BMJ. 2018 Sep 12;362:k3218.
18. Harris PNA, Tambyah PA, Lye DC, Mo Y, Lee TH, Yilmaz M, et al. Effect of Piperacillin-Tazobactam vs Meropenem on 30-Day Mortality for Patients With E coli or Klebsiella pneumoniae Bloodstream Infection and Ceftriaxone Resistance: A Randomized Clinical Trial. JAMA. 2018 Sep 11;320(10):984–94.
19. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. Stat Sci. 2001 May;16(2):101–33.

20. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. Am Stat. 1998;52(2):119–126.

21. Royston P. Profile Likelihood for Estimation and Confidence Intervals. Stata J Promot Commun Stat Stata. 2007 Sep;7(3):376–87.

11Annexe 4 : équivalence des tests d'hypothèses du Poisson bivarié et binomial univarié

11.1 Maximum de vraisemblance dans un modèle de Poisson bivarié

Considérons $Y_1 \sim \mathcal{P}(\lambda_1)$ et $Y_2 \sim \mathcal{P}(\lambda_2)$ deux variables aléatoires indépendantes suivant des lois de Poisson. Considérons le problème d'estimation de $\frac{\lambda_1}{\lambda_2}$ dans un modèle linéaire généralisé à fonction de lien log et distribution de Poisson :

On définit la matrice de modèle $x = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ et $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ le paramètre du modèle.

Le modèle s'exprime par $\log(E[Y]) = x\beta$ où $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$

soit $\log(E[Y_1]) = \beta_0$ et $\log(E[Y_2]) = \beta_0 + \beta_1$ de telle sorte que $\beta_1 = \log\left(\frac{E[Y_2]}{E[Y_1]}\right)$

La fonction de log-vraisemblance de β_0, conditionnelle à une réalisation y et un β_1 donnés, s'exprime comme :

$$L_{\beta_0}(\beta_0; y, \beta_1) = y_1\beta_0 - \exp(\beta_0) + y_2(\beta_0 + \beta_1) - \exp(\beta_0 + \beta_1) - \log(y_1!) - \log(y_2!)$$

Cette fonction se dérive facilement selon β_0 et la racine (zéro) de la dérivée se trouve simplement, ce qui permet d'en déduire le $\widehat{\beta_0}$ du maximum de vraisemblance conditionnel à un β_1 fixé :

$$\widehat{\beta_0} = \frac{y_1 + y_2}{\exp(\beta_1) + 1}$$

11.2 Équivalence des problèmes binomial et de Poisson conditionnel

Posons $Y_1 \sim \mathcal{P}(\lambda_1)$ et $Y_2 \sim \mathcal{P}(\lambda_2)$ deux variables indépendantes suivant des lois de Poisson et définissons une variable Z égale à la somme des deux :

$$Z = Y_1 + Y_2 \sim \mathcal{P}(\lambda_1 + \lambda_2)$$

Pour un $n \in \mathbb{N}$, Notons $\widetilde{Y_1}$ et $\widetilde{Y_2}$ les lois de Y_1 et Y_2 conditionnelles à $Z = n$.

Nous cherchons à prouver que la loi de $\widetilde{Y_1}$ est une loi binomiale $\mathcal{B}\left(n; \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$.

11.2.1 Approche intuitive

Intuitivement, lorsque Y_1 et Y_2 sont générées par deux processus de Poisson indépendants, on peut comprendre que, conditionnellement à un $y_1 + y_2$ donné, chacun des événements ayant participé à cette somme a une cote de λ_1/λ_2 d'appartenir au premier processus de Poisson et λ_2/λ_1 d'appartenir au second processus. Pour simplifier le schéma, on peut imaginer que les deux processus de Poisson surviennent sur la même période de temps, mais restent indépendants. On numérote alors chacun des événements survenant dans l'un ou l'autre des processus, de 1 à n dans l'ordre de leur apparition. On peut alors assigner à chacun des événements une variable binaire dont les modalités sont l'appartenance au premier ou au second processus. En recodant cette variable comme une variable de Bernoulli qui vaut 1 lorsque l'événement appartient au premier processus de Poisson, 0 sinon, on en arrive intuitivement à des variables de Bernoulli indépendantes et identiquement distribuées de probabilité $\frac{\lambda_1}{\lambda_1+\lambda_2}$. En conséquence, leur somme, représentant le nombre d'événements issus du premier processus de Poisson, devrait suivre une loi binomiale $\mathcal{B}\left(n; \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$. Une démonstration plus rigoureuse suit ci-dessous. Le cas où $n = 0$ et $n = 1$ n'ont rien de particulier sinon qu'ils ont participé de base de réflexion à l'auteur de ce document. Ainsi, vous pouvez directement vous référez au cas général pour la démonstration.

