Intervalle de confiance d'une proportion binomiale : quels enjeux et comment choisir ?

André Gillibert

FACULTÉ MIXTE DE MÉDECINE ET DE PHARMACIE DE ROUEN

ANNÉE 2017 N°

THÈSE POUR LE DOCTORAT EN MÉDECINE

(Diplôme d'État)

PAR

André GILLIBERT

NÉ LE 15 NOVEMBRE 1984 À EU (76)

PRÉSENTÉE ET SOUTENUE PUBLIQUEMENT LE 13 OCTOBRE 2017

Intervalle de confiance d'une proportion binomiale : quels enjeux et comment choisir ?

Président du jury : *M. le Professeur* Jacques BÉNICHOU

Directeur de thèse : *M. le Professeur* Bruno FALISSARD

Membres du jury : *M. le Professeur Pierre DÉCHELOTTE*

M. le Docteur Joël LADNER

Mme le Docteur Marie-Pierre TAVOLACCI

M. le Docteur Thomas VERMEULIN

ANNEE UNIVERSITAIRE 2016 - 2017

U.F.R. DE MEDECINE ET DE-PHARMACIE DE ROUEN

DOYEN : **Professeur Pierre FREGER**

ASSESSEURS : **Professeur Michel GUERBET**

Professeur Benoit VEBER

Professeur Pascal JOLY

Professeur Stéphane MARRET

I - **MEDECINE**

PROFESSEURS DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mr Frédéric **ANSELME**	HCN	Cardiologie
Mme Isabelle **AUQUIT AUCKBUR**	HCN	Chirurgie plastique
Mr Fabrice **BAUER**	HCN	Cardiologie
Mme Soumeya **BEKRI**	HCN	Biochimie et biologie moléculaire
Mr Ygal **BENHAMOU**	HCN	Médecine interne
Mr Jacques **BENICHOU**	HCN	Bio statistiques et informatique médicale
Mme Bouchra **LAMIA**	Havre	Pneumologie
Mr Olivier **BOYER**	UFR	Immunologie
Mr François **CARON**	HCN	Maladies infectieuses et tropicales
Mr Philippe **CHASSAGNE** *(détachement)*	HCN	Médecine interne (gériatrie) – Détachement
Mr Vincent **COMPERE**	HCN	Anesthésiologie et réanimation chirurgicale
Mr Jean-Nicolas **CORNU**	HCN	Urologie
Mr Antoine **CUVELIER**	HB	Pneumologie
Mr Pierre **CZERNICHOW** *(surnombre)*	HCH	Epidémiologie, économie de la santé
Mr Jean-Nicolas **DACHER**	HCN	Radiologie et imagerie médicale
Mr Stéfan **DARMONI**	HCN	Informatique médicale et techniques de communication
Mr Pierre **DECHELOTTE**	HCN	Nutrition
Mr Stéphane **DERREY**	HCN	Neurochirurgie

Mr Frédéric **DI FIORE**	CB	Cancérologie
Mr Fabien **DOGUET**	HCN	Chirurgie Cardio Vasculaire
Mr Jean **DOUCET**	SJ	Thérapeutique - Médecine interne et gériatrie
Mr Bernard **DUBRAY**	CB	Radiothérapie
Mr Philippe **DUCROTTE**	HCN	Hépato-gastro-entérologie
Mr Frank **DUJARDIN**	HCN	Chirurgie orthopédique - Traumatologique
Mr Fabrice **DUPARC**	HCN	Anatomie - Chirurgie orthopédique et traumatologique
Mr Eric **DURAND**	HCN	Cardiologie
Mr Bertrand **DUREUIL**	HCN	Anesthésiologie et réanimation chirurgicale
Mme Hélène **ELTCHANINOFF**	HCN	Cardiologie
Mr Thierry **FREBOURG**	UFR	Génétique
Mr Pierre **FREGER**	HCN	Anatomie - Neurochirurgie
Mr Jean François **GEHANNO**	HCN	Médecine et santé au travail
Mr Emmanuel **GERARDIN**	HCN	Imagerie médicale
Mme Priscille **GERARDIN**	HCN	Pédopsychiatrie
Mr Michel **GODIN** *(surnombre)*	HB	Néphrologie
M. Guillaume **GOURCEROL**	HCN	Physiologie
Mr Dominique **GUERROT**	HCN	Néphrologie
Mr Olivier **GUILLIN**	HCN	Psychiatrie Adultes
Mr Didier **HANNEQUIN**	HCN	Neurologie
Mr Fabrice **JARDIN**	CB	Hématologie
Mr Luc-Marie **JOLY**	HCN	Médecine d'urgence
Mr Pascal **JOLY**	HCN	Dermato - Vénéréologie
Mme Annie **LAQUERRIERE**	HCN	Anatomie et cytologie pathologiques
Mr Vincent **LAUDENBACH**	HCN	Anesthésie et réanimation chirurgicale
Mr Joël **LECHEVALLIER**	HCN	Chirurgie infantile
Mr Hervé **LEFEBVRE**	HB	Endocrinologie et maladies métaboliques
Mr Thierry **LEQUERRE**	HB	Rhumatologie
Mme Anne-Marie **LEROI**	HCN	Physiologie
Mr Hervé **LEVESQUE**	HB	Médecine interne
Mme Agnès **LIARD-ZMUDA**	HCN	Chirurgie Infantile
Mr Pierre Yves **LITZLER**	HCN	Chirurgie cardiaque
Mr Bertrand **MACE**	HCN	Histologie, embryologie, cytogénétique
M. David **MALTETE**	HCN	Neurologie
Mr Christophe **MARGUET**	HCN	Pédiatrie
Mme Isabelle **MARIE**	HB	Médecine interne
Mr Jean-Paul **MARIE**	HCN	Oto-rhino-laryngologie
Mr Loïc **MARPEAU**	HCN	Gynécologie – Obstétrique

Mr Stéphane **MARRET**	HCN	Pédiatrie
Mme Véronique **MERLE**	HCN	Epidémiologie
Mr Pierre **MICHEL**	HCN	Hépato-gastro-entérologie
M. Benoit **MISSET**	HCN	Réanimation Médicale
Mr Jean-François **MUIR** *(surnombre)*	HB	Pneumologie
Mr Marc **MURAINE**	HCN	Ophtalmologie
Mr Philippe **MUSETTE**	HCN	Dermatologie - Vénéréologie
Mr Christophe **PEILLON**	HCN	Chirurgie générale
Mr Christian **PFISTER**	HCN	Urologie
Mr Jean-Christophe **PLANTIER**	HCN	Bactériologie - Virologie
Mr Didier **PLISSONNIER**	HCN	Chirurgie vasculaire
Mr Gaëtan **PREVOST**	HCN	Endocrinologie
Mr Bernard **PROUST**	HCN	Médecine légale
Mr Jean-Christophe **RICHARD** *(détachement)*	HCN	Réanimation médicale - Médecine d'urgence
Mr Vincent **RICHARD**	UFR	Pharmacologie
Mme Nathalie **RIVES**	HCN	Biologie du développement et de la reproduction
Mr Horace **ROMAN**	HCN	Gynécologie - Obstétrique
Mr Jean-Christophe **SABOURIN**	HCN	Anatomie - Pathologie
Mr Guillaume **SAVOYE**	HCN	Hépato-gastrologie
Mme Céline **SAVOYE–COLLET**	HCN	Imagerie médicale
Mme Pascale **SCHNEIDER**	HCN	Pédiatrie
Mr Michel **SCOTTE**	HCN	Chirurgie digestive
Mme Fabienne **TAMION**	HCN	Thérapeutique
Mr Luc **THIBERVILLE**	HCN	Pneumologie
Mr Christian **THUILLEZ** *(surnombre)*	HB	Pharmacologie
Mr Hervé **TILLY**	CB	Hématologie et transfusion
M. Gilles **TOURNEL**	HCN	Médecine Légale
Mr Olivier **TROST**	HCN	Chirurgie Maxillo-Faciale
Mr Jean-Jacques **TUECH**	HCN	Chirurgie digestive
Mr Jean-Pierre **VANNIER** *(surnombre)*	HCN	Pédiatrie génétique
Mr Benoît **VEBER**	HCN	Anesthésiologie - Réanimation chirurgicale
Mr Pierre **VERA**	CB	Biophysique et traitement de l'image
Mr Eric **VERIN**	HB	Service Santé Réadaptation
Mr Eric **VERSPYCK**	HCN	Gynécologie obstétrique
Mr Olivier **VITTECOQ**	HB	Rhumatologie
Mr Jacques **WEBER**	HCN	Physiologie

MAITRES DE CONFERENCES DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mme Noëlle **BARBIER-FREBOURG**	HCN	Bactériologie – Virologie
Mme Carole **BRASSE LAGNEL**	HCN	Biochimie
Mme Valérie **BRIDOUX HUYBRECHTS**	HCN	Chirurgie Vasculaire
Mr Gérard **BUCHONNET**	HCN	Hématologie
Mme Mireille **CASTANET**	HCN	Pédiatrie
Mme Nathalie **CHASTAN**	HCN	Neurophysiologie
Mme Sophie **CLAEYSSENS**	HCN	Biochimie et biologie moléculaire
Mr Moïse **COEFFIER**	HCN	Nutrition
Mr Manuel **ETIENNE**	HCN	Maladies infectieuses et tropicales
Mr Serge **JACQUOT**	UFR	Immunologie
Mr Joël **LADNER**	HCN	Epidémiologie, économie de la santé
Mr Jean-Baptiste **LATOUCHE**	UFR	Biologie cellulaire
Mr Thomas **MOUREZ**	HCN	Virologie
Mme Muriel **QUILLARD**	HCN	Biochimie et biologie moléculaire
Mme Laëtitia **ROLLIN**	HCN	Médecine du Travail
Mr Mathieu **SALAUN**	HCN	Pneumologie
Mme Pascale **SAUGIER-VEBER**	HCN	Génétique
Mme Anne-Claire **TOBENAS-DUJARDIN**	HCN	Anatomie
Mr David **WALLON**	HCN	Neurologie

PROFESSEUR AGREGE OU CERTIFIE

Mme Dominique **LANIEZ**	UFR	Anglais – retraite 01/10/2016
Mr Thierry **WABLE**	UFR	Communication

II - PHARMACIE

PROFESSEURS

Mr Thierry **BESSON**	Chimie Thérapeutique
Mr Jean-Jacques **BONNET**	Pharmacologie
Mr Roland **CAPRON** (PU-PH)	Biophysique
Mr Jean **COSTENTIN** (Professeur émérite)	Pharmacologie
Mme Isabelle **DUBUS**	Biochimie
Mr Loïc **FAVENNEC** (PU-PH)	Parasitologie
Mr Jean Pierre **GOULLE** (Professeur émérite)	Toxicologie
Mr Michel **GUERBET**	Toxicologie
Mme Isabelle **LEROUX - NICOLLET**	Physiologie
Mme Christelle **MONTEIL**	Toxicologie
Mme Martine **PESTEL-CARON** (PU-PH)	Microbiologie
Mme Elisabeth **SEGUIN**	Pharmacognosie
Mr Rémi **VARIN** (PU-PH)	Pharmacie clinique
Mr Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

MAITRES DE CONFERENCES

Mme Cécile **BARBOT**	Chimie Générale et Minérale
Mr Jérémy **BELLIEN** (MCU-PH)	Pharmacologie
Mr Frédéric **BOUNOURE**	Pharmacie Galénique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Camille **CHARBONNIER (LE CLEZIO)**	Statistiques
Mme Elizabeth **CHOSSON**	Botanique
Mme Marie Catherine **CONCE-CHEMTOB**	Législation pharmaceutique et économie de la santé
Mme Cécile **CORBIERE**	Biochimie
Mr Eric **DITTMAR**	Biophysique
Mme Nathalie **DOURMAP**	Pharmacologie
Mme Isabelle **DUBUC**	Pharmacologie
Mme Dominique **DUTERTE- BOUCHER**	Pharmacologie
Mr Abdelhakim **ELOMRI**	Pharmacognosie

Mr François **ESTOUR**	Chimie Organique
Mr Gilles **GARGALA** (MCU-PH)	Parasitologie
Mme Nejla EL **GHARBI-HAMZA**	Chimie analytique
Mme Marie-Laure **GROULT**	Botanique
Mr Hervé **HUE**	Biophysique et mathématiques
Mme Laetitia **LE GOFF**	Parasitologie – Immunologie
Mme Hong **LU**	Biologie
Mme Marine **MALLETER**	Toxicologie
Mme Sabine **MENAGER**	Chimie organique
Mme Tiphaine **ROGEZ-FLORENT**	Chimie analytique
Mr Mohamed **SKIBA**	Pharmacie galénique
Mme Malika **SKIBA**	Pharmacie galénique
Mme Christine **THARASSE**	Chimie thérapeutique
Mr Frédéric **ZIEGLER**	Biochimie

PROFESSEURS ASSOCIES

Mme Cécile **GUERARD-DETUNCQ**	Pharmacie officinale
Mr Jean-François **HOUIVET**	Pharmacie officinale

PROFESSEUR CERTIFIE

Mme Mathilde **GUERIN**	Anglais

ASSISTANT HOSPITALO-UNIVERSITAIRE

Mme Sandrine **DAHYOT**	Bactériologie

ATTACHES TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE

Mr Souleymane **ABDOUL-AZIZE**	Biochimie
Mme Hanane **GASMI**	Galénique
Mme Caroline **LAUGEL**	Chimie organique
Mr Romy **RAZAKANDRAINIBE**	Parasitologie

LISTE DES RESPONSABLES DES DISCIPLINES PHARMACEUTIQUES

Mme Cécile **BARBOT**	Chimie Générale et minérale
Mr Thierry **BESSON**	Chimie thérapeutique
Mr Roland **CAPRON**	Biophysique
Mme Marie-Catherine **CONCE-CHEMTOB**	Législation et économie de la santé
Mme Elisabeth **CHOSSON**	Botanique
Mr Jean-Jacques **BONNET**	Pharmacodynamie
Mme Isabelle **DUBUS**	Biochimie
Mr Loïc **FAVENNEC**	Parasitologie
Mr Michel **GUERBET**	Toxicologie
Mr François **ESTOUR**	Chimie organique
Mme Isabelle **LEROUX-NICOLLET**	Physiologie
Mme Martine **PESTEL-CARON**	Microbiologie
Mme Elisabeth **SEGUIN**	Pharmacognosie
Mr Mohamed **SKIBA**	Pharmacie galénique
Mr Rémi **VARIN**	Pharmacie clinique
Mr Philippe **VERITE**	Chimie analytique

III – MEDECINE GENERALE

PROFESSEUR

Mr Jean-Loup **HERMIL**	UFR	Médecine générale

PROFESSEURS ASSOCIES A MI-TEMPS

Mr Emmanuel **LEFEBVRE**	UFR	Médecine Générale
Mme Elisabeth **MAUVIARD**	UFR	Médecine générale
Mr Philippe **NGUYEN THANH**	UFR	Médecine générale

MAITRE DE CONFERENCES ASSOCIE A MI-TEMPS

Mr Pascal **BOULET**	UFR	Médecine générale
Mr Emmanuel **HAZARD**	UFR	Médecine Générale
Mme Lucile **PELLERIN**	UFR	Médecine générale
Mme Yveline **SEVRIN**	UFR	Médecine générale
Mme Marie Thérèse **THUEUX**	UFR	Médecine générale

ENSEIGNANTS MONO-APPARTENANTS

PROFESSEURS

Mr Serguei **FETISSOV** (med)	Physiologie (ADEN)
Mr Paul **MULDER** (phar)	Sciences du Médicament
Mme Su **RUAN** (med)	Génie Informatique

MAITRES DE CONFERENCES

Mr Sahil **ADRIOUCH** (med)	Biochimie et biologie moléculaire (Unité Inserm 905)
Mme Gaëlle **BOUGEARD-DENOYELLE** (med)	Biochimie et biologie moléculaire (UMR 1079**)**
Mme Carine **CLEREN** (med)	Neurosciences (Néovasc)
M. Sylvain **FRAINEAU** (phar)	Physiologie (Inserm U 1096)
Mme Pascaline **GAILDRAT** (med)	Génétique moléculaire humaine (UMR 1079)
Mr Nicolas **GUEROUT** (med)	Chirurgie Expérimentale
Mme Rachel **LETELLIER** (med)	Physiologie
Mme Christine **RONDANINO** (med)	Physiologie de la reproduction
Mr Antoine **OUVRARD-PASCAUD** (med)	Physiologie (Unité Inserm 1076)
Mr Frédéric **PASQUET**	Sciences du langage, orthophonie
Mme Isabelle **TOURNIER** (med)	Biochimie (UMR 1079)

CHEF DES SERVICES ADMINISTRATIFS : Mme Véronique **DELAFONTAINE**

HCN - Hôpital Charles Nicolle	*HB - Hôpital de BOIS GUILLAUME*
CB - Centre Henri Becquerel	*CHS - Centre Hospitalier Spécialisé du Rouvray*
CRMPR - Centre Régional de Médecine Physique et de Réadaptation	*SJ – Saint Julien Rouen*

Par délibération en date du 3 mars 1967, la faculté a arrêté que les opinions émises dans les dissertations qui lui seront présentées doivent être considérées comme propres à leurs auteurs et qu'elle n'entend leur donner aucune approbation ni improbation.

Remerciements

Je remercie mon directeur de thèse, le Pr Bruno FALISSARD, avec qui les échanges sont toujours très constructifs et sans lequel ce travail n'aurait pas abouti.

Je remercie le Pr Jacques BÉNICHOU qui m'a guidé le long de mon internat et me fait l'honneur de présider à ma thèse.

Je remercie le Pr Pierre DÉCHELOTTE d'avoir accepté de juger mon travail, après m'avoir donné sa confiance lorsque je côtoyais son service.

Je remercie le Dr Marie-Pierre TAVOLACCI de prendre part à ce jury et d'avoir su me stimuler dès que j'ai joint son service.

Je remercie le Dr Joël LADNER, de prendre part à ce jury et de la bienveillance qu'il a toujours eu à mon égard.

Je remercie le Dr Thomas VERMEULIN, de prendre part à ce jury et de toutes les conversations statistiques, méthodologiques et sociologiques que nous avons partagées convivialement.

Je remercie Lina MUSTAPHA de sa disponibilité, sa bienveillance et son travail de relecture minutieux.

Je remercie le Dr Alexandra ROUQUETTE pour son travail de relecture et ses commentaires pertinents.

Je remercie le Pr Stefan DARMONI qui sait me stimuler avec des méthodes efficaces.

Je remercie le Dr Mourad Ould-Slimane avec qui c'est toujours un plaisir de travailler.

Je remercie le Dr Mher JOULAKIAN avec qui les conversations sont toujours passionnantes, même s'il s'éloigne un peu.

Je remercie le Dr Sacha SCHUTZ pour son support et son amitié.

Je remercie Siré N'Diaye qui m'a motivé pour suivre la voie de l'enseignement et dont l'optimisme est contagieux.

Merci à Thibaut LAFFOUILLHERE pour son amitié et son goût partagé des statistiques et de la sociologie.

Merci à Nathalie DI MARCO qui sait toujours écouter.

Merci à Mikaël DUSENNE, Josselin DIOT, Arthur SPILLEBOULT, Akpene FRED, Kristell HARDY, Anca VASILIU, Costin ZAINEA, Aurélien ZHU-SOUBISE, Lucile BRUYÈRE, Andrea FIORDALISO, Maggie LE BOUHRIS, Ludivine BOULET, Damiano CERASUOLO, Sorina MIHALESCU pour leur bonne humeur, leur support et les moments passés ensemble.

Merci à Lamisse BOUTI qui m'a appris beaucoup plus qu'elle ne pense.

Merci à Marie-Laure BARANNE de toutes nos discussions.

Merci à Caroline THILL et Estelle HOUIVET qui m'ont permis d'apprendre à utiliser SAS.

Merci à Caroline BARRY, Christine HASSLER et Juliette GUEGUEN pour leur compagnie et leur confiance.

Merci à Charlène BOULAY, Henri GONDÉ, Charlotte ORSINI avec qui discuter est toujours un plaisir.

Merci à Serena, Anaïs, Valentina, Anne-Laurène, Cherifa et Fañch.

Merci à mes parents Françoise et Guy, qui ont su m'élever, me transmettre le goût des mathématiques et de la médecine et m'ont permis de réunir les deux.

Merci à mes frères et sœurs pour tout ce que nous avons partagé, partageons et partagerons : Jean, Georges, Luc, Pierre, Florence, Myriam, Raymond, Catherine, Félicia et Noël GILLIBERTs, et merci à mes beaux-frères et belles sœurs Nadia, Adeline, Delphine, Gabriele.

Et à tous les autres dont la liste est trop longue pour que je les mentionne.

Table des matières

1 Introduction

Ce travail est principalement destiné aux usagers occasionnels ou réguliers de logiciels statistiques dans le cadre de travaux en recherche biomédicale. Il est aussi adapté aux lecteurs réguliers d'articles scientifiques cherchant à avoir une compréhension fine des statistiques présentées. Ce travail traite d'un des problèmes les plus simples en apparence : calculer l'intervalle de confiance d'une proportion, telle que la prévalence d'une maladie. Ce problème paraît tellement simple que la méthode utilisée par le logiciel statistique n'est pas toujours indiquée (manuel R (85)). Généralement, le logiciel offre une option « approximative » et une méthode « exacte » pour le calcul d'un intervalle de confiance d'une proportion (R, SAS, Stata). Si on définit p la proportion réelle dans la population, $\hat{p}$ la proportion observée sur l'échantillon et n la taille de l'échantillon, la méthode approximative est presque toujours la formule bien connue que l'on retrouve dans tous les livres de statistique élémentaire :

$$\hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (1)$$

Les conditions de validité habituelles telles que $np \geq 5$ et $n(1-p) \geq 5$, décrite par Millot (75) sont très insuffisantes, d'après Brown (21). Nous montrerons que cette méthode n'est acceptable que lorsque le nombre de succès et d'échecs dépassent tous les deux 40. Devant des effectifs insuffisants, un statisticien peut sélectionner la méthode « exacte », se disant qu'un résultat qualifié d'exact ne doit pas être statistiquement critiquable. On pourrait s'attendre à ce qu'un intervalle de confiance exact à 95% garantisse que l'intervalle contienne la vraie proportion dans 95% des expériences et ne la contienne pas dans 5% des expériences, mais ce n'est pas le cas d'après Brown (21). D'après Agresti (4) les méthodes approximatives pourraient même être meilleure que les méthodes exactes. Plus de 50 méthodes d'estimation différentes sont présentées dans ce travail, dont 11 se vantent d'être exactes, optimales ou proches de l'optimum. De nouvelles méthodes ont encore été proposées ces trois dernières années telles que celles proposées par Schilling et Doi (93) en 2014, Wang (109) en 2014, et Lecoutre et Poiteniveau (66) en 2016.

Le développement incessant de nouvelles méthodes est témoin d'un problème insoluble. Le problème est l'impossibilité, du fait d'une loi discrète, de maîtriser parfaitement la *couverture*, c'est-à-dire, la probabilité que l'intervalle de confiance contienne la proportion de la population. Les solutions proposées à ce problème en créent de nouveaux. Thulin (100) mentionne le déséquilibre entre les risques de surestimer et de sous-estimer la valeur, apportés par des méthodes exactes dont l'objectif est de minimiser la largeur des intervalles. Vos (107) mentionne des paradoxes liés à ces mêmes intervalles, tels que l'intervalle de confiance à 90% qui n'est pas contenu dans l'intervalle de confiance à 95%, conduisant à rejeter une hypothèse au risque 5% mais à l'accepter au risque 10%.

Nous verrons que certains problèmes soulevés par l'analyse du comportement des intervalles de confiance d'une proportion, s'appliquent à d'autres statistiques, dont les coefficients de régressions linéaires, de régressions logistiques ou de modèles de Cox. Certains éléments nous aiderons à juger des méthodes générales de calcul des intervalles de confiance : Wald, Score et rapport de vraisemblance.

Nous présenterons des aspects théoriques généraux dans un 1er temps (chapitre Théorie statistique générale), puis aborderons plus précisément l'état des connaissances sur le problème d'estimation d'une proportion (chapitre Problématique), puis nous présenterons notre travail original (Matériel & Méthodes, Résultats) et le discuterons.

Les objectifs de ce travail sont de présenter une revue de la littérature sur l'estimation des intervalles de confiance d'une proportion binomiale, définir des critères d'évaluation pertinents des estimateurs d'intervalles de confiance, d'appliquer ces critères de manière systématique aux intervalles existants, puis de fournir des conseils pratiques sur le ou les intervalles à utiliser, selon le logiciel. En objectif secondaire, des conditions de validité de l'estimateur d'intervalle de Wald sont définis.

Afin de diffuser l'usage de l'intervalle de confiance que nous conseillons d'utiliser de manière prioritaire, nous fournissons en Annexe 2 les codes sources permettant d'implanter cette méthode dans divers logiciels utilisés en biostatistique pour lesquels la procédure n'existe pas nativement : SAS, SPSS, Stata, Python, SYSTAT, Minitab, HTML+JavaScript et Microsoft Excel et les calculatrices Texas Instruments Ti 83/84. Le logiciel R n'est pas concerné car il bénéficie de paquets complémentaires très bien supportés. La diffusion de ce travail en anglais est prévue sous forme de l'article joint en Annexe 4.

1.1 Théorie statistique générale

La statistique est la science concernant la collecte et l'analyse de l'information numérique selon Brillinger (19). Elle a une longue histoire, le recensement en étant un des exemples les plus anciens. Le premier enregistré en Chine aurait eu lieu environ il y a quatre mille ans selon *The Canadian Encyclopedia* (29). Il a pour objectif de compter les citoyens, mais aussi de les caractériser. Comme le nombre d'observations est trop grand pour être accessible à la compréhension humaine, elles peuvent être exprimées sous forme de nombres synthétiques, tels que l'âge moyen ou la proportion de sujets de sexe féminin. Ces synthèses numériques sont appelées statistiques. Lorsque les populations sont trop grandes, voire virtuellement infinies, l'échantillonnage, idéalement aléatoire, permet d'estimer, à partir d'un échantillon, les statistiques de la population. Il existe toujours un écart entre la statistique évaluée sur un échantillon aléatoire et la statistique de la population complète. Cet écart, dû au hasard, est nommé *fluctuation d'échantillonnage*. Les fluctuations d'échantillonnage sont prévisibles et d'autant plus petites que l'échantillon est grand. Cela permet d'estimer précisément des statistiques d'une grande population, voire d'une population virtuellement infinie, à partir d'un échantillon.

Le hasard, qu'il s'agisse de phénomènes non déterministes ou de phénomènes déterministes trop complexes et instables pour être prédits, se retrouve dans toutes les sciences expérimentales. Même dans des conditions de laboratoire maîtrisées, la réaction de plusieurs souris de laboratoire de même souche, élevées dans le même environnement et soumises à la même expérience ne sera pas toujours la même. La partie non reproductible de l'expérience sera appelée hasard. Plutôt que d'appliquer l'expérience à une seule souris, on multipliera les expériences jusqu'à ce que les fluctuations d'échantillonnage deviennent suffisamment faibles pour qu'une statistique s'appliquant à toutes les souris de cette même souche soit estimée précisément. Par exemple, on pourra estimer la proportion de souris, modèles de myopathie, s'améliorant en présence ou en l'absence d'un traitement, puis la différence des deux proportions. Dans des conditions expérimentales contrôlées, la statistique permet, en présence de phénomènes imprévisibles suivant les lois du hasard, de retrouver la reproductibilité nécessaire à la démarche scientifique. Dans des conditions contrôlées, bien qu'une même cause ne produise pas toujours les mêmes conséquences, la statistique permet de retrouver la causalité dans une forme affaiblie : une même cause ne produira pas toujours une même conséquence, mais produira, en moyenne, ou à un certain taux, une même conséquence.

Nous ne nous intéresserons que peu aux plans expérimentaux, car ce travail concerne l'aspect statistique pur de l'estimation d'une proportion binomiale. Les difficultés, notamment, de l'échantillonnage représentatif, nécessaire à l'inférence, seront mentionnées mais ne seront pas détaillées.

1.1.1 Variable aléatoire, unité statistique

Les statistiques ne sont calculables que sur des valeurs numériques. Comme nous traiterons de biostatistique, ces valeurs numériques seront toujours issues d'objets d'un monde physique, généralement des êtres vivants. Les nombres étant des objets mathématiques abstraits, il y a donc forcément une étape de codage. La réalité étant trop complexe, il y a une part de simplification et de standardisation. L'*unité statistique* est l'élément d'une population à laquelle des caractéristiques, appelées variables aléatoires, sont rattachées. L'unité statistique peut être le patient, le sujet d'expérience, mais peut aussi être une visite médicale, un soignant, un établissement, un dispositif. La *variable aléatoire* est une caractéristique de l'*unité statistique*. L'ensemble des valeurs possibles d'une variable aléatoire constitue son *support*. Plusieurs types de variables aléatoires sont distingués :

Variable aléatoire quantitative discrète : c'est une variable dont le support est un ensemble numérique fini ou dénombrable (Wikipedia (70)). Le plus souvent il s'agit de nombres entiers.

Exemples : nombre d'enfants d'une femme, nombre de lits d'une unité d'hospitalisation, nombre de culots globulaires reçus par un patient.

Variable aléatoire quantitative continue : c'est une variable aléatoire pour laquelle on ne peut pas individualiser une valeur de probabilité non nulle entourée d'un voisinage de valeurs de probabilité nulle, comme on en trouve dans les variables quantitatives discrètes.

Exemples : poids en kilogrammes, glycémie en concentration millimolaire, âge en minutes depuis la naissance

Variable aléatoire catégorielle nominale : aussi dite variable aléatoire qualitative nominale. C'est une variable dont le support est constitué de valeurs appartenant à un ensemble non ordonné. Les valeurs du support sont appelées modalités de la variable.

Exemples : profession, sexe, motif de consultation, mode d'exercice d'un médecin (libéral, salarié, mixte)

Variable aléatoire catégorielle ordinale : aussi dite variable aléatoire qualitative ordinale ou qualitative ordonnée. Les valeurs du support (modalités) appartiennent à un ensemble fini mais il existe une relation d'ordre sur ces modalités.

Exemples : échelle de Likert (pas du tout d'accord, plutôt d'accord, tout à fait d'accord), exposition au tabac (aucune, ancienne, actuelle), lourdeur de prise en charge (ambulatoire, hospitalisation en service conventionnel, hospitalisation en service de réanimation).

Variable binaire : c'est une variable dont le support ne contient que deux valeurs.

Exemples : sexe, présence d'une sclérose en plaques, statut décédé ou vivant, appartenance au groupe de traitement actif ou placébo.

Variable de Bernoulli : c'est un cas particulier de variable binaire quantitative discrète. Le support de la variable est constitué des deux éléments 0 et 1 seulement.

Toute variable binaire peut être associée à une variable de Bernoulli par une recodage de son support (via une bijection). Ainsi, on pourra décider de recoder la variable décès à 1 pour les sujets décédés et

0 pour les sujets survivants. Pour le sexe, le choix est arbitraire, puisqu'on pourra coder 0 pour les hommes et 1 pour les femmes, ou, inversement, coder 1 pour les hommes et 0 pour les femmes.

1.1.2 Échantillon

Une population est un ensemble d'unités statistiques. Souvent les populations sont infinies ou suffisamment grandes pour y être approximables. Le tirage d'unités statistiques, que l'on appellera observations une fois le tirage réalisé, constituera un ensemble d'observations appelé échantillon. L'échantillonnage aléatoire simple, par tirage au sort d'observations indépendantes est le seul qui nous intéressera. Nous ne nous intéresserons pas aux plans d'échantillonnage aléatoires en grappes ou à plusieurs niveaux, dans lesquels les observations ne sont pas indépendantes mais les lois du hasard restent prévisibles.

Nous mentionnons l'échantillonnage de convenance, consistant à sélectionner les sujets qui se présentent spontanément. L'échantillonnage de convenance fournit des statistiques dont les lois ne sont pas bien connues. Des connaissances supplémentaires sur l'aspect non aléatoire de l'échantillonnage permettent parfois d'interpréter les résultats sans toutefois maîtriser complètement l'incertitude. Le redressement d'échantillon, méthode statistique consistant à rééquilibrer l'échantillon *a posteriori* sur certaines variables d'intérêt, reste limité aux variables collectées. En biostatistique, l'échantillonnage de convenance est plus la règle que l'exception. C'est une des raisons de l'hétérogénéité retrouvée dans les méta-analyses. C'est-à-dire, d'une étude à l'autre, il existe des écarts statistiques plus grands que ceux prévisibles par les lois du hasard. Les autres problèmes limitant la reproductibilité des expériences sont l'inconstance de tous les paramètres expérimentaux : intervention, variables mesurées et méthodes de mesures, voire analyse statistique. Les problèmes de choix des méthodes d'estimation statistique, en comparaison de ceux-là, sont bien maigres. Cela reste néanmoins un problème additionnel qui a l'avantage d'être en grande partie maîtrisable. Si la solution est implémentée de manière transparente dans tous les logiciels, toutes les statistiques concernées peuvent être très légèrement améliorées.

1.1.3 Loi binomiale

Pour une variable quantitative discrète, la *loi de probabilité* détermine à la fois les valeurs possibles et leur probabilité. Dans le contexte biostatistique, la réalisation d'une variable aléatoire consiste en un tirage au sort d'une observation de la population, suivie de la mesure de la variable sur l'observation sélectionnée.

Une loi de Bernoulli est la loi de probabilité correspondant à une variable de Bernoulli. Son support est l'ensemble $\{0, 1\}$. Il existe une probabilité p qu'une réalisation de la variable soit égale à 1. La probabilité complémentaire $1 - p$ est la probabilité qu'une réalisation de la variable soit égale à zéro. Cette probabilité p est le paramètre de la loi de Bernoulli. En bref, une variable X suivant une loi de Bernoulli de paramètre p vérifie :

$$\begin{cases} P(X = 1) = p \\ P(X = 0) = 1 - p \end{cases} \quad (2)$$

La loi de Bernoulli, au sens strict, n'est pas une loi de probabilité, c'est une famille de lois de probabilité, paramétrée par un nombre réel $p \in [0,1]$. Par convention, on parlera de succès pour une réalisation d'une variable de Bernoulli égale à 1 et d'échec pour une réalisation égale à 0.

La loi binomiale est une famille de lois de probabilité dépendant de deux paramètres n et p correspondant à la somme de n réalisations de variable de Bernoulli indépendantes de même paramètre p. Cela correspond au nombre de succès d'une variable binaire dans un échantillon aléatoire d'observations indépendantes identiquement distribuées, l'unité statistique étant l'échantillon.

La planche de Galton, dont un modèle est présenté en Figure 1 a été inventée par Sir Francis Galton (1822 – 1911). Ce dispositif illustre la loi binomiale de paramètre n égal au nombre de rangées de clous ($n = 8$ dans la Figure 1) et p égal à $\frac{1}{2}$. L'intervalle entre deux clous est presque égal au diamètre d'une bille et deux rangées successives sont décalées d'un demi-diamètre. Le dispositif est placé verticalement dans un champ gravitationnel. Les billes sont introduites au sommet. L'entonnoir dirigera toutes les billes à la même position horizontale. Chaque bille tombera, en position centrée, sur le clou unique de la 1ère rangée et s'orientera aléatoirement à droite ou à gauche du clou, avec une probabilité de 0,50 (1ère expérience de Bernoulli). La bille tombera au milieu d'un clou de la 2ème rangée, et, encore une fois elle s'orientera aléatoirement à droite ou à gauche du clou avec une probabilité de 0,50 (2ème expérience de Bernoulli, indépendante de la 1ère), et ainsi de suite. La succession de ces n expériences de Bernoulli indépendantes identiquement distribuées conduira la bille à l'une des $n + 1$ positions finales possibles. Si on compte le mouvement à gauche du clou comme le 0 d'une expérience de Bernoulli et le mouvement à droite comme le 1, alors la position finale, numérotée de gauche à droite de 0 à n représente la somme des n expériences de Bernoulli, c'est-à-dire, la réalisation d'une loi binomiale. En introduisant successivement de multiples billes, un échantillon représentatif de billes suivant une même loi binomiale se retrouvera en bas du dispositif. Si les billes sont suffisamment nombreuses, la distribution empirique de cet échantillon se rapprochera de la distribution théorique de la loi binomiale. Sur la Figure 1, une courbe en cloche est représentée. Cette courbe représente la loi de densité de probabilité d'une loi normale. Le théorème de Moivre-Laplace (Wikipedia (99)) prouve en effet que la loi binomiale centrée réduite de paramètres n et p, tend vers une loi normale centrée réduite lorsque n tend vers $+\infty$. Cela a pour conséquence qu'une loi binomiale est approximable à une loi normale de même moyenne et de même variance lorsque son paramètre n est suffisamment grand. Nous verrons que la précision de cette approximation ne dépend pas seulement de n mais aussi de p, l'optimum étant pour $p = 0{,}50$. Ce théorème a été généralisé à la somme de n observations indépendantes, identiquement distribuées, issues d'une loi de moyenne et écart-type finis. La généralisation s'appelle le théorème central limite.

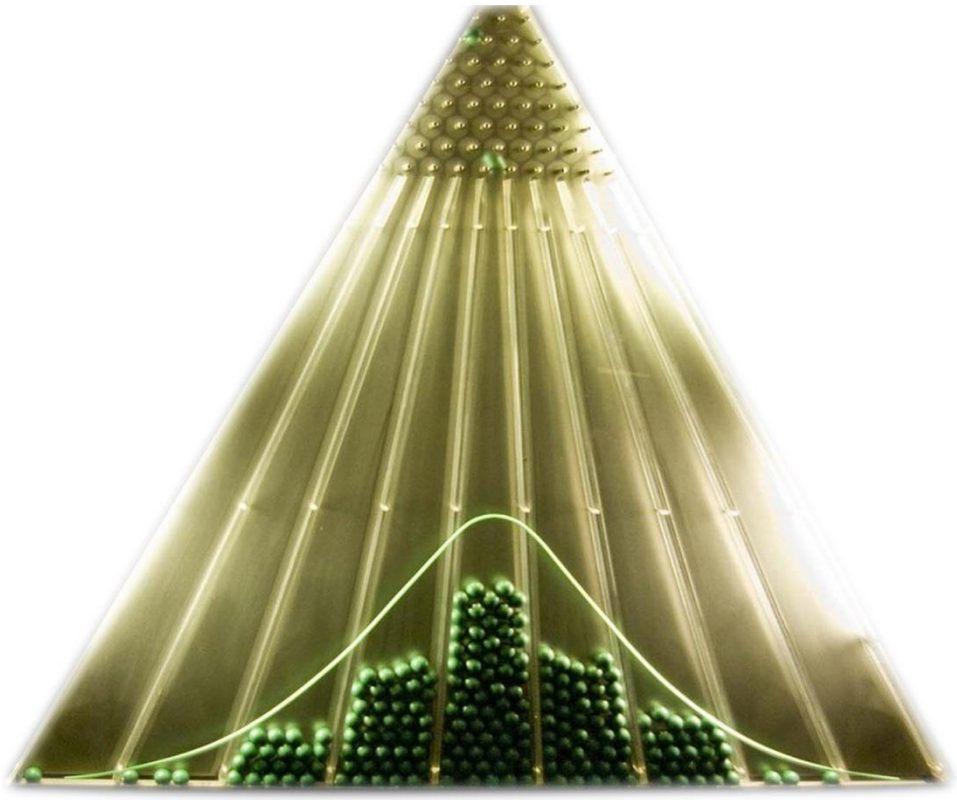

Figure 1 : planche de Galton. Auteur : Antoine Taveneaux (© juin 2008). Licence : CC-BY-SA 4.0.

1.1.3.1 Notations et propriétés de la loi binomiale

On note $B(n;p)$ la loi binomiale de paramètres n et p. Pour définir une variable X suivant une loi binomiale on note $X \sim B(n;p)$ ce qui se lit X suit une loi binomiale de paramètres n et p. La fonction de masse f de la loi binomiale est définie comme suit :

$$f(x) = P(X = x) = BPF(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3}$$

où $x \in \{0, \dots, n\}$ et le nombre de combinaisons de x éléments parmi n est défini par :

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} \tag{4}$$

Pour rappel $n! = 1 \times 2 \times 3 \times \dots \times n$ représente factorielle n. Enfin $0! = 1$.

La fonction de masse de la loi binomiale $B(10\,;0{,}20)$ est présentée graphiquement par un diagramme en barres sur la Figure 2. C'est une loi discrète et asymétrique. Contrairement à la loi normale, les fluctuations d'échantillonnage de sa moyenne et de sa variance sont corrélées.

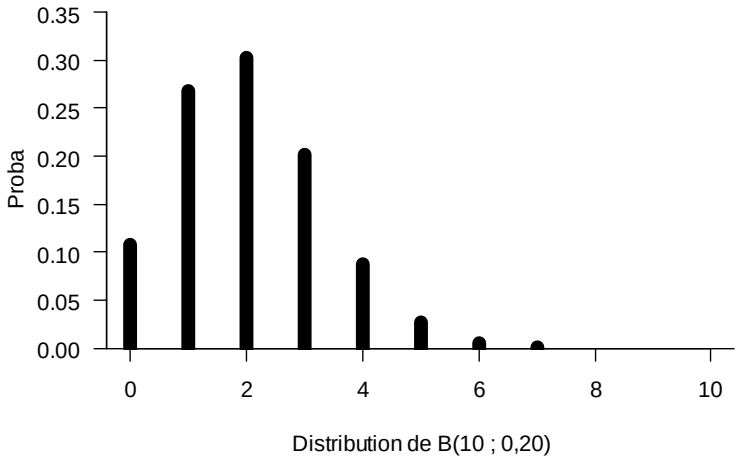

Figure 2 : fonction de masse de la loi binomiale $\boldsymbol{B(10\,;0,20)}$

La fonction de répartition, aussi appelée fonction cumulative de probabilités est définie comme suit :

$$F(k) = P(X \leq k) = BCDF(k;n,p) = \left(\sum_{x=0}^{k} P(X=x)\right) = \sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x} \qquad (5)$$

On définit la fonction des quantiles, ou fonction inverse cumulative de probabilités comme :

$$F^{-1}(q) = BiCDF(q;n,p) = \min\{k|F(k) \geq q\} \qquad (6)$$

Les propriétés basiques de la loi binomiale sont bien connues :

Statistique	**Formule**	
Espérance	$E(X) = np$	(7)
Variance	$VAR(X) = np(1-p)$	(8)
Écart-type	$SD(X) = \sqrt{np(1-p)}$	(9)
Coefficient d'asymétrie (skewness)	$SKEW(X) = \dfrac{1-2p}{\sqrt{np(1-p)}}$	(10)
Coefficient d'aplatissement (Kurtosis normalisé)	$EKUR(X) = \dfrac{1-6p(1-p)}{np(1-p)}$	(11)

Tableau 1 : premiers moments d'une variable X suivant une loi binomiale $B(n;p)$.

Pour rappel, la variance d'une variable aléatoire X est :

$$VAR(X) = E\left(\left(X - E(X)\right)^2\right) \tag{12}$$

son coefficient d'asymétrie est :

$$SKEW(X) = E\left(\left(\frac{X - E(X)}{\sqrt{VAR(X)}}\right)^3\right) \tag{13}$$

et son kurtosis normalisé est :

$$EKUR(X) = E\left(\left(\frac{X - E(X)}{\sqrt{VAR(X)}}\right)^4\right) - 3 \tag{14}$$

1.1.3.2 Proportion binomiale

Si $X \sim B(n; p)$ est une variable aléatoire suivante une loi binomiale, telle que le nombre de sujets atteints d'une maladie dans un échantillon aléatoire issu d'une population, alors $\hat{P} = \frac{X}{n}$ est la proportion observée de malades dans l'échantillon. C'est aussi un estimateur non biaisé de p la prévalence de la maladie dans la population, ce qui signifie que l'espérance de $\hat{P}$ est égale à p. Il n'y a pas de biais d'estimation qui tendrait à systématiquement surestimer ou sous-estimer la prévalence de la maladie.

En bref, la loi binomiale représente le nombre de malades alors que la proportion binomiale représente la proportion de sujets malades. À une constante près (la taille de l'échantillon), les deux suivent la même loi. Cette constante, change l'espérance et la variance.

Statistique	**Formule**	
Espérance	$E(\hat{P}) = p$	(15)
Variance	$VAR(\hat{P}) = \frac{p(1-p)}{n}$	(16)
Écart-type	$SD(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$	(17)
Coefficient d'asymétrie (skewness)	$SKEW(\hat{P}) = \frac{1-2p}{\sqrt{np(1-p)}}$	(18)
Coefficient d'aplatissement (Kurtosis normalisé)	$EKUR(\hat{P}) = \frac{1-6p(1-p)}{np(1-p)}$	(19)

Tableau 2 : premiers moments d'une proportion binomiale $\hat{P}$ telle que $n\hat{P} \sim B(n; p)$.

1.1.3.3 Lien avec la loi de Poisson

La loi de Poisson peut être vue comme un cas limite de la loi binomiale dans lequel le dénominateur (taille de l'échantillon) tend vers l'infini, alors que le numérateur np reste constant.

Cette loi permet de décrire l'incidence d'une maladie. Un médecin peut compter le nombre de sujets dont il fait le diagnostic dans son cabinet sur un mois, mais il ne connaît pas forcément exactement le bassin de population susceptible de le consulter en cas de maladie. Pour une maladie aiguë de courte durée ou une maladie chronique rare, on peut négliger le fait que les sujets malades ne peuvent pas faire une deuxième fois la maladie. Dans ce cas, le bassin de population susceptible de consulter ce

médecin en cas de maladie est constitué de n individus avec chacun une probabilité p de développer la maladie dans le mois et le nombre X de patients consultant ce médecin pour cette maladie suit une loi binomiale $B(n;p)$. Le médecin ne connaît pas n mais il peut estimer np par le nombre de sujets qu'il aura observé sur le mois. Dans ce cas, la loi binomiale sera approximable à une loi de Poisson d'espérance np. Pour une maladie d'incidence stable, son estimation de l'espérance de la loi de Poisson lui permettra à la fois de calculer l'espérance du nombre de sujets qu'il diagnostiquera le mois suivant, mais aussi sa variance, car la variance d'une loi de Poisson est égale à son espérance.

Dans ce travail, la loi de Poisson pour des espérances inférieures à 64, sera souvent approximée à une loi binomiale correspond à 2048 expériences de Bernoulli.

1.1.3.4 Rapidité des calculs

Les algorithmes naïfs décrits ci-dessus sont particulièrement peu performants pour des grandes valeurs de n. Notamment, le calcul de factorielle 171 (égal à $1{,}24 \times 10^{309}$), nécessaire au calcul de la fonction de masse pour $x = 171$, dépasse les capacités de calcul des logiciels statistiques qui ne gèrent généralement pas les nombres plus grands que $1{,}8 \times 10^{308}$; c'est la limite de capacité des nombres à virgule flottante 64 bits selon la norme IEEE-754. La fonction de masse f peut être calculée rapidement en s'aidant des premiers termes de la série de Stirling-Moivre (Loader (69)) qui approximent la fonction factorielle ou la fonction log-factorielle (ou même log-gamma) en un nombre fixe d'opérations. L'usage de transformations logarithmiques simplifie les calculs et évite les débordements de capacité des nombres à virgule flottante. Ainsi, l'algorithme proposé par Loader, utilisé par le logiciel R, calcule précisément la fonction de masse en une trentaine d'opérations (multiplications, divisions, additions, soustractions) de nombre à virgule flottante, plus une racine carrée et une exponentielle. Le temps de calcul ne dépend plus de n et x.

Des simplifications permettent aussi au logiciel R d'estimer la fonction de répartition F de la loi binomiale en s'aidant de la loi béta et les quantiles F^{-1} en s'aidant du développement de Cornish-Fisher (33) puis d'un algorithme de recherche sur F.

Tout ceci, associé aux améliorations des performances des ordinateurs rendent les calculs exacts sur la loi binomiale extrêmement rapides.

1.1.4 Théorie de l'estimation

1.1.4.1 Estimateur, estimation

De manière schématique, un *estimateur* est une méthode (par exemple, une formule mathématique) qui, à partir d'un échantillon, calcule une valeur numérique qui se rapproche d'une statistique de la population. Pour un échantillon donné, la valeur numérique précise s'appelle l'*estimation*. Chaque échantillon a une estimation différente de la statistique de la population. On peut alors définir une variable aléatoire dont l'unité statistique est l'échantillon, associant à chaque échantillon l'estimation correspondante. On peut appeler *estimateur* cette variable aléatoire. La population concernée par cette variable aléatoire n'est pas la même que la population des individus. En effet, c'est une population d'échantillons, constituée de tous les échantillons imaginables. Comme l'estimateur est une variable aléatoire, toutes les statistiques classiques, telles que la moyenne, la variance, le coefficient d'asymétrie, sont calculables sur cet estimateur.

Par convention, les statistiques de la population d'individus se présentent sous forme de lettres grecques minuscules. La moyenne est notée μ et l'écart-type est noté σ. Il y a des exceptions. Pour la

loi binomiale, on note souvent p la moyenne. D'autres auteurs notent π la moyenne d'une loi binomiale, à ne pas confondre avec la constante mathématique $\pi = 3{,}14159265\ldots$ Dans la théorie fréquentiste, ces statistiques sont fixes mais généralement inconnues. Par exemple, la prévalence exacte du Pemphigus en France en 2017 existe, pourrait être notée p ou π mais n'est pas connue. Les estimateurs sont notés par des lettres majuscules grecques surmontées d'un chapeau. Par exemple, un estimateur de l'écart-type sera noté $\hat{\Sigma}$. Les estimations sont soit notées par des lettres minuscules latines (s pour l'écart-type), soit notées par des lettres minuscules grecques surmontées d'un chapeau ($\hat{\sigma}$ pour l'écart-type).

Voici quelques exemples dans le cadre la loi binomiale. Supposons qu'une population contienne des sujets diabétiques et qu'on définisse une variable de Bernoulli D valant 1 pour les sujets malades et 0 pour les sujets sains. Des échantillons aléatoires de taille n sont tirés au sort dans cette population. Notons $\{D_1, \ldots, D_n\}$ les n variables aléatoires, dont l'unité statistique est l'échantillon, correspondant respectivement, à chacune des observations, numérotées de la 1ère à la nème. Alors, nous définirons :

p la prévalence du diabète dans la population et X la variable aléatoire égale au nombre de sujets diabétiques dans l'échantillon, égale à la somme des D_i pour $i = 1, \ldots, n$ comme défini ci-dessous :

$$X = \sum_{i=1}^{n} D_i \tag{20}$$

L'unité statistique de X est l'échantillon.

L'échantillonnage étant supposé simplement randomisé, alors $X \sim B(n; p)$ car X est la somme de n variables de Bernoulli indépendantes identiquement distribuées. La variable aléatoire $\hat{P}$ définie ci-dessous, dont l'unité statistique est l'échantillon, est un estimateur de p :

$$\hat{P} = \frac{X}{n} = \frac{\sum_{i=1}^{n} D_i}{n} \tag{21}$$

Pour une réalisation de l'échantillon, c'est-à-dire, un tirage donné d'un échantillon, on note $\{d_1, \ldots, d_n\}$ les n réalisations des D_i. Par exemple, si $n = 4$, on pourra tirer $\{d_1, \ldots, d_n\} = \{0,0,1,0\}$. La réalisation x de la variable aléatoire X vaudra alors $x = \sum_{i=1}^{n} d_i = 0 + 1 + 0 + 0 = 1$. L'estimation de la prévalence du diabète sera alors $\hat{p} = \frac{x}{n} = \frac{1}{4}$.

La variance de l'estimateur $VAR(X)$ est égale à $np(1-p)$ car X suit une loi binomiale. Par contre, la variance de x n'a pas de sens, car x, en tant que réalisation d'une variable aléatoire, est une constante égale à 1 dans notre cas. De même, $\hat{p} = \frac{1}{4}$ n'a pas de variance associée, c'est une constante. Enfin, la variance de $\hat{P}$ est bien définie car $\hat{P}$ en tant qu'estimateur, est une variable aléatoire. Cette variance est connue :

$$VAR(\hat{P}) = \frac{p(1-p)}{n} \tag{22}$$

Lorsque p est inconnu, deux estimateurs de la variance peuvent être définis :

$$\widehat{VAR}_1(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n} \tag{23}$$

$$\widehat{VAR}_2(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n-1} \quad (24)$$

L'estimateur $\widehat{VAR}_1$ est biaisé, son espérance étant égal à $\frac{n-1}{n}VAR(\hat{P})$, alors que l'estimateur $\widehat{VAR}_2$ n'est pas biaisé, son espérance étant égale à $VAR(\hat{P})$. En effet, la variance observée sur un échantillon d'observations indépendantes identiquement distribuées est toujours un estimateur biaisé de la variance de la population, quel que soit la distribution de la variable. La loi binomiale n'y fait pas exception. Ceci est dû au fait que la variance de l'échantillon est égale à la moyenne des carrés des écarts à la moyenne de l'échantillon plutôt que la moyenne des carrés des écarts à la moyenne de la population. La moyenne de l'échantillon a une position plus centrale dans l'échantillon que la moyenne de la population, et minimise mieux les carrés des écarts à cette position. Asymptotiquement (lorsque $n \to +\infty$) les deux estimateurs sont équivalents, mais il existe une différence sur les très petits échantillons.

1.1.4.2 Intervalle de fluctuation

Notons A une variable aléatoire de loi de distribution parfaitement caractérisée dans une population. Par exemple, notons A la variable de Bernoulli égale à 1 pour les femmes et 0 pour les hommes dans la population française. Grace au recensement, l'INSEE connaît suffisamment précisément cette variable de Bernoulli pour que l'incertitude soit négligée. Au 1er janvier 2017 l'INSEE (57) estimait la population française, incluant Mayotte, à 32 455 859 hommes et 34 534 967 femmes. La variable de Bernoulli a donc un paramètre $p = \frac{34\,534\,967}{32\,455\,859+34\,534\,967} = 0{,}516$. Le nombre de femmes dans un échantillon aléatoire de taille $n = 30$ suit alors une loi binomiale $X \sim B(30; 0{,}516)$. En reprenant les notations définies dans la section « Estimateur, estimation » on définit aussi $\hat{P} = \frac{X}{n}$ un estimateur du taux de féminité de la population française et $\hat{p} = \frac{x}{n}$ l'estimation de ce taux sur un échantillon donné. C'est aussi le taux observé de femmes sur cet échantillon. L'espérance de X est égale à $np = 30 \times 0{,}516 = 15{,}48$.

Un intervalle de fluctuations à 95% du taux de féminité aussi appelé intervalle de fluctuations au risque 5% et noté $IF_{0,95}$, est un intervalle construit autour du taux de féminité $p = 0{,}516$ de telle sorte qu'en multipliant les expériences d'échantillonnage aléatoire, 95% des estimations seraient contenues dans l'intervalle.

Formellement :

$$Proba(\hat{P} \in IF_{0,95}) = 0{,}95 \quad (25)$$

L'intervalle est fixe, l'estimateur $\hat{P}$ varie d'un échantillon à l'autre. Ce principe est illustré sur la Figure 3.

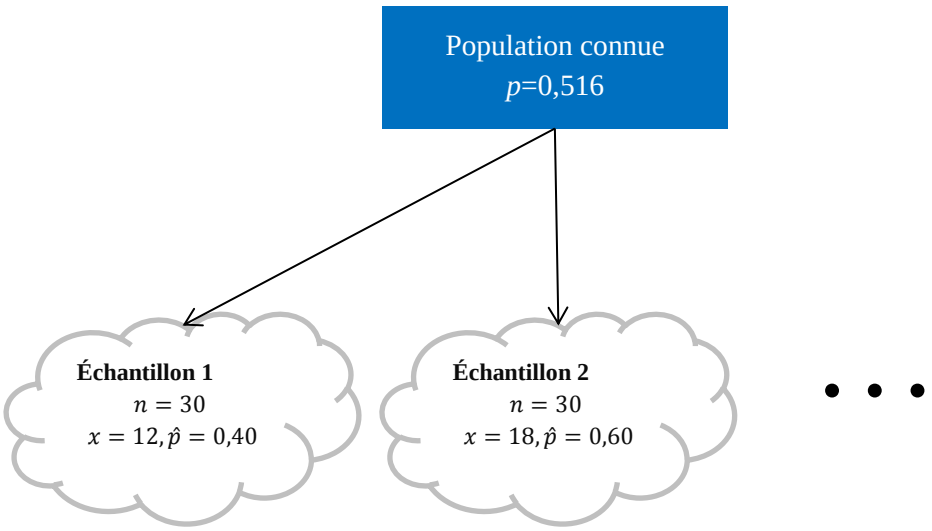

Figure 3 : illustration des données connues (carré bleu) et des données théoriques (nuages blancs) dont les probabilités sont calculables par la connaissance de la loi binomiale $B(30\ ;\ 0,516)$. Ces connaissances théoriques permettent le calcul d'un intervalle de fluctuations.

Dans le cas présent où $p = 0{,}516$ et $n = 30$, aucun intervalle ne vérifie parfaitement cette propriété. Les deux intervalles s'en rapprochant le plus sont $[10\ ;\ 20]$ et $[11\ ;\ 21]$.

$$Proba\left(\hat{P} \in [10\ ;\ 20]\right) = 0{,}954 \tag{26}$$

$$Proba\left(\hat{P} \in [11\ ;\ 21]\right) = 0{,}953 \tag{27}$$

On remarquera aussi que

$$Proba\left(\hat{P} \in [10{,}9\ ;\ 21{,}4]\right) = 0{,}953 \tag{28}$$

Dans un contexte d'intervalle de fluctuation, la probabilité effective que l'estimateur soit contenu dans l'intervalle, est appelé *couverture réelle* de l'intervalle. La couverture de $[10\ ;\ 20]$ est $0{,}954$. La couverture cible de 0,95 est appelée *couverture nominale*. La différence entre la couverture réelle et la couverture nominale est le *biais de couverture* de l'intervalle de fluctuation. Par raccourci, la *couverture*, sans précision, désigne la couverture réelle.

