Tests d'adéquation à la loi de Newcomb-Benford comme outils de détection de fraudes

Komlavi Vovor-Dassu

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Biostatistique

École doctorale I2S – Information, Structures, Systèmes

Unité de recherche UMR 5149 – IMAG – Institut Montpelliérain Alexander Grothendieck

Tests d'adéquation à la loi de Newcomb-Benford comme outils de détection de fraudes.

Présentée par Komlavi Innocent Credo Vovor-Dassu

Le 06 décembre 2021

Sous la direction de Gilles Ducharme

Devant le jury composé de

Marcel Ausloos	Professeur	Université de Leicester	Rapporteur
Jean-Paul Delahaye	Professeur émérite	Université de Lille	Rapporteur
Gilles Ducharme	Professeur	Université de Montpellier	Directeur
Elise Janvresse	Professeur	Université de Picardie Jules Verne	Examinatrice
Jean-Michel Marin	Professeur	Université de Montpellier	Examinateur
Catherine Trottier	Maître de Conférences	Université Paul Valéry - Montpellier 3	Examinatrice

UNIVERSITÉ DE MONTPELLIER

Résumé

Je vous propose le pari suivant tiré de Delahaye (1999) : ouvrons le journal, choisissons une page au hasard et notons le premier nombre que nous rencontrons ; si le premier chiffre significatif de ce nombre est supérieur à 3, je vous donnerai 100 euros, sinon c'est vous qui me donnerez 100 euros. La proposition vous est, semble-t-il, nettement favorable : il n'y a en effet que trois chiffres qui me font gagner $(1, 2, 3)$, alors qu'il y en a six pour vous $(4, 5, 6, 7, 8, 9)$; le 0 ne compte pas, car il ne peut pas être un premier chiffre significatif. Vous pensez donc gagner environ deux fois sur trois. Serais-je idiot de vous proposer un tel pari ? Eh bien non : si vous acceptez, je gagnerai dans plus de 60 pour cent des cas. Aussi étonnant que cela paraisse, le premier chiffre significatif d'un nombre rencontré dans un article de journal n'a pas autant de chances d'être un 1, un 2, un 3, ..., ou un 9 (la probabilité serait alors 1/9, ou 11, 11 pour cent). La loi de Newcomb-Benford indique que, dans un contexte général comme celui d'un article de journal, les probabilités p de rencontrer les différents chiffres comme premier chiffre significatif sont, exprimées en pourcentage : $p(1) = 30, 1$; $p(2) = 17, 6$;$p(3) = 12, 5$; $p(4) = 9, 7$;$p(5) = 7, 9$;$p(6) = 6, 7$;$p(7) = 5, 8$; $p(8) = 5, 1$;$p(9) = 4, 6$. Puisque $30, 1 + 17, 6 + 2, 5 = 60, 2$, je gagnerai mon pari dans $60, 2$ pour cent des cas. Comment tester la qualité de l'ajustement des données à cette loi si contraire à l'intuition ?
Wong (2010) apporta des réponses à cette question en utilisant le premier et deuxième chiffre significatif basées sur les travaux de Lesperance *et al.* (2016) sur le premier chiffre significatif. Mais Cerioli *et al.* (2018) fournissent une motivation pour de nouveaux tests de conformité à la loi de Newcomb-Benford. Le but de ce travail est donc de proposer de nouveaux tests d'adéquations basés sur les tests lisses de Neyman (1937). Nous étudions la puissance de nos tests sous différentes alternatives et nous arrivons à la conclusion que notre test est globalement préférable aux tests existants.

I propose to you the following bet taken from Delahaye (1999) : let us open the newspaper, choose a random page, and note the first number which we see. If the first significant figure of this number is higher than 3, I will give you 100 euros, if not you will give me 100 euros. It seems to you that the proposal is clearly favorable to you : there are indeed only three numbers that make me win $(1, 2, 3)$, while there are six for you $(4, 5, 6, 7, 8, 9)$; the 0 does not count, because it cannot be a first significant number. So, you think you'll win twice out of three times. Would I be a fool to offer you such a bet ?

Well, no : if you accept, I will win more than 60 percent of the time. Surprisingly, the first significant digit of a number you see in a newspaper article is not as likely to be a 1, a 2, a 3, ..., or a 9(the probability would be 1/9, or 11.11 percent). The Newcomb-Benford law indicates that, in a general context like a newspaper article ones, the probabilities "p" in percent of encountering the various digits as the first significant digit are : $p(1) = 30.1$; $p(2) = 17.6$; $p(3) = 12.5$; $p(4) = 9.7$; $p(5) = 7.9$; $p(6) = 6.7$; $p(7) = 5.8$; $p(8) = 5.1$; and $p(9) = 4.6$. Since $30.1 + 17.6 + 2.5 = 60.2$, I will win my bet 60.2 percent of the time. How to test the quality of the adjustment of the data to this law which is so contrary to the intuition ?

Wong (2010) provided answers to this question using the first and second significant figures based on the work of Lesperance *et al.* (2016) on the first significant figure. On the other hand, Cerioli *et al.* (2018) provides motivation for new tests of compliance with Newcomb-Benford law. The goal of this work is therefore to propose new adequate tests based on the smooth tests of Neyman (1937). We study the power of our tests under different alternatives, and we concluded that our test is globally preferable to existing ones.

Remerciements

« Le fruit le plus agréable et le plus utile au monde est la reconnaissance. »

Ménandre

« La gratitude est la clé qui ouvre les portes du vrai savoir »

Omraam Mikhael Aivanhov

« La reconnaissance est la mémoire du cœur »

Hans Christian Andersen

Je suis à fleur de mots!!!
A fleurs de lettres!!!
4 ans de tribulations, dont plusieurs moments de pure satisfaction grâce à vous Pr Gilles Ducharme. Je me rappellerai toujours de ce jeune un peu perdu qui est rentré dans votre bureau après une séance de 3 heures de statistiques computationnelles vous parlant de son désir de faire un doctorat. Votre réponse fut la réponse d'un père à son fils. Jean de la Bruyère écrivait : « Le plaisir le plus délicat est de faire celui d'autrui ». Vous avez fait le mien en m'honorant de votre réponse. Durant ces 4 ans passés à vos côtés, permettez-moi de vous appeler « Papa ». Vous n'avez pas été qu'un directeur de thèse surtout dans les moments difficiles de l'année 2019. Vous avez su à chaque étape me relever avec des mots forts et me donner la force de continuer. Merci de m'avoir appris à être moins « bon élève » et plus autonome tout au long de ce travail de recherche. Nonobstant votre relecture méticuleuse de chaque section m'a sans aucun doute permis de pouvoir rendre ce travail. Veuillez trouver ici l'expression de toute ma reconnaissance et gratitude.

Les plus grandes leçons ne sont pas tirées d'un livre mais des enseignants. Je ne vous remercierai jamais assez, vous qui m'avez encadré, orienté, aidé et conseillé en particulier Xavier Bry, Jean-Michel Marin, Jean-Noël Bacro et Catherine Trottier. La passion de chacun de vous pour son travail a su enflammer ma curiosité et garder ma motivation au fil du temps.

Mes remerciements vont également aux membres du jury Pr Marcel Ausloos , Pr Jean-Paul Delahaye et Pre Elise Janvresse pour avoir accepté d'évaluer mon travail. C'est un

réel honneur de vous présenter mes travaux.

Maman, ma chérie, mon amour , ma femme, ma colombe, tu as été la première femme de ma vie. Avant que je ne dépose les yeux sur toi, toi tu me connaissais déjà, tu connaissais mon nom et tu rêvais déjà du moi d'aujourd'hui. Quel que soit le sens dans lequel souffle le vent dans ma vie, tu as toujours été présente. Merci de m'avoir réveillé chaque matin à 5 heures du matin.
Mon vieux père Komlan Tobie, « Credo de Papa », un problème usuel en Mathématiques consiste à déterminer une fonction connaissant son comportement et une condition initiale. C'est ce qu' on appelle un « problème de Cauchy ». L'exemple le plus simple est la recherche de la primitive d'une fonction prenant une valeur donnée en un point donné. Il s'énonce comme suit : Etant donné une fonction f et des nombres a et b, déterminer F sachant que $F' = f$ et $F(a) = b$. On montre qu'un tel problème a une (seule) solution. Mon interprétation est que nous pouvons savoir ce que furent nos ancêtres (F) en analysant nos propres qualités physiques et intellectuelles (f) et les traces $(F(a) = b)$ qu'ils ont laissées en nous. Je ne suis qu'une projection de toi, en espérant un jour être pour mes enfants un bon père comme tu l'es pour nous. Tu es mon exemple. Merci.
A mes chers frères et sœurs, Fabrice, Serge, Danielle, Roland et Laudamus, pour leur soutien et attention. Vous m'avez permis de réaliser que la famille est sacrée. Vous étiez pour moi, une vraie source d'inspiration et avez été toujours à mes côtés durant les moments difficiles. Mes mots ne seraient jamais à la hauteur de l'amour et de l'affection que vous m'avez témoignés tout au long de mes études. J'aimerais vous exprimer toute ma gratitude et reconnaissance. Cette dédicace serait pour moi, la meilleure façon de vous honorer et vous montrer à quel point vous avez été magnifiques.

A toi mon binôme Manon Culioli, pour ton soutien moral et de ne m'avoir jamais laissé sombrer.
Mes remerciements vont aussi à mes amis qui, avec cette question récurrente, « quand est-ce que tu la soutiens cette thèse ? », bien qu'angoissante en période fréquente de doutes, m'ont permis de ne jamais dévier de mon objectif final.

Que toutes celles et ceux qui m'ont aidé d'une manière ou d'une autre à la réalisation de ce travail soient ici remerciés.

A Nathanael Andrew Kossi VOVOR DASSU,

Mon fils

Sommaire

Introduction

Chapitre 1

Introduction

1.1 Contexte général

La notion de « bien » est incluse à la racine même de l'humanité. Elle est le pilier central de notre évolution. Elle génère une guerre incessante entre les partisans d'un ordre moral profitable à tous (ce qu'on pourrait appeler les « bons » faisant le « bien »), à ceux qui sont à la recherche de leur profit personnel au détriment de l'intérêt collectif ou d'autrui (les « méchants » faisant le « mal »). Les champs de bataille où se livre cette guerre sont nombreux et les victoires des uns ne sont jamais pérennes, les avancées technologiques procurant de nouveaux moyens aux autres.

Un de ces champs de bataille est la détection et la prévention des fraudes. Produire une définition précise du terme « fraude » (*« $XIII^e$ siècle, emprunté du latin fraus, fraudis »*) est complexe, car sur le plan légal par exemple, il existe dans le droit français deux concepts de fraudes liés, mais suffisamment différents pour être géré par des codes de lois différents : la fraude civile et la fraude pénale.

Le dictionnaire Merriam Webster's Dictionary of Law (cité par Manurung et Hadian (2013, p 4)) offre une définition très générale de la fraude :

> *"Any act, expression, omission, or concealment calculated to deceive another to his or her disadvantage, specifically, a misrepresentation or concealment with reference to some fact material to a transaction that is made with knowledge of its falsity. And or in reckless disregard of its truth or falsity and worth the intent to deceive another and that is reason- ably relied on by the other who is injured thereby"*

« *... a misrepresentation ... with reference to some fact material to a transaction that is made with knowledge of its falsity* » ; c'est-à-dire : *falsification d'information effectuée dans le but d'obtenir des avantages indu*s. C'est ce type de fraude, la falsification de l'information, qui est l'objet de notre étude.