11.2.2 Démonstration

$$\widetilde{Y_2} = n - \widetilde{Y_1}$$

Le support de $\widetilde{Y_1}$ est $\{0, \dots, n\}$

Pour $k \in \{0, \dots, n\}$, $\mathbb{P}\left(\widetilde{Y_1} = k\right) = \mathbb{P}(Y_1 = k\ et\ Y_2 = n - k | Y_1 + Y_2 = k) = \frac{\mathbb{P}(Y_1 = k\ et\ Y_2 = n-k)}{\mathbb{P}(Y_1+Y_2=n)}$

$$\mathbb{P}(Y_1 + Y_2 = n) = \mathbb{P}(Z = n) = \frac{\exp\left(-(\lambda_1 + \lambda_2)\right)(\lambda_1 + \lambda_2)^n}{n!}$$

car $Z \sim \mathcal{P}(\lambda_1 + \lambda_2)$

Par ailleurs, par indépendance de Y_1 et Y_2 on a :

$$\mathbb{P}(Y_1 = k\ et\ Y_2 = n - k) = \mathbb{P}(Y_1 = k)\,\mathbb{P}(Y_2 = n - k)$$

$$= \frac{\exp(-\lambda_1)\,\lambda_1^k \exp(-\lambda_2)\,\lambda_2^{n-k}}{k!\,(n-k)!} = \exp\left(-(\lambda_1 + \lambda_2)\right)\frac{\lambda_1^k \lambda_2^{n-k}}{k!\,(n-k)!}$$

On en déduit :

$$\mathbb{P}(\widetilde{Y_1} = k) = \exp\big(-(\lambda_1 + \lambda_2)\big)\frac{\lambda_1^k \lambda_2^{n-k}}{k!\,(n-k)!}\left(\frac{\exp\big(-(\lambda_1 + \lambda_2)\big)\,(\lambda_1 + \lambda_2)^n}{n!}\right)^{-1}$$

$$= \frac{n!\,\lambda_1^k \lambda_2^{n-k}}{k!\,(n-k)!\,(\lambda_1 + \lambda_2)^n} = \binom{n}{k}\frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n} = \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k}$$

$$= \binom{n}{k}\pi_1^k (1 - \pi_1)^{n-k}$$

Où $\pi_1 = \frac{\lambda_1}{\lambda_1+\lambda_2}$, ce qui correspond à la fonction de masse d'une loi binomiale $\mathcal{B}\left(n; \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$

11.3 Équivalence des tests du rapport de vraisemblance et du score de Rao

Pour montrer l'équivalence de l'intervalle de confiance, à transformation près, il suffit de montrer l'équivalence du test du rapport de vraisemblance de Poisson au test du rapport de vraisemblance binomial.

Étant donné $y_1, y_2 \in \mathbb{N}$ strictement positifs tous deux et $\rho \in \mathbb{R}^{+*}$ le rapport de risques correspondant au $\frac{\lambda_2}{\lambda_1}$ de l'hypothèse nulle que l'on cherche à rejeter si y_1 et y_2 sont interprétés comme des réalisations de variables $Y_1 \sim \mathcal{P}(\lambda_1)$ et $Y_2 \sim \mathcal{P}(\lambda_2)$.

Définissons un modèle linéaire généralisé de Poisson bivarié avec pour matrice de modèle $x = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, $Y_1 \sim \mathcal{P}(\lambda_1)$ et $Y_2 \sim \mathcal{P}(\lambda_2)$, $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ et $\log(E[Y]) = x\beta$ où $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ est le paramètre du modèle.

Définissons un modèle linéaire généralisé logit-proportion-binomiale univarié avec $Z \sim \frac{1}{y_1+y_2}\mathcal{B}(y_1 + y_2; \pi)$ et $\text{logit}(E[Z]) = \gamma$ où γ est le paramètre du modèle. La matrice du modèle de dimension 1×1 est une constante égale à 1.

Supposons que nous estimions, par la méthode fréquentiste du maximum de vraisemblance, $\hat{\beta}$ et $\hat{\gamma}$, en utilisant, pour la régression de Poisson, $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ comme réalisation de Y, et pour la régression logit-proportion-binomiale $z = \frac{y_2}{y_1+y_2}$ comme réalisation de Z.

Il est assez aisé de voir que l'estimateur du maximum de vraisemblance (ajustement parfait, aucune différence entre valeur observée et prédite par l'estimation du maximum de vraisemblance) arrive aux

estimations $\hat{\beta} = \begin{pmatrix} \log(y_1) \\ \log(y_2/y_1) \end{pmatrix}$ et $\hat{\gamma} = \log(y_2/y_1)$. Ainsi $\widehat{\beta_1} = \hat{\gamma}$. Nous allons montrer l'équivalence de l'inférence sur β_1 et de l'inférence sur γ par les méthodes du rapport de vraisemblance et du score.