La loi binomiale étant quantitative discrète, les intervalles contiendront forcément un nombre entier de valeurs du support de la loi binomiale. C'est-à-dire, sur les 31 valeurs possibles (0 à 30), chacune sera, soit contenue, soit non contenue dans l'intervalle. La probabilité que $\hat{P}$ soit inclus dans l'intervalle sera égale à la somme des probabilités de réalisation de chacune des valeurs possibles contenues dans l'intervalle. Par exemple :

$$\begin{aligned} Proba\left(\hat{P} \in [10\ ;\ 12]\right) &= Proba\left(\hat{P} = 10\right) + Proba\left(\hat{P} = 11\right) + Proba\left(\hat{P} = 12\right) \\ &= 0{,}02001 + 0{,}03878 + 0{,}06546 = 0{,}1243 \end{aligned} \tag{29}$$

Le nombre de choix est limité et on ne peut obtenir exactement la probabilité 0,95.

Certains intervalles pourtant très différents, ont des couvertures proches.

$$Proba\left(\hat{P} \in [0\,;20]\right) = 0{,}968 \tag{30}$$

$$Proba\left(\hat{P} \in [11\,;30]\right) = 0{,}966 \tag{31}$$

Même si ces deux intervalles ont presque la même couverture, la probabilité complémentaire $Proba\left(\hat{P} \notin [0\,;20]\right)$, se décompose de manière différente.

$$Proba\left(\hat{P} \notin [0\,;20]\right) = Proba\left(\hat{P} < 0\right) + Proba\left(\hat{P} > 20\right) = 0 + 0{,}0321 \tag{32}$$

$$Proba\left(\hat{P} \notin [11\,;30]\right) = Proba\left(\hat{P} < 11\right) + Proba\left(\hat{P} > 30\right) = 0{,}0339 + 0 \tag{33}$$

Ces deux intervalles de fluctuations, n'ayant chacun qu'un risque d'un seul côté de l'intervalle, sont appelés intervalles de fluctuations *unilatéraux*. L'intervalle $[0\,;20]$ est unilatéral à droite alors que l'intervalle $[11\,;30]$ est unilatéral à gauche.

Les intervalles qui ne sont pas unilatéraux sont appelés intervalles *bilatéraux*. Lorsque le risque que l'estimateur soit en-dessous de l'intervalle est égal au risque qu'il soit au-dessus, on parle d'intervalle bilatéral à *risques symétriques* ou à *risques équilibrés*. Le terme peut aussi s'appliquer à des procédures de construction d'intervalle visant à équilibrer les risques.

Même si l'intervalle de fluctuation à 95% a été présenté, le principe est applicable à n'importe quelle couverture nominale comprise entre 0 et 1.

1.1.4.3 Intervalle de confiance

L'intervalle de fluctuations est un outil théorique qui n'est pas manipulé dans la pratique, car sa construction nécessite de connaître parfaitement la loi de probabilité de la statistique dans la population (cf Figure 3). Il sert à prédire ce qu'on observera dans les échantillons connaissant la population. La situation inverse est beaucoup plus fréquente. Disposant d'un échantillon dont les caractéristiques sont bien connues, le statisticien souhaiterait connaître une statistique de la population. À défaut de pouvoir exactement calculer la statistique de la population, il peut l'approcher par une estimation ponctuelle (cf section Estimateur, estimation) et y associer un intervalle dans lequel il parie que la valeur de la population se situera.

Un *intervalle de confiance* à 95% est un intervalle construit par une méthode garantissant que la répétition du même protocole expérimental, à la limite du nombre d'expériences infini, produise 95% d'intervalles chanceux, contenant la statistique de la population, et 5% d'intervalles malchanceux, ne contenant pas la statistique de la population. La répétition des expériences est théorique. Chaque expérience produit un intervalle de confiance différent alors que la statistique de la population est fixe. On distingue l'intervalle de confiance obtenu au cours d'une expérience donnée, de l'*estimateur d'intervalle de confiance* qui est la méthode de calcul de l'intervalle ; ce dernier est aussi interprétable comme une variable aléatoire dont l'unité statistique est l'échantillon. Le terme intervalle de confiance est parfois ambigu.

Pour lever toute ambiguïté, on peut distinguer la *réalisation d'un intervalle de confiance* de l'*estimateur d'intervalle de confiance*. Cette distinction est importante car, la propriété selon laquelle

l'intervalle de confiance a 95% de chances de contenir la statistique de la population, est une propriété de l'estimateur mais pas de la réalisation. Une fois une expérience réalisée, l'intervalle de confiance calculé contiendra ou ne contiendra pas la valeur théorique. Ce problème peut être illustré par un cas exceptionnel dans lequel la loi théorique est connue. Reprenons l'exemple du taux de féminité de la population française au 1er janvier 2017, décrit dans la section « Intervalle de fluctuation » en page 26. Le nombre de femmes observé dans un échantillon de taille 30 est décrit par une variable aléatoire suivant une loi binomiale $X \sim B(30\ ;\ 0{,}516)$. En employant l'estimateur d'intervalle de confiance de Clopper-Pearson mid-P (11), la distribution de ses réalisations sur des expériences de tirage au sort d'échantillons aléatoires est présentée dans le Tableau 3.

Lorsque le nombre de femmes dans un échantillon aléatoire de taille 30 est compris entre 11 et 20, l'intervalle de confiance réalisé sur cet échantillon contiendra forcément le taux de féminité de la population française, c'est-à-dire 0,516. Par contre, lorsque le nombre de femmes observé sur l'échantillon est inférieur ou égal à 10 ou supérieur ou égal à 21, l'intervalle de confiance ne contiendra jamais la valeur du taux de féminité de la population française. Avant de lancer l'expérience, on peut parier qu'on observera un nombre de femmes compris entre 11 et 20 et qu'en conséquence, l'intervalle de confiance contiendra le taux de féminité de la population française. Les chances de gagner ce pari sont égales à $\left(\sum_{k=11}^{20} Proba(X = k)\right) = 0{,}934$, ce qui n'est pas très éloigné du taux visé égal à $0{,}95$.

Imaginons maintenant que l'on ait réalisé une expérience et qu'on ait observé 13 femmes sur les 30. L'intervalle de confiance du taux de féminité sera $[0{,}27\ ;\ 0{,}61]$. L'interprétation selon laquelle cet intervalle a 95% de chances de contenir le taux de féminité de la population est erronée puisqu'il contient manifestement ce taux (égal à 0,516). L'interprétation selon laquelle il y a 5% de chances que le taux de féminité de la population française est inférieur à 0,27 ou supérieur à 0,61 prête à sourire.

Dans la théorie fréquentiste, la statistique de la population (taux de féminité) est constante, bien que souvent inconnue, alors que l'intervalle de confiance est variable bien qu'on en observe un seul dans une expérience. Dans la théorie bayésienne que nous ne détaillerons pas, les observations sont fixes et l'incertitude sur la population est modélisée par une crédibilité de chaque valeur théorique. Cette dernière théorie est plus intuitive au sens où elle permet de directement fournir des probabilités concernant les valeurs théoriques. Cette théorie présente des problèmes, le principal étant la nécessité d'obtenir une information sur la crédibilité des valeurs théoriques avant même de débuter l'expérience (probabilités *a priori*). Cette information est subjective de telle sorte que les résultats fournis par les analyses bayésiennes sont eux-mêmes subjectifs sauf lorsque l'étude est suffisamment grande pour que la part de subjectivité soit négligeable en comparaison de l'information collectée par l'expérience. Le problème de subjectivité peut être résolu par l'usage d'hypothèses *a priori* neutres (distribution *a priori* non informatives). Dans ce cas, la théorie bayésienne devient presque équivalente à la théorie fréquentiste mais simplifie certains calculs dans les modèles les plus complexes. Quelques-uns des intervalles de confiance présentés dans cette étude sont issus de la théorie bayésienne.

Valeur de X réalisé	Probabilité de cette réalisation	Réalisation d'intervalle de confiance	L'intervalle de confiance contient-il la réalisation
0	3,51E-10	[0,00 ; 0,10]	NON
1	1,12E-08	[0,00 ; 0,15]	NON
2	1,74E-07	[0,01 ; 0,20]	NON
3	1,73E-06	[0,03 ; 0,25]	NON
4	1,24E-05	[0,04 ; 0,29]	NON
5	6,89E-05	[0,06 ; 0,33]	NON
6	3,06E-04	[0,09 ; 0,37]	NON
7	1,12E-03	[0,11 ; 0,41]	NON
8	3,43E-03	[0,13 ; 0,44]	NON
9	8,94E-03	[0,16 ; 0,48]	NON
10	2,00E-02	[0,18 ; 0,51]	NON
11	3,88E-02	[0,21 ; 0,55]	OUI
12	6,55E-02	[0,24 ; 0,58]	OUI
13	9,66E-02	[0,27 ; 0,61]	OUI
14	1,25E-01	[0,30 ; 0,64]	OUI
15	1,42E-01	[0,33 ; 0,67]	OUI
16	1,42E-01	[0,36 ; 0,70]	OUI
17	1,25E-01	[0,39 ; 0,73]	OUI
18	9,61E-02	[0,42 ; 0,76]	OUI
19	6,47E-02	[0,45 ; 0,79]	OUI
20	3,80E-02	[0,49 ; 0,82]	OUI
21	1,93E-02	[0,52 ; 0,84]	NON
22	8,40E-03	[0,56 ; 0,87]	NON
23	3,12E-03	[0,59 ; 0,89]	NON
24	9,69E-04	[0,63 ; 0,91]	NON
25	2,48E-04	[0,67 ; 0,94]	NON
26	5,08E-05	[0,71 ; 0,96]	NON
27	8,03E-06	[0,75 ; 0,97]	NON
28	9,17E-07	[0,80 ; 0,99]	NON
29	6,74E-08	[0,85 ; 1,00]	NON
30	2,40E-09	[0,90 ; 1,00]	NON

Tableau 3 : distribution de la loi binomiale $B(30\,;0,516)$ et description de l'estimateur d'intervalle de confiance de Clopper-Pearson mid-P de la proportion binomiale, par ses réalisations à tous les échantillons possibles. En vert, les intervalles de confiance contenant la proportion binomiale $0,516$.

D'une manière plus générale, pour une valeur α choisie entre 0 et 1 (souvent égal à 0,05 en biostatistiques), appelée *risque nominal*, un *estimateur d'intervalle de confiance au risque* α d'une statistique θ est une variable aléatoire $IC_{1-\alpha}$ dont l'unité statistique et l'échantillon, qui vérifie la propriété :

$$\Pr oba(\theta \in IC_{1-\alpha}) = 1 - \alpha + \varepsilon \qquad (34)$$

Où ε est le biais de couverture, $1-\alpha$ la *couverture nominale*, $1-\alpha+\varepsilon$ la *couverture réelle*, α le *risque nominal* et $\alpha-\varepsilon$ le *risque réel*. Un bon estimateur a un biais de couverture ε proche de zéro. Dans cette formule, θ est constant et $IC_{1-\alpha}$ est la variable aléatoire. Un estimateur d'intervalle de confiance au risque α est encore *appelé estimateur d'intervalle de confiance au niveau de confiance* $1-\alpha$, ou plus directement *estimateur d'intervalle de confiance à* $1-\alpha$.

Un estimateur d'intervalle de confiance est unilatéral à droite s'il est construit de telle sorte que l'intervalle ne puisse jamais être au-dessus de la statistique θ de la population. Cela est possible si les bornes basses des intervalles sont toujours inférieures ou égales à la valeur théorique minimale imaginable pour cette statistique. Par exemple, pour une proportion, la borne basse pourra être 0. De la même manière, un estimateur d'intervalle de confiance est unilatéral à gauche si l'intervalle ne peut jamais être en-dessous de la statistique θ. En d'autres termes, le risque α d'un estimateur d'intervalle de confiance unilatéral est complètement déporté d'un côté de l'intervalle. Le risque est nul d'un côté, et égal à $\alpha-\varepsilon$ de l'autre côté.

Définissons un estimateur d'intervalle de confiance $IC_{1-\alpha}$ par les estimateurs correspondant à sa borne basse $L_{1-\alpha}$ et sa borne haute $U_{1-\alpha}$.

$$IC_{1-\alpha} = [L_{1-\alpha}\ ;\ U_{1-\alpha}] \tag{35}$$

Un estimateur d'intervalle de confiance est bilatéral s'il n'est pas unilatéral. Lorsque le risque que l'intervalle soit entièrement au-dessus de θ est construit afin qu'il soit le plus proche possible du risque que l'intervalle soit entièrement en-dessous de θ, l'estimateur d'intervalle bilatéral est à *risques symétriques*, ou à *risques équilibrés*, ou tout *bilatéral équilibré*. Dans ce cas, les propriétés suivantes sont vérifiées :

$$\Pr oba(L_{1-\alpha} > \theta) = \frac{\alpha}{2} - \varepsilon_1 \tag{36}$$

$$\Pr oba(U_{1-\alpha} < \theta) = \frac{\alpha}{2} - \varepsilon_2 \tag{37}$$

expressions dans lesquelles ε_1 et ε_2 sont les biais partiels de couverture de telle sorte que $\varepsilon = \varepsilon_1 + \varepsilon_2$. Au sens strict, les risques ne sont équilibrés que lorsque $\varepsilon_1 = \varepsilon_2$. En pratique nous parlerons de risques équilibrés dès que l'estimateur vise à minimiser à la fois ε_1 et ε_2.

1.2 Problématique

Comme vu dans l'introduction de cette thèse, le calcul d'un intervalle de confiance n'est pas aussi simple qu'il n'y paraît. Les problèmes rencontrés lors de l'estimation d'une proportion sont décrits ci-dessous.

1.2.1 Biais dus à la loi binomiale

1.2.1.1 Distribution discrète : biais de couverture obligatoire et discontinu

Les oscillations de couverture dues à l'aspect discret de la loi binomiale ont été décrites par Brown (21). Aucun estimateur d'intervalle de confiance ne peut garantir une couverture réelle exactement égale à la couverture nominale. Il existera toujours un biais de couverture ε non nul. L'intervalle « exact » utilisé par la plupart des logiciels statistiques ne déroge pas à la règle. Cet intervalle « exact » est presque toujours l'intervalle de Clopper-Pearson (Clopper (31)). Cet intervalle étant extrêmement répandu et ayant des propriétés intéressantes il sera décrit en détail dans la section Matériel & méthodes. La couverture réelle, non seulement diffère de la couverture nominale, mais présente aussi des discontinuités selon p la proportion théorique (pour une taille d'échantillon n constante). Par exemple, sur un échantillon de taille 30 et pour une proportion théorique $p = 0{,}3471$, l'estimateur d'intervalle de confiance à 95% de Clopper-Pearson présente une couverture réelle égale à $0{,}9805$ alors que pour une proportion théorique très légèrement supérieure $p = 0{,}3473$, la couverture réelle descend brutalement à $0{,}9637$. Il existe une discontinuité dans la couverture au point $p = 0{,}3472$. Ce point correspond précisément à la borne haute de l'intervalle de confiance réalisé pour une proportion observée de $5/30$. En effet, l'intervalle de confiance correspondant est égal à $[0{,}0564, 0{,}3472]$. Il contient $0{,}3471$ mais pas $0{,}3473$. Les échantillons correspondant à 5 succès sur 30 observations participeront à la couverture pour une proportion théorique $p = 0{,}3471$ mais pas pour une proportion théorique $p = 0{,}3473$. Il existe une discontinuité aux deux bornes de chacun des 31 intervalles de confiance correspondant aux nombres de succès de 0/30 à 30/30. Ces discontinuités sont d'autant plus marquées que l'échantillon est de petite taille. Pour une proportion théorique p constante, il existe aussi des oscillations de couverture selon n, encore décrites par Brown (21).

Le problème des oscillations et du biais de couverture a pu être résolu, d'une certaine manière par Stevens (97) qui a proposé d'ajouter une part de hasard dans le calcul des bornes de l'intervalle de

confiance, de telle sorte que deux expériences conduisant à des échantillons identiques ne produisent pas le même intervalle de confiance. Il n'existe alors plus de discontinuité aux bornes des intervalles, car ces bornes ne sont plus constantes. L'application des intervalles randomisés est difficile dans la pratique biostatistique quotidienne. Cette difficulté est bien résumée par Stevens lui-même :

> "When any experiment has been performed (or series of observations taken), the investigator is allowed once and once only to select, at random, his value of x; the distribution thus determined will be called the fiducial distribution of n, and neither he nor anyone else is permitted another drawing of the number x".

Ce qui peut se traduire par :

> « Quand une expérience a été réalisée (ou une série d'observations a été faite), l'investigateur est autorisé une fois et une fois seulement à sélectionner, aléatoirement, sa valeur de x; la distribution ainsi déterminée sera appelée la distribution fiduciale de n, et ni lui ni personne d'autre n'a le droit de tirer un autre nombre x ».

Une alternative a été proposée par Geyer (50). Il s'agit d'*intervalles flous*, aussi appelés *intervalles randomisés abstraits*. Plutôt que de choisir une constante de randomisation, toutes les possibilités de randomisation sont présentées, avec la densité de probabilité des bornes de l'intervalle de confiance. Même si l'auteur de cette thèse admire l'intelligence de cette solution théorique, elle lui paraît difficilement communicable. Les intervalles randomisés de Stevens sont peut-être plus simples à communiquer mais nécessitent une grande rigueur de travail, et, de préférence pas de conflit d'intérêt.

1.2.1.2 Distribution asymétrique : biais des approximations

La distribution binomiale $B(n\,;\,p)$ est asymétrique sauf pour $p = 0{,}50$. Le coefficient d'asymétrie (skewness) tend vers $+\infty$ lorsque p tend vers 0 et vers $-\infty$ lorsque p tend vers 1. Pour un nombre constant de succès $\lambda = np$, l'asymétrie de distribution binomiale augmente avec n et s'approche de celui de la distribution de Poisson lorsque $n \to +\infty$. Comme la loi normale est symétrique, pour une constante λ, l'erreur de l'approximation normale s'aggrave lorsque n augmente. En conséquence les estimateurs d'intervalles de confiance basés sur la loi normale, tels que le Wald et le Wilson, sont plus fortement biaisés pour un nombre attendu de 5 succès sur un échantillon de taille 15, que pour un nombre attendu de 5 succès sur un échantillon de taille 1000. Le cas limite où le nombre attendu de succès serait égal à 5 mais l'échantillon serait de taille infinie correspond à la loi de Poisson et reste toujours biaisé, ce qui a conduit Reiczigel (87) à conclure que certains estimateurs d'intervalle de confiance ne sont pas asymptotiquement exacts. En réalité, tous les estimateurs d'intervalle présentés seront asymptotiquement exacts car le scénario asymptotique d'une proportion binomiale correspond à un p constant, $n \to +\infty$ et donc, $np \to \infty$. Lorsqu'on augmente indéfiniment la taille de l'échantillon, le nombre de succès augmente aussi indéfiniment et les biais s'atténuent. Les estimateurs d'intervalles de confiance basés sur la loi binomiale exacte ou prenant en compte l'asymétrie sont résistants, voire immunisés à ce problème.

1.2.1.3 La relation entre variance et moyenne

La moyenne d'une loi binomiale $B(n;p)$ est égale à p alors que sa variance est égale à $\frac{p(1-p)}{n}$. Quand np est proche de 0 ou de 1, les fluctuations de la variance de l'échantillon $\frac{\hat{P}(1-\hat{P})}{n}$ sont grandes et sont dépendantes des fluctuations de $\hat{P}$. Ceci est particulièrement problématique pour les nombreuses méthodes qui approximent la variance de la population à la variance de l'échantillon. Ce problème a mo-

tivé l'usage des transformations stabilisatrices de variances telles que celle décrite par Bartlett (10) ou celles revues par Yu (112). Les estimateurs d'intervalle de confiance par inversion de tests, consistant à analyser toutes les valeurs théoriques de p compatibles avec la valeur observée sur l'échantillon, sont immunisés à ce problème.

1.2.2 Enjeux dans l'estimation des intervalles de confiance

Certaines propriétés des estimateurs d'intervalles de confiance sont désirables mais toutes ne sont pas compatibles. Optimiser une propriété conduira à une perte sur une autre propriété. Un estimateur peut être mauvais sur toutes les propriétés mais aucun estimateur ne peut être optimal sur toutes les propriétés.

1.2.2.1 Maîtrise du risque conditionnel, du risque moyen ou du risque moyen local

La minimisation du biais de couverture est souhaitable. On peut alors parler de maîtrise de la couverture, puisqu'il s'agit de rapprocher autant que possible la couverture réelle de la couverture nominale. En analysant la probabilité complémentaire à la couverture, ou risque α, on peut parler de maîtrise du risque α. Un estimateur d'intervalle dont la couverture réelle est supérieure au risque nominal est dit *conservatif* alors qu'un intervalle dont la couverture réelle est inférieure au risque nominal est dit *libéral* ou *anti-conservatif*.

Comme le biais de couverture ne peut être annulé, deux approches de sa maîtrise ont été proposées. Certains intervalles tels que celui proposé par Clopper et Pearson, Blaker ou encore Blyth et Still (12,15,31) garantissent une couverture réelle supérieure ou égale à la couverture nominale, quels que soient n et p, supposés tous les deux fixes. Ces intervalles sont donc strictement conservatifs, au sens où ils ne sont jamais libéraux. En conséquence ils sont plus larges que des intervalles plus libéraux. Leur trop fort conservatisme a été critiqué par Agresti (4) qui proposait de rechercher un équilibre entre le conservatisme (pour certaines valeurs de p) et le libéralisme (pour d'autres valeurs de p) afin de minimiser le biais de couverture moyen. Les biais positifs (conservatisme) et des biais négatifs (libéralisme) peuvent se compenser pour donner un biais moyen proche de zéro. Dans ce cas, la couverture réelle moyenne se rapproche de la couverture nominale.

La couverture moyenne proposée par Agresti se base sur l'hypothèse d'une proportion théorique p variable d'une expérience à l'autre, selon une loi prédéfinie. Agresti a proposé la loi uniforme sur l'intervalle $]0\,;1[$. Ainsi, on peut comprendre le plan expérimental comme la succession de deux étapes : tirage au sort d'une proportion théorique p selon une loi uniforme sur $]0\,;1[$, puis tirage au sort d'un échantillon aléatoire dont le nombre de succès suit une loi $B(n\,;\,p)$. D'un point de vue bayésien, cela veut dire, que la répartition *a priori* de la proportion théorique est uniforme, ou, encore, que toutes les proportions sont aussi crédibles les unes que les autres *a priori*. Sous cette hypothèse, un intervalle libéral pour les proportions théoriques inférieures à 10% et conservateur pour les proportions théoriques supérieures à 90% pourra avoir une couverture moyenne satisfaisante, le libéralisme et le conservatisme se compensant bien que situés à des positions opposées de l'intervalle. Le plus souvent, on sait à l'avance si la proportion estimée sera faible ou forte. Le taux d'effets indésirables graves (EIG) d'un traitement sera généralement inférieur à 10% alors que le taux de réponse sera souvent supérieur à 70%. Dans ce cas, à l'avance on peut prévoir que cet estimateur d'intervalle de confiance sera libéral pour les effets secondaires. L'incertitude concernant le taux d'EIG sera sous-estimée et le risque que le taux soit sous-estimé, c'est-à-dire, que la borne haute de l'intervalle de confiance soit inférieure au taux réel d'EIG, risque d'être élevé. En d'autres termes, avec cette approche du risque moyen, un estimateur d'intervalle de confiance qui sous-estime les effets secondaires et surestime l'efficacité des traitements aura l'air de bien maîtriser les risques puisqu'une erreur compense

l'autre. Agresti a aussi proposé une loi béta de moyenne 0,10 et écart-type 0,05, correspondant donc à l'estimation d'un risque moyen au voisinage de 10%, plus approprié à l'analyse des effets indésirables. Les tirages au hasard de la valeur théorique p sont alors majoritairement situés près de 10% et un biais de couverture pour une proportion de 90% ne peut plus compenser un biais de couverture pour une proportion basse. Agresti n'est pas le seul à s'être posé la question de la distribution *a priori* lorsqu'on analyse les risques moyen, Newcombe (77) ayant aussi mentionné ce problème mais n'ayant présenté que les résultats concernant la loi uniforme. Bien que la conscience du problème de l'information *a priori* sur la proportion soit là, les analyses d'Agresti (4) ou de Newcombe (77) restent insuffisantes pour apprécier le comportement des intervalles au voisinage de chaque proportion théorique.

En poussant plus loin les analyses d'Agresti, nous analyserons le *risque moyen local* ou *risque moyen de voisinage* autour d'une valeur p_0 fixe, et pour un n fixe, correspondant à un plan expérimental en deux étapes : tirage au sort d'une valeur théorique p proche de p_0 dans une loi logit-normale de faible variance, puis tirage au sort d'un échantillon aléatoire avec estimation dont le nombre de succès suit une loi binomiale $B(n\,;p)$. L'analyse de toutes les valeurs de p_0 permettra de décrire le comportement de l'estimateur dans tous les voisinages.

Nous définissons les *risques α conditionnels* comme les risques que l'intervalle ne contienne pas la proportion théorique pour une proportion théorique fixée p et une taille d'échantillon fixée n. Ce risque correspondrait à la répétition d'expériences parfaitement maîtrisées, dans lesquelles le nombre exact de sujets inclus est connu et l'échantillonnage aléatoire s'applique à une même population stable. Ce sont les risques présentés dans les articles de revue systématique de Brown (21) ou de Pires (83).

L'analyse du risque moyen local peut être justifiée par le constat d'une hétérogénéité dans les méta-analyses. C'est-à-dire, pour des protocoles expérimentaux similaires, les proportions observées diffèrent plus fortement que l'on ne peut l'expliquer par les fluctuations d'échantillonnage. Les raisons de cette hétérogénéité sont multiples : la population diffère (recrutement dans des régions différentes, dans des contextes différents), l'échantillonnage est rarement aléatoire et la participation incomplète biaise encore l'échantillonnage, les interventions et les mesures diffèrent car elles sont faites par des opérateurs différents et les protocoles ne sont jamais parfaitement standardisés. Une part de l'hétérogénéité est due à des biais, telles que le biais de sélection alors qu'une autre partie est due à de réelles différences, notamment de population et d'intervention. Dans ce cas il est légitime de considérer que la proportion p réelle dans la population, diffèrera d'une étude à l'autre, bien que les protocoles expérimentaux soient très proches voire identiques. Ceci a conduit aux méta-analyses à effets aléatoires dans lesquels on suppose que la statistique de la population est une variable aléatoire dont l'unité statistique est l'étude. Pour l'estimation d'une proportion, on supposera qu'il existe une proportion théorique moyenne p_0 égale à l'espérance des proportions réelles p de toutes les études. On suppose que les proportions p suivent une loi aléatoire fluctuant autour de p_0. Dans un modèle de régression logistique à effets mixtes, adapté à l'analyse des proportions, on supposera que la distribution des proportions réelles p est logit-normale, centrée autour de la proportion réelle moyenne p_0. C'est-à-dire, on suppose que $\log\left(\frac{P}{1-P}\right)$ suit une loi normale. Un autre paramètre est mal maîtrisé en recherche biomédicale : le nombre de sujets de l'échantillon. Même lorsque le nombre de sujets planifié est établi dans un protocole, le nombre de sujets réellement inclus dans l'analyse est légèrement inférieur ou supérieur, car le recrutement est rarement arrêté brutalement, les erreurs d'inclusion et les données manquantes (sauf en analyse en intention de traiter) font disparaître quelques observations de l'analyse finale. En d'autres termes, le nombre de sujets planifié est souvent mal maîtrisé. On peut alors considérer que la taille de l'échantillon est elle-même une variable aléatoire dépendant de l'expérience. Ces fluctuations de taille d'échantillon seront analysées dans un deuxième temps.

L'analyse des risques conditionnels est justifiée pour les comparaisons de proportion observée à théorique puisque dans ce cas, la proportion théorique p définie dans le protocole n'est pas une variable aléatoire. Certes, la proportion réelle diffèrera de la proportion théorique, et est susceptible de fluctuer d'une expérience à l'autre. En analyse unilatérale, l'enjeu n'est plus de s'assurer que 95% des intervalles contiennent la proportion réelle (variable), mais il s'agit de s'assurer que si la proportion réelle est supérieure ou égale à la proportion théorique p, l'estimateur d'intervalle maîtrise le risque que l'intervalle de confiance soit entièrement en dessous de la proportion théorique. Si la taille de l'échantillon est mal maîtrisée, alors la modélisation des risques pour taille d'échantillon aléatoire restera intéressante.

Le risque moyen local semble un critère de jugement souvent plus pertinent que le risque conditionnel en biostatistique dans le contexte où la maîtrise des conditions expérimentales n'est pas parfaite. Nous présenterons néanmoins le risque moyen local pour p variable et n fixe, le risque moyen local pour p fixe et n variable, et les risques conditionnels pour p et n fixes.

1.2.2.2 Intervalle bilatéral à risques symétriques

Le problème de symétrie des risques a déjà été mentionné dans la section « Intervalle de confiance » du chapitre « Théorie statistique générale » en page 31. Cet enjeu mérite qu'on s'y attarde un peu plus.

Pour un estimateur d'intervalle de confiance d'une proportion au risque nominal α, le risque réel $\alpha - \varepsilon$ diffèrera peu ou prou du risque nominal. Ce risque réel est la probabilité, avant d'avoir réalisé l'expérience, que l'intervalle de confiance ne contienne pas la proportion réelle. Ce risque peut se décomposer en deux risques. Le premier, que nous appellerons risque à gauche, et le deuxième que nous appellerons risque à droite. Le *risque à droite* est le risque de sous-estimer la proportion réelle, la borne haute de l'intervalle de confiance étant inférieure à la proportion réelle. Le *risque à gauche* est le risque de surestimer la proportion réelle, la borne basse de l'intervalle de confiance étant supérieure à la proportion réelle. Un estimateur d'intervalle de confiance unilatéral aura un des deux risques nuls, et l'autre risque proche du risque nominal α. Tout le risque sera déporté d'un côté de l'intervalle.

Tout estimateur d'intervalle non unilatéral sera appelé bilatéral. Certains intervalles bilatéraux sont construits afin de rapprocher chacun des deux risques de $\frac{\alpha}{2}$; ce sont les intervalles *bilatéraux à risques équilibrés*. D'autres estimateurs d'intervalles sont nettement plus déséquilibrés et pourraient se rapprocher d'estimateurs unilatéraux. Par exemple l'estimateur d'intervalle du rapport de vraisemblance exact (Sakakibara (89)), pour un échantillon de taille $n = 2048$ et une proportion théorique $p = \frac{3{,}1}{2048}$ a des risques à droite et à gauche (conditionnels à n et p) égaux, respectivement à $0{,}0449$ et $0{,}0046$. Dans ce contexte, il se comporte presque comme un intervalle unilatéral à droite. Pour le même n et le même p l'estimateur d'intervalle de Blaker (12) a des risques à droite et à gauche respectivement égaux à $0{,}0000$ et $0{,}0387$. Il se comporte donc comme un intervalle unilatéral à gauche. En bref, à moins de bien connaître les propriétés de l'estimateur d'intervalle de confiance, on peut craindre qu'un estimateur déséquilibré fournisse un intervalle unilatéral sans qu'on sache de quel côté il est unilatéral.

L'équilibre des risques sera apprécié par l'analyse séparée du risque à droite et du risque à gauche. Pour les proportions inférieures à 50%, ce concept est équivalent aux défauts de couverture médial et distal décrits par Newcombe (77). L'analyse de la somme des deux risques, c'est-à-dire le risque bilatéral, ne sera faite que pour les risques conditionnels, afin de rapprocher les résultats de ceux qui ont déjà été comme ceux de Brown (21).

1.2.2.3 Précision de l'estimateur d'intervalle

Un intervalle de confiance peut être faiblement biaisé mais peu précis. Dans ce cas, l'intervalle de confiance sera large et instable. Un exemple caricatural de mauvais estimateur d'intervalle de confiance serait un mélange entre un estimateur d'intervalle à 90% pour les nombres pairs de succès sur un échantillon (p.e. 0/10, 2/10, 4/10, etc.) et un estimateur d'intervalle à 99,99% pour les nombres impairs de succès (1/10, 3/01, 5/10, etc.). La couverture serait assez proche de 95%, mais une expérience sur deux conduirait à un intervalle excessivement large (nombre impair de succès) qui ne serait pas compensé par l'intervalle un peu plus étroit obtenu pour les nombres pairs de succès. Dans ce cas, la largeur moyenne d'intervalle serait particulièrement large. L'intervalle serait instable et deux expériences dont les résultats bruts sembleraient proches (40/100 et 41/100) fourniraient pourtant des intervalles de confiance très différents. Donc, à erreur de couverture égale, un intervalle moins large est préféré.

La largeur de l'intervalle est classiquement analysé dans les articles de revue systématique tels que ceux d'Agresti, de Brown et de Pires (4,21,83). On peut parler de largeur moyenne d'intervalle ou de *largeur attendue* d'intervalle. D'autres mesures de précision existent, telles que le risque β d'une comparaison de valeur observée à théorique lorsque l'estimateur intervalle de confiance est employé comme un test d'hypothèse. On pourra aussi mentionner l'étroitesse de Neyman (78) qui correspond à la probabilité que l'intervalle de confiance contienne une valeur p'différente de p. Cette mesure n'est pas facilement applicable aux scénarii envisagés, car elle fait appel à une valeur p' supplémentaire, ou alors, elle nécessiterait une simplification qui la rendrait équivalente à la mesure de la largeur attendue.

De la même manière que les risques α à droite et à gauche des intervalles de confiance seront tous les deux analysés afin d'apprécier la symétrie des risques, les demi-largeurs à droite et à gauche seront aussi analysées. La *demi-largeur à droite* d'un intervalle de confiance est égale à la différence entre la borne haute de l'intervalle et l'estimation ponctuelle de la proportion, c'est-à-dire, la proportion observée sur l'échantillon. De même, la *demi-largeur à gauche* est la différence entre la proportion observée et la borne basse de l'intervalle. La *demi-largeur attendue à droite* est égale à l'espérance de la demi-largeur à droite dans un contexte expérimental défini. De même on définit la *demi-largeur attendue à gauche*.

1.2.2.4 Autres propriétés souhaitables

1.2.2.4.1 Cohérence avec un test d'hypothèse

Vos et Hudson (107) définissent la *p*-confiance comme un critère de jugement des estimateurs d'intervalle. Cette *p*-confiance est maximale lorsque les valeurs situées en dehors de l'intervalle seraient toutes fortement rejetées (P-valeur proches de zéro) par un test d'hypothèse. Le test d'hypothèse analysé par Vos et Hudson est le test binomial exact strictement conservatif à risques équilibrés, qui, une fois inversé, définit l'estimateur d'intervalle de Clopper-Pearson (31). De manière prévisible, les intervalles se rapprochant le plus de l'intervalle de Clopper-Pearson ont la meilleure *p*-confiance. D'une manière plus générale, lorsqu'un estimateur d'intervalle de confiance est construit par inversion d'un test d'hypothèse, il est souhaitable que les valeurs contenues dans l'intervalle soient acceptées par le test d'hypothèse et les valeurs en dehors de l'intervalle soient rejetées par le test d'hypothèse. Selon Vos et Hudson (107), cette propriété n'est pas assurée par les intervalles de Sterne (96), ou de Blaker (12). Elle n'est pas non plus assurée par les estimateurs d'intervalles obtenus par inversion de test exact du score ou de par inversion de test du rapport de vraisemblance exact décrits par Sakakibara (89). En effet, ces quatre estimateurs d'intervalle sont basés sur des tests fondés sur une fonction de P-valeur qui n'est pas bimonotone le long de la proportion théorique, comme expliqué par Fay (43), Blaker (13), Klaschka (60) ou Thulin (100). La Figure 4 montre les P-valeurs de comparaison propor-

tion théorique à proportion observée pour 4 succès sur un échantillon de 180 observations selon Sterne. Le taux observé est 4/180 = 2,22%. Au risque $\alpha = 0{,}05$, la proportion théorique 5,48% est acceptée (P-valeur = 0,0683), la proportion théorique 5,49% est rejetée (P-valeur = 0,0495) mais la proportion théorique 5,71% est acceptée (P-valeur = 0,0509). La région de valeurs théoriques acceptées (région de confiance) ne forme donc pas un intervalle, car elle n'est pas connexe. Sterne et Blaker ont donc défini leur intervalle comme le plus petit intervalle contenant la région de confiance. En conséquence, certaines valeurs rejetées par le test sont contenues dans l'intervalle de confiance. Ce comportement non monotone de la P-valeur peut être expliqué par sa méthode de calcul. Pour une proportion théorique p et un nombre de succès observé x sur un échantillon de taille n, la P-valeur de Sterne est égale à la probabilité d'obtenir un nombre de succès aussi probable ou moins probable que x sous l'hypothèse que x est la réalisation d'une variable binomiale $X \sim B(n\,;\,p)$. La formule de la P-valeur de Sterne est la suivante :

$$psterne(x,n,p) = \sum_{y=0}^{n} BPF(y;n,p) \times 1_{d(y,n,p)\leq d(x,n,p)} \tag{38}$$

où BPF représente la fonction de masse de la loi binomiale $B(n\,;p)$ comme décrit dans l'équation (3) en page 21 et $1_{propriété}$ est une fonction qui vaut 1 lorsque la propriété est vraie et 0 lorsqu'elle est fausse.

$$1_{propriété} = \begin{cases} 1\ si\ propriété \\ 0\ sinon \end{cases} \tag{39}$$

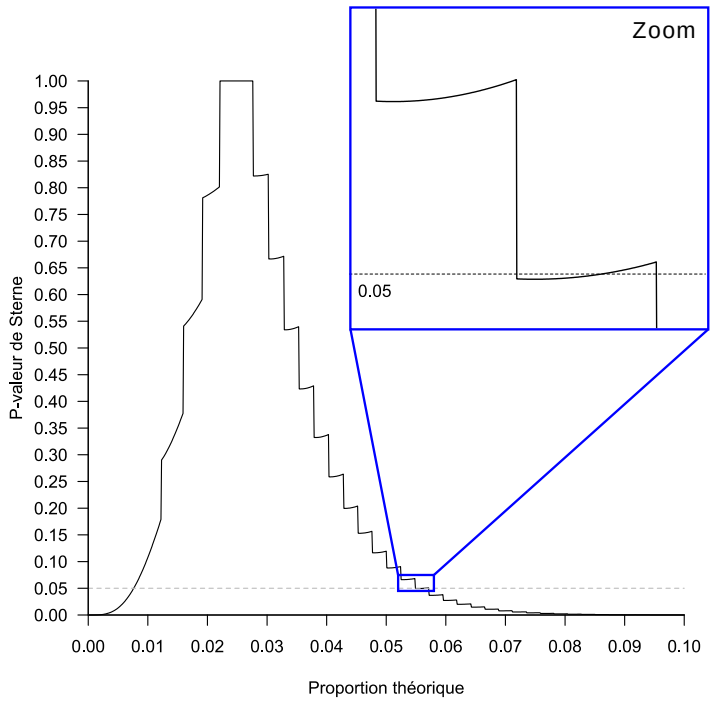

Figure 4 : P-valeur de la comparaison de proportion observée 4/180 à une proportion théorique, selon Sterne (96).

La Figure 5 illustre les paradoxes de la P-valeur de Sterne. Alors que la proportion théorique 0,803 (Figure 5B) est plus éloignée de la valeur observée 15/16=0,938 que la proportion théorique 0,818 (Figure 5A), la P-valeur correspondante est plus grande. Ceci est dû au fait que la somme des petites

flèches noires de la Figure 5B est positive. Alors que la proportion théorique diminue, la proportion observée s'éloigne, en probabilités, de la proportion théorique (flèches noires descendantes de la Figure 5B), mais s'en rapproche par le côté opposé (flèches noires montantes). La P-valeur de Clopper-Pearson qui est construite en doublant la P-valeur unilatérale, ne présente pas ce paradoxe. La Figure 5C en comparaison de la Figure 5B permet d'identifier une discontinuité dans la P-valeur de Sterne selon la proportion théorique. La proportion de succès 11/16 est un tout petit peu moins probable que la proportion de succès 15/16 si la proportion théorique est égale à 0,803, et participe donc à la P-valeur (cf Figure 5B). Par contre, la proportion 11/6 est un tout petit peu plus probable que la proportion de succès 15/16 si la proportion théorique est égale à 0,802 (Figure 5C), de telle sorte qu'elle ne participe plus à la P-valeur, qui descend brutalement de 0,3400 à 0,2234 (discontinuité). La P-valeur de Clopper-Pearson, quant à elle, ne présente pas de discontinuité.

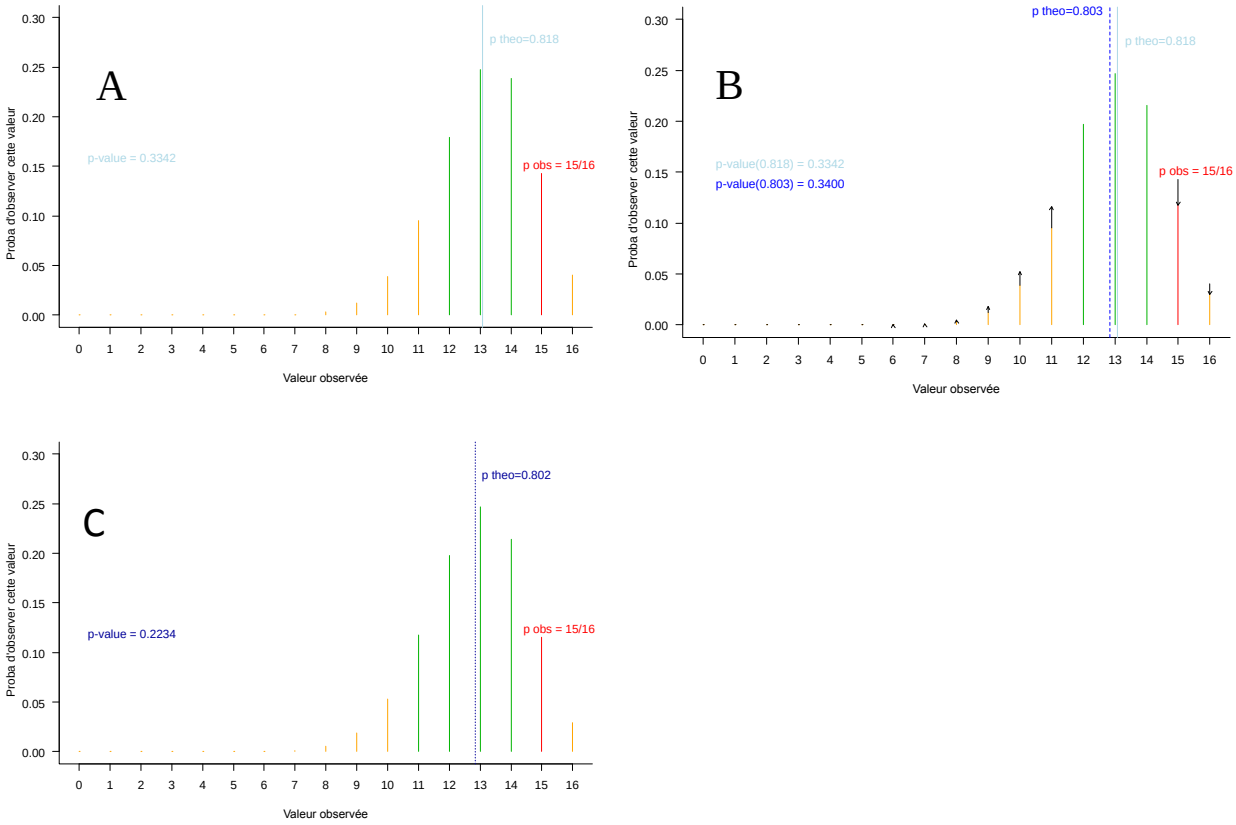

Figure 5 : construction de trois P-valeurs selon la méthode de Sterne, pour la comparaison des proportions théoriques p=0,818 (panneau A), $p = 0,803$ (panneau B) et $p = 0,802$ (panneau C) à la valeur observée 15/16=0,938. La barre rouge montre le nombre observé de succès. Les barres oranges, additionnée à la barre rouge, montrent quelles probabilités sont additionnées pour former la P-valeur. La droite verticale bleue décrit la proportion théorique p. Les flèches noires verticales du panneau B décrivent l'évolution de la distribution en relation au panneau A, lorsque la valeur théorique évolue de 0,818 à 0,803.

1.2.2.4.2 Généralisation aux modèles bivariés et multivariés

L'usage d'une même méthode d'estimation pour toutes les analyses, des plus simples au plus complexes, favorise la cohérence des résultats d'analyses de complexité croissante. De plus, les développements de la théorie statistique bénéficient simultanément aux modèles simples et complexes. Nous analyserons les régressions logistiques, régressions log-binomiales et régressions de Poisson avec les intervalles de confiance de Wald et d'inversion de test du rapport de vraisemblance.

1.2.2.4.3 Simplicité théorique et existence d'une solution analytique

Si l'intervalle de confiance de Wald, fondé sur une approximation normale (cf équation (1) en page 16) est encore largement enseigné dans les ouvrages statistiques, c'est certainement en raison de sa simplicité théorique. De plus, le calcul est faisable avec une simple calculatrice. L'existence d'une solution analytique simple a motivé Agresti et Coull (4,5) à définir leur intervalle, qui se calcule comme un intervalle de Wald après avoir virtuellement ajouté 2 succès et 2 échecs à l'échantillon. Cette simplicité est importante pour l'enseignement en cours de statistique élémentaire mais les ordinateurs étant largement diffusés, les algorithmes itératifs, bien que nécessitant nettement plus de calculs, ne posent pas de problèmes au quotidien du statisticien.

1.2.2.4.4 Équivariance

Un estimateur d'intervalle de confiance est équivariant si l'analyse des succès est équivalente à l'analyse des échecs. L'intervalle de confiance de x succès sur un échantillon de taille n est alors le miroir de l'intervalle de $n-x$ succès sur n échecs. Si on note $L_{1-\alpha}(x,n)$ et $U_{1-\alpha}(x,n)$ respectivement la borne basse et la borne haute de l'intervalle de confiance au risque α pour une expérience de x succès sur un échantillon de taille n, alors l'équivariance est définie par :

$$L_{1-\alpha}(x,n) = 1 - U_{1-\alpha}(n-x,n) \tag{40}$$

Ceci, pour tout x et pour tout n.

Un corollaire immédiat en découle, par changement de variable $y = n - x$:

$$U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n) \tag{41}$$

Cette équivariance est vérifiée par presque tous les estimateurs d'intervalle. Les estimateurs basés sur les modèles log-binomiaux et les modèles de Poisson y font exception. Enfin, une imperfection de l'équation 3.2 de l'article d'Efron (41) concernant le bootstrap à biais corrigé accéléré fait perdre l'équivariance à cet estimateur. L'équivariance est restaurée si

$$\hat{G}(s) = Pr_{\hat{\theta}}\left(\hat{\theta}^* < s\right) \tag{42}$$

(Efron 3.2)

est remplacé par

$$\hat{G}(s) = Pr_{\hat{\theta}}\left(\hat{\theta}^* < s\right) + \frac{1}{2} Pr_{\hat{\theta}}\left(\hat{\theta}^* = s\right) \tag{43}$$

où $\hat{G}(s)$ représente un estimateur de la fonction de répartition de l'estimateur de la statistique de θ, empiriquement établie à partir des ré-échantillonnages. Les deux estimateurs seront analysés.

1.2.2.4.5 Monotonie des bornes de l'intervalle selon $\boldsymbol{x}$, $\boldsymbol{n}$ et $\boldsymbol{\alpha}$

Des écarts à cette règle sont décrits par Vos et Hudson (107) notamment pour l'estimateur d'intervalle de Sterne (96). Avec la méthode de Sterne, un intervalle de confiance à 95% peut ne pas être inclus dans un intervalle de confiance à 90% de telle sorte que l'estimateur d'intervalle employé comme test d'hypothèse, rejetterait certaines valeurs au risque 5% en les acceptant au risque 10%.

Avec les intervalles de Sterne ou de Blaker, il est possible que l'ajout d'un succès (x remplacé par $x+1$ et n remplacé par $n+1$) abaisse la borne basse de l'intervalle de confiance bien que la proportion observée augmente et que la taille d'échantillon augmente. En conséquence, il est possible qu'une valeur soit rejetée mais qu'on sache que l'ajout d'une observation, quel que soit son statut d'échec ou de succès, fasse accepter la valeur.

Les bornes des estimateurs d'intervalles sont généralement croissantes selon x pour un n fixé. Les procédures mixtes, consistant à choisir l'estimateur d'intervalle selon certaines conditions, peuvent ne plus respecter cette règle. Par exemple, la procédure consistant à utiliser l'intervalle de Wald lorsque le nombre de succès et d'échecs sont tous deux supérieurs ou égaux à 10, et à utiliser l'intervalle de Clopper-Pearson autrement, a une borne supérieure non monotone. La borne haute de l'intervalle de confiance correspondant à 9 succès sur 64 tirages est égale à 0,250, alors que pour 10/64, elle est égale à 0,245 et pour 11/64, elle est égale à 0,264.

1.2.2.4.6 Procédure déterministe

La plupart des estimateurs d'intervalles fournissent toujours le même intervalle pour un nombre de succès x donné sur une taille d'échantillon n donnée. L'intervalle de Stevens (97) déjà mentionné en page 31, contient une part de hasard indépendante de l'échantillon, générée par ordinateur. Cela rend la pratique de double analyse plus difficile puisque deux statisticiens analysant les mêmes données avec la même méthode utiliseront deux nombres aléatoires différents et produiront donc des résultats légèrement discordants.

1.2.3 Conditions de validité de l'intervalle de Wald

Vollset (106) en 1993 notait déjà que les conditions de validité de l'intervalle de Wald décrits dans les manuels statistiques étaient très insuffisantes. Il remarquait que pour $\min(x, n-x) < 30$ le biais de couverture pouvait dépasser 5% et pour $\min(x, n-x) < 173$ le biais de couverture pouvait dépasser 1% pour un risque nominal à 95%. Vollset mettait en garde contre les conditions de validité suivantes, qu'il juge insuffisantes :

$$\min(x, n-x) \geq 5 \; ou \; 10 \tag{44}$$

$$n\hat{p}(1-\hat{p}) \geq 5 \tag{45}$$

Il n'a pas étudié la condition de validité :

$$\min(nL_{1-\alpha}, nU_{1-\alpha}) \geq 5 \tag{46}$$

Où $L_{1-\alpha}$ et $U_{1-\alpha}$ sont les bornes de l'intervalle de confiance de Wald telles que définies dans l'équation (35) en page 31.

Leemis (67) en 1996 puis Newcombe (77) en 1998 notaient aussi les importants biais de l'intervalle de Wald, mais c'est Santner (90) et Agresti (4) en 1998, puis Brown en 2001 (21) qui commencèrent à déconseiller l'usage et même l'enseignement de l'intervalle en cours de statistique élémentaire, suggérant l'intervalle de Wilson ou l'intervalle d'Agresti-Coull.

> *"The poor performance of the Wald interval is unfortunate, since it is the simplest approach to present in elementary statistics courses. We strongly recommend that instructors present the score interval instead" (Agresti et Coull 1998* (4)*)*

Ce qui se traduit par :

> « Les mauvaises performances de l'intervalle de Wald sont malheureuses, puisqu'il s'agit de l'approche la plus simple présentable en cours de statistique élémentaire. Nous incitons fortement les instructeurs à présenter l'intervalle du score à la place »

L'intervalle du score est un autre nom de l'intervalle de Wilson (110).

> *"This article suggests that a third alternative to z-intervals, called q-intervals herein, should be strongly preferred in elementary courses to either t- or c-intervals" (Santner 1998* (90)*)*

Traduit par :

> « Cet article suggère qu'une troisième alternative aux z-intervalles, appelés q-intervalles ici, devraient être largement préférés dans les cours élémentaires, aux t-intervalles et c-intervalles ».

Les z-intervals sont les intervalles de Wald sans et avec correction de continuité et les t-intervals sont les intervalles basés sur la loi de Student à $n-1$ degrés de liberté alors que les q-intervals sont encore un autre nom des intervalles de Wilson (110).

> *"The erratic behavior of the coverage probability of the standard Wald confidence interval has previously been remarked on in the literature (Blyth and Still, Agresti and Coull, Santner and others). We begin by showing that the chaotic coverage properties of the Wald interval are far more persistent than is appreciated. Furthermore, common textbook prescriptions regarding its safety are misleading and defective in several respects and cannot be trusted […] Based on this analysis, we recommend the Wilson interval or the equal-tailed Jeffreys prior interval for small n and the interval suggested in Agresti and Coull for larger n." (Brown, Cai et DasGupta 2001* (21)*)*

Ce qu'on peut traduire :

> « Le comportement erratique de la probabilité de couverture de l'intervalle de confiance Wald standard a déjà été remarqué dans la littérature (Blyth et Still, Agresti et Coull, Santner et autres). Nous montrons d'abord que la couverture chaotique de l'intervalle de Wald est nettement plus constante que ce qu'il n'est connu. En outre, les recommandations habituelles des manuels concernant sa validité sont trompeuses et erronées à plusieurs égards ; on ne peut pas leur faire confiance [...] Sur la base de cette analyse, nous recommandons l'intervalle Wilson ou l'intervalle de Jeffreys équilibré pour les petites valeurs de n et l'intervalle suggéré par Agresti et Coull pour de plus grandes valeur de n »

L'article de Brown, Cai et DasGupta reçut de nombreux commentaires, dont ceux de Ghosh, Agresti et Coull et Santner, qui avaient déjà participé à l'évaluation de l'intervalle de Wald par le passé (4,51,90). Ces auteurs suggéraient l'usage des intervalles de Wilson, de Jeffreys modifié ou d'Agresti-Coull. Ce dernier intervalle est une approximation de l'intervalle de Wilson construit en ajoutant deux succès et deux échecs à l'échantillon avant d'appliquer la procédure standard de Wald (4). L'article de Brown, Cai et DasGupta, selon Google Scholar au 20/09/2017, a été cité 1777 fois par d'autres articles et livres universitaires ; il a indéniablement été influent.

Certains livres tels que ceux d'Agresti (1) en 2003, DasGupta (36) en 2008, Olsen (80) en 2011, Nikolaidis (79) en 2011, Hollander (56) en 2013, Ott et Longnecker en 2015 (81), Meeker en 2017 (74) décrivent toujours l'intervalle de Wald mais mettent en garde contre son usage sans toutefois préciser de condition de validité. D'autres livres, donnent des conditions de validité. Vidakovic (105) en 2011 décrit la condition suivante :

$$np, n(1-p) \geq 10 \quad (47)$$

Vidakovic précise pourtant que ses performances sont médiocres ; il propose une correction de continuité. Sauro et Lewis (92) en 2016 précisent la condition suivante :

$$n\hat{p}, n(1-\hat{p}) \geq 15 \quad (48)$$

Tsuang (101) en 2011 cite l'article de Brown (21) mais ensuite précise la condition de validité suivante :

$$n\hat{p}, n(1-\hat{p}) \geq 5 \quad (49)$$

De même Fritz et Berger (47) en 2015 citent Agresti et Coull mais donnent la même condition (équation (49)). Herson (54) en 2016 fournit la condition de l'équation (48). Gerstman (49) en 2013 proposait l'équation (49). Razdolsky (86) en 2014 précise la condition suivante :

$$np, n(1-p) > 5 \quad (50)$$

Enfin, certains comme Forthofer (45) en 2006 restent flous ; Forthofer précise que n doit être grand ou p proche de 50%. Certains ne mettent pas en garde contre l'intervalle de Wald lorsqu'une correction de continuité est ajoutée, comme Fagerland (42) en 2017.

Ainsi, l'intervalle de Wald est toujours présenté dans les ouvrages pédagogiques, même par ceux d'Agresti ou de DasGupta ; sa simplicité théorique, et son intérêt historique le rendent difficile à omettre. Par contre, même en connaissance des articles d'Agresti ou de Brown, Cai et DasGupta, les conditions de validité ne semblent pas avoir été mises à jour. Un des objectifs secondaires de cette thèse est de définir les conditions de validité de l'intervalle de Wald, en se basant sur les critères de jugements qui seront donnés. Selon la littérature existante, cet intervalle semble inutilisable dans la pratique ; fournir des conditions de validité permettrait de formaliser ce fait.

1.3 Objectifs

Les objectifs de cette étude étaient les suivants :

1) Présenter une revue de la littérature sur l'estimation des intervalles de confiance d'une proportion binomiale ;
2) Définir des critères d'évaluation pertinents des estimateurs d'intervalles de confiance ;
3) Appliquer ces critères de manière systématique aux intervalles existants ;
4) Fournir des conseils pratiques sur le ou les intervalles à utiliser ;
5) Définir des conditions de validité de l'estimateur d'intervalle de Wald.

2 Matériel & méthodes

Des estimateurs d'intervalles de confiance seront analysés selon divers critères de jugements.

2.1 Critères de jugements

Les intervalles de confiance seront jugés selon leur maîtrise du risque α, par l'analyse des risques réels conditionnels et des risques réels moyens locaux, ainsi que sur leur précision évaluée par les demi-largeurs attendues. De manière plus formelle, nous allons définir les risques.