Cette catégorie peut être décrite comme une « population » de fraudes donc la cardinalité est l'infinie. Comme individu membre de cette population, on peut penser bien sûr à la fraude fiscale, commise par le contribuable (une personne lambda) envers le fisc ; mais aussi, à la fraude bancaire ou comptable définie par l'Association of Certified Fraud Examiner (ACFE) comme étant « un acte accompli dans l'illégalité qui consiste à tromper

délibérément, à soutirer de l'argent contre la volonté de quelqu'un ou à falsifier intentionnellement un document afin de porter atteinte aux droits ou aux intérêts d'autrui ». Évidemment, les domaines où s'exercent des fraudes sont très nombreux et les « astuces » développées pour les perpétrer ne sont limitées que par l'imagination des fraudeurs.

En réaction, les organismes voués au bien du public (par exemple l'OLAF, Office européen de Lutte Anti-Fraude `https://ec.europa.eu/anti-fraud/`) soutiennent et favorisent la mise en place d'outils pour détecter les fraudes et prévenir leur utilisation, ainsi qu'un arsenal, notamment législatif, pour protéger le public. Ces outils et cet arsenal législatif sont en constante évolution, car l'imagination et la vénalité des fraudeurs sont sans limites.

Selon un article récent (Abdullahi et Mansor (2015)), le nombre de fraudes, notamment en comptabilité, est en augmentation constante tel qu'en font foi les scandales de Enron, WorldCom, Global Crossing et Tyco aux États-Unis. L'Europe n'est pas épargnée avec par exemple, l'affaire Cahuzac en France et plus récemment en Juillet 2020 le scandale financier Wirecard en Allemagne. La vision générale des organismes publics est que la prévention de la fraude est de loin préférable à sa détection : lorsque celle-ci a été perpétrée, il est souvent très difficile de récupérer les sommes dérobées. En outre, il est très complexe ou coûteux d'investiguer des fraudes de grande ampleur.

Pour mettre en place des politiques de prévention de la fraude, il convient de se mettre dans la peau d'un « méchant » pour tenter de comprendre les éléments pouvant favoriser la fraude. Toujours selon Abdullahi et Mansor (2015), il existe deux théories principales, soit la théorie du triangle de fraude et celle, emboîtant la première , du losange (diamond) de fraude, recensant les facteurs pouvant mener à un comportement frauduleux. Parmi ceux-ci, on peut noter, dans la théorie triangulaire, 1) la pression, 2) l'opportunité et 3) la rationalisation. Ainsi, un individu se percevant sous pression financière, estimant avoir très peu de chance d'être pincé et qui arrive à auto-justifier son acte est un candidat-fraudeur potentiel. Et concernant le Point 2), l'opportunité, selon les auteurs (ibid. p. 33) « *In most cases, the lower the risk of being caught, the more likely it is that fraud will take place* ». Ainsi, les outils de détection de fraudes sont aussi des outils pour leur prévention.

On voit poindre ici trois tâches pour juguler un type de fraude : 1) le développement de méthodes de détection de la fraude ; 2) la consolidation des éléments de la preuve en un argumentaire à soumettre aux instances judiciaires ou législatives ; 3) lesquelles s'appuient sur un arsenal législatif en constante évolution. Ceci mène éventuellement à des peines ou à la réparation des dommages causés.

Le cadre du présent travail s'inscrit dans les tâches 1) et 2) que sont la détection de la fraude et la consolidation de la preuve. Plus précisément, le but de ce travail est de développer une boîte à outils pour détecter des fraudes (tâche 1) en minimisant les coûts de cette opération et en quantifiant et contrôlant les risques d'erreur (tâche 2) de

Type I , soit d'alerter qu'une fraude a été potentiellement commise quand il n'y en a pas eu, ou de Type II, soit ne pas détecter qu'une fraude s'est produite. S'il est possible de créer une telle boîte à outils avec des risques d'erreur très faibles, un fraudeur potentiel, selon la citation plus haut (ibid. p. 33), sera moins tenté de commettre son méfait. Nous nous inscrivons ici dans la logique des spécialistes en sécurité numérique : un « méchant » voulant faire le mal ne peut pas être arrêté. À nous de faire en sorte que l'équation coût/bénéfice ne lui soit pas favorable et qu'il aille commettre ses méfaits ailleurs. Étant donné qu'à ce stade, le jugement des instances judiciaires n'a pas été encore formulé, il convient parfois d'utiliser un langage prudent : nous utiliserons alors parfois le terme détection d'anomalies. L'introduction du terme anomalie achète la paix avec les partisans du « politiquement correct » mais mène tout de suite à la difficulté principale de ce genre d'entreprise, car ce qui est anormal se définit par opposition à ce que serait une certaine « normalité », si tant est qu'une telle chose existe. Or ce qui caractérise dans un contexte donné la « normalité » est généralement très difficile à définir, l'établissement de normes étant une tâche complexe, autant pour les législateurs que pour les scientifiques.

Il existe malheureusement de trop nombreux types de fraudes (l'imagination humaine n'a pas de limites quand il s'agit de faire facilement de l'argent) mais on peut établir un certain continuum concernant la fraude perpétrée en falsifiant des données. D'un coté de ce continuum, on trouve les cas où chacun des individus d'un échantillon doit produire un nombre représentant sa situation (e.g. son revenu annuel), lesquels nombres sont colligées par une entité au dessus de tout soupçon (un centre des impôts régional). Dans ce cas de figure, chaque individu peut falsifier sa donnée pour avantager sa situation. Dans ce genre de contexte, on estime généralement que la majorité des individus sont honnêtes. Les données observées sont donc majoritairement des données saines contaminées par un faible pourcentage de données frauduleuses. Ce type de fraude apparaît dans les fraudes fiscales ou dans les transactions financières à l'échelle d'un pays ou d'une communauté de pays. À l'autre bout de l'échelle, on peut identifier le cas où un seul individu décide de falsifier les résultats de son échantillon de données de façon à obtenir la conclusion qu'il souhaitait. Un exemple de ce cas est décrit dans le Chapitre 17 de l'ouvrage Benford's Law Theory and Applications (Miller (2015)) où on raconte l'histoire du scientifique Eric Poechlman de l'université de Vermont. En particulier, Poechlman voulait montrer que le niveau de lipides chez certains sujets se détériorait avec l'âge. Un de ses étudiants (DeNimo) trouva que les données ne soutenaient pas cette hypothèse, Poechlman altéra les données en ajoutant des patients fictifs et en modifiant la valeur d'autres données. DeNino devient méfiant quand les nouvelles données, supposément corrigées de certaines "erreurs", se mirent clairement à soutenir l'hypothèse de Poechlman. En passant au travers de centaines de dossier de patients dans son laboratoire et dans les archives de son CHU, DeNimo y trouva les preuves de falsification des données par Poechlman. Ce genre de fraude apparaît notamment dans le contexte d'expérience scientifiques et peut avoir des conséquences

graves sur l'évolution de la science et du traitement des patients.

On peut imaginer de nombreuses situations entre ces deux extrêmes, par exemple celle où les données des revenus des individus sont d'abord agrégées à un niveau régional, puis au niveau national. Si un individu indélicat dans un des centres régionaux modifient ses données, on se retrouve dans une situation intermédiaire à ces deux extrêmes.

Le contexte que nous considérons dans le présent travail est celui où des données $x_1, ..., x_n$ ont été recueillies pour être par la suite utilisée dans la prise de la décision ou la direction d'actions à mener selon un algorithme qui n'est pas secret. Nous craignons cependant qu'avant leur passage dans l'algorithme, ces données aient été modifiées par un (ou des) fraudeur(s) pour mener à une décision ou une action qui l'avantage indûment. Si tel est le cas, nous espérons que ce faisant, le fraudeur a perturbé le caractère « normal » des données. L'« anormalité » résultante constitue une prise dont nous disposons pour détecter la possibilité d'une fraude. Cette « normalité » doit maintenant être définie et pour ce faire, nous supposons que, s'il n'y a pas eu de fraude, les données $x_1, ..., x_n$ sont les réalisations d'une suite de variables aléatoires iid $X_1, ..., X_n$. Dans la prochaine section, nous décrivons brièvement l'histoire de la découverte d'une certaine « normalité» dans le cas simple de données continues, comme celles qui sont souvent observées dans les sciences expérimentales.

1.2 A la recherche d'une certaine normalité

1.2.1 La normalité des données

Nous nous concentrons donc sur le problème de déterminer ce qui pourrait constituer un caractère de « normalité » lorsque les données $x_1, x_2, \ldots, x_n$ sont supposées être les réalisations d'une suite de variables aléatoires iid $X_1, X_2, \ldots, X_n$. La découverte et la preuve de l'existence d'une certaine « normalité » furent le résultat de nombreuses observations empiriques et de recherches théoriques qui s'étalèrent sur plusieurs siècles, témoignant de la difficulté de l'entreprise. Ce qui suit est partiellement extrait et traduit du livre de l'historien Bernstein (1998).

Les fils de la trame que nous déroulons s'imbriquent avec ceux d'une autre trame concernant la formalisation de la notion de « probabilité » d'un événement. La théorie des probabilités a débuté entre 1400 et 1700 (environ) essentiellement pour répondre à des questions se rapportant à des jeux de hasard du type : A et B jouent à un jeu de hasard. Ils s'accordent pour jouer jusqu'à ce qu'un joueur gagne six parties. Mais le jeu doit s'arrêter prématurément alors que A a gagné cinq parties et B trois. Comment la mise des joueurs doit-elle être répartie entre eux ? À cette époque, la notion de probabilité était associée au concept d'équité du jeu. Puis avec la découverte de la loi des grands nombres par Jacques Bernoulli, cette notion évolua, en fréquence limite d'occurrence

d'un événement, puis avec le théorème de Bayes, en une mesure du degré d'incertitude, en passant par la curieuse théorie du Dr Arbuthnot (MEUSNIER (1999)) (les probabilités sont des signaux faibles envoyés par Dieu pour prouver son existence), pour terminer à ce jour avec la notion philosophique de résonance développée par Burdzy (2016). Pour éviter de confondre le lecteur, on va ici s'appuyer sur la notion intuitive de probabilité qui prévaut généralement aujourd'hui dans l'esprit d'un citoyen lambda.

Les bases de la théorie des probabilités ont été développées par des mathématiciens comme Luca Paccioli (1445-1517)(Bernstein, 1998, p.41), Giorolamo Cardono (1500-1571) , (Bernstein, 1998, p.45, 49 -53) dans son livre *Liber de Ludo Aleae* (1525, mais réécrit, semble-t-il, en 1565 ; notons le mot aleae qui réfère au jeu de dés !), Galilée (dans le livre *Sopra le Scoperte dei Dadi -On playing dice* , 1623) (Bernstein, 1998, p.55), Christian Huygens, et surtout Blaise Pascal, Pierre de Fermat et le Chevalier de Méré (Bernstein, 1998, Chap. 4). Suite à leurs travaux, vers la fin du $17^{\text{ième}}$ siècle, de nombreux problèmes relatifs aux jeux de hasard avaient été résolus. Il restait maintenant à sortir de ce cadre et à appliquer leurs méthodes aux problèmes autrement plus difficiles que pose la nature.