La vraisemblance d'un γ quelconque dans le modèle binomial univarié (pour la réalisation $z = \frac{y_2}{y_1+y_2}$) est égale à $L_{binom}(\gamma) = \mathbb{P}\left(\mathcal{B}\left(y_1 + y_2; \frac{\exp(\gamma)}{1+\exp(\gamma)}\right) = y_2\right)$.

Nous voudrions démontrer que la vraisemblance d'un β_1 quelconque dans le modèle de Poisson univarié, conditionnellement au $\widehat{\beta_0}$ maximisant la vraisemblance de β_1, est égale à $L_{poisson}(\beta_1) = L_{binom}(\gamma) \times \mathbb{P}(\mathcal{P}(y_1 + y_2) = y_1 + y_2)$, ce qui rendra alors équivalentes les inférences sur les deux fonctions de vraisemblance car elles ne diffèreraient que d'un facteur constant, conduisant à une inférence du rapport de vraisemblance identique, la constante disparaissant dans le rapport de vraisemblance, ainsi qu'une inférence du score identique, le score et l'information de Fisher étant inchangées par un facteur multiplicateur constant sur la vraisemblance.

Pour un β_1 fixé, le β_0 du maximum de vraisemblance dans le modèle de Poisson est égal à

$$\widehat{\beta_0} = \frac{y_1 + y_2}{\exp(\beta_1) + 1}$$

Ceci avait été montré en section 11.1.

On obtient alors :

$$L_{poisson}(\beta_1) = \mathbb{P}\left(\mathcal{P}\left(\frac{y_1 + y_2}{\exp(\beta_1) + 1}\right) = y_1\right) \times \mathbb{P}\left(\mathcal{P}\left(\frac{y_1 + y_2}{\exp(\beta_1) + 1} \times \exp(\beta_1)\right) = y_2\right)$$

Si on note $a = \frac{y_1+y_2}{\exp(\beta_1)+1}$ et $b = \frac{y_1+y_2}{\exp(\beta_1)+1} \times \exp(\beta_1)$ on observe $a + b = \frac{(y_1+y_2)}{\exp(\beta_1)+1}(\exp(\beta_1) + 1) = y_1 + y_2$ et $L_{poisson}(\beta_1) = \mathbb{P}(\mathcal{P}(a) = y_1) \times \mathbb{P}(\mathcal{P}(b) = y_2)$

Note : a et b représentent les valeurs prédites du modèle sous l'hypothèse nulle de β_1

Nous avions montré dans la section 11.1 que cette distribution est une loi binomiale de telle sorte que :

$$\frac{L_{poisson}(\beta_1)}{\mathbb{P}(\mathcal{P}(a + b) = y_1 + y_2)} = \frac{\mathbb{P}(\mathcal{P}(a) = y_1) \times \mathbb{P}(\mathcal{P}(b) = y_2)}{\mathbb{P}(\mathcal{P}(a + b) = y_1 + y_2)}$$

$$= \mathbb{P}\left(\mathcal{B}\left(a + b; \frac{b}{a + b}\right) = y_2\right)$$

Or $a + b = y_1 + y_2$ et $\frac{b}{a+b} = b(a + b)^{-1} = \frac{y_1+y_2}{\exp(\beta_1)+1} \times \exp(\beta_1) \times (y_1 + y_2)^{-1} = \frac{\exp(\beta_1)}{\exp(\beta_1)+1}$

$$= \mathbb{P}\left(\mathcal{B}\left(y_1 + y_2; \frac{\exp(\beta_1)}{\exp(\beta_1)+1}\right) = y_2\right) = L_{binom}(\beta_1)$$

donc pour y_1 et y_2 non nuls, $L_{poisson}(\beta_1) = L_{binom}(\beta_1) \times \mathbb{P}(\mathcal{P}(y_1 + y_2) = y_1 + y_2)$.

Les fonctions de vraisemblance étant égales à un facteur constant près, l'inférence sur le paramètre β_1 dans le modèle binomial univarié et l'inférence sur le paramètre γ dans le modèle de Poisson bivarié sont équivalentes : $\hat{\gamma} = \widehat{\beta_1}$ pour l'estimateur du maximum de vraisemblance. Pour tout β_{H1}, les statistiques des tests d'hypothèses du rapport de vraisemblance et des tests du score pour l'hypothèse nulle $\gamma = \beta_{H1}$ dans le modèle de Poisson bivarié sont égales à celles des tests d'hypothèses pour l'hypothèse nulle $\beta_1 = \beta_{H1}$. On en déduit alors l'équivalence des intervalles de confiance par ces mêmes méthodes.