2.1.1 Risques conditionnels

Soit x la réalisation d'une variable $X_{n,p}$ suivant une loi binomiale de paramètres n et p :

$$X_{n,p} \sim B(n\,;p) \tag{51}$$

Ainsi n représente la taille de l'échantillon, x le nombre de succès observé et p la proportion réelle dans la population. On note $\hat{p}$ la proportion observée de succès :

$$\hat{p} = x/n \tag{52}$$

On note λ l'espérance du nombre de succès :

$$\lambda = np \tag{53}$$

On note $IC_{1-\alpha}(x,n)$ l'intervalle de confiance au risque nominal α d'un échantillon de taille n contenant x succès :

$$IC_{1-\alpha}(x,n) = [L_{1-\alpha}(x,n); U_{1-\alpha}(x,n)] \tag{54}$$

Pour n et p constants, on note $\alpha'_l(p,n,\alpha)$ le risque réel que l'intervalle de confiance soit entièrement au-dessus p :

$$\alpha'_l(p,n,\alpha) = Proba\big(L_{1-\alpha}(X_{n,p},n) > p\big) \tag{55}$$

De même, on note $\alpha'_u(p,n,\alpha)$ le risque réel que l'intervalle soit entièrement en-dessous de p :

$$\alpha'_u(p,n,\alpha) = Proba\big(U_{1-\alpha}(X_{n,p},n) < p\big) \tag{56}$$

On note $\alpha'(p,n,\alpha)$le risque réel que l'intervalle ne contienne pas p :

$$\alpha'(p,n,\alpha) = \alpha'_u(p,n,\alpha) + \alpha'_l(p,n,\alpha) \tag{57}$$

Les risques α'_l, α'_u et α' seront appelés *risques conditionnels* ou *risques réels conditionnels*, car il s'agit de probabilités conditionnelles à des valeurs données de p et de n. Le risque α'_l sera appelé *risque conditionnel à gauche* ou *risque conditionnel de la borne inférieure*, le risque α'_u sera appelé *risque conditionnel à droite* ou *risque conditionnel de la borne supérieure* et le risque α' sera appelé *risque conditionnel bilatéral* ou simplement *risque conditionnel*. Les risques conditionnels à droite et à gauche seront nommés *risques conditionnels unilatéraux* ou *risques conditionnels directionnels*. Comme moyen mnémotechnique, la lettre l se réfère à la « lower bound » ou borne basse de l'intervalle alors que la lettre u se réfère à la « upper bound » ou borne haute de l'intervalle.

2.1.2 Demi-largeurs attendues

Définissons la *demi-largeur attendue à gauche* $w_l'(p,n,\alpha)$, l'espérance de la distance entre la proportion observée et la borne inférieure de l'intervalle de confiance :

$$w_l'(p,n,\alpha) = E\left(\frac{X_{n,p}}{n} - L_{1-\alpha}\left(X_{n,p},n\right)\right) \tag{58}$$

Définissons, de même la *demi-largeur attendue à droite* $w_u'(p,n,\alpha)$:

$$w_u'(p,n,\alpha) = E\left(U_{1-\alpha}\left(X_{n,p},n\right) - \frac{X_{n,p}}{n}\right) \tag{59}$$

Les valeurs w_l' et w_u' seront appelées *demi-largeurs attendues conditionnelles*, *demi-largeurs attendues* ou *demi-largeurs* sans précision. Le *w* peut se lire « width ».

2.1.3 Risques moyens locaux et demi-largeurs moyennes locales

Notons P une proportion aléatoire dont le logit suit une distribution normale d'écart type $\log(OR_S)$ telle que l'espérance de P soit p_0.

$$\text{logit}(P) \sim \mathcal{N}(\mu_P, (\log(OR_S))^2) \tag{60}$$

$$p_0 = E(P) \tag{61}$$

Définissons le *risque moyen local à droite* comme :

$$\alpha_u''(p_0,n,\alpha) = E\left(\alpha_u'(P,n,\alpha)\right) \tag{62}$$

C'est la probabilité qu'un intervalle de confiance autour de $\hat{p}$ contienne p dans une expérience en deux étapes. Dans la première étape, une proportion p est réalisée à partir de la variable aléatoire P. Dans la deuxième étape, une proportion de succès $\hat{p}$ est réalisée dans un échantillon aléatoire de taille n issu d'une population dans laquelle la proportion réelle serait égale à p. La taille de l'échantillon n est maintenue constante dans toutes les expériences. De même, nous définissons $\alpha_l''(p_0,n,\alpha)$, $\alpha''(p_0,n,\alpha)$, $w_l''(p_0,n,\alpha)$ et $w_u''(p_0,n,\alpha)$.

$$\alpha_l''(p_0,n,\alpha) = E\left(\alpha_l'(P,n,\alpha)\right) \tag{63}$$

$$\alpha''(p_0,n,\alpha) = E\left(\alpha''(P,n,\alpha)\right) \tag{64}$$

$$w_u''(p_0,n,\alpha) = E\left(w_u'(P,n,\alpha)\right) \tag{65}$$

$$w_l''(p_0,n,\alpha) = E\left(w_l'(P,n,\alpha)\right) \tag{66}$$

La constante OR_S sera égale à 1.20 sauf dans des analyses de sensibilité menées avec d'autres valeurs OR_S.

2.1.4 Risques moyens à effectifs aléatoire

De même, définissons $\alpha_u'''(p,n_0,\alpha)$ et $\alpha_l'''(p,n_0,\alpha)$ les risques moyens pour un N aléatoire ayant une distribution discrète telle que l'espérance de N soit n_0. Cette variable N discrète suit une loi log-normale latente d'écart-type géométrique SR_S, arrondie au nombre entier le plus proche.

$$\log(Q) \sim \mathcal{N}\left(\mu_Q, (\log(SR_S))^2\right) \quad (67)$$

$$N = \lfloor Q \rfloor \quad (68)$$

$$n_0 = E(N) \quad (69)$$

$$\alpha_u'''(p, n_0, \alpha) = E\left(\alpha_u'(p, N, \alpha)\right) \quad (70)$$

$$\alpha_l'''(p, n_0, \alpha) = E\left(\alpha_l'(p, N, \alpha)\right) \quad (71)$$

L'équation (68) définit N comme une variable aléatoire égale à la partie entière de la variable aléatoire latente Q. L'écart-type géométrique SR_S sera fixé à $1{,}20$ sauf en analyse de sensibilité. Les risques α_l''' et α_u''' seront appelés *risques moyens à effectifs aléatoires*.

2.1.5 Demi-largeurs relatives attendues

Définissons la *demi-largeur relative attendue à gauche* $v_l'(p_0, n, \alpha)$ égale au rapport entre la *demi-largeur attendue à gauche* d'un estimateur d'intervalle et la *demi-largeur attendue à gauche* de l'estimateur d'intervalle de confiance de Clopper-Pearson mid-P pour les même paramètres p, n et α. De même, définissons v_u' la *demi-largeur relative attendue à droite*, v_l'' la *demi-largeur relative moyenne locale à gauche* et v_u'' la *demi-largeur relative moyenne locale à gauche*, en prenant toujours l'estimateur d'intervalle de confiance de Clopper-Pearson mid-P comme référence.

2.2 Paramètres et méthodes de calcul

Le risque α nominal sera fixé à $0{,}05$. Les risques réels conditionnels α_u' et α_l' seront estimés à partir de la distribution binomiale exacte $B(n\,;\,p)$. Les risques moyens locaux $\alpha_l''(p_0, n, \alpha)$ et $\alpha_u''(p_0, n, \alpha)$ et les demi-largeurs moyennes locales $w_l''(p_0, n, \alpha)$ et $w_u''(p_0, n, \alpha)$ seront approximés par intégration numérique sur un histogramme à 512 barres de la distribution logit-normale. Les intervalles d'intégration numérique étaient réguliers sur le logit de la loi de probabilité logit-normale, c'est-à-dire, sur la loi normale. Pour chaque barre de l'histogramme, la densité de probabilité multipliée par la largeur de l'espace était encore multipliée par le risque conditionnel ou la demi-largeur conditionnelle α_l', α_u', w_l' ou w_u' selon le résultat souhaité. La somme de ces 512 résultats partiels était une approximation des risques moyens locaux et demi-largeurs moyennes locales. La précision excédentaire gagnée par une augmentation du nombre de barres d'histogrammes était graphiquement invisible. L'usage de pseudo-aléatoire ne fût pas nécessaire.

À titre indicatif, des lignes horizontales seront tracées sur les figures pour les risques $0{,}025$, $0{,}025 \times 1{,}50 = 0{,}0375$ et $0{,}025 \,/\, 1{,}50 = 0{,}016667$. Ils peuvent servir de référence aux limites de risques acceptables, même si le lecteur de cette thèse peut lui-même définir les limites qu'il accepte.

Les analyses porteront sur des échantillons de taille 32, 64 et 2048. Ces trois nombres sont des puissances de deux. La plus grande taille est presque équivalente au scénario asymptotique de la loi de Poisson. Pour l'analyse des demi-largeurs ou des risques moyens locaux, nous invitons le lecteur, intéressé à la loi de Poisson, à interpréter ce scénario comme tel. Une différence plus notable pourrait exister dans l'analyse des risques conditionnels pour lesquels des discontinuités existent. Pour les principaux intervalles la loi de Poisson sera alors analysée pour les risques conditionnels.

Les nombres de succès attendus $\lambda = np$ les plus petits seront présentés car les différences principales entre les estimateurs se font lorsque le nombre de succès ou d'échecs est petit. Presque tous les estimateurs d'intervalles analysés étant équivariants, l'autre extrême, où le nombre d'échecs est presque nul, est déductible de l'analyse des succès.

2.3 Recherche et implémentation des estimateurs intervalles

Une recherche bibliographique avec les mots clés « binomial », « confidence » et « interval » a été menée sur l'outil Google® Scholar en Juillet 2017, à la recherche d'articles définissant des intervalles de confiance d'une proportion binomiale. Les articles étaient triés par pertinence selon l'algorithme de Google et les 400 premiers résultats ont été screenés sur leur titre puis leur résumé. Les références des articles de revue systématique ont été suivies afin de retrouver les références originales.

Les méthodes considérées comme redondantes ne sont pas présentées, telles que les approximations de Molenaar (76), de Pratt (14) et l'équation C de Blyth améliorant l'approximation de Molenaar (14) ou l'intervalle de Chen (30) qui sont toutes des approximations du Clopper-Pearson (31). Toutes ces approximations sont précises et les résultats seraient indistinguables de ceux du Clopper-Pearson sur les figures. Les estimateurs d'intervalles de Zhou (114) de Bolboaca (16), de Brenner (18), de Lang (65) et de Crow (35) n'ont pas été implémentés, du fait d'une complexité d'implémentation, ou d'une redondance *a priori* de ces intervalles à un estimateur déjà décrit.

Cinquante-cinq estimateurs d'intervalles seront présentés. Pour chacun d'entre eux, l'estimateur a été implémenté à partir de l'algorithme défini dans l'article original, avec des correctifs mineurs pour les cas limites, comme suit :

- lorsqu'une borne basse était inférieure à 0, elle était remplacée par 0 ;
- lorsqu'une borne haute était supérieure à 1, elle était remplacée par 1 ;
- lorsque l'estimation ponctuelle n'était pas contenue dans l'intervalle, la borne de l'intervalle la plus proche de l'estimation ponctuelle était déplacée jusqu'à l'estimation ponctuelle.

On peut s'attendre à ce qu'un statisticien vérifiant ses résultats applique spontanément ces corrections.

Certains intervalles étaient indéfinis lorsque le nombre de succès ou d'échecs était trop petit, comme l'intervalle logit-normal qui a une division par zéro lorsque le nombre de succès x ou le nombre d'échecs $n - x$ est nul. Dans ce cas, l'intervalle de Clopper-Pearson était employé lorsque $\min(x, n - x) \leq seuil$. Le seuil sera toujours précisé.

Depuis l'article original, la formule a été recopiée dans ce document, en y ajoutant les correctifs mineurs, une notation standardisée et des simplifications de formule. Une première implémentation a été faite dans le logiciel R (version 3.4.0, R Foundation for Statistical Computing, Vienne, Autriche). La validité de l'implémentation a été vérifiée par comparaison aux procédures de paquetages R implémentant le même algorithme et/ou de tables de calculs présentées dans l'article original lorsque ces implémentations ou données étaient disponibles. Lorsqu'aucune implémentation n'était disponible, les propriétés attendues de l'intervalle étaient vérifiées. Par exemple, lorsqu'un intervalle était construit pour être strictement conservatif, cette propriété était vérifiée. De même, la monotonie des bornes selon x et l'équivariance étaient validées lorsqu'elles étaient théoriquement vraies. Enfin, les courbes de couverture conditionnelle étaient comparées à celles de la littérature, validant encore l'implémentation.

Un mois plus tard, l'auteur de ce document a ré-implémenté chacune de ces procédures en ne se basant que sur la formule décrite dans ce document. Les discordances entre la 1ère et la 2ème implémentation étaient investiguées. Les deux implémentations étaient comparées sur tous les intervalles pour $n = 32, 64, 128\ et\ 2048$ et pour tous les nombres de succès $x = 1 \ldots n$. En cas de discordance d'au moins une borne d'au moins un intervalle, le problème était investigué. Le plus souvent la 1ère implé-

mentation était correcte mais la formule présentée dans ce document ne prenait pas bien en compte les cas limites ou contenait une erreur de recopie.

Quelques estimateurs d'intervalle font exception à ce processus de double implémentation à partir de l'article original : les intervalles de Schilling-Doi (93) et de Wang (109) ont été implémentés en se basant sur le code source R fourni en annexe des articles originaux. Ces algorithmes ayant des performances trop faibles pour une taille d'échantillon $n = 2048$, ils ont été améliorés. Pour des tailles d'échantillons plus faibles ($n = 32$ et $n = 64$), la concordance entre la version originale et la version améliorée a été vérifiée. L'estimateur d'intervalle de Blyth-Still-Casella (15,27) (encore appelé intervalle de Blyth-Still) a d'abord été décrit par Blyth et Still, puis l'algorithme de calcul a été amélioré par Casella afin d'obtenir une meilleure précision décimale sans augmenter le temps de calcul sur ordinateur. Cet estimateur étant particulièrement complexe à implémenter, l'implémentation de Winstein's écrite en langage C++ a été utilisée (111). Elle a été comparée à l'implémentation de StatXact par Winstein, mais cette vérification n'a pas été refaite. Par contre, les résultats de l'algorithme de Winstein ont été comparés à la table de l'article de Blyth et Still (15).

2.4 Définition des principaux estimateurs d'intervalles

Comme le nombre d'estimateurs d'intervalles est grand (cinquante-cinq), et que la plupart n'a pas d'intérêt car rarement ou jamais utilisé et présentant des performances très médiocres à la fois en termes de maîtrise de la couverture moyenne et de la couverture conditionnelle, un sous-ensemble de neuf estimateurs a été sélectionné. Ces estimateurs ont un intérêt historique (Wald, Wilson modifié, Clopper-Pearson), ont montré des performances intéressantes sur des revues systématiques (Wilson modifié, Jeffreys équilibré modifié), se sont montrés particulièrement performants sur notre étude (Arc-Sinus de Bartlett, Clopper-Pearson mid-P), sont généralisables aux modèles bivariés ou multivariés (Logit-normal modifié, rapport de vraisemblance modifié) ou illustrent un phénomène intéressant (Blaker). Cette sélection n'est pas aléatoire et s'est faite en partie sur les résultats observés. Les autres estimateurs d'intervalles sont aussi présentés dans un deuxième temps, mais leur discussion est plus succincte.

Les quarante-six autres estimateurs sont présentés en Annexe 1. Les résultats de l'Annexe 1 sont commentés. Ces résultats concernent les diverses approximations normales et de Student, les variantes du bootstrap, les approximations normales après transformation, les intervalles bayésiens à distribution *a priori* non informative, les estimateurs « exacts » et les intervalles par formule explicite avec correction du coefficient d'asymétrie.

Dans une expérience binomiale, notons x le nombre de succès, n la taille d'échantillon (ou nombre de tentatives), $\hat{p} = \frac{x}{n}$ la proportion observée sur l'échantillon et p la proportion théorique dans la population telle que $X \sim B(n\,;p)$. Notons k le plus petit du nombre de succès ou d'échecs :

$$k = \min(x, n - x) \tag{72}$$

Le Tableau 4 décrit les bornes inférieures des neuf estimateurs de l'intervalle de confiance qui seront analysés. Les bornes hautes sont définies par équivariance. Pour rappel $L_{1-\alpha}(x,n)$ et $U_{1-\alpha}(x,n)$ désignent, respectivement, les borne basse et borne haute de l'intervalle de confiance comme présentée dans l'équation (54) en page 43.

Nom de l'estimateur	Borne basse $L_{1-\alpha}(x,n)$	
(4,21,77,83,106) Wald[a]	$\max\left(0, \frac{x}{n} - \kappa\sqrt{\frac{x(n-x)}{n^3}}\right)$	(73)
(21) Wilson 1927 modifié par Brown en 2001[abc]	$\begin{cases} \frac{1}{2n}\chi^2_{\alpha,2x}\ si\ 1 \le x \le x^* \\ \frac{x + \frac{\kappa^2}{2} - \kappa\sqrt{\frac{x(n-x)}{n} + \frac{\kappa^2}{4}}}{n+\kappa^2}\ autrement \end{cases}$	(74)
(8,10) Arc-sinus de Bartlett 1936[a]	$\sin^2\left(\max\left(0, \arcsin\left(\sqrt{\frac{x+\frac{1}{2}}{n+1}}\right) - \frac{\kappa}{2\sqrt{n+\frac{1}{2}}}\right)\right)$	(75)
(21) Logit-normal modifié[ad]	$\begin{cases} \text{logitinv}\left(\log\left(\frac{x}{n-x}\right) - \kappa\sqrt{\frac{n}{x(n-x)}}\right)\ si\ 0 < x < n \\ \sqrt[n]{\alpha/2}\ si\ x = n \\ 0\ si\ x = 0 \end{cases}$	(76)
(106) Rapport de vraisemblance modifié[a]	$\begin{cases} \inf\left\{q \middle\| \log\left(\left(\frac{x}{nq}\right)^x\left(\frac{n-x}{n(1-q)}\right)^{n-x}\right) \le \frac{1}{2}\kappa^2\right\}\ si\ x < n \\ \sqrt[n]{\alpha/2}\ si\ x = n \end{cases}$	(77)
(21) Jeffreys équilibré modifié par Brown[e]	$\begin{cases} \beta iCDF(\alpha/2; x+1/2, n-x+1/2)\ si\ 2 \le x < n \\ \sqrt[n]{\alpha/2}\ si\ x = n \\ 0\ si\ x \le 1 \end{cases}$	(78)
(12) Blaker 2000[fg]	$\inf\{q \mid \text{bpval}(q,x,n) \le \alpha\}$	(79)
(21,31) Clopper-Pearson[e]	$\beta iCDF\left(\frac{\alpha}{2}; x, n-x+1\right)$	(80)
(11,63,64) Clopper-Pearson mid-P[g]	$\inf\left\{q \mid \min\left(BCDF(x;n,q), 1 - BCDF(x-1;n,q)\right) - \frac{1}{2}BPF(x;n,q) \le \frac{\alpha}{2}\right\}$	(81)

Tableau 4 : définition des bornes basses des intervalles de confiance, les bornes hautes étant définies par équivariance. L'équivariance est définie en page 39 par $U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)$.

[a]Notons $\kappa = z_{1-\alpha/2}$ le quantile $1-\alpha/2$ de la distribution normale centrée réduite $\mathcal{N}(0,1)$

[b] Le seuil x^* est défini comme suit :

$$x^* = \begin{cases} 2\ si\ n \le 50 \\ 3\ si\ n > 50 \end{cases} \quad (82)$$

[c]$\chi^2_{q,df}$ est le quantile q de la distribution du χ^2 à df degrés de liberté

[d]La réciproque de la transformation logistique est définie par :

$$\text{logitinv}(t) = \frac{\exp(t)}{1+\exp(t)} \quad (83)$$

[e] $\beta iCDF(q;\alpha,\beta)$ est le quantile q de la distribution beta dont les paramètres de forme sont α et β

[f] La fonction bpval (Blaker P-valeur) est définie comme suit :

$$\text{bpval}(q,x,n) = \begin{cases} \min\left(1, BCDF(x;n,q) + 1 - BCDF\left(BiCDF\left(1 - BCDF(x;n,q);n,q\right);n,q\right)\right)\ si\ q \ge x/n \\ \text{bpval}(1-q,n-x,n)\ si\ q < x/n \end{cases} \quad (84)$$

[g]Notons $BPF(x;n,p)$ la fonction de masse de la loi binomiale, $BCDF(x;n,p)$ la fonction de répartition (ou fonction cumulative de probabilités) et $BiCDF(q;n,p)$ la fonction des quantiles de la loi binomiale (ou fonction inverse cumulative de probabilités) comme décrit par les équations (3), (5) et (6) en pages 21 et 22.

2.4.1 Intervalle de Wald

L'intervalle de Wald (équation (74) est basé sur deux approximations : l'approximation de la loi binomiale à la loi normale et l'approximation de la variance de la loi binomiale à la variance observée sur l'échantillon. En effet, la proportion théorique p, nécessaire au calcul de la variance n'étant pas connue, on y substitute $\hat{p}$, la proportion observée de l'échantillon. Une conséquence immédiate, c'est la nullité de la largeur de l'intervalle lorsque la proportion observée est 0% ou 100%. Cet intervalle est souvent présenté dans les ouvrages statistiques sous la forme suivante :

$$\hat{p} \pm z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{85}$$

où $z_{1-\alpha}$ représente le quantile $1-\alpha$ de la loi normale centré réduite.

Le cas limite où le nombre de succès égale 1 présente généralement une borne inférieure négative. Cette aberration est corrigée par la formule présentée dans le Tableau 4 car, il est difficile d'imaginer un statisticien ne corrigeant pas la sortie du logiciel.

2.4.2 Intervalle de Wilson 1927 modifié par Brown en 2001

Cet intervalle est basé sur l'intervalle de Wilson (110), modifié par Brown (21) pour les petites valeurs de x. L'intervalle de Wilson est obtenu par inversion de test de Wald (ou du χ^2) d'une proportion observée à théorique. Ainsi, l'intervalle contient l'ensemble des valeurs théoriques compatibles avec la valeur observée, c'est-à-dire, les valeurs théoriques qui ne seraient pas rejetées dans un test d'hypothèse de comparaison de valeur observée à valeur théorique. Comme l'intervalle de Wald, l'intervalle de Wilson est basé sur une approximation normale de la loi binomiale, par contre, il n'approxime plus la variance de la théorique à la variance observée, puisque chacune des valeurs théoriques est comparée, avec sa propre variance, à la valeur observée. Il existe une solution analytique à l'intervalle de Wilson. Il s'agit de trouver les solutions d'une équation polynomiale du second degré. L'intervalle de Wilson est aussi appelé intervalle du Score car dans le cas simple d'une proportion binomiale, le test du χ^2 est équivalent au test du score de Rao.

L'intervalle de Wilson, reposant sur moins d'approximations que l'intervalle de Wald, il est moins biaisé selon Brown (21). Néanmoins, Brown a constaté des défauts de couverture conditionnelle bilatérale majeurs lorsque np est proche de zéro. Brown a alors proposé une modification (page 112 de son article (21)) de l'intervalle, approximant la loi binomiale a une loi de Poisson lorsque x est en-dessous d'un certain seuil x^*. Ainsi, il s'agit d'une substitution de l'intervalle de Wilson par l'intervalle de Garwood (48) pour les petites valeurs de x. L'intervalle de Garwood est un intervalle strictement conservatif à risques équilibrés d'une espérance de la loi de Poisson, équivalent à l'intervalle de Clopper-Pearson lorsque $n \to +\infty$. L'intervalle de Garwood bénéficie d'une solution analytique exacte sous forme du quantile d'une loi du χ^2. Le choix du seuil x^* par Brown est empirique. Brown a défini ce seuil à $x^* = 2$ pour $n \leq 50$ et $x^* = 3$ pour $51 \leq n \leq 100$ mais ne précise pas le seuil pour $n > 100$. Cette omission est due au fait que Brown n'a pas analysé le comportement de l'intervalle au-delà de cette taille d'échantillon (communication personnelle avec l'auteur). La convergence de la loi binomiale à la loi de Poisson suggère que le comportement est peu différent pour

$n > 100$. Ceci sera confirmé par la similarité des scénario $n = 64$ et $n = 2048$. En conséquence, le seuil $x^* = 3$ a été conservé pour $n > 100$.

2.4.3 Intervalle Arc-Sinus de Bartlett 1936

Cet intervalle est une amélioration de l'intervalle Arc-Sinus classique défini ci-dessous :

$$\sin^2\left(\text{asin}\left(\sqrt{\hat{p}}\right) \pm \frac{\kappa}{2\sqrt{n}}\right) \tag{86}$$

Il est basé sur une transformation stabilisatrice de variance qui réduit la corrélation entre les fluctuations de la variance observée et de la proportion observée. La transformation de Bartlett (10) est obtenue par ajout de 0,5 succès et 0,5 échec avant de passer à la racine carrée puis à l'arc-sinus. Cette transformation stabilise la variance et réduit aussi le coefficient d'asymétrie de la loi binomiale, améliorant alors l'approximation normale. Il existe d'autres variantes de cette transformation telles que celle d'Anscombe (8) ou de Tukey-Freeman (46). Ces transformations ont été comparées par Yu (112) et par Foi (44). La transformation de Bartlett ayant des résultats moins biaisés que les autres a été présentée en priorité, mais les autres transformations sont présentées en Annexe 1.

2.4.4 Intervalle logit-normal modifié

L'intervalle logit-normal est présenté par Brown (21). Il s'agit d'une approximation normale après transformation logistique. Cet intervalle est équivalent à l'intervalle de Wald que l'on obtiendrait lors de l'estimation d'une régression logistique à intercept seul, c'est-à-dire une régression logistique sans covariable avec seulement l'ordonnée à l'origine. On peut obtenir ce type d'intervalle sous le logiciel R avec la commande `confint.default(glm(family=binomial,success~1)` ou sous le logiciel SAS avec l'option `WALDCI` de la commande `MODEL` de `PROC GENMOD`.

L'estimation de la variance du logit d'une variable binomiale est facilement estimable par la méthode du delta :

$$\begin{aligned}\widehat{VAR}\left(\text{logit}\left(\hat{P}\right)\right) &\approx \left(\text{logit}'\left(\hat{P}\right)\right)^2 \times \widehat{VAR}\left(\hat{P}\right) = \left(\frac{1}{\hat{P}\left(1-\hat{P}\right)}\right)^2 \times \frac{\hat{P}\left(1-\hat{P}\right)}{n} \\ &= \frac{1}{n\hat{P}\left(1-\hat{P}\right)} = \frac{n}{X(n-X)}\end{aligned} \tag{87}$$

L'approximation normale est appliquée au logit de la loi normale. L'intervalle logit-normal n'est pas défini pour $x = 0$ ni pour $x = n$. L'intervalle logit-normal modifié consiste en l'application de l'intervalle de Clopper-Pearson pour $x = 0$ ou $x = n$. Le Clopper-Pearson étant obtenu par inversion d'un test binomial exact qui se simplifie beaucoup pour $x = 0$, on arrive à la borne $\sqrt[n]{\alpha/2}$. Cette modification avait aussi été proposée par Brown en page 119 (21).

2.4.5 Intervalle du rapport de vraisemblance modifié

Cet intervalle est obtenu par inversion d'un test du rapport de vraisemblance, comparant une proportion théorique à la proportion observée. Ce test du rapport de vraisemblance est fait par approximation à un χ^2 de la différence des déviances, définies comme $-2 \times \log(vraisemblance)$. La vraisemblance est définie par la fonction qui à une proportion p théorique associe la probabilité binomiale $\binom{n}{x}p^x(1-p)^{n-x}$. Le nombre de combinaisons de x éléments parmi n disparaît du rapport, car il est indépendant de p. La statistique de déviance d'une proportion théorique p dans un modèle binomial sans transformation est égale à la déviance de $\text{logit}(p)$ dans un modèle binomial logistique (fonction de lien cano-

nique). En conséquence, cet intervalle est équivalent à celui obtenu dans une régression logistique sous R avec la commande `confint(glm(family=binomial, success~1)` ou sous SAS avec l'option `LRCI` de la commande `MODEL` de `PROC GENMOD`. Cet intervalle est aussi connu sous le nom de « profile likelihood » ou d'intervalle de profilage de la vraisemblance. En effet, l'inversion du test peut se faire par un algorithme de dichotomie assez lent (l'erreur étant égale à $\frac{1}{2^{iter}}$ où $iter$ est le nombre d'itérations de l'algorithme) ou par une amélioration de l'algorithme de Newton-Raphson par Venzon (104) dont l'itération est basée sur un développement limité à l'ordre deux de la log-vraisemblance dont la convergence est plus rapide mais fournissant le même résultat si le nombre d'itérations adéquat est appliqué. Bien que le rapport de vraisemblance soit encore défini lorsque $x = 0$ ou $x = n$, la log-vraisemblance est indéfinie pour $x = 0$ ou $x = n$ et $0 < p < n$. Ainsi, les algorithmes de R ou SAS échouent. Afin de rapprocher cet intervalle de confiance de celui produit par R ou SAS pour la régression logistique et de le rendre comparable à l'intervalle logit-normal modifié, dans le cas limite $x = 0$ ou $x = n$ il a été remplacé par l'intervalle de Clopper-Pearson.

2.4.6 Jeffreys équilibré modifié par Brown

L'intervalle de Jeffreys équilibré présenté par Brown (21) semblait faiblement biaisé à l'exception de quelques pics de défaut de couverture. C'est un intervalle de crédibilité bayésienne basé sur le *prior* (distribution *a priori*) non informatif de Jeffreys $Beta(0{,}50\,;0{,}50)$. Il est analysé comme un intervalle de confiance fréquentiste. L'intervalle de crédibilité est basé sur les quantiles $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la distribution Beta *a posteriori* $Beta(0{,}50+x\,;0{,}50+n-x)$. C'est un intervalle à risques équilibrés. Brown a proposé une modification *ad hoc*, pour éliminer les pics de défaut de couverture et garantir que l'estimation ponctuelle appartienne à l'intervalle de confiance. Cette modification est basée sur une borne basse nulle pour $x = 0$ ou $x = 1$ et la borne basse de l'intervalle de Clopper-Pearson pour $x = n$. La borne haute est obtenue par équivariance.

2.4.7 L'intervalle de Blaker 2000

L'intervalle décrit par Blaker en 2000 (12) avait été proposé par Cox et Hinkley (34) en 1974, mais cet ouvrage n'ayant pas été consulté, seul l'article de Blaker a servi de référence. L'intervalle de Blaker est un intervalle strictement conservatif bilatéral à risques déséquilibrés strictement inclus dans l'intervalle de Clopper-Pearson. C'est-à-dire, pour tous α, x et n, l'intervalle de confiance de Blaker pour ces trois paramètres est inclus dans l'intervalle de confiance de Clopper-Pearson pour ces trois mêmes paramètres. L'intervalle de Blaker est donc toujours plus étroit que l'intervalle de Clopper-Pearson contrairement à l'intervalle de Sterne qui chevauche parfois l'intervalle de Clopper-Pearson. L'intervalle de Blaker est obtenu par inversion d'un test d'hypothèse de comparaison de proportion théorique à proportion observée. Pour une proportion observée x/n comparée à une proportion théorique p telle que $p < x/n$ (l'équivariance permet de calculer la P-valeur lorsque $p > x/n$), la première partie de la P-valeur de ce test d'hypothèse est obtenue en calculant la probabilité pour une loi binomiale $B(n\,;p)$, d'obtenir un nombre de succès supérieur ou égal à x (probabilité de la queue à droite). La deuxième partie de la P-valeur de ce test d'hypothèse est obtenue en additionnant la queue de probabilité à gauche de la loi binomiale (0 succès, 1 succès, 2 succès, etc) en incrémentant progressivement la longueur de la queue à gauche jusqu'à obtenir la plus grande queue à gauche de probabilité inférieure ou égale à la queue à droite.

Pour information, dans le même scénario, la P-valeur de Clopper-Pearson est obtenue en multipliant par deux la probabilité de la queue à droite, de telle sorte que la P-valeur de Blaker est toujours inférieure ou égale à la P-valeur de Clopper-Pearson. La P-valeur de Sterne, quant à elle, est obtenue en incrémentant la longueur de la queue à gauche tant que la probabilité du nombre de succès exactement

égal au maximum de succès dans la queue à gauche, est inférieur ou égal à la probabilité d'un nombre de succès exactement égal à x.

Dans la Figure 5 (page 38), la P-valeur de Sterne est obtenue en additionnant toutes les barres de hauteur inférieure à la barre rouge ; cette dernière correspond au nombre de succès observé. La P-valeur de Blaker, par contre, est obtenue en additionnant, dans un premier temps les barres correspondant à 15/16 et 16/16, puis en ajoutant progressivement les barres orange de 0/15 à y/15 tant que cette somme à gauche ne dépasse pas la somme des barres de 15/16 et 16/16. Cela fait une différence pour la Figure 5B dans laquelle la barre 11/15 participe à la P-valeur de Sterne mais ne participe pas à la P-valeur de Blaker. L'intervalle de Sterne est présenté en Annexe 1, son analyse n'apportant pas grand-chose par rapport à l'analyse de l'intervalle de Blaker.

L'intervalle de Blaker, comme l'intervalle de Sterne, est basé sur une région de confiance non connexe (cf Cohérence avec un test d'hypothèse, page 36). En d'autres termes, cette région de confiance peut avoir des trous, une valeur étant rejetée, une valeur plus grande acceptée et une valeur encore plus grande étant de nouveau rejetée. La solution proposée par Blaker (comme par Sterne) est de boucher les trous, puisque l'intervalle de Blaker (ou de Sterne) est défini comme le plus petit intervalle contenant entièrement la région de confiance. Ceci garantit le strict conservatisme de l'intervalle. Cela complique l'algorithme de recherche des bornes, puisqu'un algorithme par dichotomie n'est adapté qu'à une fonction de P-valeur monotone. En effet, il peut y avoir plusieurs solutions à l'équation $\mathrm{bpval}(p, x, n) = \alpha/2$ pour $p < x/n$. L'algorithme de dichotomie permet de trouver une solution à cette équation, mais ce n'est pas forcément la plus petite solution. Blaker avait proposé un algorithme particulièrement peu performant, en approximant les nombres réels à des nombres décimaux avec un nombre fixe (suffisamment grand) de décimales, de telle sorte qu'il est possible d'itérer sur l'ensemble des nombres réels, du plus petit au plus grand, et ainsi, de trouver le plus petit nombre vérifiant la propriété. Cet algorithme était peu performant et fournissait parfois des résultats erronés puisque l'algorithme pouvait sauter par-dessus une solution si le nombre fixe de décimales n'était pas suffisamment élevé. L'algorithme de Klaschka (60) est une amélioration importante. Il est nettement plus rapide à calculer et ne risque plus de sauter par-dessus une solution car il est basé sur la connaissance théorique d'une zone dans laquelle il n'y a qu'une solution à l'équation $\mathrm{bpval}(p, x, n) = \alpha/2$. L'algorithme de dichotomie est alors possible. Nous avons donc utilisé l'algorithme de Klaschka.

2.4.8 L'intervalle de Clopper-Pearson

L'intervalle de Clopper-Pearson (31) décrit en 1934, est à notre connaissance, le 1er intervalle de confiance d'une proportion publié se basant sur la loi binomiale exacte. C'est l'intervalle, décrit sous le nom « binomial exact », le plus souvent retrouvé dans les logiciels statistiques (R, SAS, Stata). C'est un intervalle à risques équilibrés strictement conservatif. Comme les intervalles de Blaker ou de Sterne, c'est un intervalle par inversion de test d'hypothèse. À la différence des intervalles de Sterne ou de Blaker, la fonction de P-valeur de Clopper-Pearson est continue et bimonotone, comme décrit par Fay (43) et présenté sur la Figure 6. Ceci est une première simplification puisque l'algorithme de dichotomie est possible et l'intervalle de confiance est toujours cohérent avec le test d'hypothèse (cf Cohérence avec un test d'hypothèse, page 36). La fonction de P-valeur de Clopper-Pearson pour une proportion théorique p supérieure ou égale à la proportion observée x/n est égale au double de la probabilité pour une loi binomiale $B(n\,;p)$ d'observer un nombre de succès supérieur ou égal à x. Lorsque $p < x/n$, la P-valeur est égale au double de la probabilité d'observer un nombre de succès inférieur ou égal à x. Lorsque la P-valeur dépasse 1, la valeur 1 lui est substituée.

Deux solutions analytiques au calcul des bornes de l'intervalle de Clopper-Pearson sont présentées par Blyth (14) à partir des quantiles de la loi Béta (équation 8 de Blyth) ou des quantiles de la loi *F* de Fisher (équation 9 de Blyth). Les approximations de Moleenar (76) ou de Pratt (84) sont d'une époque où les ordinateurs n'étaient pas encore largement diffusés et ne présentent plus d'intérêt de nos jours.

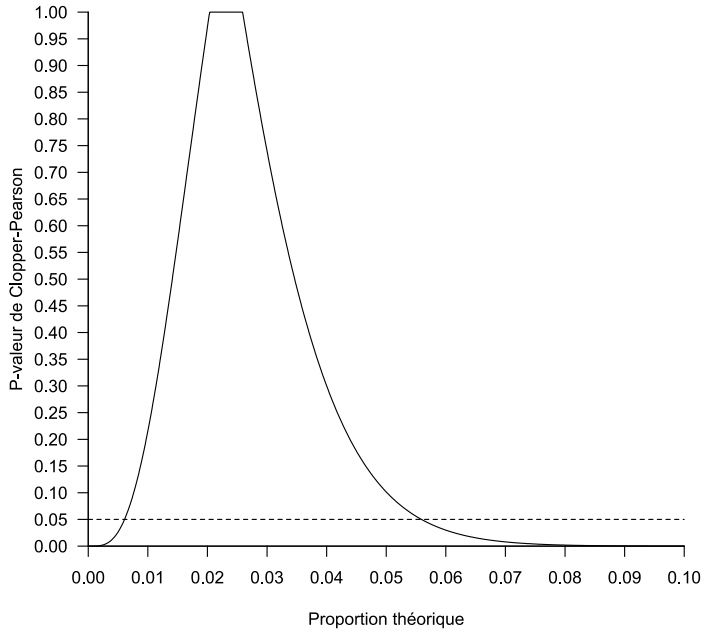

Figure 6 : fonction de P-valeur de la comparaison de proportion observée 4/180 à une proportion théorique, selon le test de Clopper-Pearson (31). Cette fonction de P-valeur est à comparer à celle de Sterne présentée en Figure 4 en page 37.

L'intervalle de Clopper-Pearson peut être interprété comme l'intersection de deux intervalles unilatéraux strictement conservatifs, chacun au risque nominal $\frac{\alpha}{2}$. Il garantit un contrôle des risques conditionnels unilatéraux $\alpha'_u \leq \frac{\alpha}{2}$ et $\alpha'_l \leq \frac{\alpha}{2}$. Bien que l'intervalle soit souvent appelé intervalle exact, il ne garantit pas que le risque nominal α soit égal au risque réel. Le risque réel est toujours inférieur ou égal au risque nominal. Brown (21) a montré que les oscillations du risque α' réel se font toutes en-dessous du risque α nominal avec des pics touchant le risque α nominal mais ne le dépassant pas (figure 11, page 113 de Brown (21)).

2.4.9 L'intervalle de Clopper-Pearson mid-P

L'intervalle de Clopper-Pearson mid-P est une variante du Clopper-Pearson discutée par Berry (11). Si X suit une distribution quantitative absolument continue $Proba(X \geq x) = Proba(X > x)$, car $Proba(X = x) = 0$ pour tout x. Les choses sont différentes avec les lois discrètes. Pour la comparaison d'une proportion observée x/n à une proportion théorique p, dans le cas où $p < x/n$, la P-valeur unilatérale de Clopper-Pearson est égale à $Proba(X \geq x)$ où $X \sim B(n\,;p)$. L'inégalité large, garantit le conservatisme strict du test (α' réel toujours inférieur ou égal au α nominal). À l'opposé, une inégalité stricte $Proba(X > x)$ créerait un test strictement libéral (α' réel toujours supérieur ou égal au α nominal). Un équilibre peut être obtenu par une relation intermédiaire « supérieur ou à moitié égal à », définissant alors la P-valeur unilatérale de Clopper-Pearson mid-P. Plus précisément, la P-valeur unila-

térale d'un test de Clopper-Pearson mid-P de comparaison d'une observée x/n à une proportion théorique p, dans le cas où $p < x/n$, est égale à :

$$Pvaleur_{CPmidP} = Proba(X > x) + \frac{1}{2} Proba(X = x) \quad (88)$$

où $X \sim B(n\,;p)$. Le cas où $p > x/n$ est obtenu par équivariance. La P-valeur bilatérale est obtenue en multipliant par deux la P-valeur unilatérale. On peut ainsi comprendre que les oscillations de couverture de l'intervalle de Clopper-Pearson mid-P se feront autour du risque nominal. Cet intervalle a été construit afin de maîtriser le risque α moyen plutôt que le risque maximal. Reiczigel qualifie l'intervalle de Clopper-Pearson mid-P comme équivalent à l'intervalle de Clopper-Pearson ajusté par augmentation volontaire du risque α nominal pour que le risque moyen soit maîtrisé (pour une distribution *a priori* uniforme).

2.5 Conditions de validité de l'intervalle de Wald

2.5.1 Description informelle de la méthode

Les conditions de validité de l'estimateur d'intervalle de Wald seront basées sur l'objectif de maîtrise des risques moyens locaux unilatéraux, ne devant pas dépasser 1,50 fois le risque nominal, soit 0,0375 pour un risque nominal unilatéral à 0,025. Cela correspond à un intervalle bilatéral équilibré au risque 0,05 ; c'est-à-dire à un intervalle de confiance à 95%. Les conditions porteront sur le nombre observé de succès et d'échecs car les conditions portant sur la proportion réelle sont inapplicables dans la pratique car la proportion réelle est inconnue.

Afin d'évaluer une condition de validité, il est nécessaire de préciser la procédure à appliquer lorsque la condition n'est pas vérifiée. C'est pourquoi, on supposera que l'intervalle de Clopper-Pearson mid-P sera utilisé lorsque la condition de validité de l'intervalle de Wald ne sera pas vérifiée. L'évaluation portera donc sur un estimateur hybride consistant en l'usage de l'un ou l'autre des estimateurs d'intervalle selon qu'un seuil est dépassé ou non. On définira, pour une taille d'échantillon n donnée, le seuil minimal de validité ξ'', comme le plus petit seuil de la procédure hybride qui garantisse un risque local moyen unilatéral toujours inférieur à 0,0375 pour un risque nominal unilatéral à 0,025. En bref, c'est le nombre de succès nécessaire à la validité de l'intervalle de Wald pour une taille d'échantillon n donnée. En se basant sur les méthodes de calcul de risques réels qui ont été présentées en section « Paramètres et méthodes de calcul » en page 45, et en appliquant un algorithme de dichotomie, il sera possible de définir une table empirique de seuils de validité ξ'' pour diverses tailles d'échantillon : 2^5 à 2^{11} et le cas limite $n \to +\infty$.

La table de seuils de validité étant peu pratique à appliquer, une formule approximant les valeurs de cette table sera recherchée. La formule sera recherchée empiriquement dans l'objectif de se rapprocher au mieux des valeurs de la table. En cas de discordance entre les valeurs, une formule prudente sera privilégiée, c'est-à-dire, une formule fournissant des seuils de validité légèrement plus élevés. L'élaboration de cette formule ne sera pas forcément basée sur une théorie mathématique puisqu'il s'agit seulement de se rapprocher des valeurs de la table. Des formules simplifiées seront recherchées afin de rendre leur enseignement possible dans la pratique.

Même si le seuil de risque maximum tolérable 0,0375 sera étudié en analyse principale, d'autres seuils seront aussi étudiés : 0,030, 0,035 et 0,05. En exprimant ces seuils sous la forme de risques bilatéraux équilibrés, double des risques unilatéraux, ils seront égaux à 0,060, 0,070 et 0,10.

Même si l'analyse principale portera sur les risques moyens locaux, une analyse secondaire sera faite sur les seuils de validité garantissant la maîtrise du risque conditionnel maximal pour un intervalle hybride entre le Wald et le Clopper-Pearson.

L'hybridation avec le Clopper-Pearson mid-P (pour la maîtrise des risques locaux moyens unilatéraux) ou avec le Clopper-Pearson (pour les risques conditionnels unilatéraux) est justifiée par le bon comportement de ces deux intervalles en relation aux critères de jugements concernés par l'analyse.

2.5.2 Description formelle de la méthode

Définissons $ICW_{1-\alpha}$l'estimateur d'intervalle de confiance de Wald et $ICM_{1-\alpha}$ l'estimateur d'intervalle de confiance de Clopper-Pearson mid-P. Pour un seuil entier χ, définissons l'estimateur d'intervalle de confiance hybride $ICMW_{1-\alpha}(\chi)$:

$$ICMW_{1-\alpha}(\chi) = \begin{cases} ICW_{1-\alpha}\ si\ \min(X, n-X) < \chi \\ ICM_{1-\alpha}\ si\ \min(X, n-X) \geq \chi \end{cases} \quad (89)$$

Pour une taille d'échantillon n et un risque local moyen α''_{max} nominal, définissons le seuil de validité $\xi''(\alpha''_{max}, n)$ comme le plus petit χ tel que pour toute proportion moyenne théorique aléatoire p_0, les risques $\alpha''_l(n,p,\alpha)$ et $\alpha''_u(n,p,\alpha)$ de l'estimateur $ICMW_{1-\alpha}(\chi)$ soient inférieurs ou égales à $\alpha''_{max}/2$. Pour cette analyse, OR_S sera fixée à 1,10, le logit(P) suivant une loi normale d'écart-type $\log(OR_S)$ telle que l'espérance de P soit p_0 ; cette proportion P aléatoire avait été décrite en section « Risques moyens locaux et demi-largeurs moyennes locales » en page 44.

$$\xi''(\alpha''_{max}, n) = \min\left\{\chi \middle| \begin{array}{c} \forall p_0 \in [0;1]\ Proba(ICMW_{1-\alpha}(\chi) < P) \leq \alpha''_{max}/2\ et \\ Proba(ICMW_{1-\alpha}(\chi) > P) \leq \alpha''_{max}/2 \\ \text{où logit}(P) \sim \mathcal{N}\ et\ E(P) = p_0\ et\ VAR\big(\text{logit}(P)\big) = (\log(OR_S))^2 \end{array}\right\} \quad (90)$$

De la même manière, définissons $ICCW(\chi)$ un estimateur hybride entre le Clopper-Pearson et l'estimateur d'intervalle de Wald au seuil χ. Le seuil de validité conditionnel $\xi'(\alpha'_{max}, n)$ de l'estimateur hybride $ICCW(\chi)$ est défini comme le plus petit seuil χ tel que, pour toute proportion théorique p, les risques conditionnels unilatéraux $\alpha'_l(n,p,\alpha)$ et $\alpha'_u(n,p,\alpha)$ soient tous deux inférieurs ou égaux à $\alpha'_{max}/2$.

$$\xi'(\alpha'_{max}, n) = \min\left\{\chi \middle| \begin{array}{c} \forall p \in [0;1]\ Proba(ICMW_{1-\alpha}(\chi) > p) \leq \alpha'_{max}/2\ et) \\ Proba(ICMW_{1-\alpha}(\chi) < p) \leq \alpha'_{max}/2 \end{array}\right\} \quad (91)$$

Les seuils ξ' et ξ'' pour un n donné, seront trouvés par dichotomie.

Des tables de seuils ξ' et ξ'' seront calculées pour les tailles d'échantillon n égales à 2^i où i est un nombre entier compris entre 5 et 11, puis pour le cas limite $n \to +\infty$ en se basant sur l'équivalence asymptotique de la loi binomiale à la loi de Poisson. C'est-à-dire, les tailles d'échantillon étudiées seront : 32, 64, 128, 256, 512, 1024, 2048 et $+\infty$. Les valeurs α''_{max} égales à 0,060, 0,070, 0,075 et 0,10 seront analysées pour un risque nominal bilatéral $\alpha = 0{,}05$.

Des formules empiriques approximant les seuils ξ' et ξ'' seront fournies. Les seuils approximés par la formule seront notés respectivement $\widetilde{\xi'}$ et $\widetilde{\xi''}$.

3 Résultats

3.1 Risques moyens locaux

3.1.1 Analyse principale

Les risques moyens locaux unilatéraux α_u'' et α_l'' sont présentés sur la Figure 7. Différentes tailles d'échantillon sont analysées sur cette figure, mais les résultats pour un nombre attendu de succès supérieur à 32 ne sont pas présentés. L'intervalle de Wald a un risque moyen local à droite α_u'' tendant vers 1 lorsque la proportion théorique moyenne p_0 tend vers 0 parce que la longueur d'intervalle de Wald est nulle lorsque le nombre de succès x est nul. Les courbes des risques moyens locaux sont lissées par les fluctuations de la proportion réelle P. Sur la Figure 7 l'odds ratio typique que l'on obtiendrait entre deux proportions réelles de deux réalisations de l'étude est fixé à $OR_S = 1{,}20$ (modèle à effet aléatoire présenté en page 44). Les risques locaux moyens unilatéraux sont d'autant plus biaisés que la taille de l'échantillon est grande, à nombre de succès attendu égal.

Graphiquement, pour tous les intervalles, des oscillations de risque moyen local unilatéral peuvent être visualisées pour un nombre attendu de succès np_0 inférieur ou égal à 2. De plus, pour un nombre attendu de succès inférieur ou égal à la borne haute de l'intervalle de confiance correspondant à zéro succès sur n, le risque alpha à droite est toujours nul. Cette zone a été appelée la *lèvre* (lip en anglais) par Liu et Kott (68). Cette lèvre étant inévitable, les estimateurs d'intervalles ne peuvent pas être jugés sur sa présence mais peuvent être jugés sur la largeur de cette zone ; une zone large peut être associée à une largeur moyenne d'intervalle plus grande que nécessaire. Sur cette figure, les intervalles de Wilson modifié ou logit-normal modifié sont déséquilibrés entre le risque à droite et le risque à gauche, d'autant plus que l'échantillon est de grande taille, à nombre attendu de succès égal. Pour un échantillon de taille $n = 2048$ le déséquilibre de ces deux intervalles est acceptable à partir d'un nombre attendu de succès $np_0 \geq 20$. En effet les deux risques locaux moyens unilatéraux se trouvent alors tous deux dans la l'intervalle $[0{,}016667\,;\,0{,}0375]$ qui avait été défini comme la zone tolérable en section « Paramètres et méthodes de calcul », en page 45. Ces limites sont présentées par deux barres horizontales grises. L'intervalle logit-normal a un pic de risque local moyen unilatéral atteignant $0{,}097$ pour $n = 2048$ et un nombre attendu de succès $np_0 = 0{,}11$.

L'intervalle de Clopper-Pearson est conservatif, avec un risque moyen local unilatéral toujours inférieur au risque nominal. L'intervalle de Blaker est un peu moins conservatif, surtout à gauche, pour les faibles nombres attendus de succès, et inversement, à droite pour les nombres de succès élevés. L'intervalle du rapport de vraisemblance modifié est plutôt bien équilibré, mais pas aussi bien que les intervalles Arc-Sinus de Bartlett, de Jeffreys modifié ou de Clopper-Pearson mid-P. Ces trois derniers intervalles sont presque superposables mais une analyse fine permet de distinguer le Jeffreys modifié des autres. Celui-ci a un démarrage un peu plus tardif et plus lent du risque moyen local à droite, vers $np_0 \approx 4$ plutôt que $np_0 \approx 3$ pour le Clopper-Pearson mid-P et l'Arc-Sinus de Bartlett. De plus, pour un nombre attendu de succès très faible $np_0 < 1$, l'intervalle modifié de Jeffreys est libéral pour certaines valeurs de p_0, alors que les oscillations sont décalées vers le bas pour les deux autres estimateurs d'intervalles, qui sont alors plus conservatifs dans ce scénario. Les différences entre l'Arc-Sinus de Barlett et le Clopper-Pearson mid-P sont négligeables. L'intervalle de Clopper-Pearson mid-P est juste très légèrement plus conservatif que l'Arc-Sinus de Bartlett lorsque $np_0 \leq 5$.

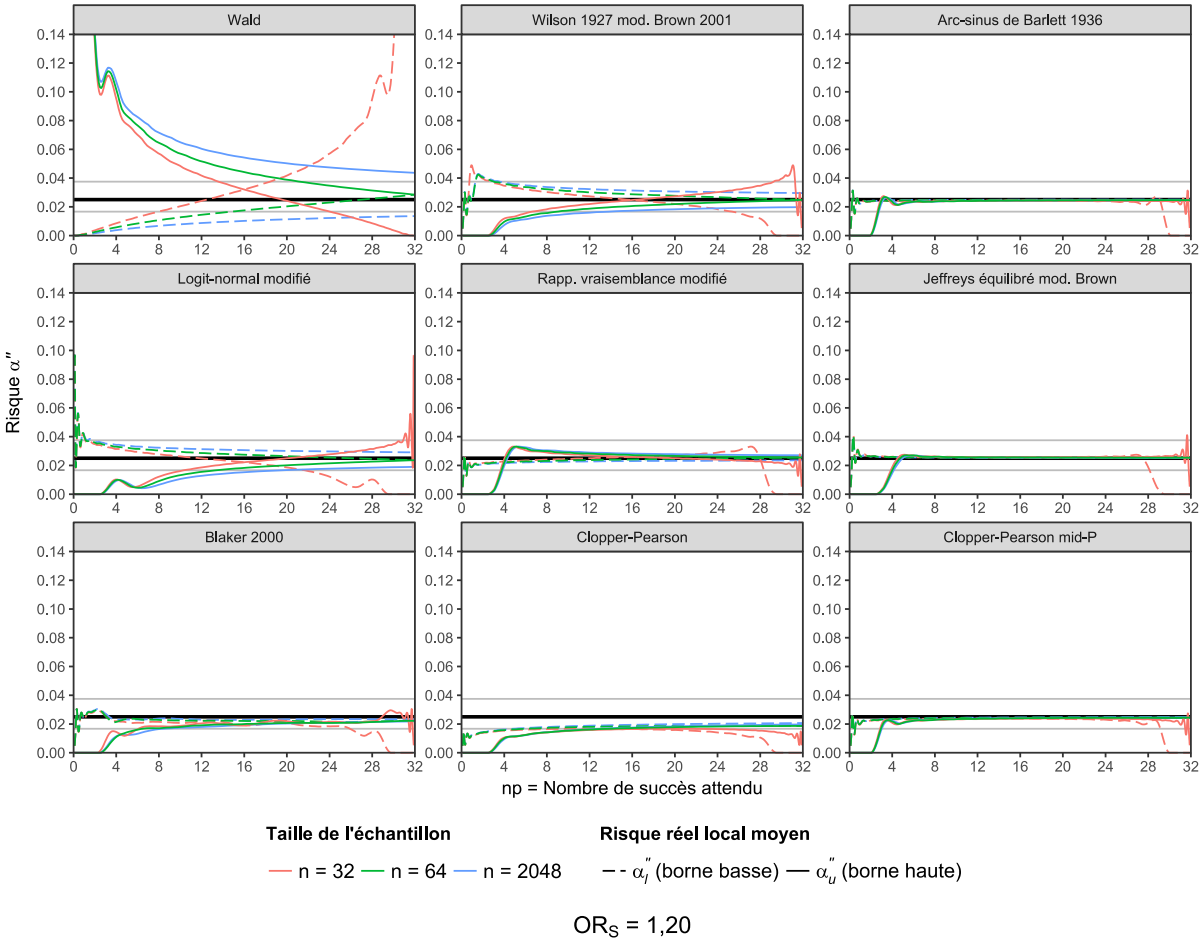

Figure 7 : risques réels moyens locaux unilatéraux (tels que définis en page 44) à gauche (pointillés) et à droite (trait plein) des neuf principaux estimateurs d'intervalle de confiance à 95%, selon différentes taille d'échantillon (rouge pour $n = 32$, vert pour $n = 64$ et bleu pour $n = 2048$), avec une proportion théorique P suivant un modèle logit-normal aléatoire avec un odds ratio typique $OR_S = 1,20$. En abscisse, le nombre attendu de succès np_0 et en ordonnée le risque que la borne basse de l'intervalle de confiance soit supérieure à la proportion réelle p (risque réel moyen local à gauche : pointillés) ou le risque que la borne haute de l'intervalle de confiance soit inférieure à la proportion réelle p (risque réel moyen local à droite : trait plein).

3.1.2 Analyse avec un niveau de confiance nominal à 90%

La Figure 8 montre les risques réels locaux moyens unilatéraux pour un risque nominal $\alpha = 0{,}10$. En comparaison de la Figure 7, les biais relatifs sont un peu réduits. On remarquera que l'échelle des ordonnées n'est pas la même sur la Figure 8 que sur la Figure 7. De même, les barres horizontales ont été placées aux risques unilatéraux 0,05 (trait épais), $0{,}05/1{,}50 = 0{,}0333$ et $0{,}05 \times 1{,}50 = 0{,}075$ (traits gris fins).