Ceci fut fait par d'autres éminents mathématiciens dans la période s'étalant de 1700 à 1900. Durant ces années, englobant le « siècle des lumières » (1715-1789), l'humanité s'engagea progressivement dans des activités soutenues de collecte de données. Un problème généré par cette activité est de faire ressortir une structure sur laquelle appuyer les notions de probabilités développées dans les jeux de hasard afin de comprendre et prédire le comportement futur de certains phénomènes. Parmi ceux qui ont œuvré à cette tâche, citons John Graunt qui développa les premières tables de mortalité à Londres (Bernstein, 1998, Chap. 5), améliorées par Edmund Halley (Bernstein, 1998, p.84), pour prédire l'évolution de la société anglaise et qui furent à l'origine du concept d'assurance et notamment de la Lloyd's of London. Suite à leurs travaux, les résultats de Cardano et Pascal ont pu être appliqués à des domaines qui n'avaient plus rien à voir avec les jeux de hasard.

De façon indépendante, Abraham de Moivre (1667 — 1754) (Bernstein, 1998, p.125) exposa la loi normale (qui ne s'appelait pas ainsi à l'époque ; il fallut attendre Francis Galton vers la fin du 19e siècle) dans la deuxième édition de son livre *The doctrine of chance* (1733). Cette découverte, qui est en fait un cas particulier du Théorème central de la limite (dont on parlera dans quelques lignes), s'avéra fondamentale, car elle montrait comment un ensemble de n tirages aléatoires indépendants dans une urne de boules vertes et rouges donne des proportions de rouges qui se répartissent selon une loi en forme de cloche, si n est grand. Et ceci ne s'appliquait pas seulement à des boules, mais à toute sorte de contextes (proportion de garçons/filles, de succès/échecs, etc.), ce qui mena de Moivre à penser qu'il devait y avoir un être supérieur ayant imposé cette structure en cloche commune. Et au fur et à mesure que les données recueillies dans les sphères scientifiques de l'époque (science de la terre, de l'univers, cosmologie, etc.)

étaient collectées et représentées sous forme d'histogrammes, on voyait apparaître cette courbe en forme de cloche. En particulier Gauss, dans le cadre de ses travaux de mesures géodésiques pour déterminer la courbure de la terre, s'étonna de voir ses données (en fait des moyennes de mesures) se distribuer selon la loi qu'il appela la « loi des erreurs », une terminologie qui perdura un petit moment. Étant le génie qu'il était, il découvrit (avec Pierre Simon de Laplace ; il n'est pas clair qui fût l'inventeur, car Gauss ne publiait quasiment jamais rien) le Théorème central de la limite qui expliquait pourquoi cette loi des erreurs apparaissait si souvent dans ce type de contexte.

À cette époque, la science des données portait essentiellement sur des données issues d'expériences scientifiques construites selon les règles de la méthode scientifique. Et la loi en cloche apparaissait quasiment partout où on regardait. Il restait à faire le passage à des données moins contrôlées, plus proches de la nature. Ceci fut le travail, vers la fin du $19^{\text{ième}}$ siècle, de deux scientifiques, Lambert Adolphe Jacques Quetelet en Belgique et Francis Galton en Angleterre, qui étendirent l'analyse des données aux sciences de l'homme.

Quetelet s'était mis en tête de définir l'homme « moyen ». Ce concept nouveau captura l'imagination populaire et le rendit rapidement célèbre. Au centre de ce concept se trouvait la loi en cloche dont le paramètre de position représentait, selon la quantité étudiée, cet « homme moyen ». Le reste de la distribution représentait les variations de toutes sortes rencontrées dans une population autour de cet individu moyen. Ainsi Quetelet était amené à voir des cloches partout. Souvent, ces dernières étaient apparemment au rendez-vous, mais pas toujours et ceci amena le statisticien Francis Ysidro Edgeworth à forger le terme « Quetelismus » pour décrire la propension croissante (Quetelet avait eu beaucoup de « followers») à voir des lois en forme de cloche partout (Bernstein, 1998, p.162), malgré l'existence de nombreuses lois de probabilité. À cette époque, l'ajustement de données à une loi de probabilité se faisait par la comparaison visuelle d'un histogramme, et Edgeworth voulait mettre en garde contre la subjectivité de cet exercice, surtout quand on veut prouver que le concept ou la théorie dont on est l'inventeur est juste.

Galton était un véritable maniaque de la mesure et quantifiait tout ce qu'il voyait, de la beauté des personnes qu'il croisait aux jugements de cours de justice (Bernstein, 1998, p.153). Il avait créé un laboratoire d'anthropométrie qui archivait des mesures sur à peu près tout ce que l'on pouvait mesurer chez les êtres humains (Bernstein, 1998, p.153). En établissant les histogrammes de ces données, lui aussi était amené à voir des lois rappelant la forme d'une cloche partout, qu'il appela éventuellement la loi « normale ». Mais il alla plus loin que Quetelet en suggérant que l'omniprésence de cette loi normale venait de l'addition à la mesure principale de nombreuses influences grandes et petites, une situation où l'effet du Théorème central de la limite peut se faire sentir. Il illustra cette idée avec un appareil appelé « la planche de Galton » qui fit sensation dans les cercles éclairés d'Angleterre et contribua à accroître le dogme de l'omniprésence de la

loi normale dans le chaos de données de toute nature. Il écrivit (Bernstein, 1998, p.141) notamment :

> *The "Law Of Frequency Of Error" (c-à-d la loi normale) ...reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob...the more perfect its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand...an unsuspected and most beautiful form of regularity proves to have been latent all along .*

Et quand il rencontrait un cas où les données n'étaient visiblement pas de loi normale, il suspectait quelques « anomalies », par exemple sur les données de la taille des conscrits de l'armée anglaise dont la non-normalité provenait, selon lui, de personnes « falsifiant » leur mesure afin d'éviter d'être appelées sous les drapeaux.

Ainsi, vers 1900, tout le monde croyait que la loi normale régissait la plupart des phénomènes aléatoires « normaux ». Ce qui ne se comportait pas comme une loi en forme de cloche pouvait être a priori considéré comme étant suspect. Le physicien Gabriel Lippmann, lauréat du prix Nobel de physique en 1908 a remarqué (rapporté par Gore et Kherdekar (2017)) : « Everybody believes in the law of errors, that is the normal curve. The experimenters believe because they think it is a mathematical theorem and mathematicians believe because they think it is an experimental fact ». Le statisticien William Youden proposa la représentation allégorique de la loi normale (voir Gore et Kherdekar (2017)) :

Bref, l'humanité avait trouvé sa « normalité » dans le contexte de données $x_1, ..., x_n$ supposées être les réalisations d'une suite de variables aléatoires iid $X_1, ..., X_n$. Certains ont même été jusqu'à dire « *this normal law of error is the law of nature* » (Pearson (1900, p.171)). Et ce dogme subsistait sans trop de challenge, car l'ajustement à la loi normale était le fruit de l'examen visuel d'une représentation graphique des données. Et la nature humaine fait que si l'on veut voir une courbe en forme de cloche, on va voir une courbe en forme de cloche.

Cet état de fait a tenu jusqu'à ce que Karl Pearson (1900), un anglais, introduise son test d'adéquation du χ^2 qui permet de valider statistiquement l'hypothèse de la normalité d'un jeu de données. Dans cet article fondateur de la théorie des tests d'adéquation, Pearson commence par développer (Sections I à V) les outils mathématiques permettant de calculer la probabilité d'observer des écarts calculés à partir de données empiriques à une distribution de probabilité postulée. Puis il présente un certain nombre d'applications à des données concernant des lancers de dés et des résultats de roulette à Monte-Carlo. Puis (Illustration VI), il signale le fait que beaucoup de scientifiques, notamment Sir George Bidell Airy et le Pr Mansfield Merriman concluent à la normalité de leurs données « *–based on no quantitative criterion–* ». Il applique son test à leurs données et montre que la p-value résultante n'est pas compatible avec la loi normale. Il se lamente que « *even today, there are those who regard it* (la loi normale) *as a sort of fetish* ». Il termine

l'article en écrivant : « *The reader may ask : is it not possible to find material which obeys within practical limits the normal law ? I reply, yes. But this law is not a universal law of nature. We must hunt for cases* ».

Le test du χ^2 de Pearson n'est pas idéal : il dépend d'une partition arbitraire des données qui permet à quelqu'un de malhonnête de faire dire aux données ce qu'il veut qu'elles disent. Winston Churchill disait « *Je ne crois aux statistiques que lorsque je les ai moi-même falsifiées* ». En outre, son traitement du cas où les paramètres de la loi normale sont estimés à partir des données est erroné. Néanmoins, ses conclusions générales tiennent et ont contribué à descendre la loi normale du piédestal qu'elle occupait. Il reste quand même qu'aujourd'hui, la loi normale garde une place spéciale dans les applications. Elle est au cœur de la plupart des systèmes de gestion des risques (Bernstein, 1998, p.144). Lorsqu'elle est présente (confirmé par un test d'adéquation), des analyses statistiques très complexes sont possibles, en général peu importe la taille des échantillons. Lorsqu'elle est réfutée, on s'interroge, on cherche des explications, on essaie de comprendre ce que les données tentent de nous dire, ce qui est l'essence même de la science des données. Bref, la loi normale permet de définir une certaine normalité, sans définir ce qui n'est pas de loi normale soit forcément anormal. Comme preuve de cette importance qui ne semble pas faiblir, les statisticiens continuent de produire des tests d'adéquation à la loi normale qui n'ont pas les défauts du test du χ^2 de Pearson. Dufour *et al.* (1998) en dénombre déjà une quarantaine et de nombreux autres se sont rajoutés depuis. Devant cette pléthore de tests, de nombreuses études ont été menées pour essayer de dégager un ou des gagnants (Farrell et Rogers-Stewart (2006) ; Yazici et Yolacan (2007) ; Yap et Sim (2011) ; Noughabi et Arghami (2011) ou encore Marmolejo-Ramos et González-Burgos (2013)). Le critère de qualité principal est la puissance des tests et sur ce point, le test du χ^2, outre ses autres défauts, ne brille pas par rapport à ses compétiteurs. Ainsi, le test du χ^2 de Pearson n'est plus en général recommandé dans le contexte de tests de normalité et de façon plus générale dans le cas de données continues. Sur les autres tests, le débat n'est pas encore clos.

1.2.2 La loi de Benford

Parallèlement, Simon Newcomb astronome, mathématicien, économiste et statisticien lors de ses travaux de calcul de la position des planètes et de constantes astronomiques, dont la vitesse de la lumière remarquait de manière empirique une autre forme de « normalité ». En effet dans le cadre de ses travaux, il était amené à utiliser les outils de calculs de l'époque, notamment un usage intensif des tables de logarithmes pour transformer les multiplications en additions (opérations beaucoup plus simples à effectuer à la main). En compulsant ces ouvrages, il s'est rendu compte que les premières pages (celles qui donnent les logarithmes des nombres commençant par 1,) étaient souvent plus abîmées

	1	2	3	4	5	6	7	8	9
Fréq. Intuitive	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111
Fréq. Réelle	0.3010	0.1760	0.1249	0.0969	0.0791	0.0669	0.0579	0.0511	0.0457

TABLE 1.1 – Table des fréquences intuitives et réelles des PCS,Newcomb (1881).
Les « fréquences réelles » sont données par la formule $\log_{10}\left(1+\frac{1}{d}\right), d = 1, \ldots, 9$ qu'il justifie par un argument informel.

que les dernières (donnant les logarithmes des nombres commençant par 9) et que, notamment, la tranche des livres était plus salie par les doigts au début du livre que vers sa fin.