Si on généralisait les statistiques d'Hirji [42] au modèle de Poisson, dans le cas limite où le dénominateur des lois binomiales tend vers l'infini, on se ramènerait aussi au même problème. La reparamétrisation faite par Hirji est conditionnelle à la marge $y_1 + y_2$, de telle sorte qu'elle reste constante dans tous les calculs, et que les fonctions de vraisemblance sont juste modifiées d'un facteur constant. On devrait donc arriver à la même inférence.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Critères d'évaluation de la validité des estimateurs d'intervalle

Mots clés : estimateur, intervalle de confiance, test d'hypothèse, directionnel, interprétation

Résumé :
INTRODUCTION : l'interprétation des intervalles de confiance et des tests d'hypothèses est généralement directionnelle, c'est-à-dire, distinguant la supériorité de l'infériorité. Par ailleurs, de nombreux outils statistiques conditionnent sur des statistiques qui sont variables, tel que la matrice d'exposition d'un modèle linéaire généralisé ou la taille d'échantillon. L'objectif de cette thèse est de proposer des outils d'évaluation des estimateurs statistiques en adéquation avec l'usage de ces outils.
MÉTHODES : les conséquences de l'absence de prise en compte de l'interprétation directionnelle dans l'évaluation des outils statistiques, notamment les tests d'hypothèses concernant les analyses de survie, ont été évaluées à partir de revues de la littérature et de considérations théoriques. Des critères de jugement, basés sur les moyennes locales ont été proposés afin de s'abstraire du conditionnement sur des statistiques variables, notamment la taille d'échantillon, lors de l'évaluation des outils statistiques. Les risques de sous-estimation et surestimation ont été distingués dans l'évaluation des défauts de couverture des intervalles de confiance. Des estimateurs classiques ont été réévalués à la lumière de ces nouveaux critères de jugement : estimateurs d'une proportion binomiale et estimateurs de régressions de Poisson.
RÉSULTATS : l'interprétation directionnelle de tests d'hypothèses concernant les différences de courbes de survie peut conduire à un risque de conclure à une différence dans le sens opposé de la différence réelle, qui approche parfois 50 %. Les nouveaux critères de jugement, notamment les risques alpha unilatéraux moyens locaux, ont mis en évidence la supériorité du test du rapport de vraisemblance et des intervalles de confiance associés, par rapport aux méthodes du score de Rao et de Wald, pour les régressions logistiques et de Poisson. Les méthodes visant le strict conservatisme s'avèrent pertinentes dans certains contextes très spécifiques ; autrement les méthodes basées sur les mid-P-valeurs devraient leur être préférées. Les intervalles de confiance visant à un déséquilibre des risques de surestimation et de sous-estimation afin d'en rétrécir la largeur totale perdent des propriétés d'interprétation directionnelle.
CONCLUSION : l'interprétation directionnelle qui sera faite des tests d'hypothèses et estimateurs d'intervalles de confiance devrait être prise en compte lors de leur conception et évaluation.

Title : Assessment criteria for interval estimators

Keywords : estimator, confidence interval, hypothesis test, directional, interpretation

Abstract:
INTRODUCTION: the interpretation of confidence intervals and hypothesis tests is usually directional, drawing a distinction between superiority and inferiority. On another note, numerous statistical tools are conditional to statistics that are actually variable such as the model matrix of a generalized linear model or sample size. The aim of this thesis is to propose methods for the assessment of statistical estimators that are consistent with the use that will be done of these tools.
METHODS: the consequences of ignoring the directional interpretation in the assessment of statistical tools, especially for survival analysis, were evaluated from reviews of the literature and theoretical considerations. Assessment criteria Assessment criteria based on local averages were proposed to free the assessment of statistical tools from conditioning on variable statistics, such as sample size even when such tools are conditional. Underestimation and overestimation risks were differentiated in the assessment of coverage flaws of confidence intervals. Usual estimators were reassessed with these new criteria : binomial proportion and Poisson regression estimators
RESULTS: the directional interpretation of hypothesis tests about differences in survival curves can lead to a risk of concluding in the opposite direction to the actual difference approaching 50% in some cases. New assessment criteria, in particular, one-sided local average type I error rates, showed the better behavior of the likelihood ratio test and confidence intervals compared to Rao's score and Wald's methods for logistic and Poisson regressions. Strictly conservative methods are relevant in some very specific contexts ; otherwise, methods based on mid-P-values should be preferred. Confidence intervals aiming at an imbalance between risks of overestimation and underestimation in order to shorten their overall width lose directional interpretation properties.
CONCLUSION: the directional interpretation of hypothesis tests and confidence interval estimators should be taken into account in their design and assessment.

Maison du doctorat de l'Université Paris-Saclay
2ème étage aile ouest, Ecole normale supérieure Paris-Saclay
4 avenue des Sciences,
91190 Gif sur Yvette, France