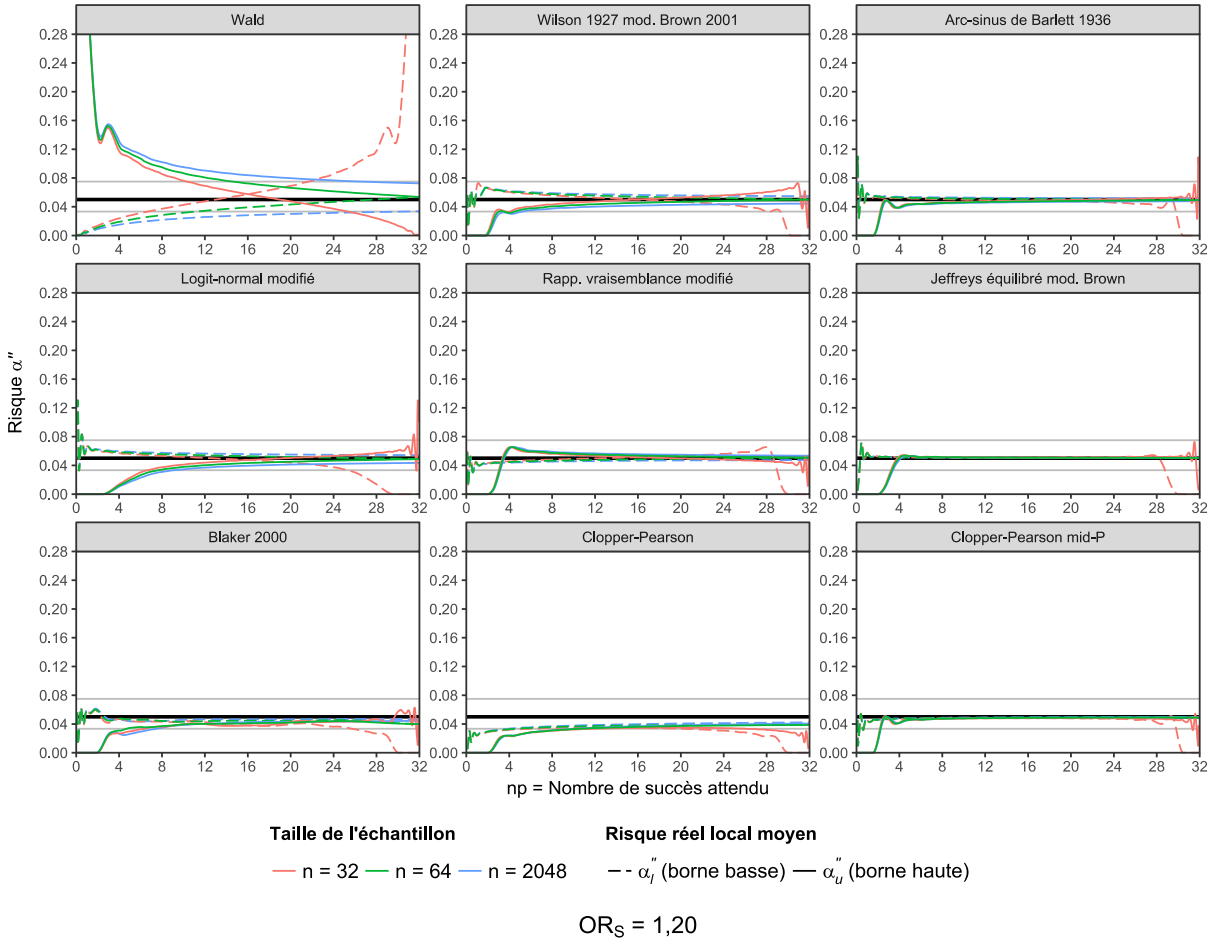

Figure 8 : risques réels moyens locaux unilatéraux (tels que définis en page 44) à gauche (pointillés) et à droite (trait plein) des neuf principaux estimateurs d'intervalle de confiance à 90%, selon différentes taille d'échantillon (rouge pour $n = 32$, vert pour $n = 64$ et bleu pour $n = 2048$), avec une proportion théorique P suivant un modèle logit-normal aléatoire avec un odds ratio typique $OR_S = 1,20$. En abscisse, le nombre attendu de succès np_0 et en ordonnée le risque que la borne basse de l'intervalle de confiance soit supérieure à la proportion réelle p (risque réel moyen local à gauche : pointillés) ou le risque que la borne haute de l'intervalle de confiance soit inférieure à la proportion réelle p (risque réel moyen local à droite : trait plein). Le niveau de confiance nominal diffère de celui de la Figure 7 ainsi que l'échelle des ordonnées.

3.1.3 Analyse de sensibilité avec variance aléatoire réduite

Lorsque l'odds ratio aléatoire de la proportion réelle P est réduite à $OR_S = 1{,}10$, des oscillations de grande amplitude sont visibles lorsque le nombre attendu de succès est inférieur à 3 ou 4 (cf Figure 9). Au-delà les résultats sont peu différents de ceux de la Figure 7, notamment l'intervalle de Clopper-Pearson mid-P a un risque local moyen unilatéral proche du risque nominal de 0,025.

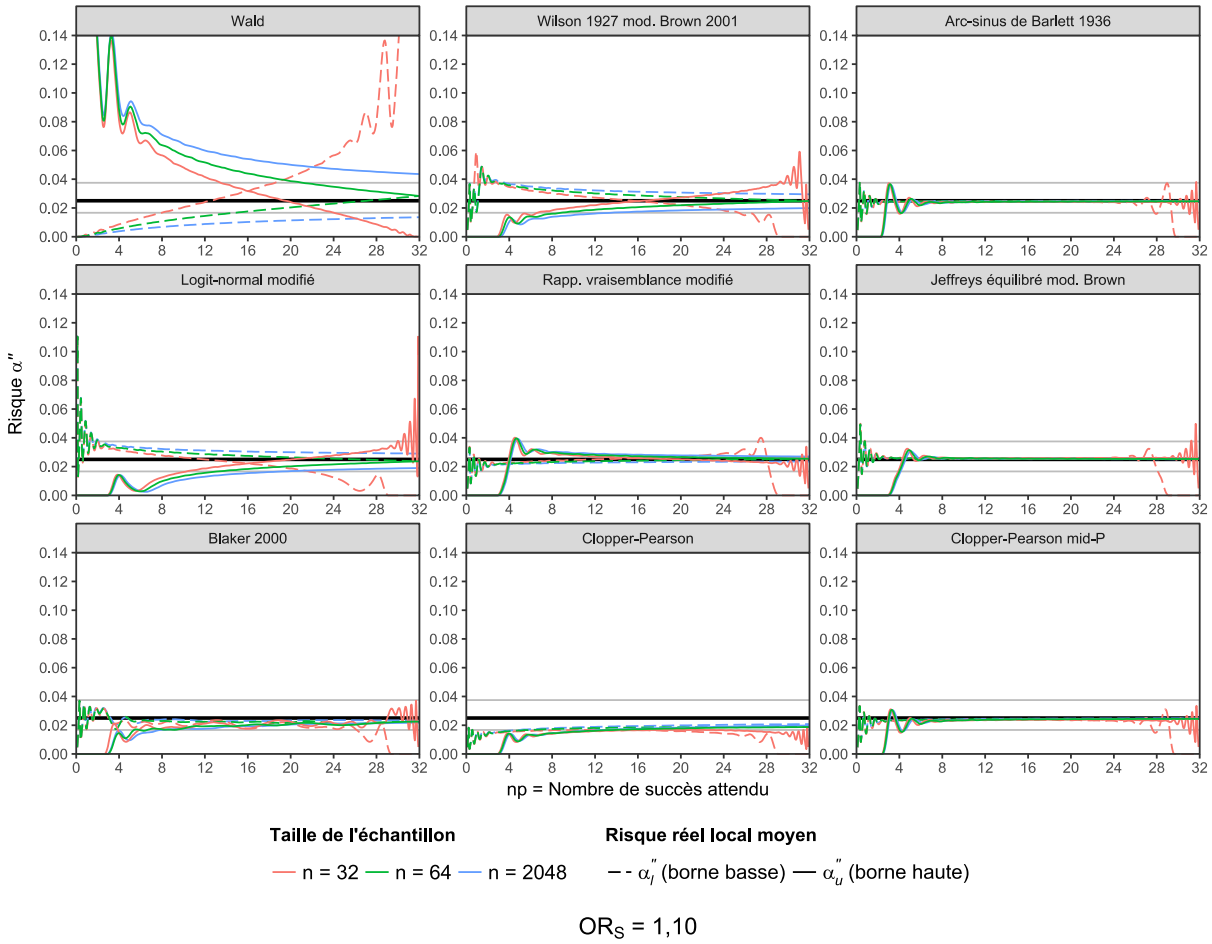

Figure 9 : risques réels moyens locaux unilatéraux (tels que définis en page 44) à gauche (pointillés) et à droite (trait plein) des neuf principaux estimateurs d'intervalle de confiance à 95%, selon différentes taille d'échantillon (rouge pour $n = 32$, vert pour $n = 64$ et bleu pour $n = 2048$), avec une proportion théorique P suivant un modèle logit-normal aléatoire avec un odds ratio typique $OR_S = 1,10$. La seule différence avec la Figure 7, c'est une variabilité moindre de la proportion théorique P puisque OR_S égale $1,10$ plutôt que $1,20$.

Lorsque l'odds ratio aléatoire de la proportion réelle P est réduite à $OR_S = 1{,}05$ (effet aléatoire minime), des oscillations de grande amplitude sont visibles lorsque le nombre attendu de succès est inférieur à 8 (cf Figure 10). L'intervalle de Clopper-Pearson présente alors un risque local moyen unilatéral maximal égal à 0,0381, plus grand que le risque de 0,025 nominal et très légèrement supérieur au risque de 0,0375 qui avait été considéré comme le maximum tolérable (la limite étant subjective) dans la section matériel et méthodes.

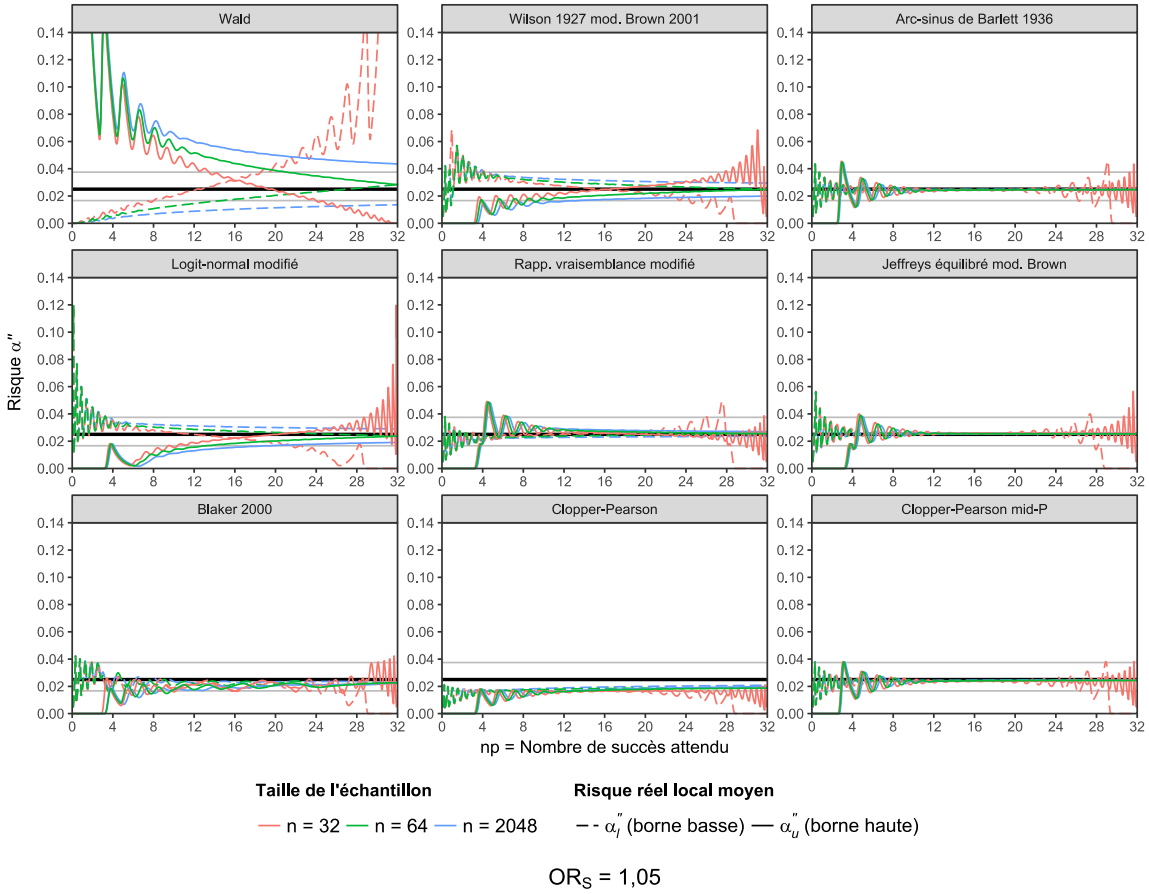

Figure 10 : risques réels moyens locaux unilatéraux (tels que définis en page 44) à gauche (pointillés) et à droite (trait plein) des neuf principaux estimateurs d'intervalle de confiance à 95%, selon différentes taille d'échantillon (rouge pour $n = 32$, vert pour $n = 64$ et bleu pour $n = 2048$), avec une proportion théorique P suivant un modèle logit-normal aléatoire avec un odds ratio typique $OR_S = 1,05$. La seule différence avec la Figure 7 ou la Figure 9, c'est une variabilité moindre de la proportion théorique P puisque OR_S égale $1,05$ plutôt que $1,10$ ou $1,20$.

3.1.4 Cas limite de la loi de Poisson

La loi de Poisson étant un cas limite de la loi binomiale lorsque la taille d'échantillon est infinie, ce scénario est comparé à celui où $n = 2048$ sur la Figure 11. Cette figure montre, en rouge, l'échantillon de taille 2048 et en bleu, l'échantillon limite infini. Les deux lignes sont graphiquement indistinguables suggérant que le cas où $n = 2048$ est déjà très proche du cas limite.

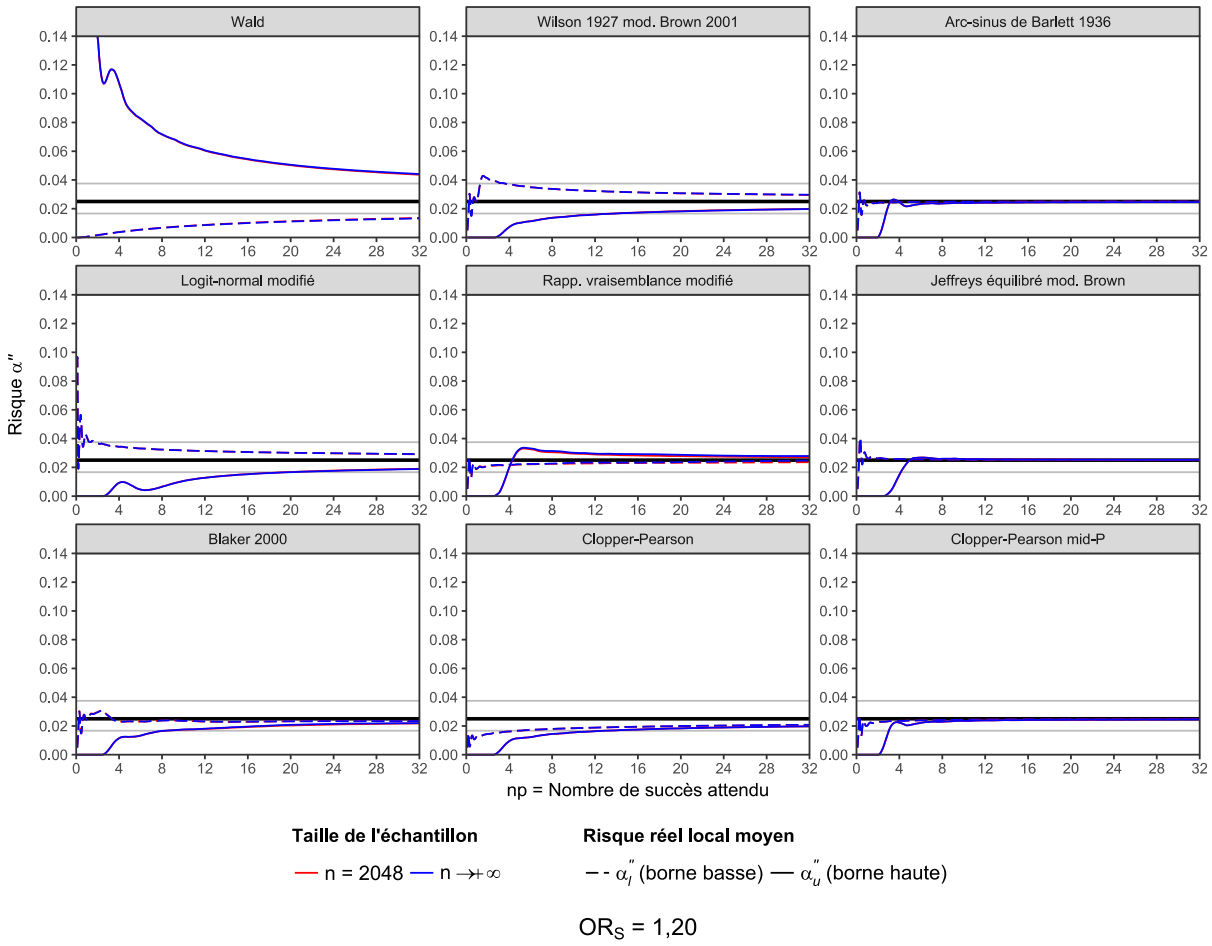

Figure 11 : comparaison des risques locaux moyens unilatéraux d'estimateurs d'intervalles de confiance à 95% pour des tailles d'échantillon $n = 2048$ et $n \to +\infty$. Le cas $n \to +\infty$ est modélisé par la loi de Poisson.

3.2 Risques moyens à effectifs aléatoires

La Figure 12 présente les risques moyens unilatéraux sur un modèle à effectif aléatoire mais à proportion fixe sur l'échantillon Arc-sinus de Bartlett. Ceci est à comparer aux risques moyens locaux unilatéraux d'un modèle à proportion aléatoire mais à effectif fixe tel que montré sur la Figure 7. Les résultats sont assez similaires. Pour une taille d'échantillon moyenne $n_0 = 32$ le phénomène de lissage des oscillations est un peu moins marqué pour un modèle à effectif aléatoire que pour un modèle à proportion aléatoire mais les deux modèles atténuent beaucoup les oscillations. Ceci peut être expliqué en partie par la nature discrète de la taille de l'échantillon. Si l'effectif ainsi que la proportion étaient tous deux aléatoires on pourrait s'attendre à une atténuation encore plus marquée des oscillations.

Seul l'intervalle Arc-sinus de Bartlett est présenté, les résultats étant *a priori* similaires sur tous les autres intervalles. Cet intervalle a été sélectionné en raison de ses bonnes propriétés de risque local moyen ainsi que sa rapidité de calcul.

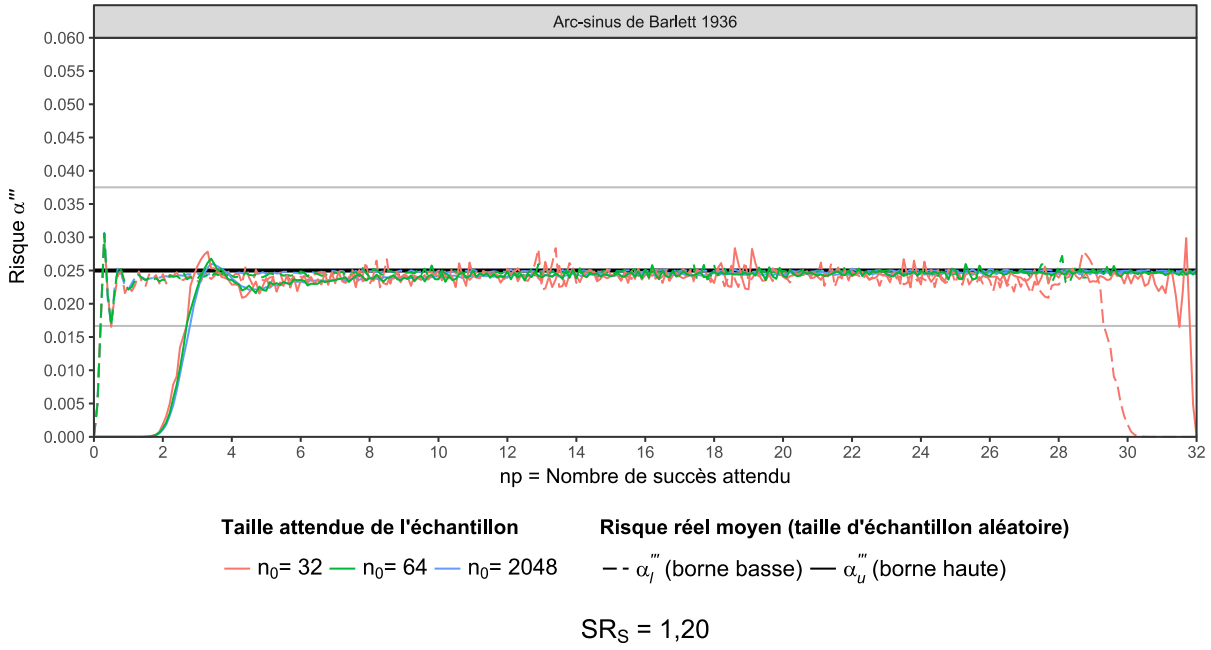

Figure 12 : risques réels moyens unilatéraux de l'estimateur d'intervalle Arc-Sinus de Bartlett à 95% dans un modèle à effectif aléatoire (cf page 44) dans lequel la proportion réelle p est constante mais la taille de l'échantillon N suit une loi log-normale de moyenne n_0 et d'écart-type géométrique $SR_S = 1,20$, arrondie au nombre entier le plus proche. En abscisse, le nombre de succès attendu $n_0 p$ et en ordonnée le risque réel à droite (borne haute, trait plein) et à gauche (borne basse, pointillés) que l'intervalle de confiance ne contienne pas la proportion réelle p.

Par rapport à la Figure 12, La Figure 13 présente une analyse de sensibilité dans laquelle l'échantillon a une taille moins aléatoire ($SR_S = 1{,}05$ plutôt que $SR_S = 1{,}20$). Les résultats sont superposables à ceux qu'on observait avec un modèle à proportion aléatoire logit-normale sur la Figure 10. Les oscillations restent assez fortes lorsque le nombre attendu de succès est inférieur ou égal à 8.

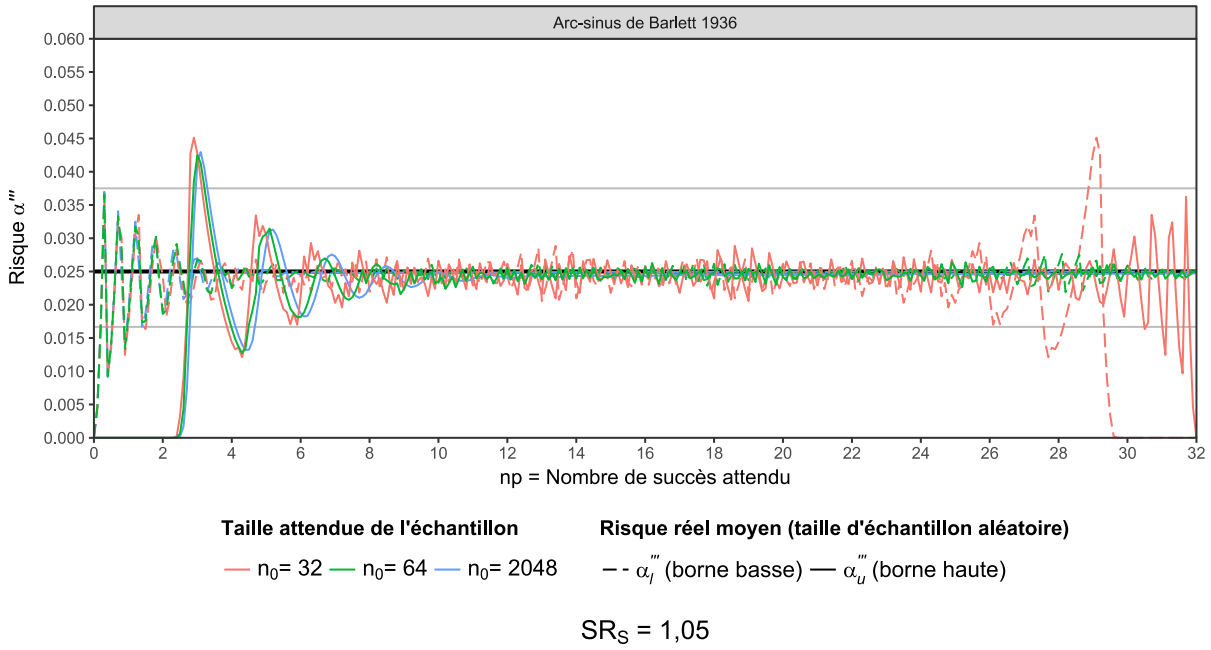

Figure 13 : risques réels moyens unilatéraux de l'estimateur d'intervalle Arc-Sinus de Bartlett à 95% dans un modèle à effectif aléatoire (cf page 44) dans lequel la proportion réelle p est constante mais la taille de l'échantillon N suit une loi log-normale de moyenne n_0 et d'écart-type géométrique $SR_S = 1,05$, arrondie au nombre entier le plus proche. En abscisse, le nombre de succès attendu $n_0 p$ et en ordonnée le risque réel à droite (borne haute, trait plein) et à gauche (borne basse, pointillés) que l'intervalle de confiance ne contienne pas la proportion réelle p. Cette figure est basée sur un effectif moins variable que la Figure 12 avec un écart-type géométrique SR_S égal à $1,05$ plutôt que $1,20$.

3.3 Risques conditionnels

3.3.1 Analyse principale

Les risques réels conditionnels à une taille d'échantillon n et une proportion théorique p fixées sont présentés sur la Figure 14. Les oscillations de risque sont visibles pour tous les estimateurs d'intervalle. Bien que la taille d'échantillon soit $n = 2048$, seuls les résultats pour un nombre attendu de succès $np < 10$, sont présentés. Le biais de l'estimateur d'intervalle de Wald dépasse largement ceux des autres intervalles.

Comme l'analyse des risques moyens locaux le montrait, les intervalles de Wilson ou logit-normal sont déséquilibrés entre le risque à droite et le risque à gauche. Plus précisément, pour une proportion théorique petite, le risque à gauche est supérieur au risque nominal alors que le risque à droite est inférieur au risque nominal $(0{,}025)$; ce biais est opposé au biais de l'intervalle de Wald qui présente un risque à droite très supérieur au risque nominal $(0{,}025)$ et un risque à gauche très inférieur. L'intervalle logit-normal présente des pics de risque conditionnel très élevés pour des nombres de succès attendus inférieurs à 1 ou 2 ; le pic conditionnel unilatéral maximal est égal à $0{,}131$ pour $np = 0{,}1399$, très largement supérieur au risque nominal de $0{,}025$. Ce pic correspond à un nombre attendu de succès très légèrement inférieur à la borne basse de l'intervalle de confiance correspond à un seul succès sur n tirages. L'intervalle du rapport de vraisemblance est plus équilibré entre les risques à droite et à gauche, et présente des pics de risque conditionnels moins élevés que l'intervalle logit-normal.

Les intervalles Arc-Sinus de Bartlett, de Clopper-Pearson mid-P et Jeffreys équilibré (modifié par Brown) présentent des oscillations de risque conditionnel unilatéral autour du risque nominal. Les pics de risques conditionnels unilatéraux peuvent être assez élevés avec un pic maximal à $0{,}0478$ pour le Clopper-Pearson mid-P pour un risque nominal à $0{,}025$. L'intervalle de Clopper-Pearson a des oscillations des risques unilatéraux conditionnels en-dessous du risque nominal à $0{,}025$ de telle sorte que l'intervalle est strictement conservatif aussi bien en analyse unilatérale que bilatérale. En comparaison, les oscillations des risques conditionnels unilatéraux du Clopper-Pearson mid-P sont décalées vers le haut afin de se situer autour du risque nominal unilatéral ; quant au risque bilatéral conditionnel, il n'est pas bien maîtrisé, les pics de risque unilatéraux pouvant se cumuler jusqu'à un pic bilatéral conditionnel maximal égal à $0{,}0833$ pour un risque nominal égal à $0{,}05$.

Les intervalles Arc-Sinus de Bartlett et de Clopper-Pearson mid-P ayant des propriétés proches. Par exemple, pour $n = 32$, pour $x = 0 \ldots 32$, les écarts relatifs entre les bornes hautes des deux intervalles sont tous inférieurs à 6% (maximum 5,5%), avec une moyenne à 0,94%. C'est-à-dire, la borne haute d'un intervalle était, en moyenne 0,94% (en relatif) plus grande que la borne haute de l'autre intervalle pour le même nombre de succès x et la même taille d'échantillon $n = 32$. Par exemple, pour 4 succès sur 32, la borne haute du Clopper-Pearson mid-P était égale à $0{,}01495$ alors que celle de l'Arc-Sinus de Bartlett était égale à $0{,}01483$ soit un écart relatif de 0,83%. La différence se fait sur le troisième chiffre significatif. Par équivariance, les rapports entre les bornes basses suivent les mêmes lois. Pour $n = 64$, les écarts relatifs entre bornes sont encore plus petits avec une moyenne à 0,57% et un maximum à 5,4%.

L'intervalle de Blaker est strictement conservatif pour le risque bilatéral conditionnel puisqu'aucun pic ne dépasse le risque bilatéral nominal égal à $0{,}05$. Par contre, les risques unilatéraux conditionnels peuvent dépasser le risque nominal $(0{,}025)$.

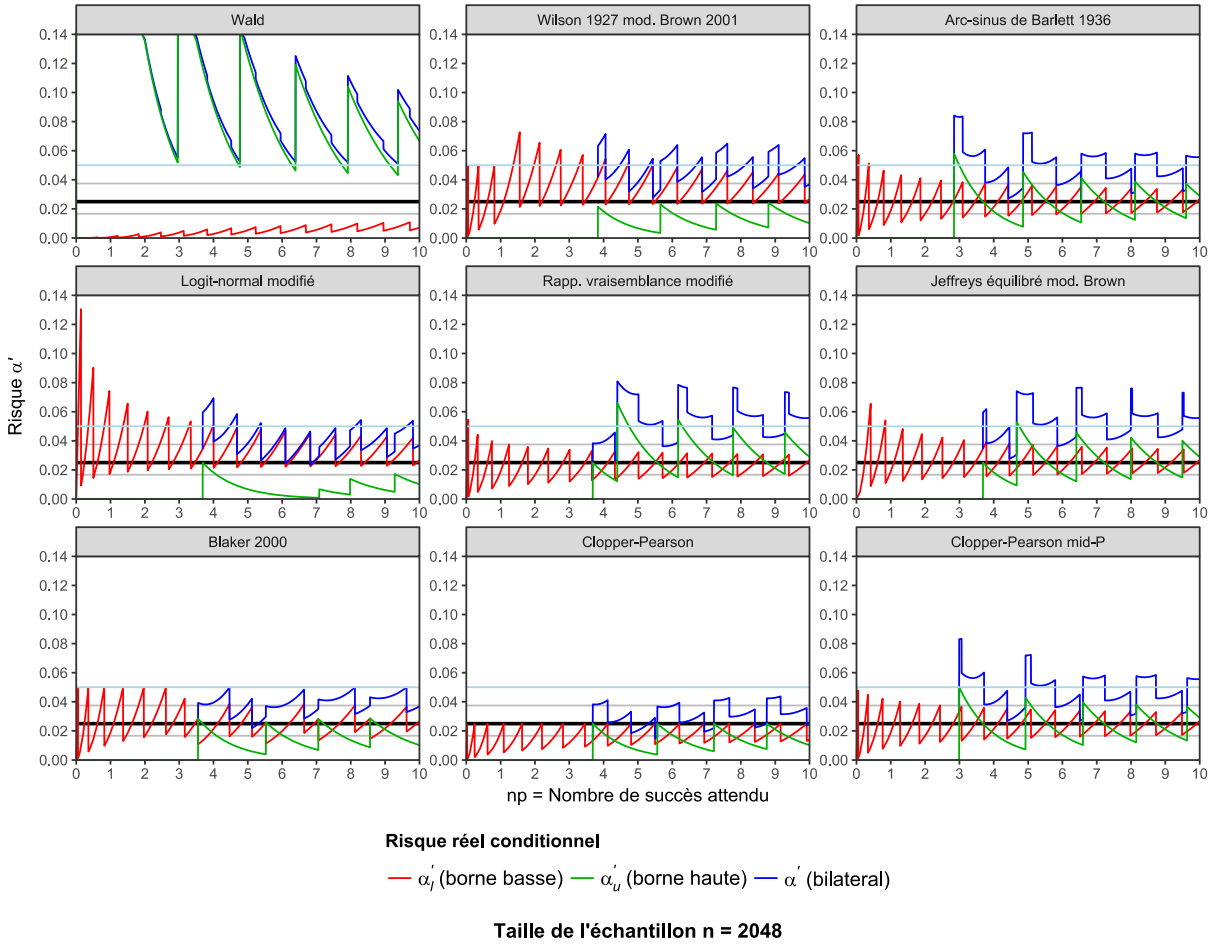

Figure 14 : risques réels conditionnels unilatéraux (tels que définis en page 43) à gauche (rouge), à droite (vert) et bilatéral (bleu) des neuf principaux estimateurs d'intervalle de confiance à 95% pour un échantillon de taille $n = 2048$.

3.3.2 Cas limite de la loi de Poisson

Les résultats obtenus en Figure 14 pour une taille d'échantillon $n = 2048$ sont presque identiques à ceux qu'on obtient en Figure 15 pour une loi de Poisson, équivalent à une loi binomiale lorsque $n \to +\infty$. On retrouve notamment, pour l'estimateur d'intervalle de Clopper-Pearson mid-P (devenant alors le Garwood mid-P (48)), lorsque $n \to +\infty$, un pic de risque conditionnel bilatéral égal à 0,0835 pour un nombre attendu de succès $\lambda = 3{,}5$.

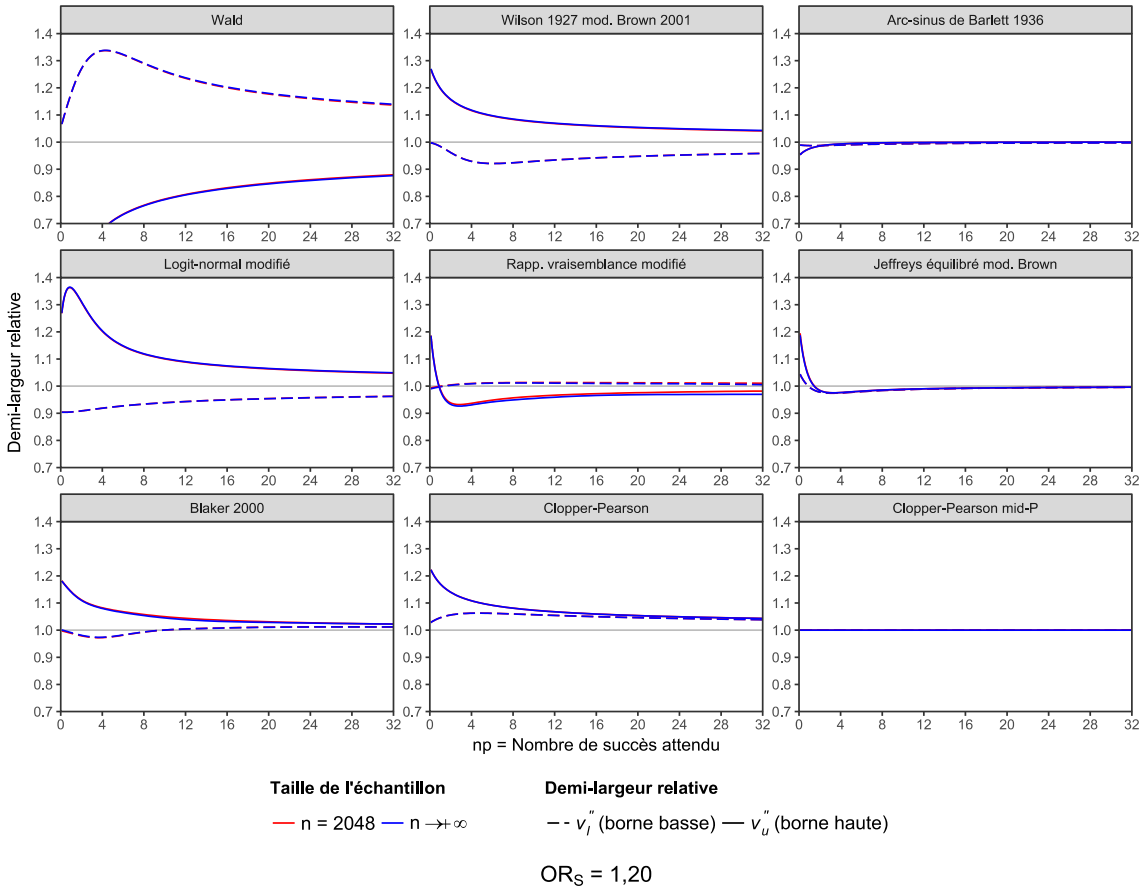

Figure 15 : risques réels conditionnels unilatéraux (tels que définis en page 43) à gauche (rouge), à droite (vert) et bilatéral (bleu) des neuf principaux estimateurs d'intervalle de confiance à 95% pour un échantillon infini correspondant au cas limite de la loi de Poisson. Les résultats sont à comparer à ceux de la Figure 14.

3.4 Demi-largeur relatives moyennes locales

Les demi-largeurs relatives moyennes locales des estimateurs d'intervalles sont présentées sur la Figure 16. Les demi-largeurs moyennes locales sont presque égales aux demi-largeurs attendues conditionnelles à des valeurs précises de n et de p car la demi-largeur attendue est une fonction continue de la proportion théorique p. Par exemple, pour l'intervalle de Wald, l'écart relatif entre demi-largeur attendue conditionnelle et demi-largeur locale moyenne, était inférieur à 1% pour toutes les valeurs de n et p montrées sur la Figure 16. Plus rigoureusement, les propriétés suivantes étaient vérifiées pour $n = 32, 64$ ou 2048 et np compris entre $0{,}1$ et $0{,}99$ pour l'intervalle de Wald :

$$\left|1 - \frac{w_u''}{w_u'}\right| < 0{,}01 \tag{92}$$

$$\left|1 - \frac{w_l''}{w_l'}\right| < 0{,}01 \tag{93}$$

En bref, la différence entre les deux mesures (conditionnelle ou locale moyenne) est négligeable pour la demi-largeur.

L'estimateur d'intervalle de référence pour les demi-largeurs relatives est le Clopper-Pearson mid-P, qui, par construction, aura toujours une demi-largeur relative égale à 1. Une valeur supérieure à 1 de la demi-largeur s'interprète comme une borne d'intervalle plus écartée, en moyenne, de l'estimation ponctuelle $\hat{p}$ que celle de l'intervalle de Clopper-Pearson mid-P.

Les résultats présentés en Figure 16 reflètent les résultats des risques moyens locaux de la Figure 7. L'intervalle de Wald présentant un risque local moyen à droite très élevé lorsque le nombre attendu de succès est bas parce qu'il a une demi-largeur d'intervalle à droite particulièrement petite ; sa borne haute est trop basse.

La similarité entre l'intervalle Arc-Sinus de Bartlett et l'intervalle de Clopper-Pearson mid-P est retrouvée ; ce dernier étant très légèrement plus conservatif pour un nombre attendu de succès très proche de zéro parce qu'il est très légèrement plus large. L'intervalle de Jeffreys équilibré modifié par Brown est un petit peu plus large que le Clopper-Pearson lorsque le nombre de succès attendu est faible, surtout sur sa borne haute. Ceci est explicable par la modification apportée par Brown qui lui substitue un intervalle assez large car strictement conservatif (celui de Garwood (48)) lorsque le nombre de succès observée est nul. Ce résultat est à superposer à la lèvre plus large sur l'intervalle de Jeffreys modifié que sur l'intervalle de Clopper-Pearson mid-P tel que décrit en page 56 et visible en Figure 7. Au contraire, l'intervalle modifié de Jeffreys est légèrement plus étroit que l'intervalle de Clopper-Pearson mid-P pour un nombre attendu de succès proche de 4.

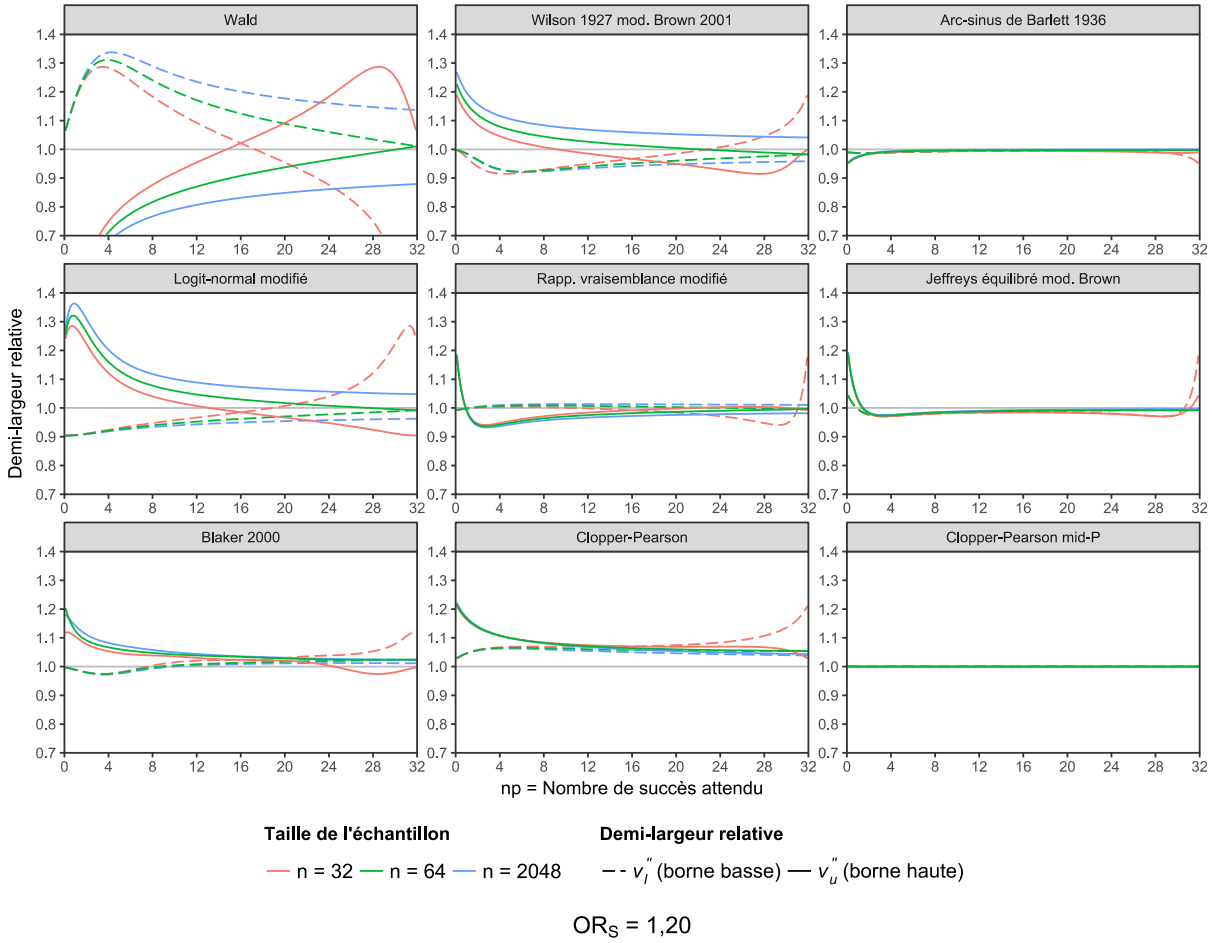

Figure 16 : demi-largeurs relatives moyennes locales (cf page 44) pour les neufs principaux estimateurs d'intervalles de confiances à 95% pour les trois tailles d'échantillon (couleur), pour une proportion *P* théorique aléatoire suivant une loi logit-normale d'odds ratio type $OR_S = 1,20$, selon le nombre de succès attendu $\lambda = np$ dans un échantillon de taille n.

En Figure 16, l'intervalle de Clopper-Pearson a une demi-largeur à droite proche de celle de l'intervalle de Blaker lorsque le nombre de succès attendu est faible, la différence entre les deux estimateurs se faisant alors principalement sur la borne basse. L'intervalle de Blaker étant toujours inclus dans l'intervalle de Clopper-Pearson, ses demi-largeurs à droite et à gauche sont toujours inférieures ou égales.

3.5 Conditions de validité de l'intervalle de Wald

3.5.1 Maîtrise des risques moyens locaux unilatéraux

Le Tableau 5 présente les seuils de validité ξ'' et leurs approximations $\widetilde{\xi''}$ obtenus par une formule empirique spécifiquement construite pour s'ajuster le mieux possible aux valeurs numériques de ξ'' observées dans le tableau, tout en restant suffisamment simple. La quatrième colonne de ce tableau ($\alpha''_{max} = 0{,}075$) correspond à une maîtrise des risques locaux moyens unilatéraux ne devant pas dépasser 1,50 fois le risque nominal, soit 0,0375 à droite et 0,0375 à gauche. Ainsi, pour un échantillon de 512 observations $\xi'' = 31$. Cela signifie qu'il faut au moins 31 succès et 31 échecs pour applique l'estimateur d'intervalle de Wald. En dessous de ce seuil, il est nécessaire d'appliquer l'estimateur d'intervalle de Clopper-Pearson mid-P. Pour le cas limite $n \to +\infty$ correspondant à la loi de Poisson,

un nombre de succès ou d'échecs atteignant ou dépassant 41 est nécessaire à la validité de l'intervalle de Wald afin de ne pas dépasser 1,50 fois le risque nominal.

n	$\alpha''_{max} = 0{,}060$ ξ'' ($\widetilde{\xi''}$)	$\alpha''_{max} = 0{,}070$ ξ'' ($\widetilde{\xi''}$)	$\boldsymbol{\alpha''_{max} = 0{,}075}$ $\boldsymbol{\xi''}$ ($\widetilde{\boldsymbol{\xi''}}$)	$\alpha''_{max} = 0{,}10$ ξ'' ($\widetilde{\xi''}$)
32	11 (14)	9 (11)	**8 (10)**	5 (6)
64	21 (25)	15 (18)	**13 (16)**	7 (8)
128	37 (43)	24 (27)	**19 (22)**	8 (9)
256	62 (70)	34 (38)	**26 (29)**	9 (10)
512	96 (107)	44 (47)	**31 (34)**	10 (11)
1024	136 (147)	51 (54)	**36 (37)**	10 (11)
2048	174 (182)	57 (58)	**38 (39)**	11 (11)
∞	245 (245)	63 (63)	**41 (41)**	11 (11)

Tableau 5 : seuils de validité ξ'' de l'estimateur d'intervalle de confiance de Wald bilatéral à 95%, exprimés en nombre de succès et d'échecs nécessaires à l'application de cet estimateur de Wald, selon le risque local moyen maximal tolérable α''_{max} précisé en colonne et la taille d'échantillon n précisée en ligne. Les seuils $\widetilde{\xi''}$ précisés entre parenthèses sont des approximations des seuils ξ'' obtenus par la condition empirique $nx(n-x) \geq \xi''(\alpha''_{max}, \infty) \times (n-2x)^2$. Les risques sont calculés en se basant sur une proportion aléatoire P pour laquelle l'odds ratio type est $OR_S = 1,10$.

En appliquant la condition de validité suivante :

$$nx(n-x) \geq \xi''(\alpha''_{max}, \infty) \times (n-2x)^2 \tag{94}$$

on trouve tous les seuils entre parenthèses du Tableau 5. Dans le cas où on accepte jusqu'à 1,50 fois le risque nominal (4ème colonne du tableau), le seuil $\xi''(\alpha''_{max}, \infty)$ est égal à 41 et la formule devient :

$$nx(n-x) \geq 41(n-2x)^2 \tag{95}$$

Cette formule correspond à un seuil de la valeur absolue du coefficient d'asymétrie (skewness) qui est égal à $\left|\frac{1-2p}{\sqrt{np(1-p)}}\right|$ pour une proportion binomiale (cf équation (18) page 23), estimé à $\left|\frac{n-2x}{\sqrt{nx(n-x)}}\right|$ sur l'échantillon. En élevant au carré ce coefficient d'asymétrie, on peut rechercher une condition de la forme $\frac{(n-2x)^2}{nx(n-x)} \leq seuil$ soit encore $nx(n-x) \geq \frac{1}{seuil} \times (n-2x)^2$. Enfin, le seuil est choisi de telle sorte que ξ'' et son approximation $\widetilde{\xi''}$ convergent lorsque $n \to +\infty$.

Présentée sous la forme de l'équation (95), cette condition nécessite 7 opérations mathématiques (application de fonction usuelle, multiplication, addition, division, soustraction), ce qui est aussi égal au nombre d'opérations nécessaires au calcul de l'intervalle de Wald lui-même $\left[p - 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; p + 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$. Cette condition complexe peut être simplifiée par restriction au scénario le pire où $n \to +\infty$. La condition devient :

$$\min(x, n-x) > 40 \tag{96}$$

Sous une forme moins formelle : le nombre de succès et d'échecs doivent tous deux dépasser strictement 40.

Les autres colonnes du Tableau 5 servent d'analyse de sensibilité, le choix de ne pas dépasser 1,50 fois le risque nominal présenté en 4ème colonne étant arbitraire. Cette analyse de sensibilité montre une

grande influence du choix de α''_{max} sur les seuils de validité, notamment dans le cas limite $n \to +\infty$ (dernière ligne du tableau). Une autre manière de présenter les α''_{max}, est de les exprimer sous forme de rapport entre le risque local moyen unilatéral maximal toléré et le risque local moyen unilatéral nominal. Ces rapports sont alors respectivement, pour les colonnes 2 à 5 du Tableau 5 : 1,20, 1,40, 1,50 et 2,00.

Une autre analyse de sensibilité a été appliquée concernant le choix de l'estimateur d'intervalle de confiance de substitution : le remplacement de l'intervalle de Clopper-Pearson mid-P par l'intervalle Arc-Sinus de Bartlett n'a changé aucun des seuils ξ'' présentés dans le Tableau 5 suggérant que l'intervalle de substitution n'a pas d'importance tant qu'il est faiblement biaisé.

3.5.2 Maîtrise des risques conditionnels unilatéraux

Le Tableau 6 présente les seuils de validité de l'intervalle de Wald pour la maîtrise du risque unilatéral conditionnel maximal. L'interprétation est la même que pour le Tableau 5 mais les risques conditionnels, plutôt que les risques moyens locaux, sont maîtrisés, et, lorsque le seuil n'est pas atteint, l'intervalle de substitution est celui de Clopper-Pearson (strictement conservatif) plutôt que le Clopper-Pearson mid-P, ce dernier maîtrisant le risque local moyen. On suppose alors que la proportion théorique et l'échantillon sont de tailles fixes lors de la répétition de l'expérience. Par exemple, pour un échantillon de taille 512, la maîtrise d'un risque inférieur ou égal à 1,50 fois le risque nominal (4ème colonne du tableau) nécessite de n'appliquer l'intervalle de Wald que lorsque le nombre de succès et d'échecs dépassent tous deux 70 (seuil ξ'). La formule empirique qui a été construite fournit un seuil légèrement différent ($\tilde{\xi}' = 72$) pour la même taille d'échantillon.

La condition empirique garantissant la maîtrise du risque conditionnel est la suivante :

$$\min(x, n-x) \geq c(\alpha'_{max}) + \frac{n}{2\xi'(\alpha'_{max}, \infty) + n} \tag{97}$$

où

$$c(\alpha'_{max}) = \begin{cases} 6 \; si \; \alpha'_{max} = 0{,}060 \\ 1 \; si \; \alpha'_{max} = 0{,}070 \\ 0 \; si \; \alpha'_{max} \geq 0{,}075 \end{cases} \tag{98}$$

où $\xi'(\alpha'_{max}, \infty)$ est donné par la dernière ligne du Tableau 6. La valeur $c(\alpha'_{max})$ est une constante d'ajustement empirique dépendante de α'_{max}. L'équation (98) n'est applicable qu'au risque nominal $\alpha = 0{,}05$.

Dans le cas où le risque nominal $\alpha'_{max} = 0{,}075$ la condition est la suivante :

$$\min(x, n-x) \geq \frac{n}{198 + n} \tag{99}$$

Si on est prêt à tolérer un risque α'_{max} dépassant légèrement 0,075, cette condition peut s'arrondir :

$$\min(x, n-x) \geq \frac{n}{200 + n} \tag{100}$$

La condition simplifiée suivante est toujours suffisante et est plus facile à enseigner :

$$\min(x, n-x) \geq 100 \tag{101}$$

Les autres seuils de risque conditionnel maximal tolérable α'_{max} présentés dans les colonnes 2, 3 et 5 du Tableau 6 fournissent des résultats différents. Le risque conditionnel maximal que l'on est prêt à tolérer a une forte influence sur le seuil de validité de l'intervalle de Wald.

n	$\alpha'_{max} = 0{,}060$ ξ' ($\tilde{\xi}'$)	$\alpha'_{max} = 0{,}070$ ξ' ($\tilde{\xi}'$)	$\boldsymbol{\alpha'_{max} = 0{,}075}$ $\boldsymbol{\xi'}$ ($\tilde{\boldsymbol{\xi}}'$)	$\alpha'_{max} = 0{,}10$ ξ' ($\tilde{\xi}'$)
32	Jamais (Jamais)	15 (16)	**14 (14)**	10 (11)
64	Jamais (Jamais)	27 (28)	**25 (25)**	14 (15)
128	62 (64)	46 (46)	**39 (39)**	19 (20)
256	110 (111)	69 (70)	**54 (56)**	22 (23)
512	181 (182)	93 (96)	**70 (72)**	25 (26)
1024	271 (274)	115 (117)	**82 (83)**	26 (27)
2048	362 (369)	130 (132)	**89 (91)**	27 (28)
∞	560 (566)	149 (150)	**99 (99)**	28 (28)

Tableau 6 : seuils de validité ξ' de l'estimateur d'intervalle de confiance de Wald bilatéral à 95%, exprimés en nombre de succès et d'échecs nécessaires à l'application de cet estimateur de Wald, selon le risque conditionnel maximal tolérable α'_{max} précisé en colonne et la taille d'échantillon n précisée en ligne. Les seuils $\tilde{\xi}'$ précisés entre parenthèses sont des approximations des seuils ξ' obtenus par la condition empirique $\min(x, n-x) > c + \frac{n}{2\xi'(\alpha'_{max}, \infty) + n}$ où $c = 6$ pour $\alpha'_{max} = 0{,}060$, $c = 1$ pour $\alpha'_{max} = 0{,}070$ et $c = 0$ pour $\alpha'_{max} \geq 0{,}075$.

4 Discussion

Les risques réels sont très variables d'un intervalle à l'autre. L'intervalle de Wald ne maîtrise ni les risques locaux moyens ni les risques conditionnels. Les intervalles de Clopper-Pearson mid-P, Arc-sinus de Bartlett maîtrisent les risques moyens locaux unilatéraux alors que l'intervalle de Clopper-Pearson maîtrise les risques conditionnels unilatéraux. À la lumière de ces résultats, nous allons discuter du problème de symétrie des risques des intervalles bilatéraux, décrire les intervalles dont les comportements nous semblent les meilleurs et discuter des résultats qui sont généralisables aux modèles plus complexes.

4.1 Intervalle bilatéral à risques symétriques

Un des éléments justifiant l'usage d'intervalles à risques déséquilibrés, c'est que pour une proportion théorique très proche de zéro, inférieure à la borne haute de l'intervalle de confiance correspondant à l'observation de zéro succès sur l'échantillon, le risque à droite est nul. Cette zone est appelée lèvre (lip en anglais) par Liu et Kott (68). Certains estimateurs comme celui de Sterne (96) ou de Blaker (12) visent un risque total (droite+gauche) le plus proche possible du risque α nominal, compensant une propension à trop rarement sous-estimer par une propension à trop souvent surestimer et vice versa. Pourtant, l'erreur de surestimation et l'erreur de sous-estimation sont rarement équivalentes, en termes de conséquences. Surestimer rarement le taux d'effets secondaires d'un traitement paraît acceptable, mais compenser ce bas taux d'erreur par une sous-estimation fréquente paraît nettement plus douteux.