Ce phénomène l'a intrigué. En effet, pour des nombres naturels, c'est à dire de nombres que l'on « croise au hasard dans la vie de tous les jours » (Newcomb se garde bien de définir avec précision ces termes), on pourrait s'attendre à voir les chiffres 1 à 9 apparaître à peu près aussi fréquemment sur le premier chiffre significatif (PCS) d'un nombre (le premier chiffre significatif d, du nombre x, $d = PCS(x)$, est le premier chiffre différent de 0. Par exemple, $PCS(2.36) = 2$; $PCS(-0.0965) = 9$, $PCS(0.004) = 4$, $PCS(\pi) = 3$, etc.), soit $11,1\%$(1 sur 9) pour chacun. Or, contrairement à l'intuition, ces salissures indiquent que le $PCS(x)$ recherché est beaucoup plus fréquemment un « 1 » (30.1% des fois) qu'un « 9 » par exemple (4.58%). Dans un article publié en 1881 (Newcomb (1881)), il entreprit de calculer une table (1.1) donnant les fréquences d'apparition du PCS dans une série de nombres « naturels ».

Newcomb s'est aussi intéressé à la fréquence d'apparition du deuxième chiffre (de 0 à 9) significatif. Il termine son court article en évoquant l'intrigante idée d'utiliser cette table pour déterminer si une série de nombres provient bien d'une collection de données « naturelles ». Mais il ne dit pas comment ceci pourrait se faire, et pour cause, il ne disposait à l'époque d'aucun outil sauf l'inspection visuelle d'un diagramme en bâton. Cette idée fut reprise par Varian (1972), mais il a fallu attendre Nigrini (1993, 1996) pour une première application formellement correcte de cette idée dans un contexte de fraude fiscale. Le travail de Newcomb resta méconnu pendant 58 ans, constituant ainsi une situation appelée dans les sciences un "sleeping beauty" (Mir et Ausloos (2017)) (notons cependant que dans son Chapitre XVII, Section 233, p. 313-320, Poincaré et Quiquet (1912) discute brièvement de la répartition des décimales dans une table numérique ; mais il passe à coté de la loi de Benford pour s'intéresser à autre chose et ne cite pas Newcomb).

C'est en 1938, que Frank Albert Benford, un américain ayant fait carrière chez General Electric apparemment à titre de chercheur en optique, redécouvre, semble-t-il par la même observation sur des tables de logarithmes, la loi de fréquence des PCS que Newcomb avait déjà trouvée et qui depuis porte son nom, la « LOI DE BENFORD » et que l'on notera

$\mathscr{B}_1(10)$. Il alla un peu plus loin que Newcomb en observant empiriquement des fréquences comparables à la loi de Benford pour le PCS d'une vingtaine de séries de nombres issues de sources diverses et apparemment indépendantes. La table suivante, tirée de Scott et Fasli (2001) résume les fréquences empiriques observées par Benford :

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Weight	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	n^{-1},$\sqrt{n}$,...	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	n^1,n^2,....n!	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Predicted		30.1	17.6	12.5	9.69	7.92	6.70	5.80	5.11	4.58	

Table 1: The distribution of leading digits in Benford's (1938) data sets expressed as percentages. The final row of the table indicates the percentages predicted by Benford's Law.

Remarque. La dernière ligne (ligne *« Predicted »*) donne les pourcentages des « fréquences réelles » de la table 1.1.

Les travaux de Benford ont permis d'établir que cette loi est un phénomène empiriquement observable (tout comme la loi normale) dans de nombreuses situations auxquelles, de façon amusante et contrairement à Newcomb, Benford réfère comme étant des « anomalous numbers ». Elle possède une certaine universalité et recale donc un certain caractère de « normalité ». Les résultats de Newcomb et Benford restèrent par la suite très confidentiels, avec quelques références sporadiques à gauche et à droite jusqu'aux travaux majeurs de Hill (1995) sur le plan théorique et Nigrini (1993, 1996) sur le plan applicatif qui relancèrent considérablement les recherches sur cette loi (Hill est considéré comme étant le "waking prince" de ce "sleeping beauty" ; il contribue d'ailleurs à maintenir un site web (*Benford online bibliography*, Berger *et al.* (2015)) recensant tout ce qui se publie sur la loi de Benford). Depuis, le nombre de publications concernant cette loi a augmenté de façon quasi exponentielle, comme en fait foi le graphique suivant tiré de Mir et Ausloos (2017).

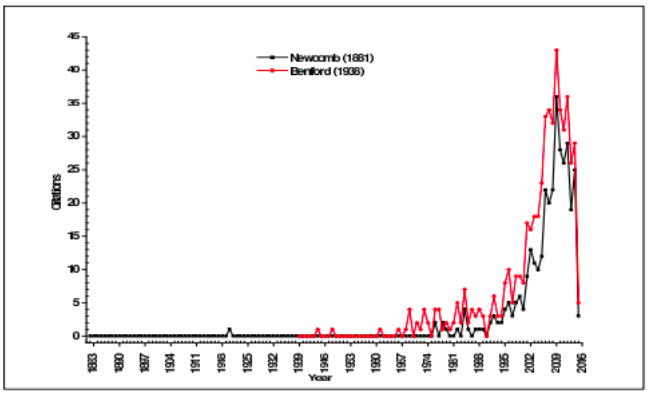

Fig. 1. Yearly citations of Newcomb (1881) and Benford (1938).

Comme on peut le voir sur ce graphique, l'explosion des recherches sur la loi de Benford s'est produite entre 1995 et 2000, quand on a découvert qu'elle pouvait servir à la détection de fraudes, comme l'évoquait déjà Newcomb en 1881.

L'intérêt ne retombe pas : A Stresa en Italie, s'est tenu en Juillet 2019, une conférence internationale sponsorisée par le Programme Anti-Fraude Hercule III de l'Union Européenne, géré par l'OLAF, l'Office européen de Lutte Anti-Fraude (`https://ec.europa.eu/anti-fraud/`) dont le titre est « Benford's law conférence » et portant spécifiquement sur ces problèmes.

1.3 État de l'art sur la loi de Benford

Après le « sleeping beauty », les recherches sur la loi de Benford peuvent être enroulées autour de 4 axes.

1.3.1 Ubiquité de la loi Benford $\mathscr{B}_1(10)$

La majeure partie des travaux porte sur son universalité, son omniprésence. Ce qui émerge d'étonnant avec la loi de Benford, à l'instar de la loi normale, c'est qu'elle apparaît dans toutes sortes de situations. Par exemple, les travaux de Seenivasan *et al.* (2016) pour concevoir un indicateur clinique permettant d'identifier les changements dans la nature de l'arythmie en se basant sur la présence de loi de Benford dans ses données, également Crocetti et Randi (2016) qui a montré que le taux d'incidence du cancer suit la loi de Benford, et Barbancho *et al.* (2015) qui montre comment la loi de Benford peut être exploitée pour la reconnaissance ou classification des données audio.

Dans le cas précis où $D_i = PCS(X_i)$ avec $X_1, \ldots, X_n$, des copies indépendantes de $X \sim F$, Leemis *et al.* (2000) ont montré que si X a pour représentation stochastique $X = 10^{\eta}$, où η est une variable aléatoire dont le support est $[a, b]$ avec a, b entiers (on peut prendre $\eta \sim U[a, b]$, où encore $\eta \sim$ Triangulaire, décentrée ou non, sur $[a, b]$ par exemple), alors $D = PCS(X)$ obéit à la loi de Benford, ce qu'on notera $D \sim \mathscr{B}_1(10)$. Notons que ce résultat reste valable dans le cas d'une base b différent de 10.

Leemis *et al.* (2000) montrent également (voir aussi Posch (2008, thm 3.2.1)) que la p-mixture de deux variables X_1 et X_2 dont les PCS sont $\sim \mathscr{B}_1(10)$ est aussi $\sim \mathscr{B}_1(10)$, avec $0 \leq p \leq 1$.

Par ailleurs, ces derniers montrent que certaines lois de probabilité courantes dans les modèles de survie, à l'instar de la loi de Weibull $(1, 0.3)$ ou la log-logistic $(1, 0.3)$, ont aussi leur PCS très proche (mais pas exactement) de la loi de Benford ; on écrira pour ces cas où la loi de Benford est une approximation de la loi exacte de la quantité aléatoire : $D \approx \mathscr{B}_1(10)$. Parallèlement, ils montrent aussi que d'autres loi de probabilité (comme la Log-Normale $(2, 0.04)$) ne sont pas, même approximativement, de la loi de Benford, détruisant ainsi l'universalité de cette loi. Par contre d'autres lois lognormales sont très proches de la loi de Benford.

La loi de Benford apparaît aussi dans des suites de nombres déterministes, ce qu'on appelle parfois des « Benford sequences » . Les « Benford sequences » sont $d_i = PCS(x_i)$ où les x_i sont une suite déterministe. Un cas intéressant est celui où F_i est le i$^{\text{ème}}$ nombre de la suite de Fibonacci. Il se trouve alors que les fréquences d'occurrence (on ne peut pas parler de probabilités, car la suite de Fibonacci est déterministe) des entiers 1,..., 9 dans la suite $\{PCS\,(F_i)\,,\}\, i \geq 1$ satisfait :

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{PCS\,(F_i) = d\} \overset{\text{n}\to\,+\infty}{\longrightarrow} \log_{10}\left(1 + \frac{1}{d}\right)$$

De nombreuses autres « Benford sequences » ont été trouvées. Par exemple, les cas $x_n = x^n$, où $\log(x)$ est irrationnel, mais aussi $x_n = n^n$ et $x_n = n!$ ainsi que les nombres de Lucas (cousine de la suite de Fibonacci avec des chiffres différents au départ et un calcul de somme pondérée). Un résumé assez exhaustif des conditions assurant qu'une suite déterministe x_n est une « Benford sequences » se trouve dans Posch (2008). Cependant certaines séquences comme $x_n = n^{10}$, $x_n = 10\,n$, ou encore les nombres premiers ne sont pas des « Benford sequences », montrant que la loi de Benford, si elle est fréquente, n'est pas universelle.

Enfin, la loi de Benford apparaît, du moins approximativement, dans une multitude de situations entre les deux cas considérés plus haut (purement stochastiques et purement déterministes). Ces cas sont importants, et il est très intéressant que la $\mathscr{B}_1(10)$ apparaisse dans ces cas, car il arrive de plus en plus souvent en pratique que les données dont on dispose aient été recueillies selon un processus qui n'est pas simple. Par exemple, elles peuvent émaner de PCS de variables aléatoires non-iid. Mais les données peuvent aussi avoir été récoltées selon un processus qui ne peut prétendre à être représentatif d'une certaine population ; dans bien des cas, leur caractère pléthorique peut rassurer sur cette représentativité, mais dans d'autres cas, on n'est même pas assuré de leur caractère aléatoire.