Certains intervalles bilatéraux, tels que l'intervalle de Blaker, sont construits en inversant des tests d'hypothèse bilatéraux. Un test d'hypothèse bilatéral de comparaison de proportion observée à une proportion théorique est basé sur une hypothèse nulle et une hypothèse alternative comme suivent :

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

En cas de test significatif, on conclut à l'hypothèse alternative H_1. Une pratique bien répandue consiste à conclure à l'existence d'une différence allant dans le sens observé, tel qu'une supériorité ou une infériorité, mais cette pratique est en dehors du cadre de la théorie. En effet, la théorie n'a pas été développée avec trois hypothèses $p = p_0$, $p > p_0$ et $p < p_0$. Si on applique la pratique de conclure dans la direction observée, les tests ne garantissent pas l'équilibre des risques de montrer $p > p_0$ et $p < p_0$ dans le cas où $p = p_0$. Supposons que l'on veuille comparer deux traitements pour lesquels il n'y a pas de raison de préférer, *a priori*, l'un par rapport à l'autre. Si on les compare par un test bilatéral, alors qu'ils sont équivalents ou presque équivalents, il est possible que l'un d'entre eux ait 4% de chances de prouver sa supériorité et l'autre ait seulement 1% de chances. Ils ne partent donc pas d'un pied d'égalité. Si on veut conclure sur le sens de la différence, il est préférable de faire successivement deux tests unilatéraux à 2,5%, de telle sorte que les deux traitements ait la même chance de prouver leur supériorité s'ils sont équivalents ou presque. Comme les deux éventualités sont incompatibles (supériorité et infériorité), la somme des deux risques unilatéraux $2{,}5\% + 2{,}5\% = 5\%$ est le risque de conclure à l'existence d'un traitement supérieur à l'autre s'ils sont équivalents. Les tests bilatéraux pourraient être renommés en « tests de non équivalence ». Ils semblent rarement utiles en médecine, même s'ils pourraient être plus utiles dans les statistiques des jeux de hasard, pour mettre en évidence, par exemple, le déséquilibre d'une machine à sous, la conclusion étant la même selon que le déséquilibre allant dans un sens ou dans l'autre : la machine doit être réformée.

Certains estimateurs d'intervalles, tels que les intervalles de Clopper-Pearson, Sterne ou Blaker (12,31,96), sont construits par inversion de test. Dans un premier temps, un test de comparaison de valeur observé à théorique est défini. Pour le test de Clopper-Pearson, c'est un test bilatéral à risques équilibrés, alors que pour les tests de Sterne ou de Blaker les risques sont déséquilibrés. Ensuite, toutes les proportions théoriques que le test ne rejetterait pas sont incluses dans l'intervalle. Toutes les valeurs à l'extérieur de l'intervalle sont des valeurs rejetées par le test. Cela illustre une dualité entre intervalles de confiance et tests d'hypothèses. Un tel intervalle peut servir à faire un test d'hypothèse non planifié. Par exemple, supposons qu'un intervalle de confiance de Clopper-Pearson ait été calculé pour un taux d'effets indésirables égal à 3/100. L'intervalle de confiance à 95% à risques équilibrés, publié dans l'article est $[0{,}006\,;0{,}085]$. Une autorité sanitaire peut considérer qu'un taux inférieur à 10% est tolérable. En observant la borne haute de l'intervalle de confiance, sans recalculer de statistique, l'autorité peut affirmer que le taux est inférieur à 10%. Le risque pris est au maximum de 2,5%, car l'intervalle de confiance de Clopper-Pearson est strictement conservatif et a des risques équilibrés. C'est un test unilatéral au risque 2,5% *a posteriori*. Si l'intervalle de Blaker avait été calculé, comme il est basé sur un test strictement conservatif à risques déséquilibrés, le risque pris serait égal à 5% dans le pire de cas. L'intervalle de Blaker deviendrait intéressant si l'autorité sanitaire craignait que le taux d'effets indésirables soit égal à 10%, mais qu'elle n'était pas gênée par un taux inférieur (p.e. 2%) ni par un taux supérieur (p.e. 70%). En effet, dans cette situation, c'est un test bilatéral qu'il faudrait faire. Le risque pris par le test basé sur l'intervalle de Clopper-Pearson, serait, dans le pire des cas, égal à 5%, de la même manière que le risque pris par l'intervalle de Blaker. L'intervalle de Blaker montrerait son intérêt par le fait qu'il est systématiquement plus précis (plus étroit) que l'intervalle de Clopper-Pearson. En fait, l'intervalle de Blaker est construit de telle sorte qu'il est toujours strictement contenu dans l'intervalle de Clopper-Pearson. Même si la dualité entre le test d'hypothèse et l'intervalle de confiance est particulièrement criante pour les estimateurs d'intervalle par inversion de test, elle existe pour tous les intervalles de confiance. En effet, on peut transformer un intervalle de confiance quelconque en test d'hypothèse en rejetant l'hypothèse nulle lorsque la valeur théorique est en dehors de l'intervalle.

Ces problèmes d'interprétation des intervalles bilatéraux ont été soulevés par un référé de l'article d'Agresti et Min (6). En effet, Agresti et Min vantaient les bénéfices d'intervalles bilatéraux déséquilibrés dans l'idée de réduire la largeur moyenne des intervalles. Agresti et Min ont répondu qu'en effet, en recherche biomédicale, le sens de la différence a souvent de l'importance et que dans ce cas un intervalle unilatéral devrait être calculé.

> "For such goals, one can argue in favor of simply calculating a one-sided confidence bound instead of a confidence interval. This may be a psychological barrier for many statisticians because most statistical texts discuss one-sided tests but few discuss one-sided confidence bounds".

Ce qui peut se traduire :

> « Dans un tel but, on peut conseiller le calcul d'une borne de confiance unilatérale plutôt que d'un intervalle de confiance bilatéral. Il peut y avoir une barrière psychologique chez beaucoup de statisticiens parce que la plupart des ouvrages statistiques parlent des intervalles unilatéraux mais peu discutent des bornes de confiance unilatérales ».

On peut comprendre une certaine réticence à ne présenter un intervalle unilatéral, parce qu'il est basé sur un choix non neutre : doit-on présenter un intervalle unilatéral à droite ou un intervalle unilatéral à gauche ? Dans certaines situations, le choix est assez évident, comme dans l'estimation de la diffé-

rence d'efficacité entre un traitement innovant et un traitement de référence, mais lorsque la position est plus neutre, il semble plus intéressant de présenter les deux intervalles de confiance unilatéraux. L'intersection de deux intervalles de confiance unilatéraux aux risques $\frac{\alpha}{2}$ forme un intervalle de confiance bilatéral à risques équilibrés au risque α. Ces intervalles permettent une interprétation orientée (supériorité ou infériorité) avec une maîtrise de tous les risques (à droite, à gauche, et bilatéral).

Ludbrook (72) mentionne que la nature unilatérale ou bilatérale des tests est rarement mentionnée. Sur 5 revues scientifiques, le taux d'articles mentionnant la nature unilatérale ou bilatérale variait de 6,7% pour le minimum à 9,6% pour le maximum. Lorsque la nature était mentionnée, elle était unilatérale dans 10,7% des cas. Il mentionne aussi que beaucoup d'auteurs et d'ouvrages statistiques déconseillent vivement les tests unilatéraux dont par exemple Lombardi et Hurlbert (71). Lombardi et Hurlbert critiquent les tests d'hypothèses unilatéraux pour lesquels, les hypothèses nulle et alternative peuvent être de la forme :

$$H_0 : p \leq p_0$$

$$H_1 : p > p_0$$

Il n'existe pas qu'une seule hypothèse nulle, mais une infinité ; autant qu'il existe de valeurs de p inférieures à p_0. Dans ce cas, le risque α est calculé sur l'hypothèse où le risque est maximal ($p = p_0$) et le risque α n'est plus un risque exact, mais un risque maximal théorique. Cette critique de Lombardi et Hubert est finalement assez rassurante. Si une proportion réelle est égale à 0,1%, le risque de conclure, à tort, qu'elle est supérieure à 30% sur un échantillon de taille 1000, sera presque nul, même si cette analyse est faite dans un test d'hypothèse unilatéral au risque 2,5%. Une autre critique de Lombardi et Hurlbert est le manque de prise en compte de la situation où on observerait une différence importante dans le sens opposé au sens attendu. Dans ce cas, le test unilatéral rendrait invisible un résultat peut-être réel ; il serait aussi assez tentant pour le chercheur, de s'écarter du protocole (s'il y en avait un) et de convertir le test unilatéral en test bilatéral. Lombardi et Hurlbert discutent aussi de la solution proposée par Kaiser (59) de combiner deux tests unilatéraux, qui permet alors, de conclure à la direction de la différence (test bilatéral directionnel). Les intervalles bilatéraux à risques équilibrés vont dans la même logique d'hypothèse *a priori* neutre tout en portant un intérêt à la position des estimations.

Lombardi et Hurlbert mentionnent le fait que l'interprétation des résultats des tests bilatéraux est presque toujours directionnelle bien que la théorie qui ait été développée à l'origine ne le soit pas. Ils précisent que la plupart du temps, la différence est minime et que ça ne vaut pas la peine de mettre à jour la théorie ; pour les citer "[...] unwilling to fight history. But as can be shown, the difference between the two approaches is both small and simple". Dans le pire des cas, l'usage d'un test bilatéral non conçu pour l'analyse unilatérale doublera le risque de conclure à l'existence d'une différence dans la mauvaise direction (erreur de type III ou erreur γ). Ce cas limite correspond à une faible différence associée à une faible puissance statistique. Le plus souvent, les déséquilibres sont plus faibles que ça.

Néanmoins, la théorie classique des tests bilatéraux n'est pas sans conséquence. D'abord, une énergie importante a été investie dans des intervalles de confiance bilatéraux strictement conservatifs à risques volontairement déséquilibrés, conçus dans l'objectif de gagner très légèrement en largeur d'intervalle au coût de biaiser un peu l'interprétation directionnelle, tels que ceux de Sterne (96), Crow (35), Blyth-Still-Casella (15,27), Blaker (12), Kabaila-Byrne (58), Cai-Krishnamoorthy (25), Zieliński (115), Wang (109), Schilling-Doi (93), Lecoutre-Poitevineau (66) ainsi que l'intervalle du score exact et du rapport de vraisemblance exact présentés par Somerville (94) ou par Sakakibara (89). Ces efforts

ne sont utiles que dans des interprétations non directionnelles, autrement l'intervalle de Clopper-Pearson (31), publié en 1934, est l'intervalle strictement conservatif à risques équilibrés optimal (le plus court) selon Wang (108). Comme décrit par selon Lombardi et Hurlbert, les tests bilatéraux équilibrés peuvent être interprétés directionnellement, avec un risque $\frac{\alpha}{2}$ alors que la plupart des tests bilatéraux déséquilibrés peuvent être interprétés directionnellement avec un risque $\leq \alpha$. Il y a quelques exceptions notables. Les tests de comparaison de courbes ROC de Venkatraman (102,103) ne comparent pas les aires sous la courbe ROC mais les courbes ROC elles-mêmes, point par point. Il est tout à fait possible que deux courbes ROC soient manifestement différentes mais aient des aires sous la courbe proches (cf Annexe 3). Dans ce cas, le test de Venkatraman ne doit pas être interprété comme un test de comparaison d'aires sous la courbe ROC, sinon, le risque d'erreur de 3ème espèce est proche de 50%. De même, les tests de Renyi (37) comparant deux courbes de survie dans des situations ou l'hypothèse des risques proportionnels n'est pas respectée, ne permettent pas de conclure dans une direction ou l'autre ; ils permettent de conclure à l'absence d'égalité stricte des courbes point à point.

La théorie bilatérale classique est trop portée sur l'hypothèse nulle d'égalité, quand bien même cette hypothèse est rarement plausible. D'après Anderson (7) sur 95 articles de la revue *The Journal of Wilwife Management* (période 1994-1998), seulement 5 contenaient au moins une hypothèse nulle plausible, bien qu'une quarantaine d'hypothèses nulles soit présentées par article en moyenne sur cette période.

4.2 Paradoxes des intervalles bilatéraux exacts à risques déséquilibrés

Un certain nombre d'intervalles de confiance bilatéraux strictement conservatifs à risques déséquilibrés ont été construits en se basant sur la loi binomiale exacte : Sterne (96) en 1954, Crow (35) en 1956, Blyth-Still (15) en 1983, Blaker (12) en 2000, Kabaila-Byrne (58) en 2001, Schilling-Doi (93) et Wang (109) en 2014, Lecoutre-Poitevineau (66) en 2016.

Tous ces intervalles sont volontairement déséquilibrés dans l'espoir de réduire la largeur moyenne d'intervalle tout en ne dépassant jamais le risque nominal. On peut mentionner deux échecs : les intervalles de Cai-Krishnamoorthy combiné (25) et le Zieliński (115). La Figure 4 de l'Annexe 1 montre des dépassements du risqué nominal bilatéral. Le premier est erroné car il est basé sur l'inversion d'un test combiné prenant le minimum de deux P-valeurs. Chacun des tests d'hypothèse est strictement conservatif, mais le fait de choisir la meilleure P-valeur induit est à l'origine d'un test libéral (biais d'inférence statistique). De même Zieliński propose une procédure le choix d'un déséquilibre volontaire de risque qui conduirait à un intervalle strictement conservatif si le déséquilibre était choisi une fois pour toutes. La constante de déséquilibre étant variable selon le nombre de succès observé, la procédure n'est plus strictement conservative.

Bien que ces procédures exactes à risques déséquilibrés aient une longue histoire, elles présentent quelques paradoxes.

On peut rappeler que les intervalles de Blaker et de Sterne sont basés sur l'inversion d'un test binomial basé sur une P-valeur discontinue et non bimonotone (43). Ces tests peuvent rejeter une proportion théorique à un risque α alors que des valeurs plus grandes ou plus petites ne sont pas rejetées (13). Les régions de confiances peuvent être la réunion de plusieurs intervalles disjoints. Ce paradoxe a déjà été mentionné dans la section « Cohérence avec un test d'hypothèse » en page 36.

Les intervalles de Crow, Blyth-Still et Kabaila-Byrne ne sont pas emboîtés (13,66), ce qui signifie qu'un intervalle de confiance à 90% n'est pas forcément inclus dans l'intervalle de confiance à 95%, conduisant à rejeter une hypothèse au risque 5% mais pas au risque 10%.

Les intervalles de Sterne et de Blaker sont emboîtés, mais l'ajout d'une observation à un échantillon est susceptible d'abaisser la borne basse de l'intervalle de confiance, quel qu'en soit son statut (succès ou échec). Un résultat statistiquement significatif peut devenir non significatif lorsqu'une observation est ajoutée quel que soit son statut. Contrairement au chat de Schrödinger dont le statut n'est pas connu avant d'ouvrir la boîte, cette fois-ci, le statut de l'expérience est connu avant de regarder l'observation. Peut-on considérer que la fonction d'onde est effondrée avant même de regarder l'observation ? Lecoutre et Poitevineau (66) ont proposé une procédure résistante à ce paradoxe. Klaschka (61) note que Lecoutre et Poitevineau n'ont pas supprimé tous les défauts de monotonie des bornes de l'intervalle de confiance selon x et selon n ; il propose alors encore une amélioration de la procédure ajustant les intervalles pour garantir cette monotonie.

Les intervalles à risques bilatéral déséquilibrés non basés sur la loi binomiale exacte, tel que l'intervalle de Wilson ne présentent pas ces paradoxes.

4.3 Le meilleur estimateur d'intervalle ?

Les estimateurs d'intervalle Arc-Sinus de Bartlett, Clopper-Pearson mid-P et de Jeffreys équilibré modifié par Brown, ont la meilleure maîtrise des risques locaux moyens unilatéraux (cf Figure 7 en page 57). Ceci est expliqué parce que ces trois intervalles ont des risques conditionnels oscillant autour du risque nominal (cf Figure 14 en page 65). L'intervalle de Jeffreys équilibré modifié est basé sur une modification *ad hoc* de Brown (21), le rendant peu généralisable et peu intéressant à enseigner en cours de statistiques élémentaires. Ces modifications, en plus de compliquer la formule, ont pour effet de fournir une borne basse d'intervalle de confiance égale à zéro pour un nombre de succès égal à 1. En conséquence, les bornes de cet estimateur d'intervalle ne sont pas strictement monotones. Par ailleurs, l'intervalle est un peu plus large que le Clopper-Pearson mid-P ou l'Arc-Sinus de Bartlett lorsque le nombre attendu de succès est très proche de zéro (cf Figure 16 en page 68). C'est pourquoi nous n'en recommandons pas l'usage.

L'intervalle Arc-Sinus de Bartlett a des propriétés de maîtrise des risques et de largeur d'intervalle presque indistinguables de l'intervalle de Clopper-Pearson mid-P. Numériquement, ces deux intervalles fournissent des résultats différant de moins d'un pourcent relatif en moyenne sur chacune des deux bornes dès que $n \geq 32$. Numériquement, il n'y a pas de raison de préférer un estimateur à l'autre. L'intervalle Arc-Sinus de Bartlett a l'avantage de la simplicité de calcul puisque l'algorithme n'est pas itératif : il s'agit d'une solution de forme fermée basée sur des fonctions usuelles disponibles sur toute calculatrice scientifique. Par contre, d'un point de vue théorique, la solution de Clopper-Pearson mid-P a un avantage.

L'intervalle Arc-Sinus de Bartlett n'est pas basé sur la meilleure fonction stabilisatrice de variance, puisqu'Anscombe (8) et Freeman-Tukey (46) ont proposé des modifications mineures de la transformation Arc-Sinus qui stabilisent mieux la variance. Pourtant, les résultats de la Figure 1 de l'Annexe 1 montrent un biais moins important pour l'intervalle Arc-Sinus de Bartlett. Selon Efron (40), la normalisation (rapprochement d'une loi normale) et la stabilisation de la variance ne peuvent être simultanément optimisées pour la distribution de Poisson ou la distribution binomiale. Les intervalles d'Anscombe et de Freeman-Tukey ont un comportement libéral (intervalle trop étroit) lorsque le nombre de succès attendu est très proche de zéro alors que la variance reste assez stable. Après trans-

formation de la loi binomiale selon la transformée Arc-Sinus de Bartlett, la variance décroit lorsque le nombre de succès attendu est très proche de zéro, ce qui semble compenser le comportement libéral par un comportement conservatif opposé, conduisant à une meilleure maîtrise du risque moyen local.

L'intervalle de Clopper-Pearson mid-P a l'avantage d'avoir un cadre théorique complet : il est naturellement associé à un test d'hypothèse, car il est construit par inversion de test. Le test est facile à confronter à celui de Clopper-Pearson (strictement conservatif), avec la théorie mid-P (Berry (11)) qui consiste à créer un compromis entre l'inégalité stricte et l'inégalité large lorsque la loi de distribution est discrète. L'inégalité large est à l'origine du Clopper-Pearson, garantissant la maîtrise des risques conditionnels, avec des oscillations ne dépassant jamais le risque nominal. L'inégalité stricte serait à l'origine d'un intervalle strictement libéral dont les oscillations ne passeraient jamais en dessous du risque nominal, et le compromis « supérieur et à moitié égal à » du Clopper-Pearson mid-P maîtrise les risques moyens locaux, en centrant les oscillations sur le risque nominal.

Doit-on préférer un intervalle strictement conservatif tel que le Clopper-Pearson ou un intervalle dont les risques moyens locaux sont au plus proche du risque nominal ?

En réponse à l'article de Brown, Cai et DasGupta (21) en 2001, Casella (28), Corcoran et Mehta (32) suggéraient qu'un intervalle strictement conservatif est préférable, alors que Ghosh (52), Santner (91) ou Agresti et Coull (5) ainsi que Brown, Cai et DasGupta privilégient plutôt le risque moyen. Les résultats de cette thèse, basés sur les hypothèses d'une proportion aléatoire ou d'un échantillon de taille aléatoire suggèrent que le risque local moyen est une meilleure mesure, car dans les conditions expérimentales en biostatistique. Comme décrit dans la section « Maîtrise du risque conditionnel, du risque moyen ou du risque moyen local » à partir de la page 33, la taille de l'échantillon est rarement bien maîtrisée et le risque réel fluctue d'une expérience à l'autre. En conséquence, le risque moyen local est un paramètre plus pertinent que le risque conditionnel.

Il existe toujours un cas où le risque conditionnel à un n et un p prédéfinis est intéressant : lorsque le protocole mentionne une comparaison de proportion observée à théorique dans des conditions expérimentales où le nombre de sujets est maîtrisé telles que des expériences de biologie fondamentale sur des souris (cf page 35). Dans ce cas, on peut privilégier l'intervalle de Clopper-Pearson (strictement conservatif) si les conséquences d'un risque α (conclure à une différence alors qu'elle n'existe pas) sont plus graves que celles d'un risque β (ne pas conclure à l'existence d'une différence qui existe). Dans le cadre d'essais cliniques régulés, le conservatisme sur les évaluations d'efficacité peut être préféré au libéralisme (ce n'est pas le cas pour les effets secondaires). Une comparaison de proportion observée à théorique prévue dans le protocole pourra se faire avec l'intervalle de Clopper-Pearson puisque la proportion théorique est fixe (donnée dans le protocole) et la taille de l'échantillon est plus ou moins maîtrisée. En présence de conflit d'intérêt on pourrait notamment craindre que les tailles d'échantillon « chanceuses », pour lesquelles le risque α est augmenté, soient visées. On peut aussi comprendre que les autorités sanitaires se méfient d'une procédure non strictement conservative.

Les intervalles de Clopper-Pearson et de Clopper-Pearson mid-P sont généralisables aux régressions logistiques exactes. La commande `exlogistic` du logiciel Stata (95) se base sur les algorithmes d'Hirji, Mehta et Patel (55) et implémente à la fois un intervalle strictement conservatif de chaque coefficient de régression mais aussi une variante mid-P. Les coefficients sont interprétables conditionnellement aux autres. D'après Derr (38), le logiciel SAS utilise aussi l'algorithme d'Hirji, Mehta et Patel. Les sorties SAS contiennent peuvent contenir les intervalles de confiance mid-P si les options `midpfactor=(0.5,0.5) cltype=midp estimate=odds` sont passées à la commande

`exact`. Ces commandes SAS ou Stata permettent de calculer les intervalles de Clopper-Pearson ou de Clopper-Pearson mid-P dans des modèles à intercept seul (aucune covariable) sauf lorsque le nombre de succès ou le nombre d'échec est nul puisque, dans ce cas, le `logit` de la proportion observée est indéfinie. La régression logistique exacte du paquet 'elrm' du logiciel R est basé sur un autre algorithme (Monte Carlo par chaînes de Markov) proposé par Zamar (113) qui ne présente que les intervalles strictement conservatifs. À notre connaissance SPSS (version 25, août 2017) ne supporte pas encore la régression logistique exacte.

On peut alors réduire le nombre d'intervalles utiles à deux : l'intervalle de Clopper-Pearson mid-P à utiliser presque toujours, et l'intervalle de Clopper-Pearson à n'utiliser que dans des conditions de comparaison de proportion observée à théorique dans un cadre expérimental bien maîtrisé ou fortement régulé.

4.4 Wald, Score et rapport de vraisemblance dans les régressions logistiques

Plusieurs intervalles d'inversion de test du rapport de vraisemblance existent. La version présentée dans l'analyse principale (Tableau 4, page 48) est basée sur l'approximation à une loi du χ^2 de la différence des déviances ; la déviance étant égale à $-2\log(vraisemblance)$. C'est la méthode '*profile likelihood*' utilisée par les logiciels statistiques (SAS, SPSS, Stata, R et bien d'autres) pour les régressions logistiques. La méthode exacte strictement conservative est présentée en annexe ; elle est basée sur la même statistique de différence des déviances mais la distribution théorique de la statistique est calculée à partir de la distribution binomiale exacte plutôt que par approximation du χ^2. On pourrait aussi créer une variante mid-P (non montrée). Le test et l'intervalle par approximation du χ^2 sont particulièrement simples à calculer, s'appliquent aux modèles statistiques les plus complexes et sont très largement disponibles dans les logiciels statistiques. Nous ne discuterons que de ce test approximatif.

En l'absence de transformation, l'intervalle de Wald a montré des biais très importants en comparaison des méthodes du score (intervalle de Wilson) ou du rapport de vraisemblance approximé à un χ^2. Cette dernière méthode semble être la moins biaisée des trois.

Les choses sont un peu différentes après transformation logistique (modèle de régression logistique). L'intervalle logit-normal est équivalent à l'intervalle de Wald d'une régression logistique. Ceci est visible en Annexe 1, Figure 4 ; les panneaux « Logit-normal » et « Wald GLM logit, CP si k=0 » montrent des courbes identiques. Par contre, l'intervalle du rapport de vraisemblance est identique avec ou sans transformation logistique, les panneaux « RV GLM logit, CP si k=0 » et « Rapp. Vraisemblance modifié » montrant les mêmes courbes sur la Figure 4 de l'Annexe 1. Ce résultat est explicable parce qu'à transformation près, la déviance du modèle de régression logistique pour l'intercept $\text{logit}(p)$ est égale à la déviance du modèle sans transformation pour la proportion théorique p. Les biais sont plus faibles et l'intervalle et mieux équilibré avec la méthode du rapport de vraisemblance qu'avec la méthode de Wald, même après transformation logistique. Ceci est cohérent avec la remarque d'Agresti dans son livre d'introduction à l'analyse catégorielle (2) (page 13):

> "When the sample size is small to moderate, the Wald test is the least reliable of the three tests. We should not trust it for such a small n as in this example (n = 10). Likelihood-ratio inference and score-test based inference are better in terms of actual error probabilities coming close to matching nominal levels".

L'inférence basée sur le score ou sur le rapport de vraisemblance serait plus fiable, sur les échantillons de petite taille que l'inférence basée sur le test de Wald, même si asymptotiquement (très grands échantillons) ils sont équivalents. Nos résultats suggèrent de plus que l'intervalle basé sur le score (intervalle de Wilson) a des risques moins équilibrés que l'intervalle du rapport de vraisemblance ; le risque bilatéral est bien conservé mais il existe un déséquilibre entre les risques de surestimation et de sous-estimation. Nous conseillons donc de privilégier le test et l'intervalle du rapport de vraisemblance dans les modèles linéaires généralisés.

4.5 Correction de continuité

Les intervalles avec corrections de continuité sont présentés en Annexe 1, Figure 1, Figure 3 et Figure 5. Ces corrections élargissent l'intervalle des deux côtés (à droite et à gauche), rendant l'intervalle plus conservatif des deux côtés. Pour un intervalle équilibré tel que l'intervalle Arc-Sinus d'Anscombe (Annexe 1, Figure 3), les oscillations des risques conditionnels unilatéraux qui se situaient autour du risque nominal, descendent au-dessous du risque nominal après application de la correction de continuité. Pour un intervalle à risques déséquilibrés comme l'intervalle de Wald, la correction de continuité réduit très légèrement le biais d'un côté mais l'aggrave de l'autre.

La correction de continuité est classiquement décrite dans le cas d'une loi binomiale $B(n\,;\,0{,}50)$, en notant que l'approximation de la fonction de répartition de la loi binomiale à la fonction de répartition de la loi normale, après centrage et réduction, et nettement plus précise après ajout d'un demi succès. Ceci s'applique à la fonction de répartition de la loi binomiale définie par $F(x) = Proba(X \leq x)$. En l'absence de correction de continuité, la fonction de répartition de l'approximation normale est plus proche de $G(x) = Proba(X < x) + \frac{1}{2} Proba(X = x)$. Ainsi, on peut rapprocher l'approximation normale avec correction de continuité de la méthode de Clopper-Pearson (strictement conservative) alors qu'en l'absence de correction de continuité, l'approximation normale se rapproche plus de la méthode de Clopper-Pearson mid-P. Par exemple, pour une loi binomiale $B(10\,;0{,}50)$, et un nombre de succès égal à 3, la fonction de répartition exacte de la loi binomiale pour $F(3) = 0{,}172$ alors que son approximation normale avec correction de continuité est égale à $F_{approx}(3) = 0{,}171$. De plus, $G(3) = 0{,}103$ et l'approximation normale sans correction de continuité vaut $G_{approx}(3) = 0{,}113$.

L'approximation de la loi binomiale par la loi normale est bonne lorsque la proportion théorique est proche de 50%, mais la loi binomiale devenant asymétrique pour les proportions proches de 0% ou de 100%, cette approximation est mauvaise. C'est pourquoi on peut exprimer la condition de validité de l'intervalle de Wald, pour la maîtrise du risque moyen local, sous la forme d'un seuil sur le coefficient d'asymétrie. À nombre attendu de succès égal, une augmentation de la taille d'échantillon aggrave l'asymétrie ainsi que le biais, le pire étant le cas limite de la loi de Poisson. La correction de continuité ne corrige donc pas le bon problème. De plus, pour un intervalle équilibré comme l'Arc-Sinus d'Anscombe, elle se contente de le rendre plus conservatif. Lorsque le strict conservatisme est souhaité, l'intervalle de Clopper-Pearson s'impose et il n'y a pas d'intérêt à utiliser des intervalles approximatifs avec correction de continuité qui ne garantissent pas théoriquement cette propriété même s'ils peuvent s'en approcher en pratique.

4.6 Bootstrap

En Annexe 1, les intervalles de bootstrap sont présentés. La méthode de bootstrap introduite par Efron (39) en 1979 est une méthode d'estimation empirique des fluctuations d'échantillonnage par simulations d'expériences virtuelles réalisées en constituant des échantillons de bootstrap par tirage au sort (par ordinateur) avec remise dans l'échantillon principal. La statistique est calculée sur chaque échan-

tillon de bootstrap et les fluctuations empiriques d'échantillonnage de bootstrap se rapprochent des fluctuations d'échantillonnage qu'on obtiendrait en tirant au sort des échantillons depuis la population complète. Les méthodes de bootstrap permettent d'estimer des intervalles de confiance pour tous types de statistiques, y compris les plus complexes, pour lesquelles on ne connaît pas de solution analytique. Le bootstrap a à la fois l'avantage d'être une méthode générique, mais aussi d'être toujours asymptotiquement correct là où les méthodes classiques peuvent ne pas l'être comme l'estimation d'un coeffîcient de corrélation de Pearson élevé avec la transformation z de Fisher (cf Brown (20)).

Il existe plusieurs variantes du bootstrap. Le principe est toujours le même : estimer la distribution de la statistique calculée sur les échantillons de bootstrap. Par contre, il y a plusieurs manières de construire un intervalle de confiance. Ces méthodes sont présentées pédagogiquement par Carpenter (26). En annexe, ce sont les méthodes dites « percentile », « basique », « studentisée » et « BC_a » qui sont présentées. Les échantillonnages par ordinateur nécessitent de la puissance de calcul, car il faut faire une dizaine de milliers d'échantillons de bootstrap pour calculer un intervalle de confiance. Pour la loi binomiale, ces fluctuations d'échantillonnage sont prévisibles et des solutions analytiques ont été utilisées. Les résultats sont équivalents à ceux qu'on obtiendrait avec le paquet standard « boot » du logiciel R pour un nombre de simulations infini.

Les résultats montrent un biais important du bootstrap percentile lorsque le nombre de succès ou d'échecs est inférieur à 32, probablement à cause de l'asymétrie de la distribution et du défaut de prise en compte du lien entre variance et moyenne. L'intervalle basique est encore pire. Le bootstrap studentisé est trop souvent incalculable (division par zéro) et autrement, est trop conservatif. Le bootstrap BC_a est très fortement biaisé mais n'est pas équivariant à cause d'un défaut de l'équation 3.2 d'Efron (41) basé sur une inégalité stricte sur une loi discrète. Une correction de cette équation, rajoutant la moitié de la probabilité d'égalité à la probabilité d'inégalité stricte, permet fournit un intervalle « Boot. BC_a modif » nettement moins biaisé, un peu conservatif lorsque le nombre de succès attendu est inférieur à 10, présentant aussi quelques oscillations même pour les risques moyens locaux. Le BC_a lissé (ajout d'un petit bruit aléatoire dans les simulations, afin de rendre la distribution de bootstrap continue) a un comportement stable mais assez conservatif, proche du Clopper-Pearson. On peut alors comprendre que les oscillations persistantes du « Boot. BC_a modif » sont dues à la nature doublement discrète des fluctuations d'échantillonnage : la distribution de $\hat{P}$ est discrète, mais les fluctuations empiriques de la statistique des ré-échantillonnages $\hat{P}^*$ sont aussi discrètes, sauf lorsqu'on les lisse par ajout de bruit aléatoire.

Le bootstrap 'normal' est strictement équivalent à l'intervalle de Wald, et, par conséquence est aussi biaisé. Pour rappel, l'intervalle de Wald présente deux problèmes principaux : l'approximation de la variance de la population à la variance de l'échantillon (1er problème) et l'approximation de la loi binomiale à la loi normale (2ème problème). Le 1er problème est corrigé par la méthode de Wilson, qui est nettement moins biaisée. L'intervalle de bootstrap 'normal' présente les deux problèmes. Les intervalles basiques et percentile présentent le 1er problème mais luttent contre le 2ème par approximation de la fonction de répartition théorique à la fonction de répartition empirique. Les intervalles studentisés et BC_a tendent à corriger les deux problèmes mais cela reste asymptotique. La statistique pivot de l'intervalle studentisé ne dépend pas de la variance, mais les fluctuations de cette statistique pivot sont estimées sur la loi binomiale de l'échantillon $B(n\,;\hat{p})$ plutôt que sur la loi binomiale de la population $B(n\,;p)$. Néanmoins, le bootstrap BC_a, moyennant une modification l'adaptant aux distributions discrètes (lissage ou correction de l'équation 3.2 d'Efron), converge assez vite ; il est satisfaisant pour un nombre attendu de succès atteignant ou dépassant environ 10.

Le bootstrap ‘normal’ n’est utile que pour des statistiques plus complexes telles que le coefficient de corrélation de Pearson ; elle n’accélère pas la convergence du théorème central limite, mais permet d’estimer la variance de l’estimateur de manière non biaisée.

Selon Carpenter (26), dans les conditions habituelles, la vitesse de convergence théorique des bootstrap ‘normal’, studentisé et percentile est de l’ordre de $\frac{1}{\sqrt{n}}$ alors que la vitesse de convergence théorique des bootstrap studentisé et BC_a sont de l’ordre de $\frac{1}{n}$. Cela signifie qu’en quadruplant la taille de l’échantillon (et par la même occasion, le nombre attendu de succès), le biais sera approximativement divisé par quatre pour le bootstrap studentisé ou le boostrap BC_a alors qu’il sera approximativement divisé par deux pour le bootstrap ‘normal’, percentile ou basique. Cette propriété de division du biais étant elle-même asymptotique, ce n’est qu’une approximation. Ainsi, la vitesse de convergence des bootstrap percentile est basique est du même ordre que celle de l’intervalle de Wald. Une observation plus fine semble montrer une hiérarchie, le biais de l’intervalle basique étant un peu plus important que celui de l’intervalle de Wald, lui-même plus important que l’intervalle percentile.

Les méthodes de bootstrap sont toutes asymptotiquement correctes mais peuvent être très biaisées sur les petits échantillons. Les méthodes sans correction de l’asymétrie (basique ou percentile) nécessitent au moins une trentaine de succès et d’échecs. La méthode BC_a prenant en compte l’asymétrie et supposé converger plus rapidement n’est correcte que lorsqu’on prend en compte, d’une manière ou d’une autre, l’aspect discret de la distribution de bootstrap. Même si le bootstrap est un ensemble de méthodes puissantes, elles ne font pas de miracle sur les petits échantillons.

4.7 Conditions de validité de l’intervalle de Wald

Le strict conservatisme n’a pas lieu d’être recherché avec l’intervalle de Wald, l’intervalle de Clopper-Pearson étant beaucoup plus adapté. La condition de validité garantissant le strict conservatisme $\min(x, n-x) \geq \frac{n}{200+n}$ (équation (100), page 71) est une formule *ad hoc* ; elle n’est pas basée sur une théorie mathématique, mais elle coïncide très bien avec les données empiriques du Tableau 6.

Les conditions de validité à retenir portent sur les risques moyens locaux. Même si la condition la plus fine peut s’exprimer sous forme d’un seuil de coefficient d’asymétrie de la loi binomiale de l’échantillon, il est plus simple de retenir une condition simple : les nombres de succès et d’échecs doivent tous deux être strictement supérieurs à 40. En dessous de ce seuil, l’intervalle de Clopper-Pearson mid-P devrait être utilisé. Si cet intervalle n’est pas disponible pour des raisons techniques, on pourra utiliser un autre intervalle faiblement biaisé, tel que l’intervalle Arc-Sinus de Bartlett.

La condition $\min(x, n-x) > 40$ (équation (96), page 69) s’applique à une tolérance d’un risque réel 1,5 fois plus grand que le risque nominal. Les conditions de validité de l’intervalle de Wald dépendent beaucoup de la tolérance que l’on porte au biais.

Lorsque les conditions de validité de Wald ne sont pas respectées, l’approche consistant à ne pas fournir d’intervalle de confiance du tout, est fortement biaisée, car il s’agit de rejeter la statistique lorsque le nombre de succès (ou d’échecs) est bas. Les intervalles présentés tendraient à systématiquement rapprocher la proportion de 50% : sous-estimer les proportions élevées et surestimer les proportions basses. On ne peut pas se passer d’un second estimateur, c’est pourquoi la validité a été définie par un estimateur hybride. L’intervalle de Wald perd complètement sa simplicité. Nous recommandons de ne jamais utiliser l’intervalle de Wald ; il est plus simple d’utiliser systématiquement l’intervalle de Clopper-Pearson mid-P.

4.8 Loi de Poisson

Certains auteurs ont analysé spécifiquement les intervalles de confiance de la distribution de Poisson comme Byrne (23). Ce qui est présenté dans cette thèse pour une taille d'échantillon $n = 2048$ est graphiquement indistinguable du cas limite de la loi de Poisson.

4.9 Risque nominal différent de 0,05

Le risque nominal bilatéral $\alpha = 0{,}05$, interprété comme un risque nominal unilatéral à $0{,}025$ de chacun des côtés de l'intervalle, a été présenté dans presque toutes les figures. Ce choix a été fait car les intervalles de confiance à 95% sont omniprésents en biostatistique, bien que le risque α soit arbitraire. D'une manière générale, les approximations asymptotiques sont plus fiables sur les quantiles centraux d'une distribution que sur les quantiles extrêmes. Le théorème central limite s'applique bien mieux au 1er quartile d'une distribution qu'à son 1er percentile. C'est pourquoi les biais relatifs sont plus raisonnables pour les intervalles de confiance à 90% que pour les intervalles de confiance à 95%. À l'opposé, les biais relatifs seraient plus importants sur les intervalles de confiance à 99%. Cela doit être pris en considération lorsqu'on applique une correction de multiplicité des tests.

4.10 Demi-largeurs attendues relatives

Le choix a été fait de présenter les demi-largeurs attendues **relatives** à la demi-largeur de l'intervalle de Clopper-Pearson mid-P. Les largeurs d'intervalle attendues, conditionnelles à une proportion théorique, sont souvent présentées graphiquement dans la littérature (Brown (21) ou Pires (83)). Les largeurs des intervalles de confiance sont beaucoup plus élevées pour des proportions proches de 50% que pour des proportions proches de 0% ou 100%. Le rapport entre la largeur d'intervalle attendue pour $p = 0{,}50$ et pour $np = 5$ augmente avec la taille d'échantillon n. Lorsque n est grand, la comparaison graphique de la largeur attendue de deux intervalles pour des petites valeurs de np est presque impossible car l'échelle des ordonnées est trop grande et les deux largeurs d'intervalle sont graphiquement trop proches de zéro. La représentation d'une largeur **relative** à la largeur d'un intervalle de référence réduit énormément la variance relative de la mesure. Par ailleurs, les analyses des risques ayant été faites séparément à droite et à gauche, les largeurs d'intervalles ont aussi été analysés séparément à droite et à gauche, sous forme de **demi-largeurs**. L'intervalle de Clopper-Pearson mid-P a été choisi comme intervalle de confiance de référence parce qu'il a des risques locaux moyens unilatéraux proches du risque nominal, ce qui facilite la comparaison de deux largeurs relatives attendues sans avoir à garder à l'esprit le biais de l'estimateur de référence.

4.11 Implémentations

Bien qu'ayant des propriétés très intéressantes, l'estimateur d'intervalle de Clopper-Pearson mid-P n'est pas disponible dans beaucoup de logiciels statistiques ; l'intervalle Arc-Sinus de Bartlett non plus. Heureusement, ces procédures sont assez faciles à implémenter.

Le logiciel R contient la procédure dans le paquet `exactci`. Cette méthode est appelée « centrale midp » et peut s'obtenir par la commande suivante :

```
exactci::binom.exact(x, n, midp=T)
```

Le logiciel SAS à partir de la version 9.4 contient l'option `MIDP` dans la commande `EXACT` de la `PROC FREQ`. L'usage de la régression logistique exacte est possible dans les versions antérieures, mais l'usage en est complexe.

Des macros ou programmes calculant l'intervalle de Clopper-Pearson mid-P sont proposées en annexe 2 pour les logiciels **SAS** (version 9.3 ou antérieur), **Stata** (version ≥ 9), **SPSS** (version 16 et 25 testées), **Python** (dépend de scipy), **SYSTAT/MYSTAT** (version ≥ 12), **Minitab** (version 18 testée), **HTML+JavaScript** (en ligne ou en local), **Microsoft Excel** (avec ou sans macro), **LibreOffice** et Texas Instruments **Ti 83/84**. La version d'évaluation de Statistica étant soumise à un contrôle trop strict (validation d'identité et d'adresse e-mail avec risque de recevoir des spams à vie), il n'a pas été développé de macro pour ce logiciel. Toutes ces macros sont librement diffusables sous licence Creative Commons CC0. Il s'agit d'une mise à disposition se rapprochant le plus possible du domaine public.

5 Conclusion

Que faut-il retenir de cette thèse ?

Les tests d'hypothèses et les intervalles de confiances bilatéraux devraient toujours être conçus dans une optique d'interprétation directionnelle. Par exemple, lorsque deux traitements sont comparés, quand bien même on n'a pas de raison de préférer l'un à l'autre *a priori*, trois conclusions sont possibles :

- le traitement A est significativement meilleur que le traitement B ;
- le traitement B est significativement meilleur que le traitement A ;
- les données ne sont pas suffisantes pour conclure à la supériorité de l'un ou de l'autre des traitements.

La théorie bilatérale usuelle n'étant basée que sur deux hypothèses (égalité ou différence), elle a conduit à certains tests ou intervalles de confiances peu utiles, qui favoriseraient indûment un traitement par rapport à l'autre (déséquilibre des risques). Idéalement, un intervalle de confiance bilatéral au risque α devrait être équivalent à l'intersection de deux intervalles de confiance unilatéraux au risque $\alpha/2$. De même, les tests d'hypothèses bilatéraux devraient être équivalents aux tests unilatéraux, à un facteur deux près.

L'intervalle de confiance par inversion de test du rapport de vraisemblance (*profile likelihood*) semble moins biaisé que l'intervalle de Wald pour l'estimation des modèles linéaires généralisés. Ceci est à confirmer sur des modèles bivariés et multivariés.

L'estimateur d'intervalle de confiance de Wald pour une proportion binomiale, est extrêmement biaisé. La condition de validité $n\hat{p} > 40\ et\ n(1-\hat{p}) > 40$ paraît suffisante pour garantir une certaine maîtrise des risques sur l'intervalle de confiance à 95%. Il est néanmoins préférable de ne jamais utiliser cet estimateur d'intervalle. Cette condition de validité peut être utilisée pédagogiquement dans l'objectif de vacciner contre la tentation d'utiliser cet estimateur d'intervalle.

Les procédures de bootstrap non paramétriques sont toutes asymptotiquement correctes mais ne peuvent pas être aussi précises qu'une méthode reposant sur la connaissance de la loi de distribution telle que l'intervalle de Clopper-Pearson mid-P. Les bootstrap basique et percentiles convergent aussi lentement que l'intervalle de Wald, alors que le bootstrap studentisé ou BC_a modifié convergent plus vite, conformément à la théorie asymptotique (Carpenter (26)).

Les procédures exactes n'ont d'exact que le nom. Aucune procédure ne peut parfaitement garantir le risque nominal d'un estimateur d'intervalle de confiance d'une proportion binomiale. Néanmoins, ces inexactitudes sont moins graves qu'on ne pourrait croire, car les proportions réelles ne sont pas constantes d'une expérience à l'autre (il y a toujours de l'hétérogénéité dans les méta-analyses) et la taille de l'échantillon est généralement mal maîtrisée ; ces phénomènes lissent la courbe de risque qui peut alors être très proche du risque nominal. Il nous paraît préférable de chercher à maîtriser ce risque moyen local, lissé par ces phénomènes, plutôt que le risque conditionnel à des hypothèses fausses.

Nous conseillons d'utiliser systématiquement l'intervalle de confiance de Clopper-Pearson mid-P à l'exception du rare scénario de comparaison de proportion observée à théorique défini dans un protocole expérimental très maîtrisé ou fortement régulé ; auquel cas l'intervalle de confiance de Clopper-Pearson est préférable.

L'intervalle de Clopper-Pearson mid-P n'est pas implémenté dans beaucoup de logiciels statistiques, mais nous mettons à disposition des macros en Annexe 2 pour les logiciels les plus souvent utilisés en biostatistique ainsi que pour Excel et LibreOffice. Ces macros sont librement diffusables, sans restriction. L'intervalle de Clopper-Pearson, plus rarement utile, est disponible dans la plupart des logiciels sous le nom « procédure exacte ».

6 Annexe 1 : analyse de 55 estimateurs d'intervalles

6.1 Définitions supplémentaires d'estimateurs d'intervalles

Les estimateurs d'intervalle ont été implémentés à partir de l'algorithme défini dans l'article original avec des correctifs mineurs pour les cas limites, comme suit :

- Lorsqu'une borne basse était inférieure à 0, elle était remplacée par 0
- Lorsqu'une borne haute était supérieure à 1, elle était remplacée par 1
- Lorsque l'estimation ponctuelle n'était pas contenue dans l'intervalle, la borne de l'intervalle la plus proche de l'estimation ponctuelle était déplacée jusqu'à l'estimation ponctuelle.

On peut s'attendre à ce qu'un statisticien vérifiant ses résultats applique spontanément ces corrections.

Certains estimateurs d'intervalles n'étaient pas calculables en-dessous d'un certain nombre de succès ou d'échecs. Dans ce cas l'estimateur de Clopper-Pearson (CP) lui était substitué. Pour x le nombre de succès et n la taille de l'échantillon, on notait $k = \min(x, n - x)$ le plus petit des deux nombres (succès ou échecs). Ainsi, l'intitulé « Boot. Studentisé, CP si k ≤ 4 » désigne un estimateur d'intervalle hybride, égal à l'intervalle du bootstrap studentisé lorsque le nombre de succès et d'échecs sont tous deux supérieurs ou égaux à 5 et égal à l'intervalle de Clopper-Pearson autrement.

Comme pour les estimateurs de l'analyse principale, il y eût double implémentation, par le même auteur, à un mois d'intervalle, la 1ère se basant sur l'article original et la seconde se basant sur la formule déclarée dans le tableau. En cas de discordance des résultats, le problème était investigué.

6.1.1 Intervalles basés sur une approximation normale ou de Student

Les formules de ces intervalles sont décrites sur le Tableau 1. Le t de Student à $n - 1$ degrés de liberté est basé sur la statistique pivot $T = \frac{\widehat{M}}{\sqrt{\widehat{VAR}}}$. Cette statistique repose sur l'approximation normale d'une variable aléatoire et l'indépendance entre les estimateurs de la moyenne et de la variance sur les échantillons issus d'une loi normale. Pan (82) a essayé de calculer les degrés de libertés de la distribution de Student qui coïncide le mieux possible avec la distribution de la statistique pivot :

$$T = \frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}} \quad (1)$$

Son calcul repose sur l'hypothèse que $\hat{P}$ et $V(\hat{P}, n) = \frac{\hat{P}(1-\hat{P})}{n}$ sont approximativement indépendants (page 145). Cette hypothèse est toujours fausse, surtout lorsque p est petit, puisque dans ce cas $\hat{P} \approx \frac{1}{n} V(\hat{P}, n)$, conduisant à un coefficient de corrélation linéaire de Pearson presque égal à 1 entre la moyenne et la variance. Les degrés de liberté calculés avec la formule de Pan conduit à une distribution de Student dont le kurtosis ne coïncide pas avec celui de la distribution binomiale. Le coefficient d'asymétrie (skewness) ne coïncide pas non plus, car il est toujours nul pour la distribution de Student. Par exemple, l'excès de kurtosis d'une distribution binomiale $B(100\ ;\ 0{,}08)$ calculé à partir de la fonc-

tion de masse exacte est 3,08 (très leptokurtique), ce qui coïncide avec une distribution t à 5,95 degrés de liberté. Pour la même distribution binomiale, l'approximation de Pan à une distribution t a un excès de kurtosis égal à 0,348 avec 21,24 degrés de liberté. C'est pourquoi un intervalle respectant le kurtosis « t à kurto égal » a été calculé à partir de la fonction de masse de la loi binomiale $B(n\,;\hat{p})$, à l'exclusion des occurrences de zéro succès ou zéro échecs. Lorsque $\min(x, n-x) \leq 3$, l'excès de kurtosis de la distribution binomiale est négatif de telle sorte qu'aucune loi de Student ne peut l'égaler. C'est pourquoi, lorsque $\min(x, n-x) \leq 3$, l'intervalle de Clopper-Pearson est utilisé.

Pan propose d'appliquer sa procédure à l'intervalle d'Agresti-Coull, aussi connu sous le nom d'intervalle de Wald ajusté. L'intervalle « t add4 selon Pan » est obtenu de la même manière que l'intervalle de Pan, mais 2 succès et 2 échecs (plus rigoureusement $\frac{(z_{1-\alpha})^2}{2}$ succès et autant d'échecs) sont rajoutés avant le calcul. Le kurtosis de l'intervalle « t add4 à kurto égal » est basé sur le calcul du kurtosis de la statistique studentisée obtenue en ajoutant $\frac{(z_{1-\alpha})^2}{2}$ succès et autant d'échecs à une loi binomiale $B(n\,;\hat{p})$. Théoriquement, cette statistique est indéfinie lorsque le nombre de succès est nul. En pratique, lorsque $\min(x, n-x) \geq 5$ on peut raisonnablement ignorer de cas limite qui a environ une chance sur 32 de survenir.

Toutes ces approximations sont douteuses : la statistique pivot est indéfinie pour zéro succès et zéro échecs et égaliser le kurtosis tout en ignorant le coefficient d'asymétrie ne garantit pas d'obtenir la meilleure approximation de Student. La loi de Student, *a priori*, ne semble pas pertinente pour approcher la loi binomiale studentisée.

Les intervalles « t à kurto égal » et « t add4 à kurto égal » définis dans le Tableau 1 Tableau ne sont pas issus de la littérature. Il s'agit de corrections apportées aux intervalles de Pan par l'auteur de cette thèse.

Pour la suite de cette annexe, notons $BPF(x;n,p) = \binom{n}{x}p^x(1-p)^{n-x}$ la fonction de masse de la loi binomiale, $BCDF(k;n,p)$ la fonction de répartition (ou fonction cumulative de probabilités) et $BiCDF(q;n,p)$ la fonction des quantiles de la loi binomiale (ou fonction inverse cumulative de probabilités) :

$$BPF(x;n,p) = \binom{n}{x}p^x(1-p)^{n-x} \tag{2}$$

$$BCDF(k;n,p) = \sum_{x=0}^{k} BPF(x;n,p) \tag{3}$$

$$BiCDF(q;n,p) = \min\{k|BCDF(k;n,p) \geq q\} \tag{4}$$

Notons $\beta iCDF(q;\alpha,\beta)$ le quantile q de la distribution beta dont les paramètres de forme sont α et β.

Nom	Borne basse $L_{1-\alpha}(x,n)$	
(4,21,77,83,106) Wald	$\max\left(0, \frac{x}{n} - \kappa\sqrt{\frac{x(n-x)}{n^3}}\right)$	(5)
(98) t de Student à n-1 ddl	$\max\left(0, \hat{p} - t_{1-\frac{\alpha}{2},n-1}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right)$	(6)
(14,106) Wald avec cc	$\max\left(0, \hat{p} - \kappa\sqrt{\frac{\hat{p}(1-\hat{p})}{n} - \frac{1}{2n}}\right)$	(7)
(22) Wald re-centré[a]	$\min\left(\hat{p}, \max\left(0, \tilde{p} - \kappa\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\right)$	(8)
(83) Wald re-centré avec cc[a]	$\min\left(\hat{p}, \max\left(0, \tilde{p} - \kappa\sqrt{\frac{\hat{p}(1-\hat{p})}{n} - \frac{1}{2n}}\right)\right)$	(9)
(82,83) t selon Pan 2002[b]	$\begin{cases} \max\left(0, \hat{p} - t_{1-\alpha/2,dfp}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) si\ x > 0 \\ Indéfini\ si\ x = 0 \end{cases}$	(10)
t à kurto égal[c]	$\begin{cases} \hat{p} - t_{1-\alpha/2,dfg}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \\ Indéfini\ si\ x \leq 3 \end{cases}$	(11)
(4) Agresti-Coull[a]	$\max\left(0, \tilde{p} - \kappa\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+\kappa^2}}\right)$	(12)
(82) t add4 selon Pan 2002[ad]	$\max\left(0, \tilde{p} - t_{1-\frac{\alpha}{2},df\tilde{p}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+\kappa^2}}\right)$	(13)
t add4 à kurto égal[ae]	$\begin{cases} \tilde{p} - t_{1-\alpha/2,df\tilde{g}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+\kappa^2-1}} \\ Indéfini\ si\ x \leq 5 \end{cases}$	(14)
(110) Wilson	$\frac{x + \frac{\kappa^2}{2} - \kappa\sqrt{\frac{x(n-x)}{n} + \frac{\kappa^2}{4}}}{n+\kappa^2}$	(15)
(21) Wilson 1927 modifié par Brown en 2001[fg]	$\begin{cases} \frac{1}{2n}\chi^2_{\alpha,2x}\ si\ 1 \leq x \leq x^* \\ \frac{x + \frac{\kappa^2}{2} - \kappa\sqrt{\frac{x(n-x)}{n} + \frac{\kappa^2}{4}}}{n+\kappa^2}\ autrement \end{cases}$	(16)
(106) Wilson avec cc	$\begin{cases} \frac{1}{2(n+\kappa^2)}\left(2x + \kappa^2 - 1 - \kappa\sqrt{\kappa^2 - 2 - \frac{1}{n} + \frac{4x}{n}(n-x+1)}\right) si\ x > 0 \\ si\ x = 0 \end{cases}$	(17)
(17) SAIFS de Borkowf 2006	$\max\left(0, p' - \kappa\sqrt{\frac{p'(1-p')}{n}}\right)$ où $p' = \frac{x}{n+1}$	(18)

Tableau 1 : définition des bornes basses des estimateurs d'intervalles de confiance basés sur une approximation à une loi normale ou à une loi de Student. Les bornes hautes sont définies par équivariance $U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)$.

[a]Le centre de l'intervalle de Wilson est défini comme :

$$\tilde{p} = \frac{x + \frac{1}{2}\kappa^2}{n + \kappa^2} \tag{19}$$

[b] Les degrés de liberté dfp sont calculés selon la méthode de Pan (82). Il y avait des erreurs dans les exposants de $\hat{p}$ dans l'article de Pan ; elles ont été corrigées dans l'article de Pires 2008 (83))

$$dfp = \frac{2\left(\frac{\hat{p}(1-\hat{p})}{n}\right)^2}{\Omega(\hat{p}, n)} \tag{20}$$

$$\Omega(p,n) = \frac{\hat{p} - \hat{p}^2}{n^3} - 2\frac{\hat{p} + (2n-3)\hat{p}^2 - 2(n-1)\hat{p}^3}{n^4} + \frac{\hat{p} + (6n-7)\hat{p}^2 + 4(n-1)(n-3)\hat{p}^3 - 2(n-1)(2n-3)\hat{p}^4}{n^5} \tag{21}$$

[c]Les degrés de liberté dfg sont égaux au nombre de degrés de liberté d'une distribution de Student dont le kurtosis est égal à $kur(x,n)$:

$$dfg = 4\frac{6}{kur(x,n) - 3} \tag{22}$$

où $kur(x,n)$ est le kurtosis de la distribution binomiale studentisée $B\left(n, \frac{x}{n}\right)$ en excluant les cas de zéro succès ou zéro échecs (BPF est défini par l'équation (2) en page 86) :

$$kur(x,n) = \frac{\sum_{i=1}^{n-1} BPF(i;n,x)(y_i - \overline{y})^4}{\left(\sum_{i=1}^{n-1} BPF(i;n,x)(y_i - \overline{y})^2\right)^2} \tag{23}$$

où y_i est la proportion studentisée pour i succès d'une distribution binomiale $B\left(n, \frac{x}{n}\right)$:

$$y_i = \frac{\frac{i}{n} - \frac{x}{n}}{\sqrt{\frac{i(n-i)}{n^3}}} \tag{24}$$

[d] $df\tilde{p}$ est calculé de la même manière que dfp mais $\tilde{p}$ remplace $\hat{p}$ et $\tilde{n} = n + \kappa^2$ remplace n.

[e] $df\tilde{g}$ est calculé de la même manière que dfg mais y_i est remplacé par $\tilde{y_i}$:

$$\tilde{y_i} = \frac{\frac{i - x}{n + 2c}}{\sqrt{\frac{(i+c)(n-i+c)}{(n+2c)^3}}} \tag{25}$$

$$\text{où } c = \frac{\kappa^2}{2} \tag{26}$$

[f] Le seuil x^* vaut 2 pour $n \leq 50$ et 3 pour $n > 50$

[g] $\chi^2_{q,df}$ est le quantile q de la distribution du χ^2 à df degrés de liberté

6.1.2 Intervalles basés sur une approximation normale après transformation

Nom	Borne basse $L_{1-\alpha}(x,n)$	
(10) Arc-sinus	$\begin{cases}\sin^2\left(\text{asin}\left(\sqrt{\hat{p}}\right)-\frac{\kappa}{2\sqrt{n}}\right) \; si\; x>0\\ 0\; si\; x=0\end{cases}$	(27)
(8,10) Arc-sinus de Bartlett 1936	$\sin^2\left(\max\left(0,\text{asin}\left(\sqrt{\frac{x+\frac{1}{2}}{n+1}}\right)-\frac{\kappa}{2\sqrt{n+\frac{1}{2}}}\right)\right)$	(28)
(9,73) Arc-sinus d'Anscombe	$\begin{cases}\sin^2\left(\text{asin}\left(\sqrt{\frac{x+\frac{3}{8}}{n+3/4}}\right)-\frac{\kappa}{2\sqrt{n+\frac{1}{2}}}\right) \; si\; x>0\\ 0\; si\; x=0\end{cases}$	(29)
(83) Arc-sinus d'Anscombe avec cc	$\sin^2\left(\text{asin}\left(\sqrt{\frac{x-\frac{1}{8}}{n+3/4}}\right)-\frac{\kappa}{2\sqrt{n+\frac{1}{2}}}\right)$	(30)
(46) Arc-Sinus de Freeman-Tukey	$\begin{cases}\sin^2\left(\frac{1}{2}\left(\text{asin}\left(\sqrt{\frac{x}{n+1}}\right)+\text{asin}\left(\sqrt{\frac{x+1}{n+1}}\right)-\frac{\kappa}{\sqrt{n+\frac{1}{2}}}\right)\right) \; si\; x>0\\ 0\; si\; x=0\end{cases}$	(31)
(21) Logit-normal modifié[a]	$\begin{cases}\text{logitinv}\left(\log\left(\frac{x}{n-x}\right)-\kappa\sqrt{\frac{n}{x(n-x)}}\right) \; si\; 0<x<n\\ \sqrt[n]{\alpha/2}\; si\; x=n\\ 0\; si\; x=0\end{cases}$	(32)
(9) Logit d'Anscombe[b]	$\max\left(0,lainv\left(la\left(\frac{x}{n},n\right)-\kappa\sqrt{\frac{(n+1)\,(n+2)}{n(x+1)(n-x+1)}},n\right)\right)$	(33)
(88) Logit+0,5 de Rubin[abc]	$\begin{cases}\text{logitinv}\left(\text{la}\left(\frac{x}{n},n\right)-\kappa\sqrt{\frac{n+1}{\left(x+\frac{1}{2}\right)\left(n-x+\frac{1}{2}\right)}}\right) \; si\; x>0\\ 0\; si\; x=0\end{cases}$	(34)

Tableau 2 : définition des bornes basses des estimateurs d'intervalles de confiance basés sur une approximation normale après l'application d'une transformation stabilisatrice de variance. Les bornes hautes sont définies par équivariance $U_{1-\alpha}(x,n)=1-L_{1-\alpha}(n-x,n)$.

[a]La réciproque de la transformation logistique est définie par :

$$\text{logitinv}(t)=\frac{\exp(t)}{1+\exp(t)} \tag{35}$$

[b]La transformation logistique d'Anscombe et sa réciproque sont définies comme suit :

$$\text{la}(\hat{p},n)=\log\left(\frac{n\hat{p}+\frac{1}{2}}{n(1-\hat{p})+\frac{1}{2}}\right) \tag{36}$$

$$\text{lainv}(t,n)=\frac{\left(n+\frac{1}{2}\right)\exp(t)-\frac{1}{2}}{n(1+\exp(t))} \tag{37}$$

[c]Cet intervalle est équivalent à ajouter un demi-succès et un demi-échec avant de calculer l'intervalle logit-normal.

6.1.3 Intervalles bayésiens

Nom	**Borne basse $L_{1-\alpha}(x,n)$**	
(21) Jeffreys équilibré[a]	$\begin{cases} \beta iCDF\left(\frac{\alpha}{2}; x+\frac{1}{2}, n-x+\frac{1}{2}\right) \; si \; x > 0 \\ 0 \; si \; x = 0 \end{cases}$	(38)
(21) Jeffreys équilibré modifié par Brown[a]	$\begin{cases} \beta iCDF(\alpha/2; x+1/2, n-x+1/2) \; si \; 2 \le x < n \\ \sqrt[n]{\alpha/2} \; si \; x = n \\ 0 \; si \; x \le 1 \end{cases}$	(39)
(83) Prior uniforme, équilibré[a]	$\beta iCDF\left(\frac{\alpha}{2}; x+1, n-x+1\right)$	(40)
Jeffreys à densité maximale[b]	$\begin{cases} \beta QJ(\inf\{\beta QJ(1-\alpha+r) - \beta QJ(r) \mid r \in [0,\alpha]\}) \; si \; x > 0 \\ 0 \; si \; x = 0 \end{cases}$	(41)

Tableau 3 : définition des bornes basses des estimateurs d'intervalles de confiance basés sur la théorie bayésienne avec *prior* non informatif. Les bornes hautes sont définies par équivariance $U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)$.

[a] $\beta iCDF(q;\alpha,\beta)$ est le quantile q de la distribution beta dont les paramètres de forme sont α et β.

[b] $\beta QJ(q) = \beta iCDF\left(q; x+\frac{1}{2}, n-x+\frac{1}{2}\right)$ est le quantile q de la distribution *a posteriori* pour une *prior* de Jeffreys.

6.1.4 Intervalles par bootstrap

Les intervalles de bootstrap sont calculés à partir de la distribution binomiale exacte de telle sorte qu'ils soient équivalents aux intervalles de bootstrap non paramétriques avec un nombre d'échantillons de bootstrap (ré-échantillonnages) infini.

La distribution de bootstrap étant la distribution binomiale $B\left(n, \frac{x}{n}\right)$ les intervalles de bootstrap percentile et basique peuvent être calculés à partir de la fonction de répartition de la loi binomiale. La studentisation étant, pour la loi binomiale, une transformation monotone, il est possible de calculer les quantiles de la statistique pivot (statistique studentisée) à partir de la fonction de répartition de la loi binomiale. Les proportions studentisées n'étant pas définies pour zéro succès et pour zéro échecs, l'intervalle studentisé n'est pas calculable lorsque $\max\left(BPF\left(0; n, \frac{x}{n}\right), BPF\left(n; n, \frac{x}{n}\right)\right) > \frac{\alpha}{2}$ où BPF est la fonction de masse de la loi binomiale. Pour les intervalles de confiance à 95%, la condition est $\min(x, n-x) \geq 5$.

L'intervalle de bootstrap à biais corrigé accéléré (BC_a) tel que défini par la méthode d'Efron (41) n'est pas équivariant. Ceci est dû à la définition de la fonction de répartition empirique $\hat{G}(s) = Pr_{\hat{\theta}}(\hat{\theta}^* < s)$ définie par l'équation 3.2 dans l'article d'Efron. Cette fonction $\hat{G}(s)$ est basée sur une inégalité stricte $\hat{\theta}^* < s$ équivalente à l'inégalité large $\hat{\theta}^* \leq s$ pour une distribution de bootstrap continue. La distribution de bootstrap étant discrète pour la loi binomiale, on peut lui substituer $\hat{G}(s) = Pr_{\hat{\theta}}(\hat{\theta}^* < s) + \frac{1}{2} Pr_{\hat{\theta}}(\hat{\theta}^* = s)$. Cette modification garantit la propriété d'équivariance $L_{1-\alpha}(x, n) = 1 - U_{1-\alpha}(n-x, n)$.

Comme pour les intervalles de bootstrap percentile, basique et studentisé les intervalles BC_a ont été calculés à partir de la distribution binomiale exacte, afin d'obtenir la distribution $\hat{\theta}^*$ exacte comme si un nombre infini de ré-échantillonnages avait été effectué sur ordinateur.

On peut noter que les bornes des intervalles basiques, percentile, BC_a et BC_a modifié (éq 3.2), sont toujours des multiples entiers de $\frac{1}{n}$.

L'intervalle de bootstrap normal calculé à partir de la distribution binomiale exacte comme si un nombre infini de ré-échantillonnages avaient été effectués, est égal à l'intervalle de Wald, et, par conséquent, n'est pas présenté séparément.

Les calculs des intervalles de bootstrap binomiaux exacts ont été comparés aux calculs du paquet 'boot' du logiciel R approchés sur un nombre de ré-échantillonnages fini. Les résultats étaient identiques sur tous les chiffres décimaux significatifs.

Nom	**Borne basse $L_{1-\alpha}(x,n)$**		**Borne haute $U_{1-\alpha}(x,n)$**	
(26) Boot. percentile[a]	$\dfrac{BiCDF\left(\frac{\alpha}{2};n,\frac{x}{n}\right)}{n}$	(42)	$\dfrac{BiCDF\left(1-\frac{\alpha}{2};n,\frac{x}{n}\right)}{n}$	(43)
(26) Boot. basique[a]	$\max\left(0,\dfrac{2x-BiCDF\left(1-\frac{\alpha}{2};n,\frac{x}{n}\right)}{n}\right)$	(44)	$\min\left(1,\dfrac{2x-BiCDF\left(\frac{\alpha}{2};n,\frac{x}{n}\right)}{n}\right)$	(45)
(26) Boot. studentisé[a]	$\hat{p}-(l-\hat{p})\sqrt{\frac{\hat{p}(1-\hat{p})}{l(1-l)}}$ avec $l=\frac{1}{n}BiCDF(1-\frac{\alpha}{2})$ $Indéfini\ pour\ \min(x,n-x)\leq 4\ si\ \alpha=0{,}05$	(46)	$\hat{p}-(u-\hat{p})\sqrt{\frac{\hat{p}(1-\hat{p})}{u(1-u)}}$ avec $u=\frac{1}{n}BiCDF(\frac{\alpha}{2})$ $Indéfini\ pour\ \min(x,n-x)\leq 4\ si\ \alpha=0{,}05$	(47)
(41) Boot. BCa [ab]	$EF\left(\frac{\alpha}{2},n,x,0\right)$ $Indéfini\ si\ x=0\ ou\ x=n$	(48)	$EF\left(1-\frac{\alpha}{2},n,x,0\right)$ $Indéfini\ si\ x=0\ ou\ x=n$	(49)
Boot. BCa modif [ab]	$EF\left(\frac{\alpha}{2},n,x,\frac{1}{2}\right)$ $Indéfini\ si\ x=0\ ou\ x=n$	(50)	$EF\left(1-\frac{\alpha}{2},n,x,\frac{1}{2}\right)$ $Indéfini\ si\ x=0\ ou\ x=n$	(51)
Boot. BCa lissé	Bootstrap BC_a non paramétrique avec bruit gaussien $\mathcal{N}\left(0,\left(\frac{1}{2}\right)^2\right)$ aléatoire ajouté à chaque observation. La constante d'accélération est calculée sur la distribution sans bruit. $Indéfini\ si\ x=0\ ou\ x=n$			

Tableau 4 : définition des bornes des estimateurs d'intervalle par bootstrap

[a] Les fonctions BPF, $BCDF$ et $BiCDF$ sont définies par les équations (2), (3) et (4) en page 86.

[b] Nous définissons les fonctions suivantes :

$$EF(q,n,x,d)=\frac{1}{n}BiCDF\left(NCDF\left(bias(x,n,d)+\frac{bias(x,n,d)+NiCDF(q)}{1-acc(x,n)\times\left(bias(x,n,d)+NiCDF(q)\right)}\right);n,\frac{x}{n}\right) \quad (52)$$

$$bias(x,n,d)=NiCDF\left(BCDF\left(x-1;n,\frac{x}{n}\right)+\frac{d}{2}BPF\left(x;n,\frac{x}{n}\right)\right) \quad (53)$$

$$acc(x,n)=\frac{n-2x}{6\sqrt{nx(n-x)}} \quad (54)$$

où $EF(q,n,x,d)$ est basée sur les équations 3.8 et 3.9 d'Efron 1987 (41) page 183, $bias(x,n,d)$ sur l'équation 4.1 page 174 et $acc(x,n)$ sur l'équation 4.4 page 174, $NCDF(z)$ est la fonction de répartition de la loi normale centré réduite $\mathcal{N}(0,1)$ et $NiCDF(q)$ est la fonction des quantiles de la loi normale centrée réduite $\mathcal{N}(0,1)$.

6.1.5 Intervalles binomiaux exacts

Ces intervalles sont basés sur la distribution binomiale exacte. Comme tous les intervalles déterministes, ils ont des oscillations de risque et ne sont donc pas vraiment exacts.

Les intervalles de Sterne et de Blaker sont basés sur des régions de confiance non connexes (réunion de plusieurs intervalles disjoints). Ces estimateurs sont définis par le plus petit intervalle contenant complètement la région de confiance (enveloppe convexe). L'intervalle de Clopper-Pearson peut être défini comme

$$\left\{q|\min\left(BCDF(x;n,q),1-BCDF(x-1;n,q)\right)\leq\frac{\alpha}{2}\right\} \tag{55}$$

cette expression étant aussi égale à $\beta iCDF\left(\frac{\alpha}{2};x,n-x+1\right)$. L'intervalle décrit par Blaker (12) en 2000 avait été précédemment proposé par Cox et Hinkley en 1974 (34) selon Lecoutre et Poitevineau (66), fait non vérifié par l'auteur de cette thèse. C'est un intervalle bilatéral déséquilibré strictement conservatif qui contient toujours l'intervalle de Clopper-Pearson contrairement à l'intervalle de Sterne qui est parfois plus large. L'intervalle de Sterne, comme l'intervalle de Blaker est emboîté : l'intervalle de confiance à 90% est toujours inclus dans l'intervalle de confiance à 95%.

L'intervalle de Schilling-Doi (93) est un intervalle bilatéral déséquilibré strictement conservatif basé sur les régions d'acceptation de Sterne. Les intervalles d'acceptation peuvent être interprétés comme des intervalles de fluctuations. Pour une proportion p théorique, une région d'acceptation AR est un intervalle de valeurs x tel que $P(X\in AR)\geq 1-\alpha$ où $X\sim B(n\,;\,p)$. Les régions d'acceptation de Sterne optimisent les critères de suivant, par ordre de priorité : largeur minimale, probabilité de couverture maximale et borne haute la plus haute possible. Schilling et Doi ont modifié les régions d'acceptation de Sterne lorsqu'une borne de région d'acceptation (basse ou haute) n'est pas monotone selon p. L'algorithme de Schilling-Doi a un temps de calcul et une utilisation de mémoire d'ordinateur exponentielle avec le nombre de chiffres décimaux demandés. Lorsqu'on souhaite une certaine précision décimale sur un nombre attendu de succès $\lambda=np$, la précision requise sur la proportion p est proportionnelle à la taille d'échantillon. Ceci rend l'algorithme de Schilling-Doi inapplicable sur des échantillons de grande taille, contenant quelques milliers d'observations. En se basant sur le fait que les bornes des régions d'acceptation de Sterne, analysés comme fonction de p, ne peuvent changer que lorsque des courbes de vraisemblance (une différente étant définis pour chaque nombre de succès x possible) croisent le niveau de confiance nominal ou se croisent l'une l'autre, un algorithme amélioré basé sur la recherche de ces points chauds a été écrit. Sur les petites valeurs de n, cet algorithme a été numériquement comparé à l'algorithme original de Schilling-Doi.

L'intervalle de Wang (109) est basé sur une compression itérative de chacun des intervalles de Clopper-Pearson tout en conservant la couverture strictement conservative du jeu d'intervalles à tout moment. Comme les intervalles de Sterne, Schilling-Doi et Blaker, c'est un intervalle bilatéral déséquilibré strictement conservatif. L'algorithme de Wang est très lent. Il déplacé les bornes des intervalles par petits pas, d'une taille inversement proportionnelle à la précision décimale demandée. À chaque pas, il recalcule la couverture réelle pour n valeurs théoriques critiques, alors qu'un seul intervalle de confiance a été changé. L'algorithme a été amélioré pour ne mettre à jour que les valeurs de couverture réelle susceptibles de changer entre deux pas. Les résultats numériques sont identiques. Les intervalles de Schilling-Doi comme celui de Wang, pour une taille d'échantillon donnée, cherchent à réduire la moyenne arithmétique des largeurs d'intervalles, mais pas la moyenne géométrique.

L'intervalle de Blyth-Still-Casella a été défini par Blyth et Still avec un algorithme basé sur une précision décimale fixe (15).Casella a amélioré l'algorithme afin d'obtenir une précision décimale dynamique (27). L'implémentation de Winstein écrite dans le langage de programmation C++ a été utilisée (111).

L'intervalle de Zieliński (115) est basé sur des changements de la formule de Clopper-Pearson, à laquelle un biais γ est rajouté dans la distribution des quantiles de la fonction beta, afin de créer un déséquilibre assumé des risques à droite et à gauche. Cet intervalle n'est pas strictement conservatif. Ceci peut être dû à la dépendance du biais γ au nombre de succès x. L'intervalle serait strictement conservatif si le biais γ était constant selon la variable aléatoire X.

L'intervalle de confiance du rapport de vraisemblance exact (Sakakibara (89)) est défini par inversion d'un test du rapport de vraisemblance exact. La région de confiance n'est pas toujours connexe car la fonction de P-valeur n'est pas bimonotone selon p. L'intervalle de confiance est définie par l'enveloppe convexe de la région de confiance, comme défini par Thulin (100).

L'intervalle de confiance exact du score, défini par Sakakibara ou Thulin (89,100), est équivalent à l'intervalle exact alternatif défini par Cai et Krishnamoorthy (25) dans leur équation 2. En effet

$$Pr\left(\frac{(X-np)^2}{np(1-p)} \geq \frac{(k-np)^2}{np(1-p)}\middle| p\right) = Pr((X-np)^2 \geq (k-np)^2|p) \quad (56)$$

où $X \sim B(n,p)$. La région de confiance n'est pas toujours connexe ; l'intervalle de confiance est défini comme son enveloppe convexe.

L'intervalle de Cai-Krishnamoorthy (25) combiné est basé sur une inversion d'un test combiné, strictement plus libéral que le test de Clopper-Pearson ou le test exact du score. La région de confiance est l'enveloppe convexe de la région de confiance. La procédure n'est pas strictement conservative, à cause du biais d'inférence statistique consistant à systématiquement choisir la P-valeur la plus petite entre deux tests qui sont chacun des deux corrects lorsqu'ils sont pris individuellement.

L'intervalle de Pratt a été décrit par Blyth (14) sous le nom d'approximation de Paulson-Camp-Pratt (équation D) ; c'est une approximation de l'intervalle de Clopper-Pearson. L'intervalle *Pratt moyen* a été proposé par Vollset (106) en 1993 comme une approximation de l'intervalle de Clopper-Pearson mid-P. Il est basé sur la moyenne de deux intervalles de Pratt, pour x et pour $x+1$. Il n'est donc pas vraiment basé sur la loi binomiale exacte mais il approxime un intervalle construit comme tel. Des ajustements minimes à la formule de Vollset ont été appliqués afin d'éviter les comportements aberrants lorsque $x = 0$ ou $x = n$. Plutôt que d'évaluer l'intervalle de Pratt en x et $x+1$, il est évalué en x et $\min(x+1, n)$. De plus, lorsque $x = 0$, la borne basse est remplacée par 0 et réciproquement, la borne haute est remplacée par 1 lorsque $x = n$.

Définissons l'*odds ratio absolu* entre deux valeurs a et b comprises toutes deux entre 0 et 1 comme $\text{logitinv}(|\text{logit}(a) - \text{logit}(b)|)$ où logitinv est la réciproque de la fonction logit (cf équation (35) page 89). C'est-à-dire, si l'odds ratio OR entre a et b est supérieur à 1, l'odds ratio absolu est égal à OR, sinon il est égal à son inverse $\frac{1}{OR}$. L'approximation du Clopper-Pearson mid-P par l'intervalle de Pratt moyen est assez bonne puisque pour $n = 32$ et $x = 0, \ldots, 32$, l'odds ratio absolu moyen entre la borne basse de l'intervalle de confiance de Clopper-Pearson mid-P et celle du Pratt moyen était égale à 1,08. Néanmoins, l'approximation était mauvaise pour $x = 1$ (Odds ratio absolu = 2,41) et médiocre pour $x = 2$ (Odds ratio absolu = 1,25) et pour $x = 32$ (Odds ratio absolu = 1,24). Deux modifications amé-

lioraient ces approximations. La 1ère modification était l'usage de la moyenne arithmétique entre les bornes de Pratt pour x et $\min(x+1, n)$, c'est-à-dire $\text{logitinv}\left(\frac{\text{logit}\left(LPratt_{1-\alpha}(x,n)\right)+\text{logit}\left(LPratt_{1-\alpha}(\min(n,x+1),n)\right)}{2}\right)$. La 2ème modification était l'usage de la borne $LPratt_{1-2\alpha}(x,n)$ lorsque $x = n$ plutôt que $LPratt_{1-\alpha}(x,n)$. Cet intervalle de *Pratt moyen modifié*, pour $n = 32$, présentait un odds ratio absolu moyen de borne basse égal à 1,027 par rapport à l'intervalle de Clopper-Pearson mid-P avec un odds ratio absolu maximal égal à 1,11. Cet intervalle Pratt moyen modifié n'est pas présenté dans les résultats mais son comportement serait très proche de celui de Clopper-Pearson mid-P.

Les intervalles bilatéraux à risques déséquilibrés basés sur la loi binomiale exacte ont été critiqués par Vos et Hudson (107) puis par Thulin et Zwanzig (100).

Nom	**Borne basse $L_{1-\alpha}(x,n)$**	
(96) Sterne 1954[ab]	$inf\left\{q \middle\| \frac{\alpha}{2} \geq \sum_{i=0}^{n} BPF(i;n,q)1_{BPF(i;n,q)\leq BPF(x;n,q)}\right\}$	(57)
(12) Blaker 2000[ac]	$\inf\{q \mid \text{bpval}(q,x,n) \leq \alpha\}$	(58)
(21,31) Clopper-Pearson[d]	$\beta iCDF\left(\frac{\alpha}{2}; x, n-x+1\right)$	(59)
(11,63) Clopper-Pearson mid-P[ae]	$\inf\left\{q \middle\| \text{cpval}(x,n,q) - \frac{1}{2}BPF(x;n,q) \leq \frac{\alpha}{2}\right\}$	(60)
(106) Pratt moyen (Vollset 1993)[f]	$\begin{cases}\frac{LPratt_{1-\alpha}(x,n) + LPratt_{1-\alpha}(\min(n,x+1),n)}{2} \; si \; x > 0 \\ 0 \; si \; x = 0\end{cases}$	(61)
(93) Schilling-Doi	Inversion des régions d'acceptation de Sterne modifiées afin d'obtenir des bornes d'acceptation monotones selon la proportion théorique p	
(109) Wang 2014	Comprimer les intervalles de confiance de Clopper-Pearson séquentiellement de $x = \frac{n}{2}$ à $x = 0$ tout en gardant le conservatisme bilatéral strict tout en utilisant l'équivariance pour compléter la définition pour $x = \frac{n}{2} + 1$ à $x = n$.	
(15,27,111) Blyth-Still-Casella	Basé sur un ensemble de régions d'acceptation de X les plus étroites possibles pour chaque proportion théorique	
(115) Zieliński 2009[g]	$\begin{cases}\beta QCL(\inf\{\beta QCU(1-\alpha+\gamma) - \beta QCL(\gamma) \mid \gamma \in [0,\alpha]\}) \; si \; x > 0 \\ 0 \; si \; x = 0\end{cases}$	(62)
(89) Rapp. vrai-semblance exact[h]	$\inf\{q \mid \text{lrpval}(x,n,q) \leq \alpha\}$	(63)
(62,89) Score exact[i]	$\inf\{q \mid \text{spval}(x,n,q) \leq \alpha\}$	(64)
(25) Cai-Krishnamoorthy combiné 2005[ie]	$\inf\left\{q \middle\| \min\left(\text{cpval}(x,n,q), \text{spval}(x,n,q)\right) \leq \alpha\right\}$	(65)

Tableau 5 : définitions des bornes basses des estimateurs d'intervalles de confiance basés sur la loi binomiale exacte. Les bornes hautes sont définies par équivariance $U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)$.

[a] Les fonctions BPF, $BCDF$ et $BiCDF$ sont définies par les équations (2), (3) et (4) en page 86.

[b] La notation $1_{prédicat}$ représente la fonction indicatrice du sous-ensemble vérifiant le prédicat. En d'autres termes, $1_{prédicat}$ vaut 1 quand le prédicat est vérifié et 0 autrement.

[c] La fonction bpval (P-valeur de Blaker) est définie comme suit :

$$\text{bpval}(q,x,n) = \begin{cases}\min\left(1, BCDF(x;n,q) + 1 - BCDF(BiCDF(1 - BCDF(x;n,q);n,q);n,q)\right) \; si \; q \geq x/n \\ \text{bpval}(1-q, n-x, n) \; si \; q < x/n\end{cases} \quad (66)$$

[d] $\beta iCDF(q;\alpha,\beta)$ est le quantile q de la distribution beta avec des paramètres de forme α et β.

[e] La fonction de P-valeur de Clopper-Pearson est définie comme suit :

$$\text{cpval}(x,n,p) = \min\left(1, 2 \times \min\left(\frac{1}{2}, BCDF(x;n,p), 1 - BCDF(x-1,n,p)\right)\right) \tag{67}$$

[f] La borne inférieure de l'intervalle de Pratt $LPratt_{1-\alpha}$ est définie par :

$$LPratt_{1-\alpha}(x,n) =$$

$$1 - \left(1 + \left(\frac{n-x+1}{x}\right)^2 \left(\frac{81x(n-x+1) - 9n - 8 - 3\kappa\sqrt{9x(n-x+1)(9n+5-\kappa^2)+n+1}}{81(n-x+1)^2 - 9(n-x+1)(2+\kappa^2)+1}\right)^3\right)^{-1} \tag{68}$$

[g] $\beta QCL(q) = \beta iCDF(q;x,n-x+1)$ et $\beta QCU(q) = \beta iCDF(q;x+1,n-x)$

[h] La fonction du rapport de vraisemblance et la P-valeur du rapport de vraisemblance exact sont définis par :

$$LIKR(x,n,p) = BPF(x,n,x/n)/BPF(p,n,p) \tag{69}$$

$$\text{lrpval}(x,n,p) = \sum_{k \in \{i | LIKR(i,n,p) \geq LIKR(x,n,p)\}} \text{BPF}(k;n,p) \tag{70}$$

[i] La P-valeur du test exact du score est définie comme suit :

$$\text{spval}(x,n,p) = \sum_{k \in \{i | (i-np)^2 \geq (x-np)^2\}} \text{BPF}(k;n,p) \tag{71}$$

6.1.6 Intervalles par approximation normale avec correction de l'asymétrie

Ces intervalles sont basés sur des approximations normales avec une correction du biais dû à l'asymétrie, dans l'objectif de construire des intervalles bilatéraux équilibrés.

L'intervalle de Kott-Liu (68) perfectionne la variance de l'intervalle de Hall mais ne change pas son centre.

L'intervalle de Cai est similaire à l'intervalle de Hall mais conserve des termes de plus petit ordre asymptotique, c'est-à-dire, des termes qui décroissent plus vite et deviennent plus rapidement négligeables.

Nom	**Borne basse $L_{1-\alpha}(x,n)$**	
(53) Hall 1982[a]	$\begin{cases} \max\left(0, \hat{p} + \delta - \kappa\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) if\ x > 0 \\ 0\ if\ x = 0 \end{cases}$	(72)
(24) Cai 2005[b]	$\begin{cases} \max\left(0, \breve{p} - \frac{\kappa}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p}) + \frac{\gamma_1\hat{p}(1-\hat{p}) + \gamma_2}{n}}\right) if\ x > 0 \\ 0\ if\ x = 0 \end{cases}$	(73)
(68) Kott-Liu[a]	$\max\left(0, \hat{p} + \delta - \sqrt{\kappa^2\frac{\hat{p}(1-\hat{p})}{n-1} + \delta^2}\right)$	(74)

Tableau 6 : définition des bornes basses des intervalles par approximation normale avec correction de l'asymétrie. Les bornes hautes sont définies par équivariance $U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)$.

[a] Définissons

$$\delta = \frac{(1+2\kappa^2)(1-2\hat{p})}{6n} \tag{75}$$

[b] Définissons

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{76}$$

$$\breve{p} = \frac{n\hat{p} + \eta}{n + 2\eta} \tag{77}$$

$$\eta = \frac{\kappa^2}{3} + \frac{1}{6} \tag{78}$$

$$\gamma_1 = -\frac{13}{18}\kappa^2 - \frac{17}{18} \tag{79}$$

$$\gamma_2 = \frac{1}{18}\kappa^2 + \frac{7}{36} \tag{80}$$

6.1.7 Intervalles basés sur des modèles linéaires généralisés

Les régressions logistiques, de Poisson et log-binomiales sont des modèles linéaires généralisés (GLM). Ils peuvent être multivariés, bivariés ou univariés. Deux méthodes de calcul des intervalles de confiance des coefficients sont largement répandues. La première est la méthode de Wald, $B \pm z_{1-\alpha/2} \times SE$ où B est l'estimation ponctuelle du coefficient brut du modèle, tel que $\log(odds\ ratio)$

pour une régression logistique et SE est l'erreur-type, obtenue à partir de la matrice de variance-covariance du modèle. Cette méthode est basée sur l'approximation normale du coefficient de régression. La seconde méthode est l'inversion d'un test du χ^2 du rapport de vraisemblance. L'inversion d'un test de rapport de vraisemblance peut être effectué par l'algorithme de profilage de vraisemblance 'profile likelihood' (104). La méthode de Wald est plus rapide à calculer car elle ne nécessite qu'une estimation ponctuelle du modèle alors que la méthode de profilage de vraisemblance nécessite quelques estimations (algorithme itératif).

Les logiciels R et SAS utilisent les intervalles de confiance par profilage de vraisemblance par défaut pour les modèles linéaires généralisés. Le logiciel R a été utilisé pour calculer les intervalles de confiance d'une proportion par des modèles linéaires généralisés à intercept seul (sans covariable).

Les estimateurs d'intervalles présentés dans le Tableau 7 ont été analysés. Comme les intervalles n'étaient pas calculables pour un nombre de succès nul, l'intervalle de Clopper-Pearson était utilisé dans ce cas. Lorsque la borne haute de l'intervalle de confiance dépassait 100%, pour la régression de Poisson ou la régression log-binomiale, elle était remplacée par 100%. Les intervalles log-binomiaux et de Poisson ne sont pas équivariants.

Nom	**Description**	**Commande R correspondant**
RV GLM logit	Rapport de vraisemblance d'une régression logistique	`confint(glm(family=binomial, cbind(x,n-x)~1))`
RV GLM Poisson	Rapport de vraisemblance d'une régression de Poisson	`confint(glm(family=poisson, x~offset(log(n))))`
Wald GLM logit	Intervalle de Wald d'une régression logistique	`confint.default( glm(family=binomial, cbind(x,n-x)~1))`
Wald GLM log binom	Intervalle de Wald d'une régression log-binomiale	`confint.default( glm(family=binomial(log), cbind(x,n-x)~1))`
Wald GLM Poisson	Intervalle de Wald d'une régression de Poisson	`confint.default( glm(family=poisson, x~offset(log(n))))`

Tableau 7 : définition des estimateurs d'intervalles de confiance basés sur les modèles linéaires généralisés univariés estimés sous le logiciel R.

L'intervalle du rapport de vraisemblance d'une régression logistique « RV GLM logit, CP si k=0 » fournissait des résultats identiques au « rapport de vraisemblance modifié » qui avait été défini dans l'analyse principale par la formule suivante :

$$L_{1-\alpha} = \begin{cases} \inf\left\{q \middle| \log\left(\left(\frac{x}{nq}\right)^x \left(\frac{n-x}{n(1-q)}\right)^{n-x}\right) \leq \frac{1}{2}\kappa^2\right\} \; si \; x < n \\ \sqrt[n]{\alpha/2} \; si \; x = n \; (i.e. \; Clopper - Pearson) \end{cases} \quad (81)$$

De plus, l'intervalle du rapport de vraisemblance non modifié est aussi présenté.

$$L_{1-\alpha} = \inf\left\{q \middle| \log\left(\left(\frac{x}{nq}\right)^x \left(\frac{n-x}{n(1-q)}\right)^{n-x}\right) \leq \frac{1}{2}\kappa^2\right\} \; si \; x < n \quad (82)$$

Ce dernier est correctement défini, même pour $x = 0$ ou $x = n$.

6.2 Résultats des intervalles supplémentaires

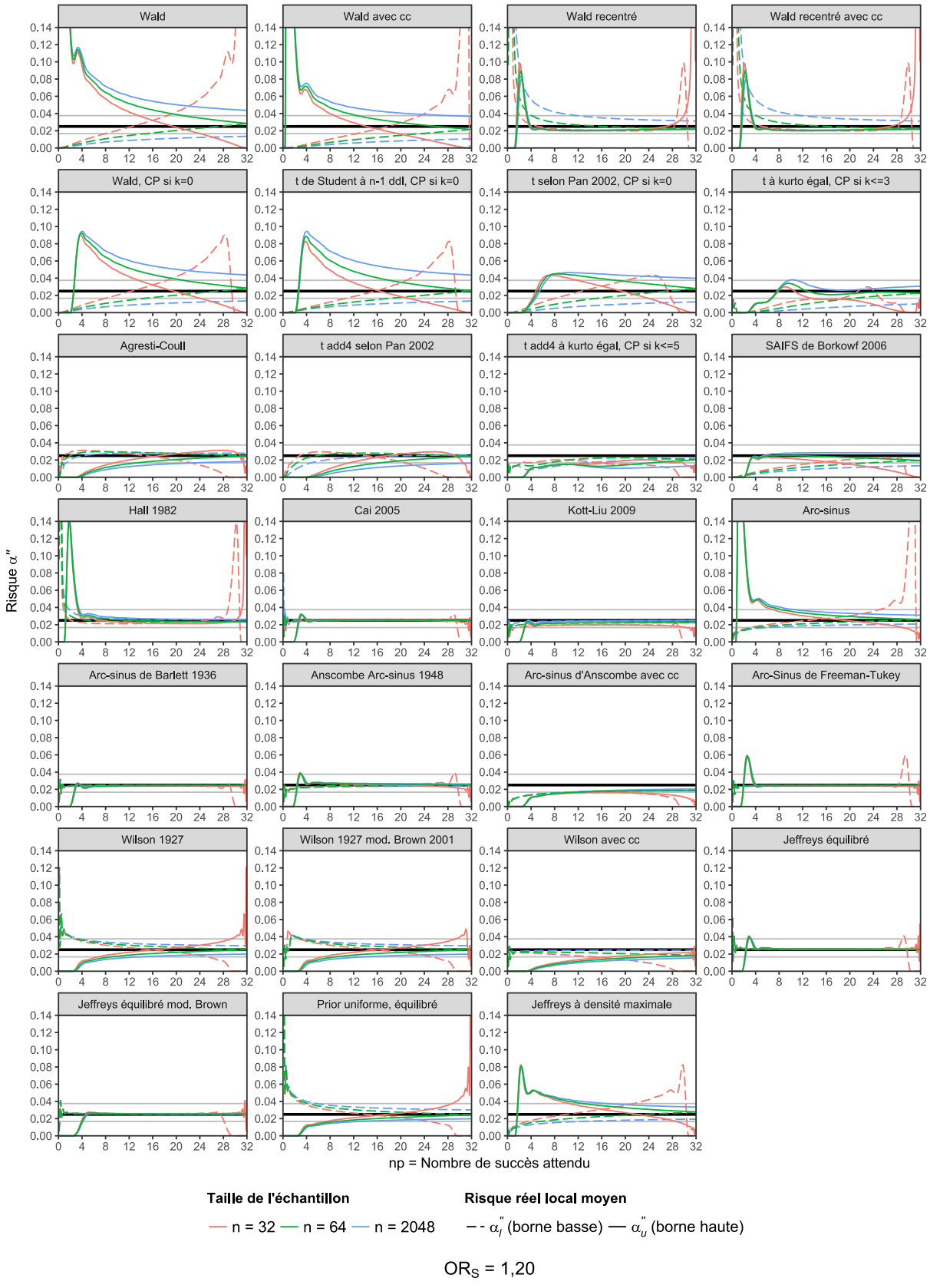

Figure 1 : risques réels moyens locaux unilatéraux de 27 estimateurs d'intervalle de confiance à 95%, selon différentes taille d'échantillon pour une proportion théorique *P* suivant un modèle logit-normal aléatoire avec un odds ratio typique $OR_S = 1,20$. La valeur k égale $min(x, n - x)$. CP signifie « Clopper-Pearson ».

Figure 2 : risques réels moyens locaux unilatéraux de 28 estimateurs d'intervalle de confiance à 95%, selon différentes taille d'échantillon pour une proportion théorique *P* suivant un modèle logit-normal aléatoire avec un odds ratio typique $OR_S = 1,20$. La valeur k égale $min(x, n - x)$. CP signifie « Clopper-Pearson ».

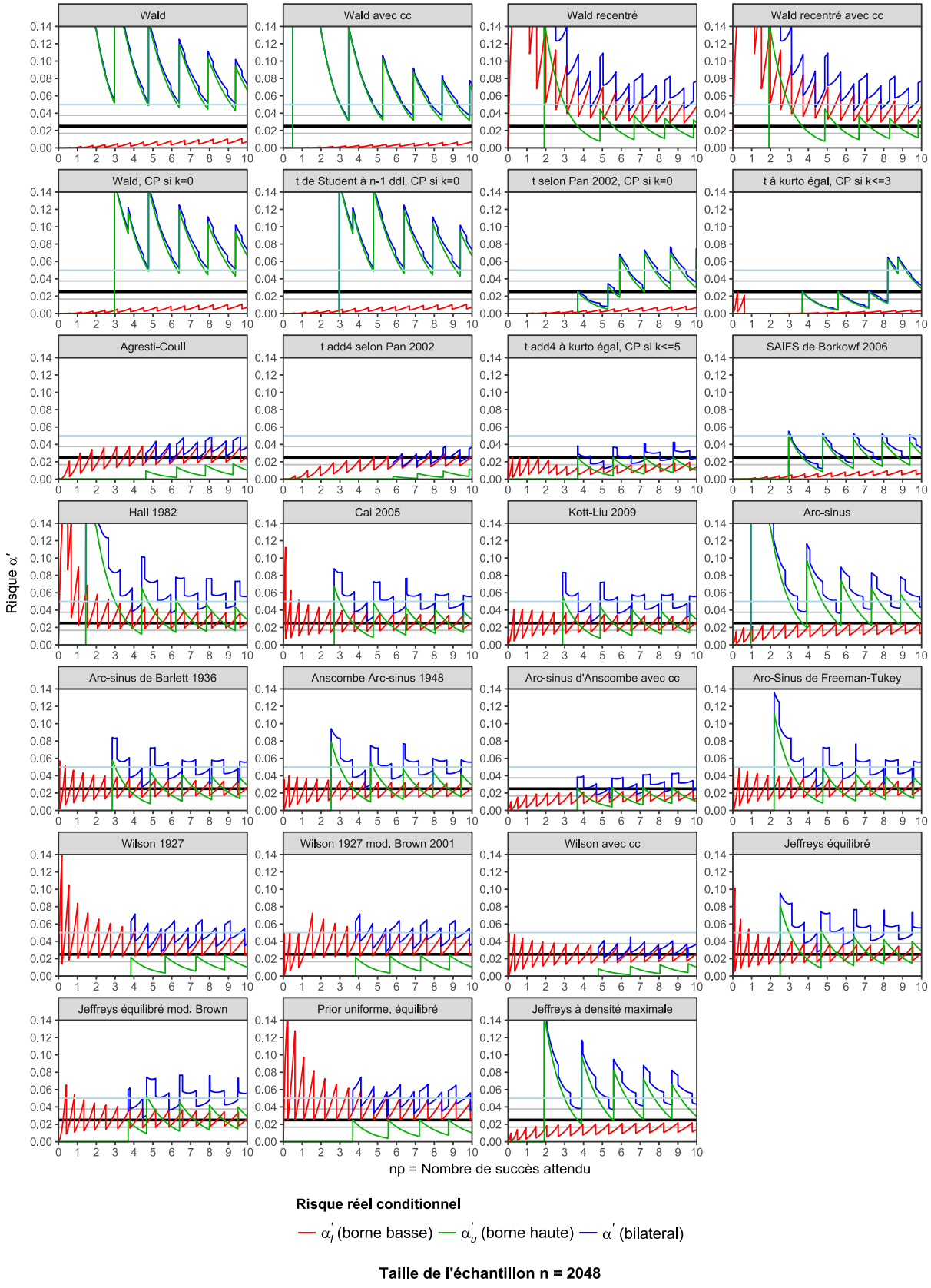

Figure 3 : risques réels conditionnels unilatéraux à gauche (rouge), à droite (vert) et bilatéral (bleu) de 27 estimateurs d'intervalle de confiance à 95% pour un échantillon de taille $n = 2048$. La valeur k égale $min(x, n - x)$. CP signifie « Clopper-Pearson ».

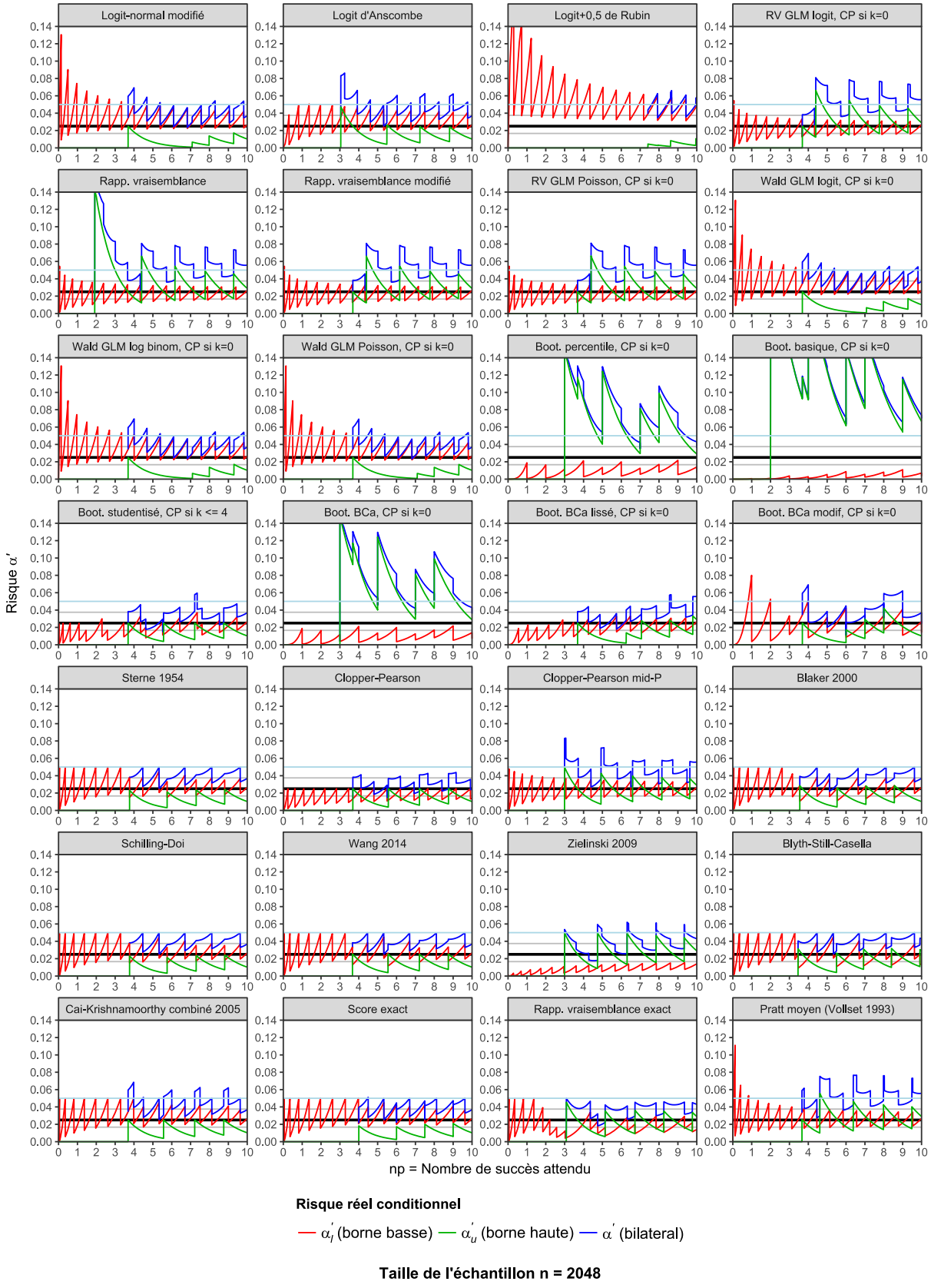

Figure 4 : risques réels conditionnels unilatéraux à gauche (rouge), à droite (vert) et bilatéral (bleu) de 28 estimateurs d'intervalle de confiance à 95% pour un échantillon de taille $n = 2048$. La valeur k égale $min(x, n - x)$. CP signifie « Clopper-Pearson ».

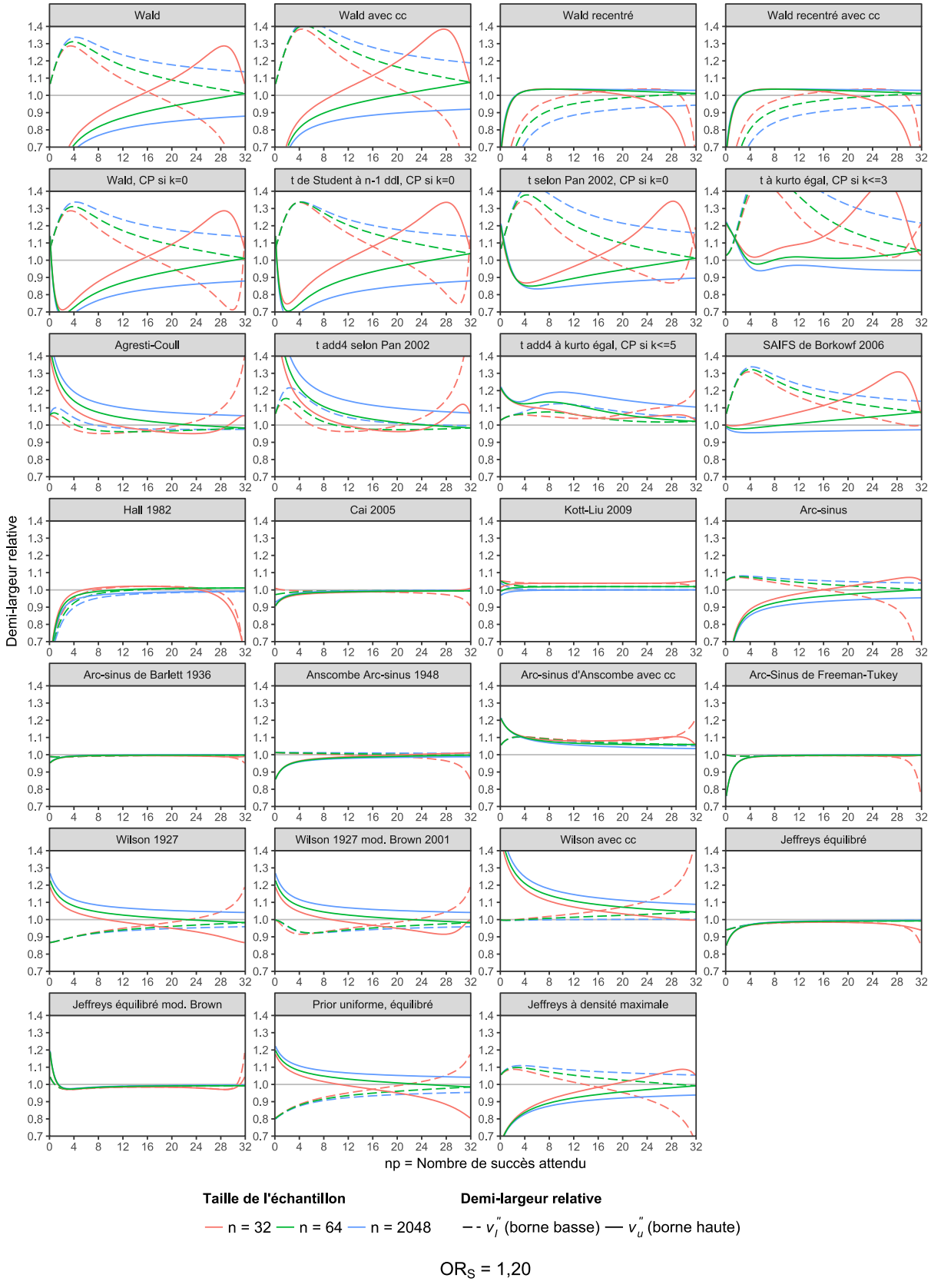

Figure 5 : demi-largeurs relatives moyennes locales pour 27 estimateurs d'intervalles de confiances à 95% pour une proportion P théorique aléatoire suivant une loi logit-normale d'odds ratio type $OR_S = 1,20$, selon le nombre de succès attendu $\lambda = np$ dans un échantillon de taille n.

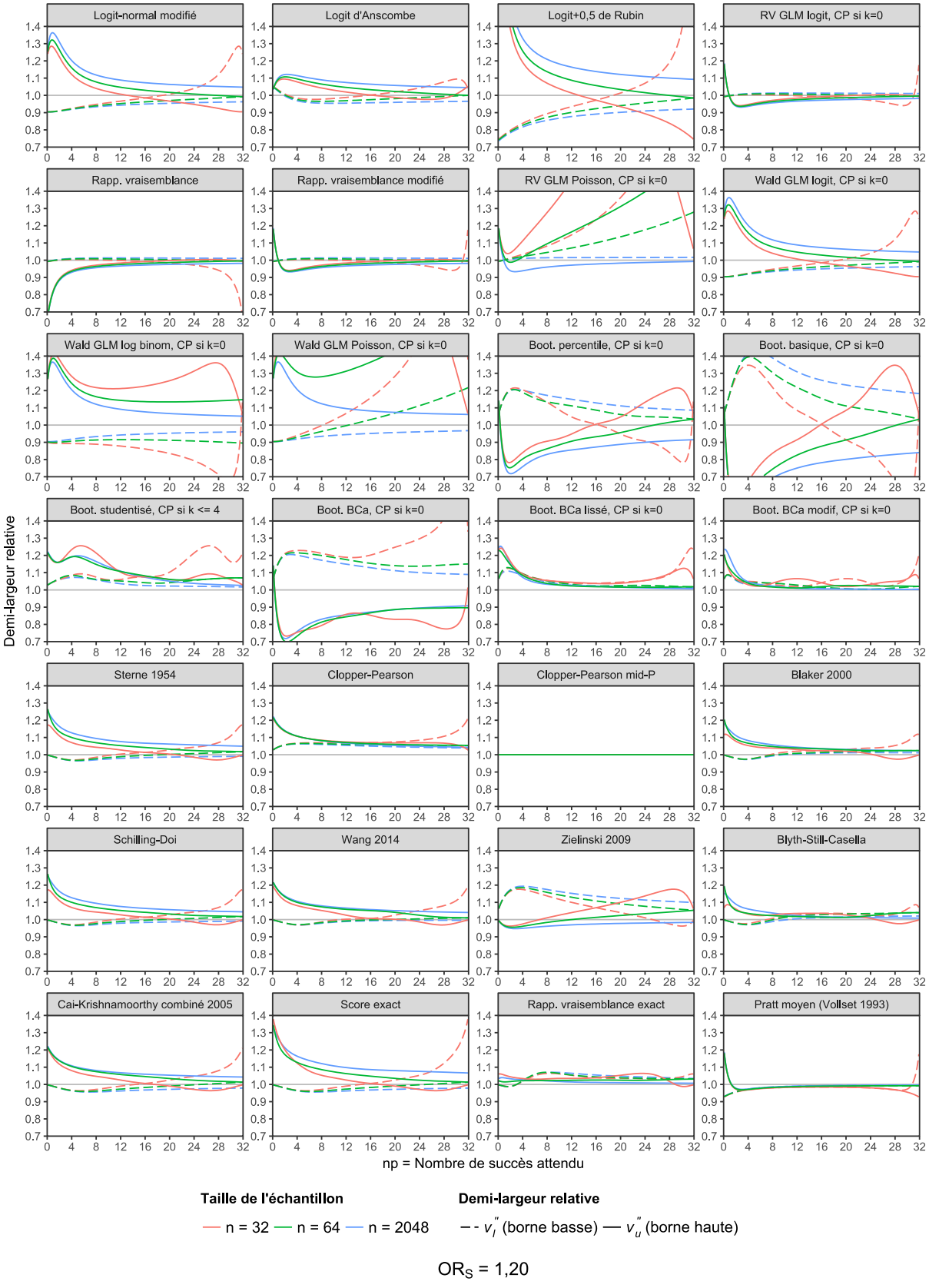

Figure 6 : demi-largeurs relatives moyennes locales pour 28 estimateurs d'intervalles de confiances à 95% pour une proportion P théorique aléatoire suivant une loi logit-normale d'odds ratio type $OR_S = 1,20$, selon le nombre de succès attendu $\lambda = np$ dans un échantillon de taille n.

6.3 Comparaison des intervalles de confiance strictement conservatifs

Les intervalles bilatéraux déséquilibrés exacts strictement conservatifs tendent à être aussi étroits que possible. Les plus récemment définis, Wang (109) et Schilling-Doi (93) nécessitent beaucoup de temps de calcul sur ordinateur mais sont supposés être plus étroits en moyenne que ceux précédemment décrits tels que celui de Sterne ou de Blaker. Ils ont été comparés sur un échantillon de 174 observations. La taille d'échantillon 174 a été sélectionnée parce qu'elle est suffisamment grande pour faire la différence entre les intervalles de Schilling-Doi et de Sterne, mais autrement, ne présente aucune particularité. Les intervalles de Wang et de Schilling-Doi optimisent la moyenne arithmétique de la largeur d'intervalle (somme des tailles des intervalles pour $x = 0, \ldots, n$ divisée par $n + 1$), mais la moyenne géométrique pourrait être plus pertinente. Rétrécir $[0\,;0{,}036]$ à $[0\,;0{,}034]$ (réduction relative de largeur de 5,6%) semble plus intéressant que rétrécir $[0{,}398\,;0{,}602]$ à $[0{,}398\,;0{,}600]$ (réduction relative de largeur de 0,98%) bien que les deux réductions absolues soient égales à $0{,}002$. Le rapport de deux moyennes géométriques est égal à la moyenne géométrique des rapports de telle sorte que l'on peut interpréter les rapports de moyennes géométriques comme des réductions relatives des largeurs d'intervalles.

Ces moyennes arithmétiques et géométriques des largeurs d'intervalles sont montrées sur le Tableau 8. L'estimateur de Clopper-Pearson a une moyenne géométrique de largeur d'intervalle 2,6% plus grande que celle de l'estimateur le plus étroit (Blaker). Tous les estimateurs d'intervalle à risques déséquilibrés sont très proches les uns des autres. La procédure de Schilling-Doi, bien que se vantant d'avoir des performances optimales dans la minimisation de la moyenne arithmétique de la largeur d'intervalle, présente une moyenne plus élevée que l'intervalle du score exact.

L'intervalle de Schilling-Doi est égal à l'intervalle de Sterne pour presque toutes les valeurs (Figure 7A). Il existe quelques pics pour lesquels les deux intervalles diffèrent, mais autrement, les intervalles sont parfaitement identiques. La différence est négligeable pour un usage pratique. L'intervalle de Blaker est légèrement plus grand que l'intervalle de Schilling-Doi en moyenne arithmétique et plus court en moyenne géométrique (Tableau 8), parce que ses largeurs d'intervalles sont plus basses pour des proportions proches de 0 et de 1 (Figure 7B).

L'intervalle de Wang a des moyennes arithmétiques et géométriques de largeur d'intervalle presque égales à celles de l'intervalle de Schilling-Doi (Tableau 8), mais est plus étroit pour des proportions proche de 0,50 et plus large pour des proportions comprises entre 0,20 et 0,40 (Figure 7C). Ceci peut être expliqué par l'algorithme de Wang, qui comprime séquentiellement les intervalles en descendant de $x = n/2$ jusqu'à $x = 0$. Les intervalles qui sont comprimés en premier ont plus de place pour le faire, et sont donc plus fortement réduits en taille.

L'analyse des demi-largeurs montre que pour des petites proportions, les intervalles de Schilling-Doi et de Wang ont des bornes hautes proches de celles de l'intervalle de Clopper-Pearson (Figure 8, panneaux A et B). Ces intervalles ont des bornes basses plus élevées pour les petites proportions. Par exemple, l'intervalle de confiance à 95% de Clopper-Pearson pour 6 succès sur un échantillon de taille 174 est égal à $[0{,}0128\,;\,0{,}0735]$ alors que l'intervalle de Schilling-Doi est égal à $[0{,}0151\,;\,0{,}0738]$. Les deux bornes sont plus élevées dans l'intervalle de Schilling-Doi, avec une augmentation du risque α_l' à gauche et une réduction du risque α_u' à droite.

Les intervalles de Blaker sont toujours contenus dans les intervalles de Clopper-Pearson (12). En conséquence, les demi-largeurs des intervalles de Blaker sont toujours inférieures ou égales à celles de

Clopper-Pearson pour toutes les proportions. Par exemple, pour 6 succès sur un échantillon de taille 174, l'intervalle de Blaker est égal à [0,0151 ; 0,0725] contre [0,0128 ; 0,0735] pour l'intervalle de Clopper-Pearson.