On parle quand même de loi $\mathscr{B}_1(10)$, mais on ne sait alors pas s'il s'agit de fréquence

d'occurrence ou de probabilités. Un exemple de ces cas est celui des constantes physiques (Knuth (1973) ou Burke et Kincanon (1991)) dont les PCS colleraient assez bien à la loi de $\mathscr{B}_1(10)$ selon ces auteurs. Dans ces articles, la taille des jeux de données n'est pas suffisante pour déterminer si les écarts entre les fréquences (ou probabilités) d'occurrence et la loi $\mathscr{B}_1(10)$ sont réelles ou non. Genest et Genest (2011a) vont beaucoup plus loin et étudient les données de « l'inverseur de Plouffe » qui contenait plus de 215 millions de constantes mathématiques diverses où la fréquence d'apparition (ou probabilités) collent de façon remarquable à la loi de Benford.

1.3.2 Justification de la loi de Benford $\mathscr{B}_1(10)$

Nous nous concentrons dans la suite sur les cas où $D_i = PCS(X_i)$, $i = 1, \ldots, n$ qui ont un « certain » caractère aléatoire.

Dans ce contexte, la plupart des personnes qui découvrent l'existence de la loi $\mathscr{B}_1(10)$ sont surpris par le fait que ses fréquences apparaissaient paradoxalement en regard de ce que l'intuition suggère, soit que les PCS d'une série de nombres devraient apparaître de façon uniforme (exemple avec probabilité $1/9 \simeq 0.111$). Ceci génère la question « Comment est-ce possible que mon intuition soit si faussée ? »

Ainsi, un des éléments ayant contribué à donner un caractère mystérieux à cette loi est qu'elle représente un paradoxe entre l'intuition et la réalité mathématique. Pour mettre en exergue ce paradoxe, considérons le jeu suivant entre des joueurs A et B : un média (journal, magazine, revue, etc.) est ouvert au hasard et on observe le premier nombre x que l'on voit (à l'exception du numéro de la page). Si $PCS(x) > 3$, le joueur A a gagné, sinon c'est B qui gagne.

Selon l'intuition, le jeu semble avantageux pour A car, comme il a 6 « chances » de gagner (en observant un $4, 5, ..., 9$), il s'attend à une probabilité de gain de 0.666. En réalité, c'est B qui devrait gagner avec une probabilité ≈ 0.60 parce que la loi de probabilité de $PCS(x)$ est très proche de la loi de Benford (ceci sera justifié plus tard). Ce paradoxe est un fait connu en psychologie sous le nom « biais d'équiprobabilité » (Lecoutre 1992), qui veut que si K événements élémentaires sont possibles, alors les humains ont tendance à penser que chacun peut se produire avec probabilité $1/K$. Par exemple, si dans le lancer de 2 dés, on considère les événements : « la somme = 11 », et « la somme = 12 », alors plus de la moitié des individus vont penser que chacun de ces événements à la même probabilité de se produire (alors que le premier est 2 fois plus probable que le second). De plus, aucun des facteurs comme le sexe, le bagage de connaissance (Licence-Master-Doctorat), le type de formation ou la pratique des jeux de hasards ne modifie sensiblement cette proportion. Ceci a fait dire aux chercheurs que « ... intuitions (correct as well as incorrect) are often very robust, being deeply rooted in the person's basic mental organisation » (Lecoutre 1992, p. 561). Ainsi, le paradoxe de la loi

de Benford est plutôt de nature psychologique que mathématique.

Mais ceci n'explique ni l'ubiquité empirique de la loi de Benford ni pourquoi les probabilités qui apparaissent dans cette loi ont l'expression $\log_{10}(1+\frac{1}{d})$ ou quelque chose d'empiriquement approchant. Ainsi, de nombreux chercheurs en mathématiques, probabilité et statistique ont tenté de fournir des réponses à ces questions. On a vu à la section précédente quelques solutions au problème : pour une séquence $X_1, \ldots, X_n$ iid $\sim F$ si F satisfait certaines conditions, on a $D_i = PCS(X_i) \sim \mathscr{B}_1(10)$. Mais à ce jour finalement peu d'autres cas ont été trouvé, et pour tenter d'expliquer l'ubiquité empirique de la $\mathscr{B}_1(10)$, on a cherché à étendre le champ des recherches au cas où $D_i = PCS(X_i) \approx \mathscr{B}_1(10)$, où le sens du symbole $\approx$ signifiant « approximativement » doit être précisé. Avec cette nouvelle formulation du problème, plusieurs autres réponses ont été trouvées.

En particulier, Morrow (2014) introduit le concept suivant :

Définition 1. Une variable aléatoire $X \sim F$ $\epsilon-$satisfait à la loi de Benford si

$$\max_{d\in\{1,\ldots,9\}} \left|\mathbb{P}\left[PCS\left(X\right)=d\right] - \log_{10}\left(1+\frac{1}{d}\right)\right| \leq \epsilon$$

Il démontre ensuite le théorème suivant :

Théorème 2. *Soit une variable aléatoire $X \sim F$ possédant une densité f. Pour tout $\epsilon > 0$, il existe un $\alpha^*(\epsilon)$ tel que pour tout $\alpha \geq \alpha^*(\epsilon)$ et pour tout $\sigma > 0$, $(d/\sigma)^\alpha$ $\epsilon-$satisfait à la loi de Benford.*

La preuve de ce résultat utilise des outils mathématiques relativement simples. Ce résultat démontre entre autres qu'il est facile de transformer des données en des lois de Benford : il suffit de transformer X en l'élevant à une certaine puissance ($\alpha = 10$ et $\sigma = 1$). Il existe quelques autres résultats de ce type qui ont permis de montrer que plusieurs situations donnent lieu à (au moins une approximation de) la loi de Benford.

Enfin, il reste à répondre à la question : dans quel cas peut-on s'attendre à observer la loi de Benford ou quelque chose s'en approchant. Durtschi *et al.* (2004) citent la règle « du pouce » suivante :

> *...the numbers in sets that conform to the Benford distribution are second generation distributions, that is, combinations of other distributions. If distributions are selected at random and random samples are taken from each of these distributions, then the significant- digit frequencies of the combined samplings will converge to Benford's distribution, even though the individual distributions may not closely follow the law (Hill, 1995, 1998). The key is in the combining of numbers from different sources. In other words, combining unrelated numbers gives a distribution of distributions, a law of true randomness that is universal.*

Ainsi , selon le Théorème de Hill (1995), une suite de valeurs obtenues en sélectionnant, selon certaines contraintes, différents échantillons dans différentes populations pour des variables diverses donne finalement une loi de Benford. C'est un équivalent, au fond, du théorème central limite : un échantillonnage bien fait doit mener à une loi particulière. Boyle (1994) montre que la multiplication entre elles de variables indépendantes conduit à la loi de Benford. Autrement dit, la loi de Benford serait naturelle si les nombreux facteurs qui expliquent telle ou telle grandeur agissent multiplicativement. Et pour citer des exemples où la loi de Benford apparaît de façon empirique, les mêmes auteurs continuent :

> *Most accounting-related data can be expected to conform to a Benford distribution, and thus will be appropriate candidates for digital analysis (Hill 1995). Such is the case because typical accounts consist of transactions that result from combining numbers. For example, accounts receivable is the number of items purchased multiplied by the price per item. Similarly, accounts payable and most revenue and expense accounts are expected to conform.*

1.3.3 Applications de la loi de Benford $\mathscr{B}_1(10)$ à la détection de fraudes

L'hypothèse fondamentale que des auditeurs veulent exploiter est que, si pour des raisons prédéterminées, comme celles évoquées à la section 1.3.2, on sait que les données devraient se conformer à la loi $\mathscr{B}_1(10)$ et que, si elles ont été manipulées, elles vont en dévier, alors l'application d'un test d'adéquation à la $\mathscr{B}_1(10)$, comme ceux de la section 1.3.4, peut être utile. Si le test rejette, il y a suspicion de manipulation de données et un audit plus sévère est recommandé. Si le test ne rejette pas, les données sont supposées n'avoir pas subies de manipulations frauduleuses. Évidemment un test d'adéquation statistique comporte des risques d'erreurs de type I et II et ceux-ci sont contrôlés par le niveau du test choisi et par la puissance du test utilisé.

Dans ce contexte, les succès de cette approche sont nombreux (mais il est probable qu'il y ait un biais de publication et que les insuccès aient été camouflés. D'ailleurs, les exemples publiés ont presque tous utilisé le test du χ^2, qui, comme on le verra à la section 1.3.4 est un des moins puissants de ces tests. Des applications réussies concernent des données de recensement (Hill 1995), des données sur le prix d'action boursières (Ley 1996), sur le débit de rivières (Nigrini et Miller 2007), sur des données de volcanologie (Geyer et Martí 2012 et évidemment sur la détection de fraudes dans le contexte d'analyse de rapports d'activités d'entreprises (Carslaw 1988; Durtschi *et al.* 2004; Guan *et al.* 2006; Kinnunen et Koskela 2003; Niskanen et Keloharju 2000; Nigrini 2005; Skousen *et al.* 2004; Thomas 1989; Caneghem 2002, 2004), de vérification fiscale (Nigrini 1996; Nigrini et Mittermaier 1997; Watrin *et al.* 2008) et d'études scientifiques (Diekmann 2007; Günnel et Tödter 2008; Hein *et al.* 2012; Tödter 2009; Schüpfer *et al.* 2012) et la manipulation de résultats

d'élections (Pericchi et Torres 2011).

Du point de vue des auditeurs, (Durtschi *et al.* 2004), les avantages de cette approche sont particulièrement évidents. Une procédure souvent suivie pour vérifier des comptes est d'en sélectionner un certain nombre au hasard et de pousser à fond leur étude par des moyens lourds. Supposons que p_F désigne la probabilité qu'une fraude ait été commise dans un compte. Cette quantité est en général inconnue et spécifique au type d'entreprise considérée, mais les auditeurs sont habitués, sur la base d'autres éléments ou d'information externe, à produire une estimation de cette quantité. Ceci leur permet de mettre en place une stratégie d'échantillonnage au hasard en regard des coûts de ces investigations lourdes. Évidemment le pourcentage des fraudes qui pourront être détectés par cette stratégie est de p_F.

Maintenant, supposons qu'un test statistique de niveau α ait une puissance de

$$\gamma(\alpha) = \mathbb{P}[signaler\ une\ fraude \mid Fraude] = \mathbb{P}[Rejeter\ la\ loi\ de\ Benford \mid Fraude]$$

alors par le théorème de Bayes

$$\mathbb{P}\left[Fraude \mid Test\ rejette\ la\ \mathscr{B}_1(10)\ \right] = \frac{\gamma\left(\alpha\right) \times p_F}{\gamma\left(\alpha\right) \times p_F + \left(1 - p_F\right)\left(1 - \gamma\left(\alpha\right)\right)}$$

Exemple 3. Si $p_F = 0.03$ et que le test a pour puissance $\gamma\left(\alpha\right) = 0.75$, alors

$$\mathbb{P}\left[Fraude \mid Test\ rejette\ la\ \mathscr{B}_1(10)\ \right] = \frac{0.75 \times 0.03}{0.75 \times 0.03 + \left(1 - 0.03\right)\left(1 - 0.75\right)} = 0.085$$

Ainsi, l'emploi d'un test d'adéquation à la loi de Benford permet presque de tripler le taux de détection de fraude par rapport à l'échantillonnage simple (unassisted random sampling) tout en contrôlant l'effort et le coût des auditions plus poussées.