Nom de l'intervalle	Moyenne arithmétique des largeurs d'intervalles	Moyenne géométrique des largeurs d'intervalles
Clopper-Pearson	0,12137	0,11470
Blaker	0,11855	0,11182
Wang	0,11848	0,11196
Sterne	0,11852	0,11216
Schilling-Doi	0,11846	0,11199
Score exact	0,11804	0,11192
Rapp. Vraisemblance (RV) exact	0,11882	0,11185

Tableau 8 : moyennes arithmétiques et géométriques des largeurs d'intervalles de confiance à 95% pour une taille d'échantillon $n = 174$, pour $x = 0, \ldots, n$, concernant des estimateurs d'intervalles bilatéraux strictement conservatifs.

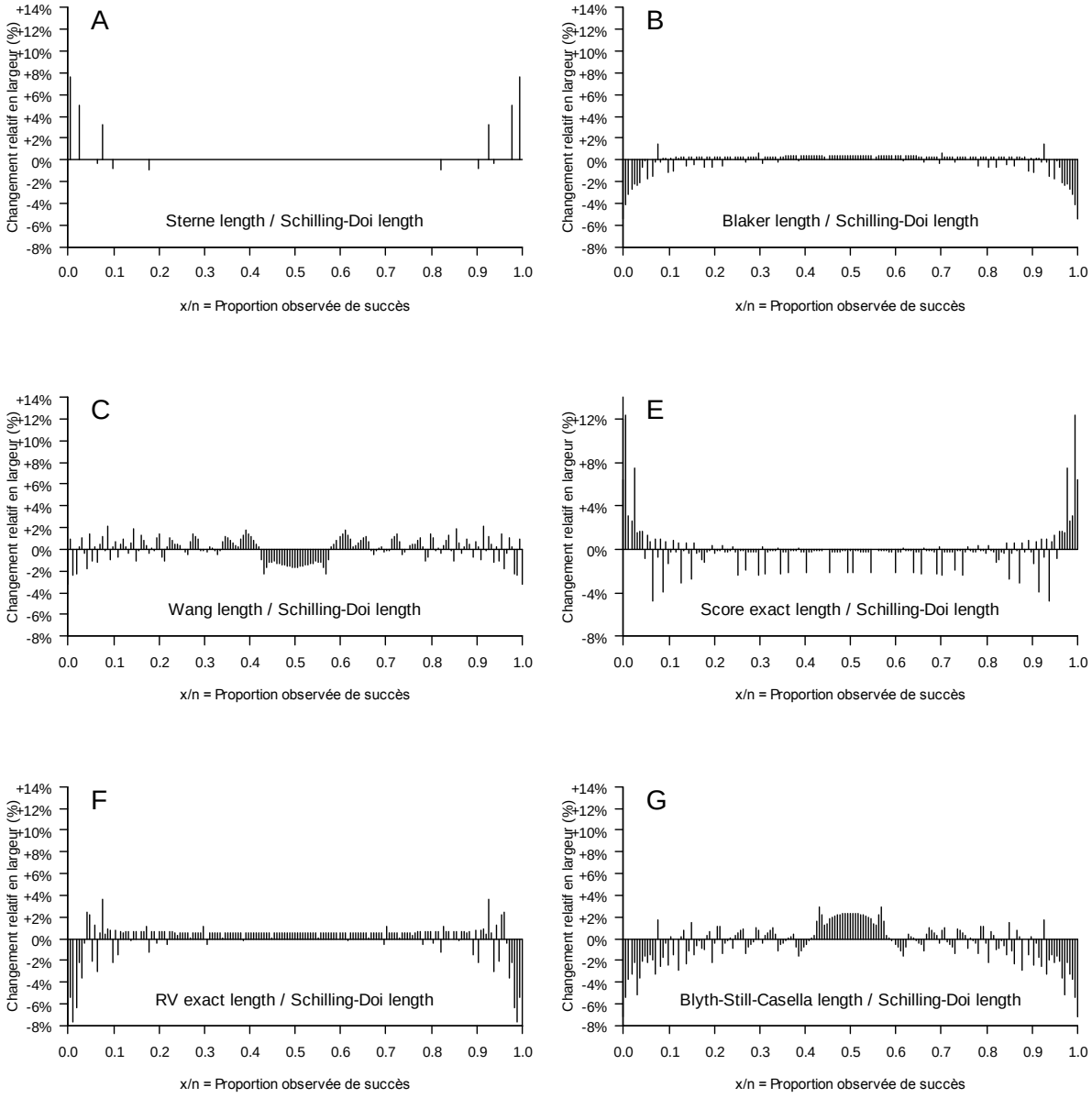

Figure 7 : largeur relatives (%) d'intervalles de confiance à 95% à risques déséquilibrés, pour une taille d'échantillon $n = 174$, pour n'importe quel nombre de succès (nombre entier) compris entre 0 et 174. Il ne s'agit pas d'expérience aléatoire ; c'est une représentation des 175 intervalles de confiance déterminés par $x = 0, \ldots, 174$ et $n = 174$.

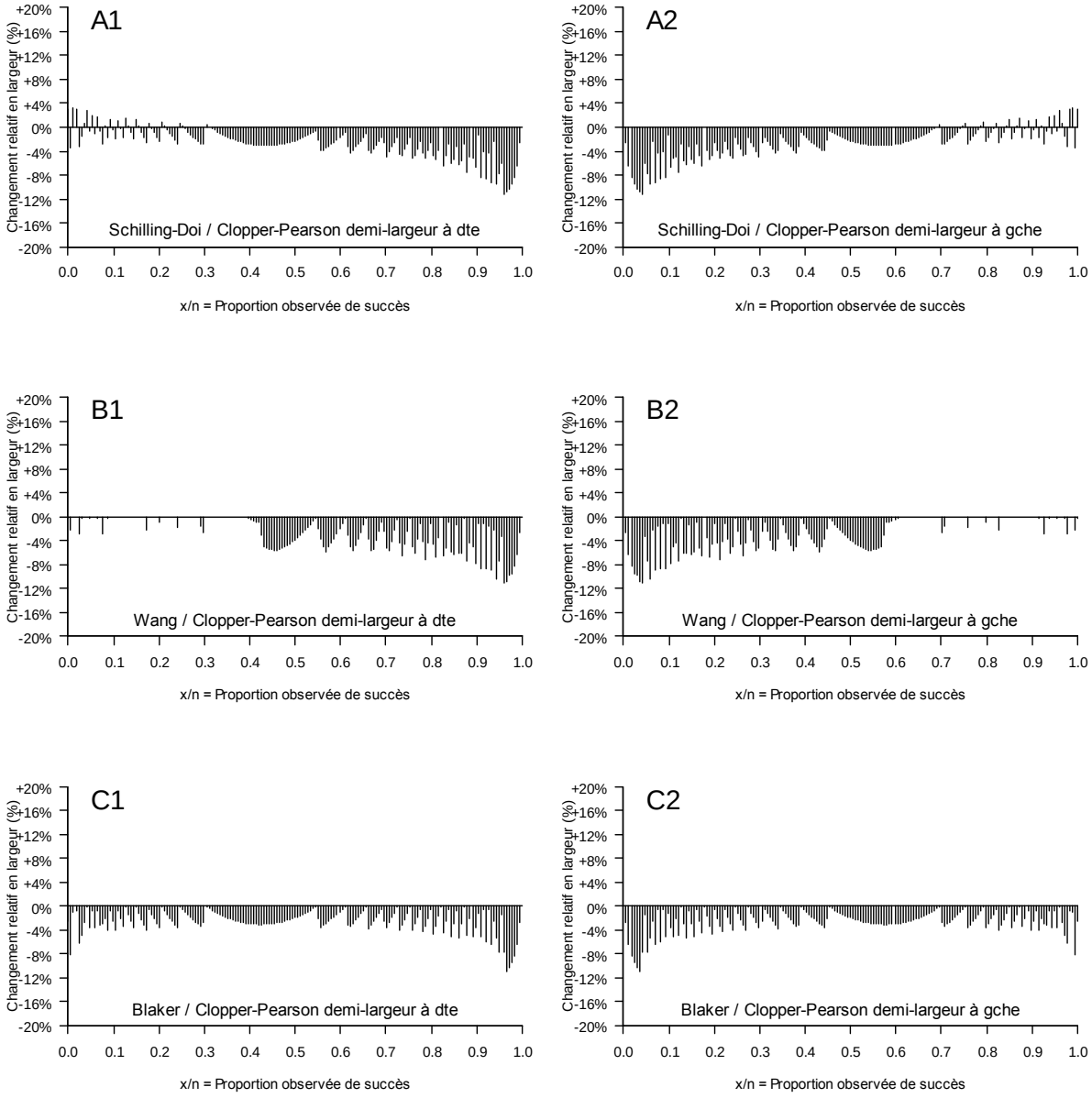

Figure 8 : demi-largeur relatives (%) d'intervalles de confiance à 95% (Schilling-Doi, Wang, Blaker vs Clopper-Pearson), pour une taille d'échantillon $n = 174$, pour n'importe quel nombre de succès (nombre entier) compris entre 0 et 174. Il ne s'agit pas d'expérience aléatoire ; c'est une représentation des 175 intervalles de confiance déterminés par $x = 0, \dots, 174$ et $n = 174$.

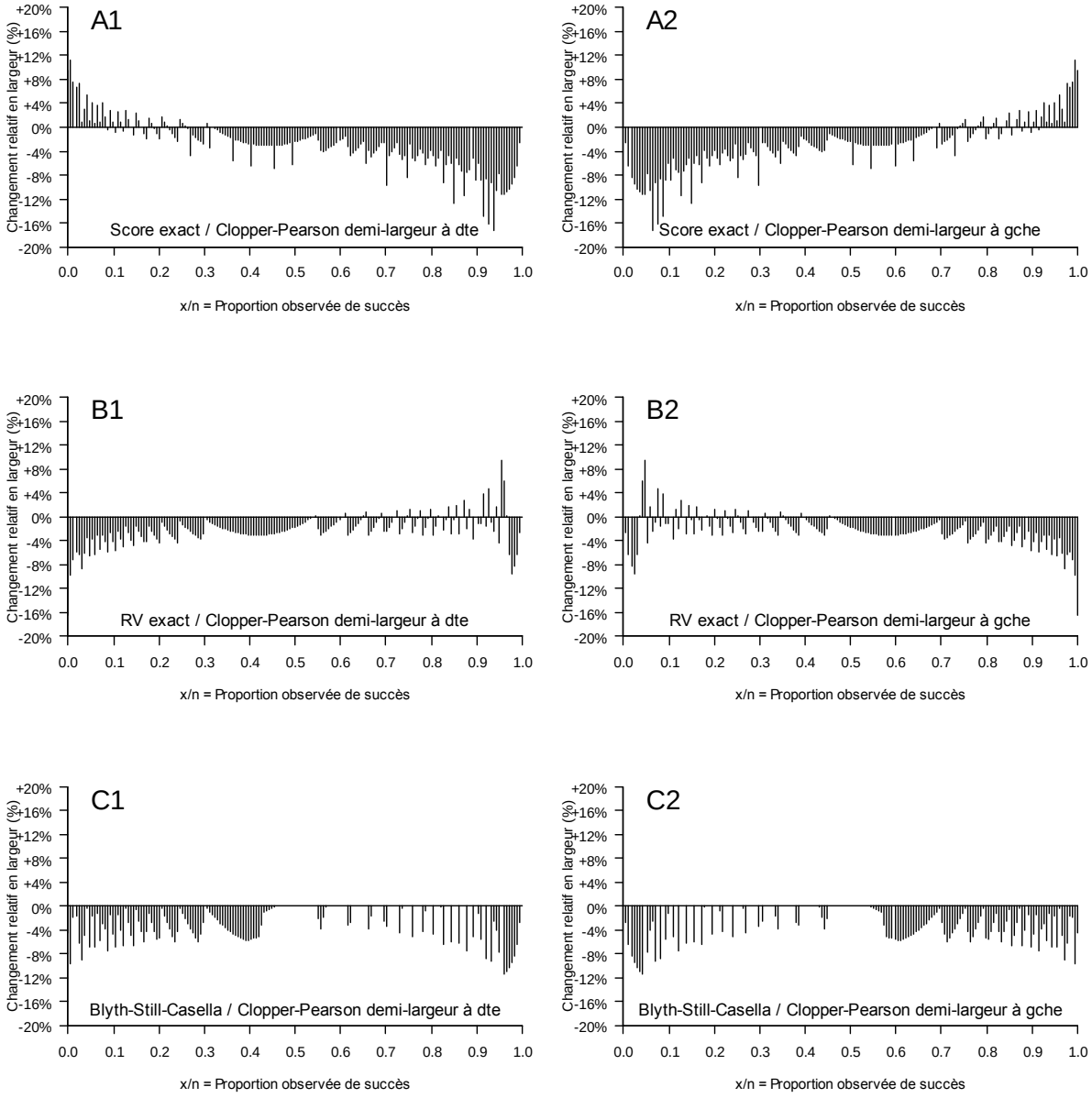

Figure 9 : demi-largeur relatives (%) d'intervalles de confiance à 95% (Score, Rapport de vraisemblance (RV) exact, Blyth-Still-Casella vs Clopper-Pearson), pour une taille d'échantillon $n = 174$, pour n'importe quel nombre de succès (nombre entier) compris entre 0 et 174. Il ne s'agit pas d'expérience aléatoire ; c'est une représentation des 175 intervalles de confiance déterminés par $x = 0, \ldots, 174$ et $n = 174$.

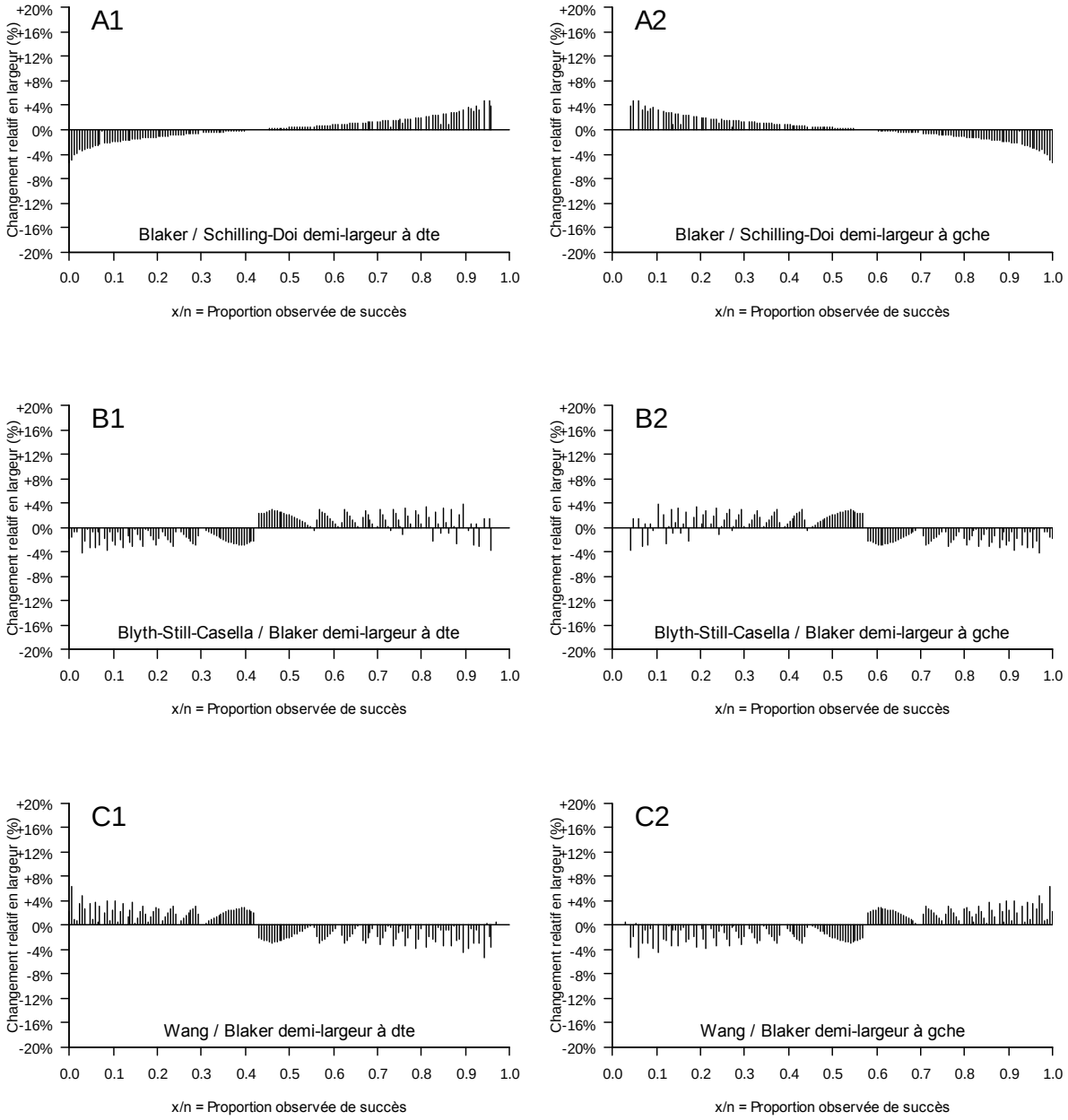

Figure 10 : demi-largeur relatives (%) d'intervalles de confiance à 95% (Blyth-Still-Casella, Blaker, Wang et Schilling-Doi), pour une taille d'échantillon $n = 174$, pour n'importe quel nombre de succès (nombre entier) compris entre 0 et 174. Il ne s'agit pas d'expérience aléatoire ; c'est une représentation des 175 intervalles de confiance déterminés par $x = 0, \ldots, 174$ et $n = 174$.

7 Annexe 2 : macros pour divers logiciels

Cette annexe contient les macros destinés aux logiciels SAS, Stata, SPSS, Python, Minitab, MYSTAT/SYSTAT, Microsoft Excel, HTML+JavaScript ainsi qu'aux calculatrices graphiques Texas Instruments Ti 83/84.

Ces fichiers, commes toutes les annexes sont librement téléchargeables à n'importe laquelle des adresses suivantes :

http://tinyurl.com/ybkqysm3

http://andre.gillibert.fr/owncloud/public.php?service=files&t=4a8a49cba9e183ed5c7be32f3087cf69

https://mega.nz/#F!o0IyAbgK!beyM7_hFDnRzWW1gyds7Vg

L'accès de la version HTML+JavaScript, utilisable en ligne ou en accès local, est disponible à l'adresse suivante :

http://andre.gillibert.fr/stats/cpmidp.html

Les codes sources sont recopiés ci-dessous :

7.1 Macro SAS

Cette macro n'est pas nécessaire à SAS 9.4. Par contre, elle peut servir pour des versions antérieures.

```
/* This file provides two macros :
   cpmidp : computes the Clopper-Pearson mid-P confidence interval from an immediate number of successes and trials
   cpmidp_varbin : computes the Clopper-Pearson mid-P confidence interval from a binary variable in a dataset
 */
%macro cpmidp(
       x,
       n,
       conf_level=0.95,
       prec=1e-8/n,
       print=1,
       out=);
/*
       AUTHOR    : André GILLIBERT (August 2017)
       LICENSE   : Creative Commons CC0
       FUNCTION  : computes the Clopper-Pearson mid-P confidence interval of a binomial proportion
       REFERENCE : Berry and Armitage. Mid-P confidence intervals: a brief review.
                           Journal of the Royal Statistical Society Series D (The Statistician) 44(4). January 1995.
                           DOI: 10.1371/journal.pone.0045723

       MANDATORY PARAMETERS:
              x = number of successes
              n = number of trials
       OPTIONAL PARAMETERS:
              conf_level     = two-sided confidence level
              prec           = Decimal precision of the confidence bounds requested.
              print          = 1 if the confidence limits should be printed to the Output Display System
```

```
               out                    = Name of a dataset containing the confidence limits,
retained after procedure exit
*/
%let out2 = &out;
%if &out2 = %then %do;
        %let out2 = temp;
%end;
data &out2;
        retain x n estimate trials lcl ucl conf_level;
        x=&x;n=&n;
        if x = 0 then out=0; else do;
                target=(0.5+&conf_level/2);
                xlower=0;xupper=x/n;
                %datastep_uniroot(upvalue_cp_midp,target,prec=&prec);
        end;
        lcl = out;
        if x = n then out=1; else do;
                target=(0.5-&conf_level/2);
                xlower=x/n;xupper=1;
                %datastep_uniroot(upvalue_cp_midp,target,prec=&prec);
        end;
        ucl = out;
        Estimate = x/n;
        conf_level=&conf_level;

        keep x n estimate lcl ucl conf_level;
run;
data &out2;
        set &out2;
        attrib lcl label="Lower confidence limit (CP mid-P)";
        attrib ucl label="Upper confidence limit (CP mid-P)";
        attrib x label="Successes";
        attrib n label="Trials";
        attrib conf_level label="Two-sided confidence level";
run;
%if &print = 1 %then %do;
proc print data=&out2 label noobs;run;
%end;
%if &out = %then %do;
        proc datasets noprint library=work;
                delete &out2;
        quit;
%end;
%mend;


%macro cpmidp_varbin(
        data,
        varbin,
        conf_level=0.95,
        print=1,
        out=);
/*
        cpmidp_varbin : it's the same internal function as cpmidp but it works from
        a dataset with a binary variable. Missing values are excluded.
        Successes are non-zero non-missing values. Trials are any non-missing values.

        MANDATORY PARAMETERS:
                data = Name of the dataset
                varbin = Name of a numeric variable in the dataset (0 values are failures, oth-
er non-missing values are successes)
        OPTIONAL PARAMETERS:
                conf_level = confidence level (from 0 to 1)
                print = 1 to print results to the Output Display System or 0 to keep the proce-
dure quiet
                out = name of an output dataset where the results will be stored
```

```
*/

data _tmpData;
	length _xvalue 3;
	set &data;
	if missing(&varbin) then delete; else _xvalue = &varbin ne 0;
	keep _xvalue;
run;

%let x=;
%let n=;
proc sql noprint;
	select sum(_xvalue) into :x from _tmpData;
	select count(_xvalue) into :n from _tmpData;
quit;
%cpmidp(&x,&n,conf_level=&conf_level,print=&print,out=&out);
proc datasets library=work noprint;
	delete _tmpData;
quit;
%mend;

%macro upvalue_cp_midp(p,context);
	xm = x - 1;
	if (xm < 0) then xm=0; else xm=PROBBNML(&p,n,xm);
	out = PROBBNML(&p,n,x)-(PROBBNML(&p,n,x)-xm)/2 - &context;
	pvalue=out + &context;
%mend;

%macro datastep_uniroot(zeromac,context,nit=10,prec=1e-6);
	/* zeromac must take an argument and output the result in the 'out' numeric variable */
	/* On entry : xlower and xupper must be set */
	%&zeromac(xlower,&context)
	flower=out;
	%&zeromac(xupper,&context)
	fupper=out;

	if (sign(flower)*sign(fupper)>0) or (xlower > xupper) then do;
		out=.;
	end; else do;
		do while ((xupper - xlower) > (&prec));
		xmid = (xlower + xupper)/2;
		%&zeromac(xmid,&context)
		fmid =out;

		if sign(fmid)*sign(flower) > 0 then do;
			xlower = xmid; flower=fmid;
		end; else do;
			xupper = xmid; fupper=fmid;
		end;
		out = (xlower + xupper)/2;
		end;
	end;
%mend;
```

7.2 Macro Stata

Cette macro requiert une version de Stata ≥ 9. Elle s'inspire des commandes Stata ci et cii.

```
/*
REQUIREMENT : STATA >= 9
AUTHOR: André GILLIBERT (August 2017)
LICENSE: Creative Commons CC0
SUMMARY: computes Clopper-Pearson mid-P confidence intervals of a binomial proportion
The procedure is described in DOI: 10.2307/2348891
```

```
SYNTAX 1: cpmidpi #trials #successes [, level(95) ]
EXAMPLE 1: cpmidpi 100 3
EXAMPLE 2: cpmidpi 100 3 , level(90)

#trials is the number of trials. It must be an immediate integer.
#successes is the number of successes. It must be an immediate integer.
level is the two-sided confidence level

SYNTAX 2: cpmidp col [, level(95) ]
EXAMPLE 3: cpmidp complications

col is a column of the current dataset.
This column must contain only 0 and 1 integer values. It may contain missing values. Missing
values are excluded.
level is the two-sided confidence level

These commands mimicks the Stata cii and ci commands.
They return scalar values:
        r(level) = confidence level
        r(ub) = upper bound of the confidence interval
        r(lb) = lower bound of the confidence interval
        r(se) = standard error of the proportion (pretty useless)
        r(mean)= estimated proportion = number of successes / number of trials
        r(N) = number of trials

The cicmidp(successes,trials,confidence_level) function can be used from MATA
*/

mata:


real scalar midpvalue(x,n,p)
{
        return(binomial(n,x,p) - (binomial(n,x,p)-binomial(n,x-1,p))/2)
}

real scalar unirootpvalue(x,n,q)
{
        if (x==0 & q>0.50) {
                return(0)
        }
        if (x==n & q<=0.50) {
                return(1)
        }
        prec=1e-8/n
        real scalar flower, fupper, pupper, plower
        plower=0;pupper=1
        flower = midpvalue(x,n,plower) - q
        fupper = midpvalue(x,n,pupper) - q

        if (sign(flower)*sign(fupper) > 0) {
                return(.)
        }
        while ((pupper-plower) > prec) {
                pmid = (plower+pupper)/2
                fmid = midpvalue(x,n,pmid) - q
                if (sign(flower)*sign(fmid) > 0) {
                        plower=pmid
                        flower=fmid
                } else {
                        pupper=pmid
                        fupper=fmid
                }
        }
        return((plower+pupper)/2)
}
```

```
real rowvector cpmidpi(real scalar x,real scalar n,real scalar conf_level)
{
	return (unirootpvalue(x,n,0.5+conf_level/2), unirootpvalue(x,n,0.5-conf_level/2))
}
end

capture program drop cpmidpi
program define cpmidpi, rclass
	version 9
	syntax anything(name=x id="Number of trials and successes") [, level(real 95) ]

	scalar n=`1'
	scalar x=`2'
	if (x > n) {
		display "{error}Number of successes must be equal or smaller to number of tri-
als"
		error 498
	}
	mata: st_matrix("outmat", cpmidpi(st_numscalar("x"), st_numscalar("n"), (`level'/100)))
	return scalar N=n
	return scalar mean=x/n
	return scalar se=sqrt((x*(n-x))/n^3)
	return scalar lb=outmat[1,1]
	return scalar ub=outmat[1,2]
	return scalar level=`level'
	display "Clopper-Pearson mid-P two-sided confidence interval"
	display "Confidence level: " `level' "%"
	display "Lower confidence limit: " outmat[1,1]
	display "Upper confidence limit: " outmat[1,2]
end

capture program drop cpmidp
program define cpmidp
	version 9
	syntax varname [, level(real 95) ]
	quietly summarize `1'
	cpmidpi r(N) r(sum) , level(`level')
end
```

7.3 Macro SPSS

```
* Encoding: UTF-8.

* TITLE : Clopper-Pearson mid-P interval.
* AUTHOR : André GILLIBERT (August 2017)
* LICENSE : Creative Commons CC0
* PROGRAMMING LANGUAGE : SPSS Syntax (tested on SPSS 16 and SPSS 25)
* SYNTAX :
* cp midp varname=<variable name of the active dataset> conf=<confidence level>
* There is no space between cp and midp, but due to the SPSS feature/bug of executing macro
inside comments, it had to be added.
* EXAMPLE :
* cp midp varname=complication conf=0.95.
* DESCRIPTION : implements the Clopper-Pearson mid-P confidence interval of a binomial propor-
tion
* This interval is described in DOI: 10.2307/2348891
* A variable name of the active dataset must be specified (varname).
* This variable must contain 0 and 1 numeric values. Missing values are excluded.
* The confidence level must be specified. It's a real number between 0 and 1.

define cpmidp(varname = !TOKENS(1) / conf = !TOKENS(1)).
dataset name OriginalData.
```

```
* We create a dataset named 'temp' with one row and two columns: successes=number of successes
and trials=number of trials.

dataset declare temp window=hidden.
temporary.
compute always1=1.
aggregate outfile=temp /break=always1 / successes=sum(!varname) / trials=N(!varname).
dataset activate temp.

* We are going to execute the procedure on the one-row dataset.
* The hard work will be done with scrap variables.

compute #conf=!conf.
loop #bound=0 to 1.
    compute #target = 0.5 - (#bound*2-1)*#conf/2.
    compute #x=successes.
    compute #n=trials.
    compute #xlower = 0.
    compute #xupper = 1.
    compute #flower=CDF.BINOM(#x,#n,#xlower)-PDF.BINOM(#x,#n,#xlower)/2 - #target.
    compute #fupper=CDF.BINOM(#x,#n,#xupper)-PDF.BINOM(#x,#n,#xupper)/2 - #target.
    compute #prec=1e-8/#n.

    loop.
    compute #xmid = (#xlower+#xupper)/2.
    compute #fmid=CDF.BINOM(#x,#n,#xmid)-PDF.BINOM(#x,#n,#xmid)/2 - #target.
    do if ((#fmid)*(#flower) > 0).
        compute #xlower = #xmid.
        compute #flower = #fmid.
    else.
        compute #xupper = #xmid.
        compute #fupper = #fmid.
    end if.
    end loop if ((#xupper - #xlower) < #prec).

    compute #xmid = (#xupper+#xlower)/2.
do if (#bound = 0).
    do if (#x = 0).
        compute #xmid=0.
    end if.
    compute lower_cl=#xmid.
else.
    do if (#x = #n).
        compute #xmid=1.
    end if.
    compute upper_cl=#xmid.
end if.
end loop.
compute conf=#conf.
execute.
formats lower_cl upper_cl (F10.8).
list / cases=1 .
dataset close temp.
dataset activate originaldata.
!enddefine.
```

7.4 Fonction Python

```
# AUTHOR : André GILLIBERT (August 2017)
# LICENSE : Creative Commons CC0
# DESCRIPTION : Provides the Clopper-Pearson mid-P confidence interval of a binomial propor-
tion
# This procedure is described in DOI: 10.2307/2348891
# EXAMPLE : cpmidp(3,100,conf=0.90)
```

```
# SYNTAX : cpmidp(successes,trials,conf=confidence level)

from scipy.stats.distributions import binom
from scipy.optimize import brentq

def cpmidp(x,n,conf=0.95):
	def midpvalue(p,level):
		return binom.cdf(x,n,p) - 0.5*binom.pmf(x,n,p) - level
	if x>0:
		lcl = brentq(midpvalue, 0, 1, args=0.5+conf/2,xtol=1e-8/n)
	else:
		lcl=0
	if x<n:
		ucl = brentq(midpvalue, 0, 1, args=0.5-conf/2,xtol=1e-8/n)
	else:
		ucl=1
	return (lcl,ucl)
```

7.5 Macro MYSTAT/SYSTAT

```
rem TITLE: Clopper-Pearson mid-P confidence interval calculator
rem AUTHOR: André GILLIBERT (August 2017)
rem LICENSE: Creative Commons CC0
rem REFERENCE ARTICLE: DOI:10.2307/2348891
rem USAGE 1: Open & run the script in MYSTAT >= 12 or SYSTAT >= 12 to get an interactive cal-
culator
rem USAGE 2:
rem Remove the interactive part of the script (delimited by comments)
rem And then, use the cpmidp_onesided and cpmidp_twosided functions
rem Their syntax is described in the interactive part

function cpmidpq(x,n,target)
{
prec=1e-8/n
xlower=0
xupper=1
flower=1-target
fupper=-target

while ((xupper - xlower) > prec)
	xmid = (xlower+xupper)/2
	fmid = NCF(x,n,xmid)-0.5*NDF(x,n,xmid)-target

	sgn=(fmid*flower) > 0

	xlower = xmid*sgn + xlower*(1-sgn)
	flower = fmid*sgn + flower*(1-sgn)
	xupper = xmid*(1-sgn) + xupper*sgn
	fupper = fmid*(1-sgn) + fupper*sgn
endwhile
return (xlower + xupper)/2
}

function cpmidp_onesided(x,n,conflvl)
{
cl = cpmidpq(x,n,1-conflvl)
lcl = cl*(x<>0)
ucl = cl*(x<>n)+(x=n)
cl  = ucl*(conflvl>0.5) + lcl*(conflvl<=0.5)
return cl
}

function cpmidp_twosided(x,n,conf)
{
lcl2 = cpmidp_onesided(x,n,0.5-conf/2)
```

```
ucl2 = cpmidp_onesided(x,n,0.5+conf/2)

PRINT lcl2, ucl2
PRINT "Two-sided Clopper-Pearson mid-P confidence interval"
PRINT cat$("Confidence level: ", str$(conf))
PRINT cat$("Lower confidence limit: ", str$(lcl2))
PRINT cat$("Upper confidence limit: ", str$(ucl2))
return 0
}

rem ###### Start of interactive part ######

print "This script computes the Clopper-Pearson mid-P confidence interval of a binomial pro-
portion"
print "You may use this script non-interactively !"
print "The key function is cpmidp_onesided(successes, trials, confidence_level)"
print "Example 1:"
print "lower = cpmidp_onesided(3,100,0.025)"
print "upper = cpmidp_onesided(3,100,0.975)"
print "print lower, upper"
print "The cpmidp_twosided function displays results but returns 0"
print "Example 2:"
print "x=cpmidp_twosided(3,100,0.95)"

FORMAT 12, 5
token &successes / type=integer, prompt='How many successes in the sample?'
token &trials / type=integer, prompt='How many trials in the sample?'
token &conflevel / type=number, prompt='Confidence level (e.g. 0.95)?'
dropit=cpmidp_twosided(&successes, &trials, &conflevel)
rem ###### End of interactive part ######
```

7.6 Macro Minitab

```
# TITLE : Clopper-Pearson mid-P macro
# AUTHOR : André GILLIBERT (August 2017)
# LICENSE : Creative-Commons CC0
# COMPATIBILITY : Tested in Minitab 18, but may work on much older versions. Maybe down to
version 9 (1993).
# SYNTAX :
# %cpmidp col;
# storecl lcl ucl;
# conf conflevel.
#
# EXAMPLE 1:
# %cpmidp C3
# EXAMPLE 2:
# %cpmidp 'complication';
# storecl K10 K11;
# conf 0.90.
#
# The storecl and conf subcommands are optional.
# col = column
# lcl and ucl are constants, written on output. The macro will store the lower confidence lim-
it in lcl and the upper confidence limit un ucl
# conflevel is a numeric constant on input. It specifies the confidence level. Its default
value is 0.95.
# The macro computes the Clopper-Pearson two-sided mid-P confidence interval of a binomial
proportion as described in DOI:10.2307/2348891
#
# Missing values are excluded. Non-zero values are seen as successes. Zero values are seen as
failures.


macro
cpmidp col;
```

```
storecl lcl ucl;
conf clevel.
mcolumn col
mconstants xx nn
mconstants lcl ucl t1 t2 clevel
mconstants succ trials lowCL uppCL
default clevel=0.95
mreset

brief 0

let xx = sum(abs(sign(col)))
let nn = n(col)

let t1=0.5-clevel/2
let t2=0.5+clevel/2
call cpmidpq xx nn t1 ucl
call cpmidpq xx nn t2 lcl

brief 2

mtitle "Clopper-Pearson mid-P two-sided Confidence-Interval";
        notitle.
let succ=xx
let trials=nn
let lowCL=lcl
let uppCL=ucl
print succ trials lowCL uppCL
endmtitle
endmacro

macro # Get the confidence limit (0.025 -> upper CL of a 95%CI and 0.0975 -> lower CL of a
95%CI. Yep, it's reversed)
# Input : xx = successes nn=trial target=0.025 or 0.975
# Output : cl=Confidence limit
cpmidpq xx nn target cl
mconstants xx nn target cl
mconstants xlower xupper xmid flower fupper fmid prec

let xlower=0
let xupper=1
let prec = 1e-8/nn

let flower = 1-target
let fupper = -target

while (xupper - xlower) > prec
        let xmid = (xupper+xlower)/2
        call cpmidpvalue xx nn xmid target fmid
        if (fmid)*(flower) > 0
                let xlower = xmid
                let flower = fmid
        else
                let xupper = xmid
                let fupper = fmid
        endif
endwhile
let cl = (xlower + xupper)/2
if target >= 0.50 and xx = 0
        let cl = 0
elseif target < 0.50 and xx = nn
        let cl = 1
endif
endmacro

macro
cpmidpvalue xx nn pp target pvalue
```

```
mconstants xx nn pp pvalue1 pvalue2 target pvalue
       cdf xx pvalue1;
              binomial nn pp.
       pdf xx pvalue2;
              binomial nn pp.
       let pvalue = pvalue1 - pvalue2/2 - target
endmacro
```

7.7 Tableurs Microsoft Excel et LibreOffice

Les fonctions d'un tableur n'étant pratiquement pas communicables sur papier, veuillez télécharger la version en ligne à l'une des adresses Web précisées en page 111.

7.8 HTML+JavaScript

```
<html>
<!-- Author : Andre GILLIBERT, 2017 -->
<!-- License : Creative Commons CC0 (except for James D. McCaffrey lgamma function -->
<title>Clopper-Pearson mid-P interval calculator</title>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
<link rel="stylesheet" media="screen" href="style.css">
<script type="text/javascript">
log=Math.log
function lgamma(x) { // log-gamma derived from C# implementation of James D. McCaffrey
// See https://jamesmccaffrey.wordpress.com/2013/06/19/the-log-gamma-function-with-c/
  var coef = [76.18009172947146,
    -86.50532032941677, 24.01409824083091,
    -1.231739572450155, 0.1208650973866179E-2,
    -0.5395239384953E-5 ];
  var LogSqrtTwoPi = 0.91893853320467274178;
  var denom = x + 1;
  var y = x + 5.5;
  var series = 1.000000000190015;
  for (i = 0; i < 6; ++i)
   {
    series += coef[i] / denom;
    denom += 1.0;
  }
  return (LogSqrtTwoPi + (x + 0.5) * log(y) -y + log(series / x));
}
function lchoose(x,n) {
       return (lgamma(n + 1) - lgamma(n - x + 1) - lgamma(x + 1))
}
function ldbinom(x,n,p) {
       return (lchoose(x,n) + x * log(p) + (n-x)*log(1-p))
}
function dbinom_fast(x,n,p) {
       if (p == 0 | p==1) {return((n*p==x)*1)}
       return (Math.exp(ldbinom(x,n,p)))
}
function dbinom_mult(x,n,p) { /* Catherine Loader algorithm */
       var f=1;
       var j0=0, j1=0, j2=0;
       if (2*x >n) {return (dbinom_mult(n-x,n,1-p));}
       while (j0<x | j1 < x | j2 < n-x) {
              if (j0 < x && f < 1) {
                     j0++;
                     f *= (n-x+j0)/j0;
              } else {
                     if (j1 < x) {j1++; f *= p;}
                     else {j2++; f*= 1-p;}
              }
       }
       return (f);
}
function dbinom(x,n,p) {
```

```
	if (n <= 100) {
		return (dbinom_mult(x,n,p));
	} else {
		return (dbinom_fast(x,n,p));
	}
}
function pbinom(x,n,p) {
	if (x > 0.50*n) {
		return (1-pbinom(n-x-1,n,1-p));
	}
	var i, xm, x0, res=0, resT;
	x0 = Math.floor(n*p);
	xm = Math.min(x, x0);
	for(i=xm; i >= 0; --i) {
		resT = res + dbinom(i,n,p);
		if (resT == res) {break;}
		res = resT;
	}
	for(i=xm+1; i <= x; ++i) {
		resT = res + dbinom(i,n,p);
		if (resT == res) {break;}
		res = resT;
	}
	res = Math.min(res, 1);
	return (res);
}
function uniroot(func,lower,upper, tol) {
	flower = func(lower);
	fupper = func(upper);
	if (flower*fupper > 0) {
		return (null);
	}
	while ((upper - lower) > tol) {
		mid = (lower + upper)/2;
		fmid = func(mid);
		if ((fmid*flower) > 0) {
			lower = mid;
			flower = fmid;
		} else {
			upper = mid;
			fupper = fmid;
		}
	}
	return ((lower+upper)/2);
}
function cpmidp_quantile(x,n,alpha,midp) {
	function rtfunc(p) {
		return(pbinom(x,n,p) - midp*dbinom(x,n,p) - alpha);
	}
	return(uniroot(rtfunc, 0, 1, 1e-7/n));
}
function compute() {
	var out = document.getElementById('out');
	var i=0;
	var x = parseInt(document.getElementById("x").value);
	var n = parseInt(document.getElementById("n").value);
	var midp=0.50;
	if (document.getElementById("scpmidp").selected) {midp=0.50;} else {midp=0;}
	if (n < 0 | x < 0 | x > n) {
		out.textContent = "Invalid number of successes or trials";
		return(0);
	}
	var conf = parseFloat(document.getElementById("conf").value);
	if (conf <= 0 | conf >= 1) {
		out.textContent = "Invalid confidence level";
		return(0);
	}
```

```
        var lower, upper;
        if (x==0) {lower=0;} else {lower = cpmidp_quantile(x,n,0.5+conf/2,midp);}
        if (x==n) {upper=1;} else {upper = cpmidp_quantile(x,n,0.5-conf/2,midp);}
        var s="";
        out.textContent = "Confidence interval : [" + lower + " , " + upper + "]\n";
}
</script>
<body>
<form>
        <table>
        <tr><td>Number of successes: <td><input type="text" id="x" name="x">
        <tr><td>Number of trials: <td><input type="text" id="n" name="n">
        <tr><td>Two-sided confidence level: <td><input type="text" id="conf" name="conf" val-
ue="0.95">
        <tr><td>Confidence interval type: <td>
        <select id="citype">
                <option value="cpmidp" selected id="scpmidp">Clopper-Pearson mid-P</option>
                <option value="cp" id="scp">Clopper-Pearson</option>
        </select>
        </table>
        <p><input type="button" value="Compute confidence interval" onclick="compute()">
</form>
<div id="out" name="out">Please enter parameters and click the [Compute confidence interval]
button.</div>
```

7.9 Texas Instruments Ti 83/84

```
Disp "Clop-Pear mid-P"
Input "Successes:",X
Input "Trials:",N
Input "Confidence:",C
If X<0 or N<X or fPart(X)!=0 or fPart(N)!=0 or N<=0
Goto 2
If C>=1 or C<0
Goto 2
1-(1-C)/2->A
If X!=0
Then
        solve(binomcdf(N,P/N,X)-0.5(binompdf(N,P/N,X))-A,P,0.5*N,{0,N})/N->L
        Else
        0->L
End
If X!=N
Then
        solve(binomcdf(N,P/N,X)-0.5(binompdf(N,P/N,X))+A-1,P,0.5*N,{0,N})/N->R
        Else
        1->R
End
Disp L
Disp R
Stop
Lbl 2
Disp "DOMAIN ERROR"
Return
Stop
```

8 Annexe 3 : description du test de Venkatraman

Cet annexe décrit le test de Venkatraman sur séries appariées et fournit un exemple illustrant la possibilité de deux courbes ROC dont les aires sous la courbe sont presque identiques alors que le test de Venkatraman détecte une différence, sans, pour autant, permettre de connaître le signe de la différence des aires sous la courbes.

Le test de Venkatraman sur séries appariées (103) compare deux courbes ROC pour une variable de classification binaire D (disease). Il est basé sur une statistique, pour un échantillon de taille n :

$$E_{\cdot} = \sum_{k=1}^{n-1} |e_{\cdot k}| \tag{1}$$

où $e_{\cdot k}$ est la différence du nombre d'erreurs de classification de D, entre les deux groupes, lorsque le point de dichotomisation correspondant à l'observation numéro k est employé. Si cette statistique $E_{\cdot}$ dépasse le seuil, estimé par la méthode des permutations, le test est significatif et on peut conclure que les deux courbes ROC ne sont pas superposées ; elles diffèrent au moins sur un point. Les différences, quelque soient leur signe, sont comptées en positif car les $e_{\cdot k}$ sont additionnés en valeur absolue.

La Figure 1 présente deux courbes ROC obtenues par échantillonnage depuis une population simulée dans laquelle les deux aires sous la courbes ROC sont presque égales. Le test de Venkatraman a une puissance statistique proche de 100%, concluant à l'existence d'une différence dans la population, mais le signe observé de la différence d'aire sous la courbe ROC est opposé au signe réel dans environ la moitié des expériences. Si le test de Venkatraman est interprété (à tort) comme un test de comparaison d'aires sous la courbe, le risque de conclure à l'existence d'une différence dans le sens opposé à la différence réelle (erreur de type III) est proche de 50%.

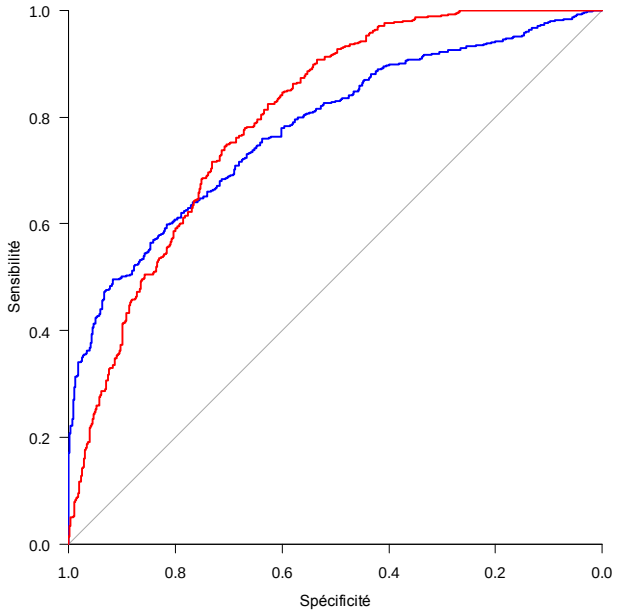

Figure 1 : deux courbes ROC obtenues par échantillonnage depuis une population simulée dans laquelle les deux aires sous la courbes ROC sont presque égales.

9 Annexe 4 : article en anglais

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

André GILLIBERT[a]*†, Jacques BÉNICHOU[b] and Bruno FALISSARD[a]
[a] INSERM UMR 1178, Université Paris Sud, Maison de Solenn, Paris, France.
[b] Department of Biostatistics, Rouen University Hospital, Rouen, France
* Correspondence to: André GILLIBERT, Department of Biostatistics, Rouen University Hospital, Rouen, France
†E-mail: andre.gillibert@chu-rouen.fr

Abstract

Introduction: Estimation of confidence intervals 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 have not been formally taken in account. Herein, we define and apply new evaluation criteria, then give recommendations for the practical use of these intervals.

Materials & methods: Google® Scholar was used for bibliographic research. The relevant judgment criteria were: One-sided local average risks assuming a random theoretical proportion or sample size, one-sided risks conditional to fixed proportion and sample sizes and expected half-lengths of confidence intervals.

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

Conclusion: While it's the most used, Wald's interval was so biased that it should never be used. We recommend the systematic use of the Clopper-Pearson mid-P interval for the estimation of a proportion except for observed-theoretical proportion comparison under controlled experimental conditions in which the Clopper-Pearson interval may be better.

KEYWORDS: binomial confidence interval, coverage bias, equal-tailed intervals, local average risks, Wald's interval, Clopper-Pearson mid-P interval.

9.1 Introduction

Estimating the confidence interval of a proportion is one of the most basic statistical problems and an everyday task for statisticians. Most statistical software provides two procedures: "approximate" and "exact". The "approximate" confidence interval estimator is usually Wald's interval

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (1)$$

where $\hat{p}$ is the observed proportion and n is the sample size. The "exact" confidence interval estimator is usually the Clopper-Pearson interval (31) (R, SAS, Stata). The word "exact" is misleading since actually, no deterministic estimator has complete control over the coverage due to the binomial distribution discreteness. Agresti and Coull (4) suggests that some approximate estimators may have better control over the average coverage than an exact estimator, because the Clopper-Pearson estimator is too conservative; its actual coverage is always equal or above the nominal coverage (strict conservatism).

The unsolvable problem of exact coverage lead to the development of many estimators: Pires presented twenty different estimators (83) and the development of new methods is still active in 2017 (115). Not all approximate estimators are equal: Wald's estimator may be the best known, but it has been criticized for its coverage bias deemed unacceptable (4,21) while Wilson (110), Agresti-Coull (4) and Jeffreys equal-tailed intervals have been recommended (5,21,52).

Some aspects are well described in systematic reviews, such as minimal coverage control, average coverage and interval length (4,21,77,83,89,106). In our opinion, two important issues have not been sufficiently addresses from biostatistician's point of view. The first issue is the balance of one-sided risks: equal-tailed intervals or unequal-tailed intervals. This is needed for a directional interpretation of confidence intervals where overestimation and underestimation are distinguished (59). The second issue is the variability of the actual proportion and sample sizes from one experiment to another. In the light of these issues, the ideal confidence interval estimator may change.

The objectives of this article are to define new evaluation criteria in the light of these issues, assess existing confidence interval estimators with these criteria and give practical advice to use the best estimator.

Issues of binomial proportion estimation and evaluation criteria rationale are first discussed in order to clarify the context, then, the usual structure of articles is followed: Materials & methods, Results, Discussion, and Conclusion.

9.1.1 Issues of binomial proportion estimation

9.1.1.1 Discreteness of distribution

Coverage oscillations due to the binomial distribution discreteness have been described by Brown, Cai and DasGupta (21). For example, the Clopper-Pearson 95% confidence interval estimator, the *nominal α risk* that the confidence interval does not contain the true proportion is 0.05 but the *actual risk* for $n = 30$ and $p = 0.3471$ equals 0.0195. For $n = 30$ and $p = 0.3473$, the actual risk equals 0.0363. This discontinuity of risk is due to the confidence interval of 5 successes for 30 trials, equal to $[0.0564\ , 0.3472]$, which contains 0.3471 but not 0.3743. For a constant number of trials n, the actu-

al α risk has discontinuities along the true proportion p, at all bounds of confidence intervals. For a constant p, oscillations of the actual α risk along n exist.

9.1.1.2 Skewed distribution

The binomial $\mathcal{B}(n, p)$ distribution is skewed except for $p = 0.50$. The skewness tends to positive infinity when p tends to 0, and to negative infinity when p tends to 1. For a constant expected number of successes $\lambda = np$, the binomial distribution skewness increases along n and approaches the Poisson distribution when $n \to \infty$. Consequently, for a constant λ, biases of normal approximations (Wald and Wilson) worsen when n increases.

9.1.1.3 Relationship between variance and expectation

The expectation of a binomial proportion equals p while its variance equals $\frac{p(1-p)}{n}$. When the expected number of successes np is close to 0 or n, the sample variance $\frac{\hat{p}(1-\hat{p})}{n}$ fluctuations are large and dependent on $\hat{P}$ fluctuations. Estimators approximating the population variance to the sample variance are biased. This motivated the use of variance stabilizing transformations (10,112).

9.1.2 Evaluation criteria rationale

9.1.2.1 Conditional or average risk control

The Clopper-Pearson (31), Blyth-Still (15) or Blaker (12) intervals are strictly conservative, meaning that the actual risk of the interval not containing the true proportion, conditional to a fixed n sample size and a fixed p proportion, is always greater than the nominal risk, usually equal to 0.05. Agresti and Coull (4) analyzed the average coverage under the assumption that the true proportion is variable, following a uniform distribution on $]0\ ,\ 1[$. Intervals controlling the average coverage have actual risk oscillations centered on the nominal risk, and so, are less conservative and have shorter lengths than strictly conservative intervals whose actual risk oscillations are always below the nominal risk.

There is often information about the proportion being analyzed. For example, a drug may be predicted to have a serious adverse event frequency of less than 10% and a response rate greater than 50%. The analysis of the mean coverage under a uniform distribution could mask a tendency to underestimate small proportions such as adverse effects by a tendency to overestimate high proportions such as efficacy. Therefore, we consider it appropriate to control separately the actual risks associated with the estimation of very different proportions. Moreover, as shown by the heterogeneity in meta-analyses, a proportion is rarely constant from one medical experiment to another. The sampled population differs in space and time and the interventions and measures depend on the operators and their environment. This is modeled in meta-analyses by random effects models. The true proportion is assumed to be a random variable of the experiment. This smoothes the actual risk that the interval does not contain the true proportion. That actual risk, in a random effect model, will be called the *local average risk*.

9.1.2.2 Equal or unequal-tailed two-sided intervals

We can decompose the actual risk α' that the confidence interval does not contain the true p proportion into two risks: The actual risk α'_l that the lower bound of the confidence interval is greater than p and the actual risk α'_u that the upper bound of the confidence interval is less than p. In other words, α'_l is the risk of overestimating the true proportion and α'_u is the risk of underestimating it. These two risks can be interpreted as the actual risks of one-sided confidence intervals built with either the lower or the upper bound. The sum of these two actual risks is equal to the actual two-sided risk $\alpha' = \alpha'_u +$

α'_l. A confidence interval is one-sided when either α'_u or α'_l is zero. All confidence intervals are two-sided. Among the two-sided confidence intervals, the ratio between α'_l and α'_u is not always the same. Equal-tailed confidence intervals such as the Clopper-Pearson interval, are constructed so that α'_l and α'_u are each close to $\frac{\alpha}{2}$. Unequal-tailed two-sided interval estimators, such as Blyth-Still or Blaker, can have are very unbalanced risks, with, for an expected number of successes np close to zero, $\alpha'_l = 0$ and $\alpha'_u = \alpha'$.

Unequal-tailed confidence intervals control the sum of the two one-sided risks without controlling them separately. In the context of adverse effects to a drug, this means that an unequal-tailed interval may compensate for the tendency to underestimate frequently the rate of adverse effects, by a tendency to overestimate it rarely. An equal-tailed interval can be transformed in an unequal-tailed interval by moving down both its upper and lower bounds. We believe that the two one-sided risks are not equivalent and it is not wise to move both bounds in a direction suggesting a lower rate of adverse effects. This is why we will analyze each one-sided risk separately, seeking for estimators controlling both risks.

9.1.2.3 Interval half-length

For an equal coverage error, a narrower interval is preferred. The expected length of the interval has been analyzed (4,21,83). As we will analyze separately each one-sided risk, we will also analyze separately the expected distance between each bound (lower or upper) and the point estimate $\hat{p}$; for Wald's interval (symmetrical), it is half the length of the interval.

9.2 Materials & methods

9.2.1 Definition of risks and interval lengths

We will first define one-sided conditional risks as actual risks that the confidence interval is completely below (or above) the true proportion p when the sample size and p are constant from one experiment to another. Second, we will define one-sided local average risks as actual risks that the confidence interval is completely below (or above) the true proportion P when the sample size is constant from one experiment to another but the true proportion P is a random variable fluctuating around an expected true proportion p_0. Third, we will define one-sided random sample risks as actual risks that the confidence is strictly below (or above) the true proportion p when the true proportion p is constant from one experiment to another but the sample size N is a random variable fluctuating around an average sample size n_0. Similarly, we define three different expected interval half-lengths (conditional, local average and random sample).

Let x be the realization of a variable $X_{n,p} \sim \mathcal{B}(n;p)$ where n represents the sample size, x the observed number of successes and $\hat{p} = x/n$ the observed proportion of successes. Let

$$IC_{1-\alpha}(x,n) = [L_{1-\alpha}(x,n); U_{1-\alpha}(x,n)] \tag{2}$$

denote the confidence interval at the nominal risk α for x successes in n trials. For constant n and p, we denote

$$\alpha'_l(p,n,\alpha) = P\left(L_{1-\alpha}\left(X_{n,p},n\right) > p|n,p\right) \tag{3}$$

the actual risk that the interval is completely above p. Similarly, we denote

$$\alpha'_u(p,n,\alpha) = P\left(U_{1-\alpha}\left(X_{n,p},n\right) < p|n,p\right) \tag{4}$$

the actual risk that the interval is completely below p. We denote

$$\alpha'(p,n,\alpha) = \alpha'_u(p,n,\alpha) + \alpha'_l(p,n,\alpha) \tag{5}$$

the actual risk that the interval does not contain p. These three risks (α'_l, α'_u and α') are called the *conditional risks*.

Let us define the *conditional expected lower half-length*

$$w'_l(p,n,\alpha) = E\left(\frac{X_{n,p}}{n} - L_{1-\alpha}\left(X_{n,p},n\right)|n,p\right) \tag{6}$$

equal to the expected distance between the point estimate and the confidence interval lower bound. Similarly, we define the *conditional expected upper half-length*

$$w'_u(p,n,\alpha) = E\left(U_{1-\alpha}\left(X_{n,p},n\right) - \frac{X_{n,p}}{n}|n,p\right). \tag{7}$$

Let us note P a random proportion whose logit is normally distributed with a $\log(OR_S)$ standard deviation. Let $p_0 = E(P)$ be the expected proportion. We define the *right local average risk*

$$\alpha''_u(p_0,n,\alpha) = E(\alpha'_u(P,n,\alpha)|n) \tag{8}$$

the *left local average risk*

$$\alpha''_l(p_0,n,\alpha) = E(\alpha'_l(P,n,\alpha)|n). \tag{9}$$

and the *two-sided local average risk*

$$\alpha''(p_0,n,\alpha) = \alpha''_l(p_0,n,\alpha) + \alpha''_u(p_0,n,\alpha). \tag{10}$$

It is the probability that a confidence interval around $\hat{p}$ does not contain p in a two steps experiment. In the first step, a p proportion is realized from the P random variable. In the second step, a $\hat{p}$ proportion of successes is realized in a random sample of size n with an actual proportion of successes p. The sample size n is held constant in all experiments. Similarly, we define $w''_l(p,n,\alpha)$ and $w''_u(p,n,\alpha)$ the *local average half-lengths*. The constant OR_S will be set at 1.20. Sensitivities analyzes will be conducted with other OR_S values.