1.3.4 Comment vérifier si elle tient dans un cas donné : Test d'adéquation à la loi de Benford $\mathscr{B}_1(10)$

Cet axe de la recherche reste crucial. En effet, savoir que la loi de Benford est régulièrement présente dans un grand nombre de situations pratiques et permet de détecter les fraudes fait rêver les partisans d'un ordre moral. Pour ce faire, il faut pouvoir vérifier la conformité, ou l'adéquation, d'un jeu de données à cette loi.

L'outil statistique de prédilection est le test d'adéquation qui permet de tester l'hypothèse nulle $H_0 : D \sim \mathscr{B}_1(10)$, où $D = PCS(X)$ avec $X \sim F$ en contrôlant les risques d'erreur de :

— Type I, soit déclarer faussement que D n'obéit pas à la $\mathscr{B}_1(10)$, ce qui mène à une fausse alerte. Il est contrôlé en fixant le niveau du test à une valeur α généralement

fixée à 1 ou 10 en fonction du nombre de situations que l'on est prêt à auditer de façon plus poussé même si in fine, aucune fraude n'est détectée.

— Type II, soit déclarer faussement que le jeu de données est conforme à la $\mathscr{B}_1(10)$ et donc de laisser filer un fraudeur. Il est contrôlé par la puissance du test d'adéquation.

Nous pourrions imaginer un grand nombre de tests d'adéquation comme dans le cas de la loi Normale (section 1.2.1) et se pose alors le problème de choisir lequel doit être utilisé. De façon générale, et notamment dans un contexte de détection de fraudes, il est normal que, à un seuil d'erreur de type I donné, on cherche à utiliser le test le plus puissant.

Malheureusement comme le bon sens l'aurait voulu, vu l'importance de cette question, très peu de recherches ont été menées sur ce problème, et celles-ci se concentrent essentiellement sur les 10 dernières années malgré le fait que les procédures pour dégager d'un ensemble de tests celui qui semble en général le plus recommandable ont une longue histoire et sont maintenant bien établies . En effet, citons Barabesi *et al.* (2017) : « *However, most of the applications in this area rely either on diagnostic checks of the data, or on informal decision rules. Formal goodness-of-fit testing of the law poses some challenging statistical problems that include both the choice of the most appropriate version of the null hypothesis, and derivation of the exact distribution of the test statistic. Nontrivial solutions to these issues are required to satisfy a crucial requirement for many antifraud exercises, that is, to ensure adequate control over the number of falsely discovered anomalies.* ». De la même façon, Joenssen (2014) déplore : *« Given the 'size-of-the-prize' in this frame of reference, when testing with a more powerful goodness-of-fit test, it seems conspicuous that evaluations of power, and comparison thereof, are lacking in literature. »*

Cette situation rend impuissant les pouvoirs publics qui ont pour mission de protéger le public des fraudes. A Stresa en Italie, s'est tenu en Juillet 2019, une conférence internationale sponsorisée par le Programme Anti-Fraude Hercule III de l'Union Européenne, géré par l'OLAF, l'Office européen de Lutte Anti-Fraude `https://ec.europa.eu/anti-fraud/`. Les motivations pour cette conférence sont décrites dans le document « detailed information [1] » de la conférence :

> *Many anti-fraud techniques fall in the class of unsupervised statistical methods, with outlier detection and (robust) cluster analysis playing a prominent role. The rationale is that the bulk of international trade data is made of legitimate transactions, possibly clustered due to unknown latent factors, and major frauds may stand out as highly suspicious anomalies.*
>
> *In this conference we focus on a different approach that has attracted considerable interest in many anti-fraud contexts, but whose potential for international trade is largely unexplored. The approach is based on testing conformance*

1. `https://ec.europa.eu/jrc/sites/jrcsh/files/brochure_benford_call4papers_final2.pdf`

> *to the Benford's Law (BL), which relies on the intriguing fact that in many natural and human phenomena the leading – i.e. the first significant – digits are not uniformly scattered, as one could naively expect.*
>
> *The general goals of the conference are to provide insight on the suitability of the BL for fraud detection, discuss the state of the art in the field and reflect on new BL-based procedures, for example goodness-of-fit testing of the law within a contamination model where frauds generate outliers. The practical target is to arrive at conformance tests of wide applicability that could be taken seriously by anti-fraud users, with empirical properties closely matching the expected nominal ones, and with good inferential ability both under the null hypothesis of no contamination and when frauds are present.*

Le troisième paragraphe dudit texte de motivation confirme l'appréciation faite plus haut concernant l'état des lieux sur l'étude de la conformité à loi de Benford : on reconnaît maintenant que la loi de Benford jouit d'une certaine ubiquité, tant empirique que théorique, pouvant être exploitée pour définir une certaine « normalité ». L'examen visuel d'une représentation graphique ne suffit pas et pour que cet outil soit pris au sérieux, on doit tester de façon formelle l'adéquation de données à cette loi.

Le même paragraphe identifie une direction de recherche : les qualités d'un test d'adéquation se mesurent par sa puissance contre les alternatives qui apparaissent pertinentes. En s'inspirant de l'argumentaire (« rationale ») décrivant les approches « classiques » de détection des fraudes, un type d'alternatives pertinentes est celui dit « contamination model where frauds generate outliers », (aussi appelée mixture). Il convient donc en premier lieu de développer des tests d'adéquation de qualités contre ces alternatives.

L'état des lieux concernant la vérification de cette conformité à la $\mathscr{B}_1(10)$ semble être le suivant. Dans les domaines d'application où les intervenants sont peu sophistiqués en statistique, cette conformité se fait par l'examen de graphiques. Mais un tel examen visuel ne contrôle aucun des risques d'erreur de façon précise. Wallace (2002) propose une approche légèrement plus formelle et suggère la procédure : si la moyenne de données est supérieure à la médiane et que le coefficient d'asymétrie est positif, alors les données pourraient suivre une loi de Benford. Encore là, aucun des risques d'erreur n'est contrôlé. Dans d'autres domaines, où la connaissance des outils de base de la statistique s'est mieux propagé, on utilise le test du χ^2 de Pearson (1900).

Dans le présent contexte, ce test prend la forme suivante. Dénotons $\pi_d = \log_{10}\left(1 + \frac{1}{d}\right)$,la $\mathbb{P}[D = d]$ quand $D \sim \mathscr{B}_1(10)$, pour $d = 1, \ldots, 9$. Pour le jeu de données $X_1, \ldots, X_n$, iid $\sim F$, soit $\hat{p}_d = n^{-1} \sum_{i=1}^{n} \{PCS(X_i) = d\}$ la proportion de PCS égaux à d. On rejette $H_0 : D \sim \mathscr{B}_1(10)$ au niveau asymptotique α quand

$$\chi^2 = n \sum_{d=1}^{9} \frac{(\hat{p}_d - \pi_d)^2}{\pi_d} > x^2_{8,1-\alpha}$$

où $x^2_{8,1-\alpha}$dénote le quantile d'ordre $1-\alpha$ de la loi χ^2_8. Ce test à l'avantage d'exister depuis plus d'un siècle ; il est bien connu et facile d'utilisation et ici, ne présente pas la part d'arbitraire qui existe dans le cas de données continues. Comme remarqué dans la section 1.2.1 dans le cas de données continues, ce test n'était pas très puissant, ce désavantage semble avoir déteint sur la perception qu'en ont certains auteurs pour le présent cas de données discrètes issues de la $\mathscr{B}_1(10)$. Par exemple, Morrow (2014) écrit (sans preuve) : *« Pearson's χ^2 test is a natural candidate for testing whether a sample satisfies Benford's Law, however, due to its low power for even moderate sample sizes it is on unsuitable »*. Mais en fait, il y a eu très peu de travaux sur la puissance du test du χ^2 pour le cas de la $\mathscr{B}_1(10)$. Parmi les plus récents, on peut citer Joenssen (2014) qui, après une étude de simulation, conclut : « *The results indicate that the current standard method, the χ^2 good- ness-of-fit test, is inferior for a wide range of alternative distributions to two other tests.* ». Dans un même état d'esprit, Lesperance *et al.* (2016) conclut aussi après une étude de simulation : *« To assess conformance with Benford's Law, investigators should perform statistical tests ; the CvM statistic U^2_d is recommended and if contamination is expected in the larger values of the first significant digit, Pearson's chi-square statistic* ». Ainsi, le test du χ^2 ne semble pas être un outil d'utilisation générale pour tester la conformité d'un jeu de données à la $\mathscr{B}_1(10)$, même s'il semble bien se comporter face à un type de contamination. Tout comme dans le cas de données continues, on a donc cherché des tests d'adéquation à la $\mathscr{B}_1(10)$ qui n'auraient pas les défauts (perçus) du test du χ^2. Certains auteurs ont proposé des statistiques de test somme toute assez proches de χ^2. Par exemple, Leemis *et al.* (2000) considèrent la statistique de test (basée sur la distance du *max*, aussi dite de Tchebycheff) :

$$m = \max_{1\leq d\leq 9} |\hat{p}_d - \pi_d|$$

alors que Cho et Gaines (2007) propose (basé sur la distance euclidienne) :

$$\mathtt{d} = \sqrt{\sum_{d=1}^{9} (\hat{p}_d - \pi_d)^2}$$

Drake et Nigrini (2000) suit cette même lignée et propose la statistique de test appelé la « Mean Average Deviation(MAD) »

$$MAD_d = \frac{1}{9}\sum_{d=1}^{9} |\hat{p}_d - \pi_d|$$

On peut noter que ces statistiques de tests, intuitivement raisonnables, ne sont pas basées sur des principes statistiques reconnus. En outre, leur loi exacte sous H_0 et leur fonction de puissance ne possèdent pas de forme explicite et doivent être simulées par Monte-Carlo.

« Partisan du moindre effort », une adaptation des tests développés pour le cas de données continues ont été effectués. Notons,

— Morrow (2014) pour la statistique de Kolmogorov :

$$D = \sqrt{n} \sup_{1 \leq d \leq 9} |S_d - T_d|$$

avec $S_d = \sum_{j=1}^{d} \hat{p}_j$, $T_d = \sum_{j=1}^{d} \pi_j$

— Lesperance *et al.* (2016)

— pour la statistique de Cramer-von Mises

$$W^2 = n \sum_{d=1}^{9} Z_d^2 t_d$$

avec $Z_d = S_d - T_d$ et $t_d = \frac{\pi_d + \pi_{d+1}}{2}$ pour $d = 1, \ldots, 8$ et $t_9 = \frac{\pi_9 + \pi_1}{2}$

— pour la statistique de Watson

$$U^2 = n \sum_{d=1}^{9} (Z_d - \bar{Z})^2 t_d$$

avec une variante proposée par Joenssen (2014), la statistique Watson modifiée par Freedman

$$U^2_{Freed} = \frac{n}{9} \left(\sum_{d=1}^{9} Z_d^2 - \frac{1}{9} \left(\sum_{d=1}^{9} Z_d \right)^2 \right)$$

— pour la statistique d'Anderson-Darling

$$A^2_d = n \sum_{d=1}^{8} \frac{Z_d^2 t_d}{T_d \left(1 - T_d\right)}$$

— Joenssen (2013a) pour la statistique Shapiro-Wilks très utilisée pour le cas de la loi Normale

$$J_p^2 = sgn\left(cor\left(\hat{\boldsymbol{p}}, \boldsymbol{\pi}\right)\right) cor\left(\hat{\boldsymbol{p}}, \boldsymbol{\pi}\right)$$

$\hat{\boldsymbol{p}}$ est le vecteur des $\hat{p}_d$ et réciproquement pour $\boldsymbol{\pi}$

— Judge et Schechter (2009) pour la statistique

$$a^* = \frac{\left|\bar{D} - \mu_e\right|}{9 - \mu_e}$$

où μ_e est l'espérance de $\mathscr{B}_1(10)$ (≈ 3.440237) et $\bar{D}$ est la moyenne empirique des données.

Le package *R* BENFORDTESTS (version 1.2.0, 2015), réalisé par Joenssen (2013b), intègre tous ces tests à l'exception de A_d^2, W_d^2 *et* U_d^2. On les retrouve via les fonctions

chisq.benftest, (χ^2), *ks.benftest* (d), *mdist.benftest* (M) , *edist.benftest* (D), *usq.benftest* (U^2_{Freed}), *jpsq.benftest* (J^2_p) et *meandigit.benftest* (A^*), auquel s'ajoute un test basé sur la statistique T de Hotelling qui n'est pas bien expliqué et semble moins intéressant car certains hyper paramètres doivent être choisis (sans critère précis à ce jour) et les seuils critiques disponibles sont les valeurs asymptotiques seulement.

Au mieux de nos connaissances, cette nomenclature couvre tous les tests étudiés à ce jour dans la littérature pour la loi de Benford, sauf le test bayesien de Geyer et Williamson (2004) dont nous ne parlerons pas car il n'offre aucun contrôle des erreurs de type I et II. Par ailleurs, comme dans bien des cas, la loi exacte de la statistique de test sous H_0 est inconnue, le package BENFORDTESTS offre la possibilité de simuler les seuils exacts par Monte-Carlo (sauf pour le test basé sur la statistique de Hotelling). Ceci est important dans les applications car dans le cas de la $\mathscr{B}_1(10)$, les seuils pour le cas de données continues ne s'appliquent pas, ce qui détruit le contrôle des erreurs et rend ces tests guère plus utiles que l'inspection visuelle.

Concernant le problème de choisir un test parmi cet ensemble, encore là au mieux de nos connaissances, quatre études seulement ont été menées (Geyer et Williamson (2004) ;Morrow (2014); Joenssen (2014); Lesperance *et al.* (2016)). Ces études n'ont pas permis l'identification parmi les tests plus haut d'un « gagnant universel » car chacun de ces tests ont leur force et leur faiblesse en regard des jeux d'alternatives différents dans chacun de ces articles. Cependant, si une recommandation doit être faite, les deux dernières études s'entendent sur le fait que le test de Watson (ou sa variante de Freedman) est globalement parmi les meilleurs. Par contre, les tests basés sur a^* et J^2_p ne doivent pas être utilisés car dans bien des cas, ils sont biaisés (puissance $<$ niveau du test).

Dans la suite de ce travail, nous allons chercher à bonifier ces résultats pour en arriver à des recommandations plus précises.

1.4 Smooth test (Test lisses)

Le « SMOOTH TEST » a été introduit pour la première fois par Neyman (1937) comme une généralisation du test χ^2 de Pearson (1900) puis étendu par Barton (1953, 1955, 1956) . L'idée principale est de mener un test où l'hypothèse alternative H_1 incorpore l'hypothèse H_0 comme cas particulier à travers une série de fonctions orthonormées. Il peut être considéré comme un compromis entre les tests « omnibus » [2] (Test de Kolomorov, Test de Cramer Von Mise, etc...), avec une puissance généralement faible dans la plupart des directions de sortie de l'hypothèse nulle, et les tests directionnels, qui concentrent leur puissance dans la détection de écarts spécifiques par rapport au modèle nul. Cette approche de Neyman (1937) permet de maîtriser les risques de type I et II et donc amène a une prise de décision concrète.

2. pouvant détecter tous les écarts à la loi postulée sous H_0

L'idée de base de Neyman est simple. Supposons $X_1, X_2, \ldots, X_n$ une suite de variables continues ou discrètes aléatoires iid , de densité $f(x, \beta)$ avec β un vecteur de taille p des paramètres de nuisance. Si f est connue et explicite, par exemple pour la loi uniforme sur $[0, 1]$ ou la loi exponentielle de paramètre λ_0, alors β est un vecteur vide. Par contre, comme dans un bon nombre de cas, si f est inconnue avec des paramètres comme la loi Normale $\mathcal{N}(\mu, \sigma)$ alors $\beta = (\mu, \sigma)^t$.
Pour construire un « MODELE LISSE » (SMOOTH MODEL), nous imbriquons f dans une famille de distribution d'ordre k notée g_k appelé « SMOOTH MODEL » :

$$g_k(x, \theta, \beta) = C(\theta, \beta) \exp\left\{\sum_{i=1}^{k} \theta_i h_i(x, \beta)\right\} f(x, \beta) \tag{1.4.1}$$

avec $C(\theta, \beta)$ une constante de normalisation et $h_i(x, \beta)$ un ensemble de fonctions orthonormales relativement à $f(x, \beta)$ tel que $h_0(x, \beta) = 1, \forall\, x$.

Définition 4. Soit X une variable aléatoire, discrète ou continue, g sa densité de probabilité et G sa fonction de distribution cumulative (*cdf*) ou de répartition. Soit f la densité de probabilité de la LOI DE BENFORD et F sa *cdf*, ayant le même support que G et F^{-1}, la fonction quantile. Supposons que si $f = 0$ alors $g = 0$.

La densité de comparaison entre G et F noté $d(u; F, G)$ est donnée par :

$$d(u; F, G) = \frac{g(F^{-1}(u))}{f(F^{-1}(u))} \tag{1.4.2}$$

avec $u = F(x)$.

La distribution de comparaison est donnée par :

$$D(u) = \int_0^u d(s; F, G)\, \delta s$$

Handcock et Martina (1999), montre que dans le cas discret, la densité de comparaison est une fonction en escalier $d(u; F, G) = \frac{g(x_r)}{f(x)_r}$ pour $F(x_{r-1}) < u < F(x_r)$. Parzen (2004, p.658), de son côté, démontre que dans le cas où X est continue $D(u) = G(F^{-1}(u))\ \forall\, u \in [0, 1]$ et dans le cas où X est discret, $D(u)$ est linéaire par morceaux entre les valeurs $u_r = F(x_r)$ où $x_0 < \cdots < x_R$ sont des points de F et $D(u_r) = G(F^{-1}(u_r)) = G(x_r)$.

Un « MODELE LISSE » (SMOOTH MODEL) en général est donc de la forme :

$$g_k(x) = g(x)\, d_m(F(x); F, G)$$

Méthodes	$d_m(u; F, G)$
Neyman (1937)	$\exp \theta_0 + \sum_{j=1}^{m} \theta_j h_j(u) - K_{\boldsymbol{\theta}}$
Barton (1956)	$1 + \sum_{j=1}^{m} \tau_j h_j(u)$
Devroye et Györfi (1985)	$\max\left\{0, 1 + \sum_{j=1}^{m} \tau_j h_j(u)\right\} / K_{\boldsymbol{\tau}}$
Gajek (1986)	$\max\left\{0, 1 + \sum_{j=1}^{m} \tau_j h_j(u) - K_{\boldsymbol{\tau}}\right\}$

TABLE 1.2 – Formes explicites de $d_m(u; F, G)$. L'ensemble des fonctions h_j forme une base orthonormée en $[0, 1]$.
$K_{\boldsymbol{\tau}}, K_{\boldsymbol{\theta}}$ sont des constantes normalisées où $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_m)$ et $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ avec $\tau_j = \int_0^1 h_j(u)\, d_m(u; F, G)\ \forall\, j = 1, \ldots, m$.

où $d_m(F(x); F, G)$ est une nouvelle forme d'écriture de l' équation (1.4.2) au moyen des séries de fonctions m de $F(x)$ noté $h_j(F(x))$ tel que $h_j(F(x))$ est un ensemble de fonction orthonormale.

Le tableau 1.2 est un récapitulatif des formes explicites proposées dans la littérature pour d_m.

Nous remarquons la correspondance entre la proposition de Neyman (1937) dans équation (1.4.2) et 1.4.1.

Notre hypothèse $H_0 : f(x) = \mathscr{B}_1(10)$, nécessitant un test non paramétrique du fait que f soit inconnue avec peut être un nombre infini de paramètre nuisant devient donc $K_0 : \theta = 0$. De même l'hypothèse alternative $H_1 : f(x) \neq \mathscr{B}_1(10)$ revient à $K_1 : \theta \neq 0$. On remarque que $K_1 \Rightarrow H_1$ mais la réciproque n'est pas vraie. Cette nouvelle hypothèse peut se résumer par la question : « Les données sont-elles bien décrites par la fonction de densité de probabilité $f(x, \beta)$, ou est-ce que l'un des éléments de la famille $\{g_k(x, \theta, \beta)\}$ donne une description nettement meilleure ? »

Le test statistique T est :

$$\mathsf{T} = \frac{1}{n} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} h_j(F(x_i)) \right)^2$$

T converge asymptotiquement vers χ^2_m sous H_0(Thas, 2010).

Ce test peut être appliqué aux données continues comme discrètes comme le montre Rayner *et al.* (2009) en donnant l'application sur les lois de poisson , binomial et géométrique dans leur chapitre 8.

Chapitre 2

Applications des tests lisses au premier chiffre significatif

2.1 Introduction

Dans ce chapitre, nous implémentons le test lisse décrit à la section 1.4 au premier chiffre significatif. Cela représente notre première contribution. Nous avons comparé les résultats obtenus aux travaux de Lesperance *et al.* (2016) et Joenssen (2013c). Nous arrivons à la conclusion que le test T_2 et le test data driven $T_{\hat{K}}$ que nous proposons pourraient être recommander dans le cadre de la détection de fraude basée sur le premier chiffre significatif.

2.2 Article : Tests d'adéquations lisses pour la loi de Newcomb-Benford

L'article a été publié dans la revue Mathématiques appliquées : déterministes et stochastiques en 2020.

Tests d'adéquations lisses pour la loi de Newcomb-Benford

Smooths Tests of Goodness-of-fit for the Newcomb-Benford distribution

Gilles R. DUCHARME[1],Samuel KACI[1], Credo VOVOR-DASSU[1]

[1]IMAG, Univ. Montpellier, CNRS, Montpellier, France

Réception : 28/02/2020
Acceptation : 19/05/2020
Publication en ligne : 22/05/2020

RÉSUMÉ. La loi de probabilité de Newcomb-Benford est de plus en plus utilisée dans les applications de la statistique, notamment en détection de fraude. Dans ces contextes, il importe de déterminer si un jeu de données est issu de cette loi de probabilité en contrôlant les risques d'erreur de Type I, soit de faussement identifier une fraude, et de Type II, soit de ne pas la détecter. L'outil statistique qui permet d'exécuter ce genre de tâche est le test d'adéquation. Pour la loi de Newcomb-Benford, le test d'adéquation le plus populaire est le test du khi-deux de Pearson dont la probabilité d'erreur de Type II est reconnue comme étant assez grande. En conséquence, d'autres tests ont été récemment introduits. Le but de ce travail est de proposer de nouveaux tests d'adéquation pour cette loi, basés sur le principe des tests lisses. Ces tests sont ensuite comparés aux meilleurs tests existants pour ce problème. Il en ressort que nos propositions sont globalement préférables aux tests existants et pourraient être utilisées dans les applications, notamment en détection de fraude. Un package de R,BENFORDSMOOTHTEST, est disponible sur le site GitHub pour effectuer nos tests.