We define a random N variable having a discrete distribution. This distribution is defined from a latent log-normal variable rounded to the nearest integer. Let us denote $n_0 = E(N)$ the expectancy of the same size N and SR_S the geometric standard deviation of the latent log-normal variable. The geometric standard deviation SR_S will be set at 1.20. We define the *random sample right average risk* as

$$\alpha'''_u(p,n_0,\alpha) = E(\alpha'_u(p,N,\alpha)|p) \tag{11}$$

the actual risk that the confidence interval is completely below p, and the *random sample left average risk* as

$$\alpha'''_l(p,n_0,\alpha) = E(\alpha'_l(p,N,\alpha)|p). \tag{12}$$

Let the *relative conditional expected left half-length* $v'_l(p,n,\alpha)$ be the ratio between the expected lower half-length of an interval estimator and the expected lower half-length of the Clopper-Pearson mid-

P confidence interval for the same set of parameters p, n and α. Similarly, we define v'_u, v''_l, v''_u taking the Clopper-Pearson mid-P reference.

The nominal α risk will be set at 0.05. The conditional risks and conditional expected half-lengths will be computed from the exact binomial distribution $\mathcal{B}(n,p)$. The local average risks and local average half-lengths will be approximated by numerical integration from 512 values the conditional risks or conditional expected half-lengths. These 512 values will be uniformly spaced on the logistic scale.

As an indication, horizontal lines are drawn on the figures at risks 0.025, $0.025 \times 1.50 = 0.0375$, and $0.025/1.50 = 0.016667$. They may serve as one-sided bias tolerance limits.

9.2.2 Bibliographic research

A bibliographic search with the keywords "binomial", "confidence" and "interval" was conducted on the Google® Scholar database in July 2017, looking for articles defining confidence intervals of a binomial proportion. The 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 the systematic review articles were followed in order to find the original references. The methods considered as redundant are not presented, such as the approximations of Molenaar (76), Pratt (14) and Blyth's equation C improving the Molenaar approximation (14) or the Chen interval (30) which are all approximations of the Clopper-Pearson interval (31). All these approximations are precise and the results would be indistinguishable from those of the Clopper-Pearson interval. The interval estimators of Zhou (114) from Bolboaca (16), Brenner (18), Lang (65) and Crow (35) have not been implemented because of complexity of implementation, or an *a priori* redundancy of these intervals to an estimator already described.

9.2.3 Interval definitions

We denote by $k = \min(x, n-x)$ the least of the number of successes or failures. Table IX describes the lower bounds of the nine confidence interval estimators that will be analyzed. Forty-six other confidence interval estimators, including bootstrap confidence intervals and closed-form skewness corrected intervals have been analyzed in Appendix I.

The modified Wilson interval, based on a Poisson approximation for small values of k, was described by Brown, Cai and DasGupta, page 112 (21) but these authors do not specify the threshold x^* for $n > 100$. Since Brown did not analyze the behavior of the interval for $n > 100$, he did not make a recommendation (personal communication of the author). The already sufficient convergence to the Poisson distribution made us keep the $x^* = 3$ threshold for $n > 100$.

The Wald logit interval is indefinite for $k = 0$; We supplement its definition by the Clopper-Pearson interval for $k = 0$. The likelihood ratio interval is defined by inversion of a χ^2 test on the deviance function; This interval is well defined even for $k = 0$, but in order to compare its performances with those of the Wald logit interval, we applied the same Clopper-Pearson substitution for $k = 0$. The unmodified likelihood ratio interval is presented in Appendix I Figures VII and IX.

Bounds outside the $[0,1]$ interval were set to nearest valid proportion, zero or one.

9.3 Results

9.3.1 General results

The one-sided local average risks are shown on Figure II. Different sample sizes are analyzed in this figure but the results for an expected number of successes above 32 are not presented. For a typical odds ratio between the true proportions of two experiments $OR_S = 1.20$, the local average risks are smoothed by true proportion fluctuations (Figure II). Graphically local average risk oscillations can be viewed for an expected number of successes below two. For a less random true proportion with $OR_S = 1.05$, large amplitudes oscillations are graphically visible when the expected number of successes is below eight (Appendix I, Figure II) with a maximum local average one-sided risk equal to 0.0381 for the Clopper-Pearson mid-P interval, greater than the nominal 0.025 one-sided risk and the 0.0375 indicative tolerance limit.

The one-sided risks are shown in Figure III for a sample size $n = 2048$. The two-sided risk, equal to the sum of the two one-sided risks (green and red) is shown in blue in Figure III except when a one-sided risk is zero; in that case the two-sided risk is equal to the other one-sided risk. In Figure III, as in Figure II, coverage bias is higher in absolute value for large values of n. The results for $n = 2048$ are very close to the asymptotic Poisson intervals (see article Figure III and Appendix I, Figure II and IV).

The random sample average risks α''' of a constant p proportion with a random N sample size are close to local average α'' risks for a random P proportion with a constant n sample size (see article Figure II and Appendix I, Figure V).

The local average expected interval lengths are shown in Figure IV. Local average half-lengths w_l'' and w_u'' are almost equal to conditional expected interval half-lengths w_l' and w_u' because the expected interval lengths have no discontinuity along p. For instance, $\left|1 - \frac{w_l''}{w_l'}\right|$ and $\left|1 - \frac{w_u''}{w_u'}\right|$ were less than 0.01 for Wald's interval for all n, p and α values shown in Figure IV. Expected interval lengths mirror the local average risks. Where an interval is shorter than another, it has a higher local average risk.

9.3.2 Specific interval results

Wald's interval is unequal-tailed. The right local average risk α_u' tends to one when the expected true proportion p_0 tends to zero because Wald's interval length is zero when the number of successes x is zero. Wald's interval has a high two-sided bias and is a very unbalanced unequal-tailed interval (see Figure II and Figure III). Modified Wilson's, Wald logit and Blaker's intervals have lower absolute bias but are not equal-tailed either. The biases of these three intervals estimators are opposite to Wald's interval estimator bias, while the modified likelihood ratio interval has a small bias in the same direction as Wald's interval. The Wald logit interval has a local average risk spike equal to 0.097 for $n = 2048$ and an average number of successes $np_0 = 0.11$. Wilson's interval has a local average risk spike equal to 0.161 for $n = 2048$ and $np_0 = 0.176$ (see Appendix I, Figure VI) but Brown's modification removes this spike (see Figure II).

Blaker's and Clopper-Pearson intervals are both conservative (Figure II) but Blaker's interval is slightly less conservative. For proportions close to zero, Blaker's interval conditional right α_u' risk is very close to Clopper-Pearson α_u' risk but Blaker's α_l' risks can get much higher with one-sided risk oscillations up to 0.05 while Clopper-Pearson's interval one-sided risk oscillations never exceed 0.025.

Risks of Bartlett's Arc-Sine interval, modified equal-tailed Jeffreys and Clopper-Pearson mid-P intervals are close to each other. For proportions close to zero, the modified equal-tailed Jeffreys interval has a local average right half-length w_u'' slightly higher than Bartlett's Arc-Sine and Clopper-Pearson mid-P intervals. The right local average α_u'' risk for an expected number of successes $np_0 = 4$ is close to the nominal risk for Bartlett's Arc-Sine and Clopper-Pearson mid-P but lower for the modified equal-tailed Jeffreys interval. Bartlett's Arc-Sine interval gets closer to the nominal risk than Anscombe and Freeman-Tukey intervals (see Figure VI in Appendix I).

The modified likelihood ratio interval has a mild one-sided local average bias, lower than the Wald, modified Wald logit and modified Wilson intervals. The unmodified approximate likelihood ratio interval shown has a high one-sided risk spike for an expected number of successes close to 2.3 (Appendix I, Figure VII).

Percentile and basic bootstrap confidence intervals (Appendix I, Figure VIII and IX) have high local average biases. Basic bootstrap has higher risk biases than Wald's interval while percentile bootstrap is slightly less biased than Wald's interval. Unmodified BC_a bootstrap is highly biased and is not equivariant; a modification of equation 3.2 of Efron (41) taking in account the discreteness of the bootstrap distribution, make the interval equivariant and reduces the bias. Modified BC_a bootstrap, smoothed BC_a bootstrap and studentized BC_a bootstrap are all conservative for local average risks. Due to division by zero errors, studentized bootstrap cannot be computed for $\min(x, n - x) \leq 4$.

9.4 Discussion

9.4.1 Originality of this work

Newcombe distinguished mesial (the confidence interval bound nearest to 0.50) and distal (the confidence interval bound nearest to 0 or 1) one-sided risks (77) but averaged them over the whole $]0\,,1[$ interval assuming a random proportion with a uniform distribution. Agresti and Coull (4) analyzed a random proportion for a Beta distribution with expectancy equal to 0.10 but did not analyze other theoretical random proportions nor analyzed one-sided risks. To our knowledge, the influence of the sample size for a fixed expected number of successes had not been graphically presented in any systematic review of binomial proportion confidence intervals. These new evaluation criteria are likely to change the recommendations concerning the use of these intervals.

9.4.2 Conditional or local average risk: which one to control?

Agresti and Coull analyzed average risks for a random proportion with an uniform distribution (4) but they also analyzed the average risk for a random proportion with a Beta distribution whose mean is 0.10. This is similar to the local average risk for an expected true proportion $p_0 = 0.10$. This may be interpreted as a Bayesian *prior* distribution for the theoretical proportion although the confidence interval and risks are frequentist. Our work analyzes separately right and left local average risks and has a different presentation of sample size and the expected true proportion, making it possible to show the risk curve for any number of expected successes np_0 below 32.

We analyzed the conditional risks, the average risks for a random proportion or a random sample size. If either the theoretical proportion *P* or the sample size *N* is random, the actual risks are close to the nominal risks as long as *P* or *N* have enough variance and the expected number of successes is not too low (Appendix I, Figures I and II). Even though the real proportion is rarely constant, a confidence interval may be used to test a hypothetical proportion defined by a protocol. For example, a protocol may seek to prove that a proportion is below 10%. The required sample size may be defined in the

protocol, but it is rarely computed the same way in different studies and observations included in the final analysis are often slightly below or above the computed sample size. If few successes are expected, the sample size variance is expected to be small and a hypothesis test is performed, a strictly conservative procedure such as Clopper-Pearson may be preferred.

9.4.3 The best confidence interval estimator

Wald's interval risk biases are very high even for an expected number of successes equal to 32. Moreover, Wald's interval is unequal-tailed. The modified Wilson interval and modified Wald logit intervals have much lower local average risk biases but are unequal-tailed.

The R and SAS statistical software use the likelihood ratio test to estimate logistic regression coefficients confidence intervals, although they cannot compute it when $\min(x, n-x) = 0$. When the Clopper-Pearson interval is substituted for $\min(x, n-x) = 0$, this method has a mild one-sided bias, lower than Wald's method after logit transformation. This suggests that likelihood ratio confidence intervals may be better for generalized linear models. This is consistent with Agresti (3) recommendation suggesting the use of the likelihood ratio interval when Wald's interval is discordant with it.

Blaker's and Clopper-Pearson intervals are both conservative. Blaker's interval is strictly conservative for the conditional two-sided risk while Clopper-Pearson is strictly conservative for each of the one-sided risk and for the two-sided risk. The Clopper-Pearson interval may be larger than Blaker's interval, but it is the shortest equal-tailed strictly conservative interval (100,108). Blaker's interval is always less conservative on both sides as Blaker's intervals are always contained in Clopper-Pearson intervals (12).

Bartlett's Arc-Sine, Clopper-Pearson mid-P and modified Jeffreys equal-tail intervals are close to each other. They properly control the local average risks and are equal-tailed. In our opinion these are the three best interval estimators in most scenarii. As the three intervals are practically equivalent, the selection may be done on the theoretical background.

The modified equal-tailed Jeffreys interval is based on *ad hoc* modifications (21). The lower bound of the confidence interval is not strictly monotone with the number of successes x. Indeed, the lower bound of the confidence interval is zero for $x = 0$ and for $x = 1$.

Bartlett's Arc-Sine intervals controls the local average risk better than Anscombe and Freeman-Tukey intervals while the latter are based on transformations that better stabilize variance (8,46,112). The faster decrease in variance in Bartlett's transform, close to zero and one, compensates the liberal behavior due to the poor normal approximation of the transformed variable. Normalization and variance stabilization are antagonist for the binomial distribution (40).

The Clopper-Pearson mid-P interval has a simple theoretical background and is associated to a hypothesis test. Its theoretical background is the same as the Clopper-Pearson interval: inversion of an equal-tailed test based on the exact binomial distribution. The Clopper-Pearson mid-P procedure has been generalized to logistic regressions: the exact logistic regression described by Hirji (55) is implemented as the exlogistic command in the Stata statistical software (95) and in proc logistic in the SAS statistical software (38). The Clopper-Pearson mid-P interval is equivalent to the confidence bounds of an intercept-only mid-P exact logistic regression, after inverse logit transform. Similarly, the Clopper-Pearson estimator can be computed from an exact logistic regression.

This theoretical background makes the Clopper-Pearson mid-P interval a bit more attractive than Bartlett's Arc-Sine and the modified Jeffreys equal-tailed interval. We suggest to use the Clopper-Pearson mid-P interval in all scenarii except when comparing an observed proportion to a theoretical proportion in strongly controlled experimental conditions or in a strongly regulated domain such as clinical trials; in that case, the Clopper-Pearson interval may be best. 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 strict conservatism of the Clopper-Pearson interval may be safer.

The Clopper-Pearson mid-P interval 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 Texas Instruments Ti 83/84 are given in Appendix II. They are licensed under the free software Create Commons CC0 license.

9.4.4 Relative interval length: rationale

Expected interval lengths, conditional to a theoretical proportion, have been graphically reported in literature (21,83). Confidence interval length is much higher for proportions around 50% than for a small expected number of successes. The rate between expected interval length for $p = 0.50$ and expected interval length for $np = 5$ gets higher as n grows. When n is large, this makes it unpractical to graphically compare expected length of two or more interval estimators for small values of np as lengths are graphically too close to zero. Representing, for every theoretical proportion, the ratio between expected half-length of the interval estimator and a reference interval estimator (v_l' and v_u'), greatly reduces the variance, making ratios between interval lengths easier to view. The Clopper-Pearson mid-P has been chosen as reference interval estimator because it has a low coverage bias, making it easier to interpret expected length ratios without having to keep in mind the coverage bias of two interval estimators.

Expected interval half-lengths (Figure II) reflect one-sided risks (Figure IV). None of the studied interval estimators are chaotic enough to get high local average risks α_l'' and α_u'' with long expected half-lengths w_l'' and w_u''.

9.4.5 Equal-tailed and unequal tailed intervals

9.4.5.1 Interpretation

Interpretation of equal-tailed intervals differs from unequal tailed intervals. An unequal tailed interval may be shorter than an equal-tailed interval but only controls the two-sided α risk. The example of the Bayesian highest posterior density credible interval with Jeffreys prior (HPD) may be instructive. It can be interpreted as a frequentist unequal-tails confidence interval. For a small observed proportion, it can be computed from an equal-tails credible interval with Jeffreys prior, by lifting down both the lower and upper bounds. Lifting down the upper bound increases the α_u' risk of under-estimating the true proportion, but this is compensated by a decreased α_l' risk of over-estimating the true proportion. The interval is shortened as the derivative of the risk is higher on the lower side than on the upper side. In Bayesian terms, the credibility is unchanged and in frequentist terms, the $\alpha_u' + \alpha_l' = \alpha'$ risk is kept close to the nominal. In biostatistics, risks of overestimating are rarely equivalent to risks of underestimating and swapping one for the other may not be wise.

Consider the example of the proportion of adverse effects to a drug. With an equal-tailed interval, one can conclude that the frequency of side effects is lower than the upper bound with a one-sided risk α'_u close to $\frac{1}{2}\alpha$. If this frequency is low enough the drug may be considered safe with a controlled $\frac{1}{2}\alpha$ risk. If the lower bound of the frequency seems too high to a drug regulating agency, it may consider that there is enough evidence of a high frequency of severe adverse effects to take action, with a controlled $\alpha'_l \approx \frac{1}{2}\alpha$ risk. An equal tailed interval is the intersection of two one-sided intervals and can be interpreted as such. The α'_l or α'_u risks in same scenario with an unequal tailed interval is unspecified between 0 and $\alpha' \approx \alpha$. Therefore, for one-sided interpretations, unequal-tailed two-sided confidence intervals may have a double risk. For low or high proportions, a two-sided unequal-tailed interval may become close to a one-sided interval, but the direction of the one-sided interval is not necessarily known. If a decision requires that a proportion be above a threshold and, at the same time, be below a second threshold, then, the interpretation is two-sided and unequal-tailed intervals may be used. This is not the common occurrence in biostatistics. When making the uncommon hypothesis test $p_0 < p < p_1$, the unequal-tailed unbalance may sometimes reduce the power when the true proportion is close to p_0 or p_1 and the unbalance tends to increase the interval length on this side.

These issues have been mentioned by Kaiser (59) who discuss the directional interpretation of two-sided tests comparing two means μ_X and μ_Y in a three hypothesis theory:

$$H_1: \mu_X < \mu_Y$$

$$H_2: \mu_X = \mu_Y$$

$$H_3: \mu_X > \mu_Y$$

Lombardi and Hurlbert (71) review the problem of one-sided, two-sided unequal-tailed and two-sided equal-tailed tests. In page 452, Lombardi and Hurlbert note that "If we insist on having P values that formally apply to directional conclusions, in most situations we can simply double the P values yielded by the two-tailed test". Indeed, but in that case, the power gain due to the voluntary tails unbalance of Blaker's or Sterne's tests or confidence interval is reversed.

9.4.5.2 Paradoxes

A number of unequal-tailed two-sided intervals have been built on the exact binomial distribution (96,35,15,12,58,93,66,109). A comparison of unequal-tailed strictly conservative exact binomial intervals is shown in Appendix I.

Blaker's and Sterne's intervals are based on the inversion of a binomial test based on a discontinuous p-value function of the theoretical proportion (43). These tests can reject a theoretical proportion at the α risk while lesser or greater values are not rejected (13). Consequently, confidence regions may be unions of several disjoint intervals. Crow's, Blyth-Still and Kabaila-Byrne intervals may not be nested (13,66) meaning that a 95% confidence interval is not always contained in the 99% confidence interval and sometimes, an theoretical proportion may be rejected at risk 0.01 but not at risk 0.05 (100,107).

Sterne's and Blaker's confidence intervals are nested (12,96) but adding an observation to a sample may reduce the lower bound of the confidence interval whatever its status. A statistically significant result can become insignificant when an observation is added whatever its status. Lecoutre and Poitevineau propose a procedure resistant to this paradox (66).

9.4.6 Other desirable properties

These few desirable properties of confidence intervals have not been mentioned:

- consistency of confidence interval inference and p-values of hypothesis tests (107);
- generalizability to bivariate and multivariate models;
- theoretical simplicity and the existence of an analytical solution. This motivated Agresti and Coull when they defined their interval (4,5);
- equivariance: The consistency of the confidence interval of successes with the confidence interval of failures (15,91,93);
- monotonicity of interval bounds along x, n and α (107);
- deterministic procedure. Randomized intervals, based on a computer-generated random number, produce different intervals for the exact same data set. Conditional risks are smoothed, but the practical use of these interval requires a rigorous analysis difficult to apply in practice (97).

The Clopper-Pearson mid-P interval verifies all these properties except property (3).

Some of these properties are described for all analyzed intervals in Appendix III.

9.4.7 Validity conditions of Wald's interval

Wald's interval has low coverage even for quite high number of success and failures, leading some authors to recommend against teaching this interval in elementary courses, favoring Agresti-Coull or Wilson's interval (4,5,21). Recent textbooks such as Fritz and Berger in 2015 (47) may mention that Wald's interval is biased, citing Agresti and Coull, but then give the old validity condition $np, n(1-p) \geq 5$. Updated validity conditions may be a pedagogical argument against use of Wald's interval. That is why validity conditions of Wald's interval are assessed in Appendix I (Tables VII to IX). The simple condition $\min(x, n-x) > 40$ is enough to control the one-sided local average risks for a 95% confidence interval provided that the Clopper-Pearson mid-P interval is used when the validity threshold is not reached. The control is not perfect as the actual one-sided risk may be 1.5 times higher (that is 0,0375) than the nominal risk (0,025).

9.4.8 Poisson distribution

The Binomial distribution $\mathcal{B}(n,p)$ is asymptotically equivalent to a Poisson distribution when $n \to \infty$ and the number of expected successes np is held constant. The binomial confidence intervals for the large sample size $n = 2048$ are indistinguishable from the asymptotic Poisson confidence intervals (Appendix I, Figure III).

9.4.9 Continuity correction

Continuity corrections make intervals more conservative on both sides. Without continuity correction, the Anscombe Arc-Sine interval (8,73) has one-sided conditional risk oscillations averaging around the nominal $\frac{\alpha}{2}$ level. The Anscombe Arc-Sine continuity correction proposed by Pires (83) moves the oscillations below the nominal level (Appendix I, Figure VIII). On Wilson's and Wald's intervals, the continuity correction (Appendix I, Figures VI and VIII) reduces, in absolute value, the α risk bias on one side, but increases it on the other side.

9.4.10 Bootstrap

According to Carpenter (26), 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 percentile bootstrap is the order $O\left(\frac{1}{n}\right)$. The 'normal' bootstrap is equivalent to Wald's interval. The asymptotical convergence rate of these intervals is reflected in the very high biases of the percentile and basic bootstrap intervals and much lower biases of the modified BC_a and studentized bootstrap intervals. No bootstrap interval controls the nominal risk as well as the Clopper-Pearson mid-P interval.

9.5 Conclusion

The binomial proportion confidence interval problem may look simple, but some aspects of the problem may be missed. The equal-tailed or unequal-tailed issue and the local average risk control or conditional risk control also applies to more complex statistics such as coefficients of multivariate logistic regressions.

Wald's interval should not be used. For a binomial proportion 95% confidence interval of x successes for n trials, the validity condition $\min(x, n-x) > 40$ partially controls the one-sided local average risks. Even under this condition, local average risks can reach 1,5 times the nominal risk.

The Clopper-Pearson mid-P interval controls the local average risk while the Clopper-Pearson interval is the shortest strictly conservative equal-tailed confidence interval. We recommend the use of the Clopper-Pearson mid-P interval in all scenarii 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, we recommend the use of the Clopper-Pearson interval. Free software implementations of the Clopper-Pearson mid-P intervals are given in Appendix II.

9.6 Table and figures

Name	Lower bound $\boldsymbol{L_{1-\alpha}(x,n)}$
Wald[a]	$\max\left(0, \frac{x}{n} - \kappa\sqrt{\frac{x(n-x)}{n^3}}\right)$
(21) Modified Wilson[abc]	$\begin{cases} \frac{1}{2n}\chi^2_{\alpha,2x}\ if\ 1 \le x \le x^* \\ \frac{x + \frac{\kappa^2}{2} - \kappa\sqrt{\frac{x(n-x)}{n} + \frac{\kappa^2}{4}}}{n+\kappa^2}\ otherwise \end{cases}$
(8,10) Bartlett Arc-sine[a]	$\sin^2\left(\max\left(0, \operatorname{asin}\left(\sqrt{\frac{x+\frac{1}{2}}{n+1}}\right) - \frac{\kappa}{2\sqrt{n+\frac{1}{2}}}\right)\right)$

(21) Modified Wald logit[ad]	$\begin{cases} \text{logitinv}\left(\log\left(\frac{x}{n-x}\right) - \kappa\sqrt{\frac{n}{x(n-x)}}\right) \; if \; 0 < x < n \\ \sqrt[n]{\alpha/2} \; if \; x = n \\ 0 \; if \; x = 0 \end{cases}$
(106)Modified likeli-hood ratio[a]	$\begin{cases} \inf\left\{q \middle\| \log\left(\left(\frac{x}{nq}\right)^x \left(\frac{n-x}{n(1-q)}\right)^{n-x}\right) \leq \frac{1}{2}\kappa^2\right\} \; if \; x < n \\ \sqrt[n]{\alpha/2} \; if \; x = n \end{cases}$
(21)Modified equal-tailed Jeffreys[e]	$\begin{cases} \beta iCDF(\alpha/2; x + 1/2, n - x + 1/2) \; if \; 2 \leq x < n \\ \sqrt[n]{\alpha/2} \; if \; x = n \\ 0 \; if \; x \leq 1 \end{cases}$
(12)Blaker[f]	$\inf\{q \mid \text{bpval}(q, x, n) \leq \alpha\}$
(21,31)Clopper-Pearson[e]	$\beta iCDF\left(\frac{\alpha}{2}; x, n - x + 1\right)$
(11,63)Clopper-Pearson mid-P[g]	$\inf\left\{q \middle\| \min\left(BCDF(x; n, q), 1 - BCDF(x-1; n, q)\right) - \frac{1}{2}BPF(x; n, q) \leq \frac{\alpha}{2}\right\}$

Table IX:Definition of the lower bounds of the confidence intervals, the upper bounds being defined by equivariance $\boldsymbol{U_{1-\alpha}(x,n) = 1 - L_{1-\alpha}(n-x,n)}$.

[a]We denote by $\kappa = z_{1-\alpha/2}$ the quantile $1-\alpha/2$ of the normal distribution $N(0,1)$

[b] The x^* threshold equals 2 for $n \leq 50$ and equals 3 for $n > 50$

[c]$\chi^2_{q,df}$ is the q quantile of the χ^2 distribution with df degrees of freedom

[d]The reciprocal of the logistic transformation is defined by $\text{logitinv}(t) = \frac{\exp(t)}{1+\exp(t)}$

[e] $\beta iCDF(q; \alpha, \beta)$ is the q quantile of the beta distribution whose shape parameters are α and β

[f] The function bpval (Blaker's p-value) is defined as follows :

$$\text{bpval}(q, x, n) = \begin{cases} \min\left(1, BCDF(x; n, q) + 1 - BCDF(BiCDF(1 - BCDF(x; n, q); n, q); n, q)\right) \; if \; q \geq x/n \\ \text{bpval}(1 - q, n - x, n) \; if q < x/n \end{cases}$$

[g]We denote by $BPF(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$ the binomial probability mass function, $BCDF(k; n, p) = \sum_{x=0}^{k} BPF(x; n, p)$ the binomial cumulative distribution function and $BiCDF(q; n, p) = \min\{k | BCDF(k; n, p) \geq q\}$ the binomial quantile function.

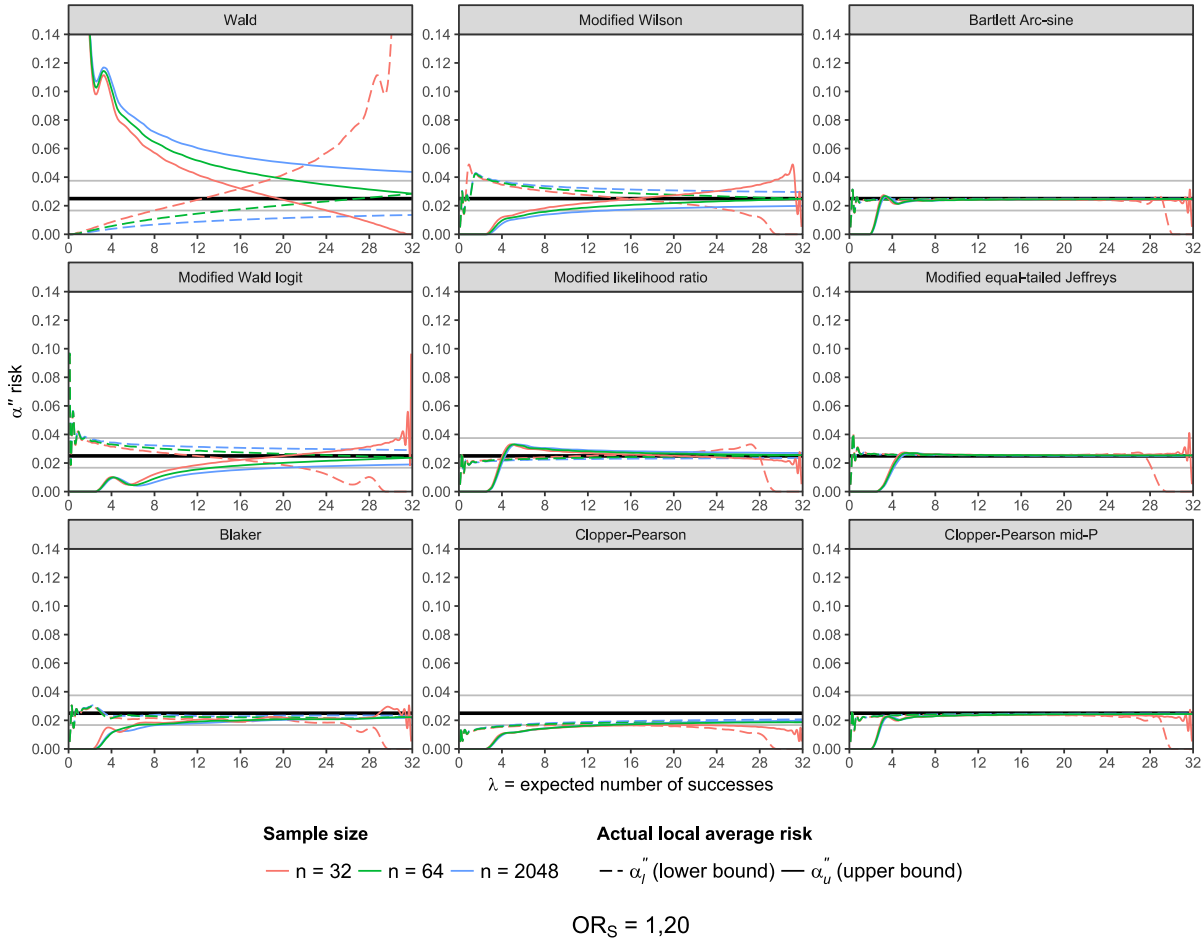

Figure II: One-sided local average risks of nine 95% confidence interval estimators according to different sample sizes (red for $n = 32$, green for $n = 64$ and blue for $n = 2048$), with a theoretical P proportion following a random logit-normal model with a typical odds ratio of the true proportion between two experiments equal to $OR_S = 1.20$. The abscissa is the expected number of successes np_0 and the ordinate is the risk that the lower bound of the confidence interval is greater than the true proportion p (left local average risk: dashed lines) or the risk that the upper bound of the confidence interval is lower than the true proportion p (right local average risk: solid line).

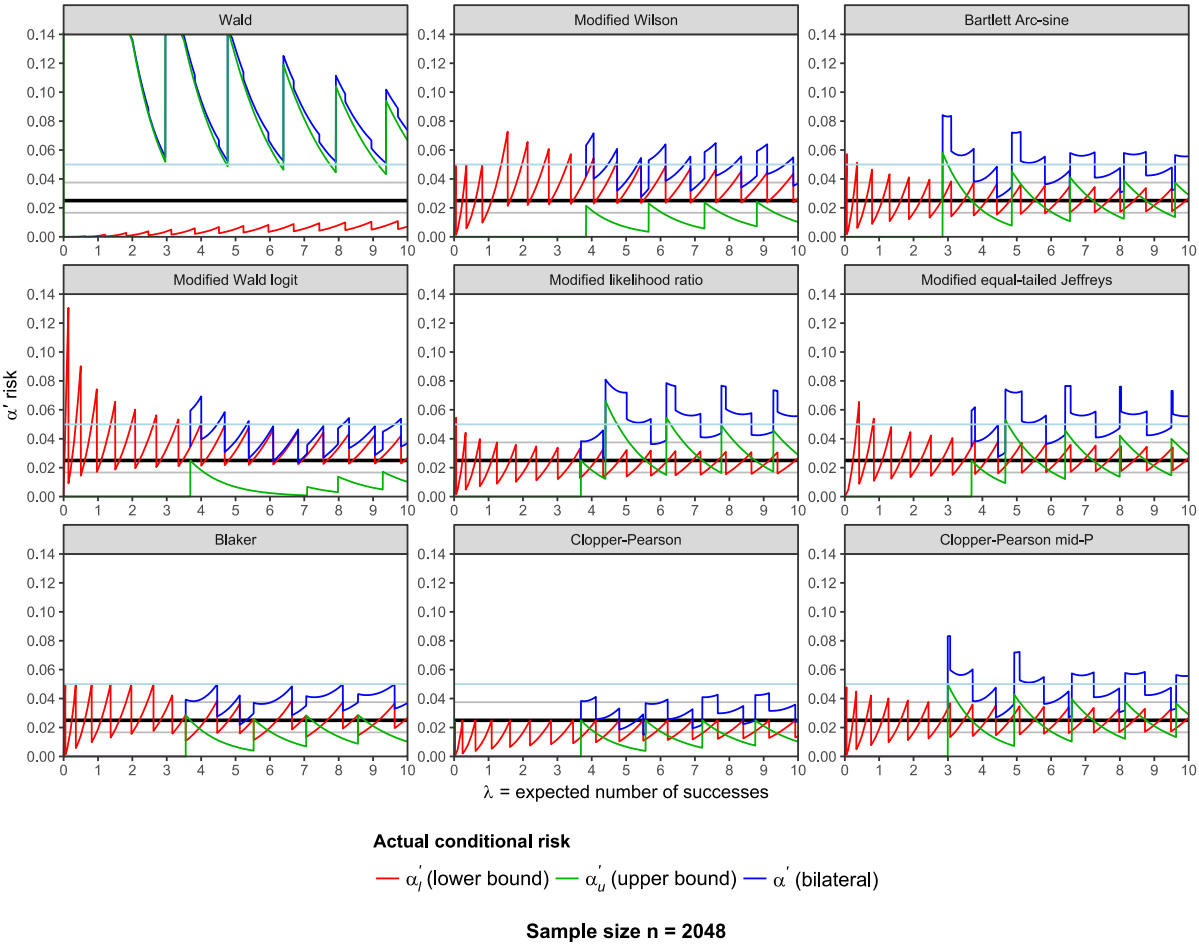

Figure III: One-sided and two-sided conditional risks of nine 95% confidence interval estimators for a sample of size $n = 2048$ and a constant theoretical p proportion. The abscissa is the expected number of successes np and the ordinate is the risk that the lower bound of the confidence interval is greater than the true proportion p (left conditional risk: red), the risk that the upper bound of the confidence interval is lower than the true proportion p (right conditional risk: green) or the risk that the confidence interval does not contain the true proportion p (two-sided risk: blue).

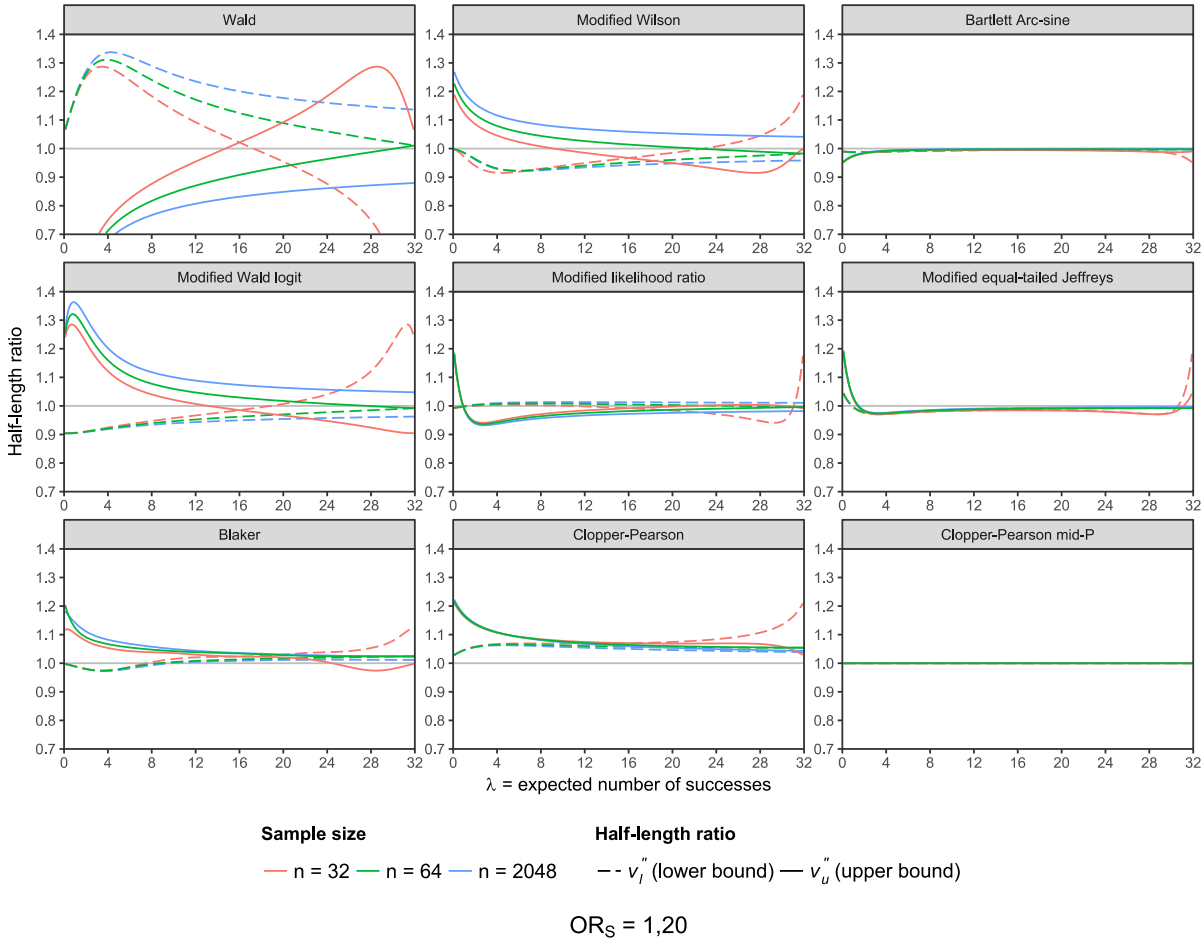

Figure IV: Relative local average half-lengths of nine 95% confidence interval estimators for a sample of size $n = 2048$ and a constant theoretical p proportion. The relative half-length is the local average half-length of one of the nine intervals divided by the local average half-length of the Clopper-Pearson mid-P interval for the same x, n and p_0 parameters. The abscissa is the expected number of successes np and the ordinate is the left relative local average half-length (dashed lines) or the right relative local average half-length (solid lines).

10 Annexe 5 : résumé des propriétés

Les propriétés des 55 estimateurs d'intervalles analysés ont été résumées dans un fichier Microsoft Excel synthétique disponible aux adresses Web spécifiées en page 111.

http://tinyurl.com/ybkqysm3

http://andre.gillibert.fr/owncloud/public.php?service=files&t=4a8a49cba9e183ed5c7be32f3087cf69

https://mega.nz/#F!o0IyAbgK!beyM7_hFDnRzWW1gyds7Vg

11 Bibliographie

1) Agresti A. Categorical Data Analysis. John Wiley & Sons 2003

2) Agresti A. An Introduction to Categorical Data Analysis. Wiley 2007

3) Agresti A. Foundations of Linear and Generalized Linear Models. John Wiley & Sons 2015

4) Agresti A. et Coull B.A. Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician – 1998 52 (2) 119–126

5) Agresti A. et Coull B.A. Comment on « Interval estimation for a binomial proportion » by Brown, Cai and DasGupta (2001). Statistical science – 2001 16 (2) 117–120

6) Agresti A. et Min Y. On Small-Sample Confidence Intervals for Parameters in Discrete Distributions. Biometrics – 2001 57 (3) 963–971

7) Anderson D.R., Burnham K.P. et Thompson W.L. Null Hypothesis Testing: Problems, Prevalence, and an Alternative. The Journal of Wildlife Management – 2000 64 (4) 912

8) Anscombe F.J. The transformation of poisson, binomial and negative-binomial data. Biometrika – 1948 35 (3/4) 246–254

9) Anscombe F.J. On Estimating Binomial Response Relations. Biometrika – 1956 43 (3/4) 461–464

10) Bartlett M.S. The Square Root Transformation in Analysis of Variance. Supplement to the Journal of the Royal Statistical Society – 1936 3 (1) 68–78

11) Berry G. et Armitage P. Mid-P confidence intervals: a brief review. The Statistician – 1995 44 (4) 417–423

12) Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. Canadian Journal of Statistics – 2000 28 (4) 783–798

13) Blaker H. et Spjøtvoll E. Paradoxes and improvements in interval estimation. The American Statistician – 2000 54 (4) 242–247

14) Blyth C.R. Approximate Binomial Confidence Limits. Journal of the American Statistical Association – 1986 81 (395) 843–855

15) Blyth C.R. et Still H.A. Binomial Confidence Intervals. Journal of the American Statistical Association – 1983 78 (381) 108–116

16) Bolboaca S.D. et Jäntschi L. Optimized confidence intervals for binomial distributed samples. International Journal of Pure and Applied Mathematics – 2008 47 (1) 1–8

17) Borkowf C.B. Constructing binomial confidence intervals with near nominal coverage by adding a single imaginary failure or success. Statistics in Medicine – 2006 25 (21) 3679–3695

18) Brenner D.J. et Quan H. Exact Confidence Limits for Binomial Proportions-Pearson and Hartley Revisited. Journal of the Royal Statistical Society Series D (The Statistician) – 1990 39 (4) 391–397

19) Brillinger D.R. Statistics. The Canadian Encyclopedia – 2013

20) Brown B.R. et Lathrop R.L. The Effects of Violations of Assumptions Upon Certain Tests of the Product Moment Correlation Coefficient. Annual Meeting of the American Educational Research Association , New York – 1971

21) Brown L.D., Cai T.T. et DasGupta A. Interval estimation for a binomial proportion. Statistical science – 2001 16 (2) 101–117

22) Brown L.D., Cai T.T. et DasGupta A. Confidence Intervals for a binomial proportion and asymptotic expansions. The Annals of Statistics – 2002 30 (1) 160–201

23) Byrne J. et Kabaila P. Comparison of Poisson Confidence Intervals. Communications in Statistics - Theory and Methods – 2005 34 (3) 545–556

24) Cai T.T. One-sided confidence intervals in discrete distributions. Journal of Statistical planning and inference – 2005 131 (1) 63–88

25) Cai Y. et Krishnamoorthy K. A simple improved inferential method for some discrete distributions. Computational statistics & data analysis – 2005 48 (3) 605–621

26) Carpenter J. et Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine – 2000 19 (9) 1141–1164

27) Casella G. Refining binomial confidence intervals. Canadian Journal of Statistics – 1986 14 (2) 113–129

28) Casella G. Comment on « Interval estimation for a binomial proportion » by Brown, Cai and DasGupta (2001). Statistical science – 2001 16 (2) 120–122

29) Census. The Canadian Encyclopedia – 2015

30) Chen X., Zhou K. et Aravena J.L. Explicit Formula for Constructing Binomial Confidence Interval with Guaranteed Coverage Probability. Communications in Statistics - Theory and Methods – 2008 37 (8) 1173–1180

31) Clopper C.J. et Pearson E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika – 1934 26 (4) 404–413

32) Corcoran C. et Mehta C. Comment on « Interval estimation for a binomial proportion » by Brown, Cai and DasGupta (2001). Statistical science – 2001 16 (2) 122–124

33) Cornish E.A. et Fisher R.A. Moments and Cumulants in the Specification of Distributions. Revue de l'Institut International de Statistique – 1938 5 (4) 307

34) Cox D.R. et Hinkley D.V. Theoretical statistics. Chapman & Hall, 1974

35) Crow E.L. Confidence intervals for a proportion. Biometrika – 1956 43 (3–4) 423–435

36) DasGupta A. Asymptotic Theory of Statistics and Probability. Springer Science & Business Media 2008

37) Davis M. et Xie S.X. Caution: Hazards Crossing! Using the Renyi Test Statistic in Survival Analysis. PharmaSUG – 20117–8

38) Derr R.E. Performing Exact Logistic Regression with the SAS System. Proceedings of the Twenty-fifth Annual SAS Users Group International Conference , Cary, North Carolina – 2000

39) Efron B. Bootstrap methods: another look at the jackknife. The annals of Statistics – 1979 7 (1) 1–26

40) Efron B. Transformation theory: How normal is a family of distributions? The Annals of Statistics – 1982 10 (2) 323–339

41) Efron B. Better bootstrap confidence intervals. Journal of the American statistical Association – 1987 82 (397) 171–185

42) Fagerland M., Lydersen S. et Laake P. Statistical Analysis of Contingency Tables. CRC Press 2017

43) Fay M.P. Two-sided exact tests and matching confidence intervals for discrete data. R journal – 2010 2 (1) 53–58

44) Foi A. Direct optimization of nonparametric variance-stabilizing transformations. 8èmes Rencontres de Statistiques Mathématiques , CIRM, Marseille – 2008

45) Forthofer R.N., Lee E.S. et Hernandez M. Biostatistics: A Guide to Design, Analysis and Discovery. Academic Press 2006

46) Freeman M.F. et Tukey J.W. Transformations Related to the Angular and the Square Root. The Annals of Mathematical Statistics – 1950 21 (4) 607–611

47) Fritz M. et Berger P.D. Improving the User Experience through Practical Data Analytics: Gain Meaningful Insight and Increase Your Bottom Line. Morgan Kaufmann 2015

48) Garwood F. Fiducial limits for the Poisson distribution. Biometrika – 1936 28 (3/4) 437–442

49) Gerstman B.B. Epidemiology Kept Simple: An Introduction to Traditional and Modern Epidemiology. John Wiley & Sons 2013

50) Geyer C.J. et Meeden G.D. Fuzzy and randomized confidence intervals and P-values. Quality Control and Applied Statistics – 2006 51 (6) 649

51) Ghosh B.K. A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter. Journal of the American Statistical Association – 1979 74 (368) 894–900

52) Ghosh M. Comment on « Interval estimation for a binomial proportion » by Brown, Cai and DasGupta (2001). Statistical science – 2001 16 (2) 124–125

53) Hall P. Improving the Normal Approximation when Constructing One-Sided Confidence Intervals for Binomial or Poisson Parameters. Biometrika – 1982 69 (3) 647–652

54) Herson J. Data and Safety Monitoring Committees in Clinical Trials. CRC Press 2016

55) Hirji K.F., Mehta C.R. et Patel N.R. Computing Distributions for Exact Logistic Regression. Journal of the American Statistical Association – 1987 82 (400) 1110

56) Hollander M., Wolfe D.A. et Chicken E. Nonparametric Statistical Methods. John Wiley & Sons 2013

57) INSEE : Population totale par sexe et âge au 1er janvier 2017, France. https://www.insee.fr/fr/statistiques/fichier/1892086/pop-totale-france.xls. Publié en 17 janvier 2017.

58) Kabaila P. et Byrne J. Theory & Methods: Exact Short Confidence Intervals from Discrete Data. Australian & New Zealand Journal of Statistics – 2001 43 (3) 303–309

59) Kaiser H.F. Directional statistical decisions. Psychological Review – 1960 67 (3) 160

60) Klaschka J. On caculation of Blaker's binomial confidence limits. , Paris – 2010

61) Klaschka J. Package 'BlakerCI'. https://cran.r-project.org/web/packages/BlakerCI/BlakerCI.pdf. Publié en 2015. Consulté le 21 septembre 2017.

62) Krishnamoorthy K., Thomson J. et Cai Y. An exact method of testing equality of several binomial proportions to a specified standard. Computational statistics & data analysis – 2004 45 (4) 697–707

63) Lancaster H.O. The combination of probabilities arising from data in discrete distributions. Biometrika – 1949 36 (3/4) 370–382

64) Lancaster H.O. Significance tests in discrete distributions. Journal of the American Statistical Association – 1961 56 (294) 223–234

65) Lang J.B. Mean-Minimum Exact Confidence Intervals for a Binomial Probability. The American Statistician – décembre 2016

66) Lecoutre B. et Poitevineau J. New results for computing Blaker's exact confidence interval for one parameter discrete distributions. Communications in Statistics-Simulation and Computation – 2016 45 (3) 1041–1053

67) Leemis L.M. et Trivedi K.S. A comparison of approximate interval estimators for the Bernoulli parameter. The American Statistician – 1996 50 (1) 63–68

68) Liu Y.K. et Kott P.S. Evaluating alternative one-sided coverage intervals for a proportion. Journal of Official Statistics – 2009 25 (4) 569

69) Loader C. Fast and accurate computation of binomial probabilities. https://lists.gnu.org/archive/html/octave-maintainers/2011-09/pdfK0uKOST642.pdf. Publié en 2000. Consulté le 12 septembre 2017.

70) Loi de probabilité. Wikipédia – 2017

71) Lombardi C.M. et Hurlbert S.H. Misprescription and misuse of one-tailed tests. Austral Ecology – 2009 34 (4) 447–468

72) Ludbrook J. Should we use one-sided or two-sided P values in tests of significance? Clinical and Experimental Pharmacology and Physiology – 2013 40 (6) 357–361

73) Matuszewski A. et Sotres D. A basic statistical problem: Confidence interval for the Bernoulli parameter. Computational Statistics & Data Analysis – 1985 3 103–114

74) Meeker W.Q., Hahn G.J. et Escobar L.A. Statistical Intervals: A Guide for Practitioners and Researchers. John Wiley & Sons 2017

75) Millot G. Comprendre et réaliser les tests statistiques à l'aide de R : Manuel de biostatistique. 2ème édition, Éditions De Boeck Université 2011

76) Molenaar W. Simple Approximations to the Poisson, Binomial, and Hypergeometric Distributions. Biometrics – 1973 29 (2) 403–407

77) Newcombe R.G. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in medicine – 1998 17 (8) 857–872

78) Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London Series A, Mathematical and Physical Sciences – 1937 236 (767) 333–380

79) Nikolaidis E., Mourelatos Z.P. et Pandey V. Design Decisions under Uncertainty with Limited Information. CRC Press 2011

80) Olsen C. Teaching Elementary Statistics with JMP. SAS Institute 2011

81) Ott R.L. et Longnecker M.T. An Introduction to Statistical Methods and Data Analysis. Cengage Learning 2015

82) Pan W. Approximate confidence intervals for one proportion and difference of two proportions. Computational statistics & data analysis – 2002 40 (1) 143–157

83) Pires A.M. et Amado C. Interval estimators for a binomial proportion: Comparison of twenty methods. REVSTAT–Statistical Journal – 2008 6 (2) 165–197

84) Pratt J.W. A normal approximation for binomial, F, beta, and other common, related tail probabilities, II. Journal of the American Statistical Association – 1968 63 (324) 1457–1483

85) R: Test of Equal or Given Proportions. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prop.test.html. Consulté le 11 septembre 2017.

86) Razdolsky L. Probability-Based Structural Fire Load. Cambridge University Press 2014

87) Reiczigel J. Confidence intervals for the binomial parameter: some new considerations. Statistics in Medicine – 2003 22 (4) 611–621

88) Rubin D.B. et Schenker N. Logit-based interval estimation for binomial data using the Jeffreys prior. Sociological methodology – 1987 17 131–144

89) Sakakibara I., Haramo E., Muto A., Miyajima I. et Kawasaki Y. Comparison of five exact confidence intervals for the binomial proportion. American Journal of Biostatistics – 2014 4 (1) 11

90) Santner T.J. Teaching Large-Sample Binomial Confidence Intervals. Teaching Statistics – 1998 20 (1) 20–23

91) Santner T.J. Comment on « Interval estimation for a binomial proportion » by Brown, Cai and DasGupta (2001). Statistical science – 2001 16 (2) 126–128

92) Sauro J. et Lewis J.R. Quantifying the User Experience: Practical Statistics for User Research. Morgan Kaufmann 2016

93) Schilling M.F. et Doi J.A. A coverage probability approach to finding an optimal binomial confidence procedure. The American Statistician – 2014 68 (3) 133–145

94) Somerville M.C. et Brown R.S. Exact likelihood ratio and score confidence intervals for the binomial proportion. Pharmaceutical statistics – 2013 12 (3) 120–128

95) Stata base reference manual release 15. Stata Press 2017

96) Sterne T.E. Some remarks on confidence or fiducial limits. Biometrika – 1954 41 (1/2) 275–278

97) Stevens W.L. Fiducial limits of the parameter of a discontinuous distribution. Biometrika – 1950 37 (1/2) 117–129

98) Student. The probable error of a mean. Biometrika – 1908 6 (1) 1–25

99) Théorème de Moivre-Laplace. Wikipédia – 2017

100) Thulin M. et Zwanzig S. Exact confidence intervals and hypothesis tests for parameters of discrete distributions. Bernoulli – 2017 23 (1) 479–502

101) Tsuang M.T., Tohen M. et Jones P. Textbook of Psychiatric Epidemiology. John Wiley & Sons 2011

102) Venkatraman E.S. A permutation test to compare receiver operating characteristic curves. Biometrics – 2000 56 (4) 1134–1138

103) Venkatraman E.S. et Begg C.B. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika – 1996 83 (4) 835–848

104) Venzon D.J. et Moolgavkar S.H. A method for computing profile-likelihood-based confidence intervals. Applied statistics – 1988 87–94

105) Vidakovic B. Statistics for Bioengineering Sciences: With MATLAB and WinBUGS Support. Springer Science & Business Media 2011

106) Vollset S.E. Confidence intervals for a binomial proportion. Statistics in Medicine – 1993 12 (9) 809–824

107) Vos P.W. et Hudson S. Problems with Binomial Two-Sided Tests and the Associated Confidence Intervals. Australian & New Zealand Journal of Statistics – 2008 50 (1) 81–89

108) Wang W. Smallest confidence intervals for one binomial proportion. Journal of Statistical Planning and Inference – 2006 136 (12) 4293–4306

109) Wang W. An iterative construction of confidence intervals for a proportion. Statistica Sinica – 2014 24 (3) 1389–1410

110) Wilson E.B. Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association – 1927 22 (158) 209–212

111) Winstein K. Efficient routines for biostatistics. https://github.com/keithw/biostat. Publié en 30 septembre 2016. Consulté le 26 juillet 2017.

112) Yu G. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. Statistics & Probability Letters – 2009 79 (14) 1621–1629

113) Zamar D., McNeney B. et Graham J. elrm: Software implementing exact-like inference for logistic regression models. Journal of Statistical Software – 2007 21 (3) 1–18

114) Zhou X.H., Li C.M. et Yang Z. Improving interval estimation of binomial proportions. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences – 2008 366 (1874) 2405–2418

115) Zieliński W. The Shortest Clopper–Pearson Confidence Interval for Binomial Probability. REVSTAT–Statistical Journal – 2017 15 (1) 141–153

André GILLIBERT Soutenance le 13/10/2017

Sous la direction du Pr Bruno FALISSARD

Intervalle de confiance d'une proportion binomiale : Quels enjeux et comment choisir ?

Thèse pour le doctorat de médecine. Diplôme d'études supérieures de santé publique et médecine sociale.

Résumé

Objectif de l'étude : Présenter une revue de la littérature des estimateurs d'intervalles de confiance d'une proportion binomiale, les enjeux de l'estimation et évaluer ces estimateurs selon des critères pertinents, puis fournir des conseils pratiques sur le ou les estimateurs utilisables.

Matériel et méthodes : Google® Scholar servit à la recherche bibliographique. Aucun intervalle ne peut maîtriser parfaitement les risques α car la distribution binomiale est discrète. Les critères de jugement pertinents étaient : Premièrement, les risques α réels moyens locaux à droite et à gauche, supposant la taille d'échantillon ou la proportion théorique comme une variable aléatoire. Deuxièmement, les risques α réels à droite et à gauche conditionnels à la taille d'échantillon et la proportion théorique. Troisièmement, la demi-largeur attendue d'intervalle à droite et à gauche reflétant la précision de l'estimateur.

Résultats : L'intervalle de Wald ne maîtrisait pas les risques, même pour de grands échantillons. L'intervalle par approximation normale de Wald dans une régression logistique était déséquilibré, dégradant l'interprétation directionnelle (supériorité ou infériorité) de l'intervalle. L'intervalle du rapport de vraisemblance (ou profile likelihood) avait un meilleur équilibre. L'intervalle de Clopper-Pearson mid-P maîtrisait bien le risque réel moyen local alors que l'intervalle de Clopper-Pearson simple était strictement conservatif sur le risque réel conditionnel. Les intervalles de bootstrap percentile et basique avaient le même ordre de biais que l'intervalle de Wald alors que les intervalles studentisés et BC_a modifié pour une distribution discrète, étaient moins biaisés sans toutefois être aussi performants que les méthodes paramétriques. Les demi-largeurs d'intervalles étaient un miroir des risques α moyens locaux réels. Même en remplaçant la condition de validité de l'intervalle de Wald $n\hat{p}, n(1-\hat{p}) \geq 5$ par $n\hat{p}, n(1-\hat{p}) > 40$ où $\hat{p}$ représente la proportion observée et n la taille de l'échantillon, tous les risques ne sont pas maîtrisés.

Conclusion : L'intervalle de Wald devrait idéalement ne jamais être utilisé. L'intervalle du rapport de vraisemblance devrait être privilégié dans les modèles linéaires généralisés bivariés ou multivariés. Nous conseillons l'usage systématique de l'intervalle Clopper-Pearson mid-P pour l'estimation d'une proportion à l'exception du cas de la comparaison de proportion observée à théorique dans des conditions expérimentales maîtrisées dans lesquelles l'intervalle de Clopper-Pearson peut lui être préféré.

Mots-clés : Loi binomiale — intervalle de confiance — biais d'estimation — risque α conditionnel — risque α moyen local — petit échantillon — échantillon fini — largeur d'intervalle — intervalle de Wald — intervalle de Clopper-Pearson mid-P — équilibre des risques

Jury

Président :	M. le Professeur Jacques BÉNICHOU (président)
Membres :	M. le Professeur Pierre DÉCHELOTTE
	M. le Docteur Joël LADNER
	Mme le Docteur Marie-Pierre TAVOLACCI
	M. le Docteur Thomas VERMEULIN