ABSTRACT. The Newcomb-Benford probability distribution is becoming very popular in many areas using statistics, notably in fraud detection. In such contexts, it is important to be able to determine if a data set arises from this distribution while controlling the risk of a Type I error, i.e. falsely identifying a fraud, and a Type II error, i.e. not detecting that a fraud occurred. The statistical tool to do this work is a goodness-of-fit test. For the Newcomb-Benford distribution, the most popular such test is Pearson's chi-square test whose probability of a Type II error is known to be large. Consequently, other tests have been recently introduced. The goal of the present work is to build new goodness-of-fit tests for this distribution, based on the smooth test principle. These tests are then compared to some of their competitors. It turns out that the proposals of the paper are globally preferable to existing tests and should be seriously considered in fraud detection contexts, among others. The R package BENFORDSMOOTHTEST is available on GitHub to compute the test statistics.

MOTS-CLÉS. Loi de Newcomb-Benford, test d'adéquation, détection de fraudes, test lisse.

KEYWORDS. Newcomb-Benford's distribution, goodness-of-fit tests, fraud detection, smooth test.

1. Introduction

La loi de Newcomb-Benford (LNB) (Newcomb, 1881; Benford, 1938) annonce que sous certaines conditions, le premier chiffre significatif (*PCS*) d'une variable aléatoire continue positive X, $D = PCS(X)$, a pour probabilité $\mathbb{P}[D = d] = \pi_d = \log_{10}[1 + 1/d]$, $d \in \{1, 2, ..., 9\}$.

L'utilisation de cette loi de probabilité connaît une popularité grandissante dans de nombreux domaines, notamment en détection de fraudes fiscales (Nigrini, 1996), financières (Cerioli et al., 2018), comptables (Durtschi et al., 2004) et scientifiques (Hein et al., 2012). Elle est aussi utilisée comme modèle statistique dans des disciplines aussi variées que l'hydrologie (Drake et Nigrini, 2000), la volcanologie (Geyer et Martí, 2012), la sismologie (Sambridge et al., 2011) et pour l'étude du trafic de données internet (Arshadi et Jahangir, 2014), entre autres applications. Cette popularité émane d'une part, du fait qu'on la rencontre empiriquement très souvent dans les jeux de données réelles (Durtschi et al., 2004) correspondant à des données dites de « deuxième génération », soit le résultat d'opérations (produits, puissances, etc.) de données brutes. D'autre part, des raisons théoriques font qu'elle apparaît

aussi dans de nombreux contextes (Posch, 2008), sinon exactement du moins en tant qu'approximation de la réalité. En outre, elle interpelle l'intuition parce que, contrairement à ce que l'on pourrait naïvement penser, D n'apparaît pas avec une probabilité de 1/9 $\simeq 0.111$. Cette dissonance cognitive vient de ce que les psychologues appellent le « biais d'équiprobabilité » (Lecoutre, 1992). Ce biais est un des facteurs faisant que les *PCS* de nombres influencés par la pensée humaine sont plus proches de la loi uniforme sur $\{1, 2, ..., 9\}$ que de la LNB (Hill, 1998; Gauvrit et al., 2017). En particulier un fraudeur voulant trafiquer un jeu de données va inconsciemment avoir tendance à uniformiser leur *PCS*. Cette discordance offre une prise permettant à des auditeurs de détecter les fraudes, et cet outil s'est avéré particulièrement efficace dans le cas de fraudes fiscales (Ausloos et al., 2017). Un site internet, le *Benford online bibliography* (Berger et al., 2015) recense la quasi-totalité des publications sur la LNB, autant les résultats théoriques que les applications.

Il est donc souvent impératif de pouvoir déterminer si un jeu de données se conforme à la LNB en contrôlant les risques d'erreurs de Type I et II. Dans le contexte de la détection de fraudes, ces erreurs correspondent à faussement suspecter une fraude (Type I) avec pour conséquence le coût d'une audition approfondie subséquente, soit de ne pas détecter un jeu de données trafiquées (Type II) qui mènera ensuite à la prise de décisions erronées. L'outil statistique pour effectuer cette tâche est un test d'adéquation statistique (*goodness-of-fit test*). Pour ce problème, le test d'adéquation de préférence (Morrow, 2014) a longtemps été le test du χ^2 de Pearson, (Lesperance et al., 2016, eq.3). Si n_d est le nombre de fois dans un jeu de n données que $PCS = d$, alors le test s'effectue en calculant la statistique de test :

$$\chi^2 = \sum_{i=1}^{9} \frac{(n_d - n\pi_d)^2}{n\pi_d}, \tag{1}$$

qui obéit approximativement, si n est grand, à la loi χ^2_8 si l'hypothèse nulle H_0 voulant que les données suivent une LNB est juste. On rejette H_0 si χ^2 dépasse le quantile de cette loi approprié au risque d'erreur de Type I souhaité. Ce test est simple d'utilisation mais sa puissance, qui vaut 1 moins la probabilité d'une erreur de Type II, n'est en général pas reconnue comme étant très élevée (Morrow, 2014). Ainsi d'autres tests d'adéquation ont été proposés, certains étant des adaptations de tests développés pour des données continues et basés sur des principes statistiques reconnus comme les tests de Carmer-von Mises ou de Watson dans (Lesperance et al., 2016), d'autres sur la base de considérations intuitives (Joenssen, 2014). Un certain nombre de ces tests (voir Section 3) sont disponibles dans le package R **BENFORDTEST** (Joenssen, 2013a).

Malgré l'importance de l'utilisation d'un « bon » test d'adéquation pour la détection de fraude, ce n'est que récemment (Morrow, 2014; Joenssen, 2014; Lesperance et al., 2016) que les premières analyses de la puissance de ces différents tests à la LNB ont été réalisées. La procédure pour comparer entre eux plusieurs tests d'adéquation est bien rodée : on détermine une liste d'hypothèses alternatives couvrant les écarts que l'on estime plausibles à H_0, on génère des pseudo-échantillons de chacune de ces alternatives, on applique les tests au même niveau, soit la probabilité d'une erreur de Type I (en général 0.05), puis on constate s'ils ont rejeté ou non H_0. En répétant ceci un grand nombre de fois, on obtient une approximation de la puissance des différents tests que l'on peut ensuite comparer entre eux. En général, aucun test n'est uniformément le plus puissant, chacun ayant ses forces et ses faiblesses en regard des alternatives considérées. Un « bon » test se range régulièrement parmi les tests les plus puissants. Dans ce contexte, l'introduction d'un nouveau test se justifie si on peut montrer qu'il se range aussi parmi

les plus puissants pour des alternatives courantes, ou s'il est performant pour de nouvelles alternatives importantes qui n'avaient pas auparavant été considérées.

En détection de fraudes, il s'ajoute à cette procédure d'évaluation de la qualité d'un test d'adéquation une dimension supplémentaire. Maintenant que l'existence d'outils de détection basés sur la LNB est bien connue, un fraudeur astucieux va chercher à trafiquer les données de façon à passer inaperçu (Gauvrit et al., 2017). En effet, la crainte d'être détecté est un puissant frein à la fraude (Abdullahi et Mansor, 2015). Si le fraudeur sait que l'auditeur de ses données trafiquées exploitera tel test d'adéquation, il essaiera de faire en sorte que ce test ne détecte pas d'écart à la LNB. Dans ce contexte, chaque nouveau test ajoute une contrainte supplémentaire augmentant la difficulté de sa tâche. En outre, avec plusieurs tests utilisés en batterie, l'auditeur a aussi le choix de moduler leur emploi d'une façon inconnue du fraudeur, augmentant ainsi les risques de détection. Dans ce contexte, l'introduction d'un nouveau test d'adéquation est justifiée non seulement par sa bonne puissance, mais aussi parce que son existence peut complexifier la tâche du fraudeur. En ce domaine, viser l'éradication de la fraude est quasi impossible ; on cherche de façon plus réaliste à la rendre difficile pour aider à sa prévention Abdullahi et Mansor (2015).

La famille des tests lisses (*smooth tests*) introduite par Neyman (Neyman, 1937) s'applique à des données autant discrètes que continues. Ces tests sont plus complexes à développer, car ils sont spécifiques à la loi de probabilité postulée en H_0 mais au fil des années, leurs grandes qualités ont été reconnues et ceci a mené Rayner et Best (1990, p.9) à la recommandation : « Don't use those other methods–use a smooth test ! ». Une version plus moderne, pilotée dans les données, a été proposée par Ledwina (1994) et a conduit Inglot et al. (1994) à affiner cette recommandation en : « Use a data-driven smooth test ! » . Mais jusqu'à présent, ces tests n'ont pas été développés pour tester l'adéquation à la LNB.

Le but du présent travail est de développer différentes variantes des tests lisses pour le cas de la LNB et d'en étudier les qualités. La Section 2 rappelle les éléments théoriques permettant de construire une stratégie de test lisse et donne l'expression des statistiques de test pour le cas de la LNB. Quelques variantes sont introduites ainsi que des résultats théoriques concernant les puissances. La suite du travail explore les avantages d'inclure ces tests lisses dans une procédure de détection de fraude. La Section 3 présente une liste assez exhaustive des tests d'adéquation compétiteurs aux tests lisses et précise ceux qui sont retenus par la suite. La Section 4 présente les alternatives qui sont considérées pour la comparaison des puissances des tests retenus. La Section 5 présente les résultats d'une expérience de simulation qui montre que l'introduction des tests lisses est tout-à-fait justifiée selon les critères évoqués plus haut.

2. Test lisse pour la LNB

Le théorème suivant explique comment construire une famille, indexée par l'entier K, de tests lisses d'adéquation pour l'hypothèse nulle $H_0 : X \sim f(\cdot)$. Ce théorème est bien connu et une référence est Thas (2010) par exemple.

Théorème 2.1. *Soit $X_1, ..., X_n$ des copies indépendantes d'une variable aléatoire X de densité $f(\cdot)$ par rapport à une mesure ν. Soit $\{h_0(\cdot) \equiv 1, h_k(\cdot), k = 1, 2, ...\}$ une suite de fonctions orthonormales par rapport à $f(\cdot)$; plus précisément, $\int h_k(x)h_{k'}(x)f(x)d\nu(x) = \delta_{kk'}$, la fonction delta de Kronecker. Soit $U_k = n^{-1/2}\sum_{i=1}^{n} h_k(X_i)$ et pour un entier $K \geq 1$, soit $T_K = \sum_{k=1}^{K} U_k^2$. Alors sous H_0, $T_K \xrightarrow{L} \chi^2_K$, la loi khi-deux à K degrés de liberté, et un test de niveau asymptotique α rejette H_0 si la valeur observée de T_K dépasse $x^2_{K,1-\alpha}$, le quantile d'ordre $1-\alpha$ de cette loi χ^2_K.*

Nous spécialisons maintenant ce théorème au cas où $f(\cdot)$ est la densité de la LNB. Pour ce faire, il faut déterminer des fonctions orthonormales $h_k(\cdot)$. Comme tous les moments de la LNB existent, nous pouvons, à l'instar de Neyman (1937) et de nombreux autres auteurs par la suite, choisir des polynômes. Le théorème suivant est aussi connu (Boulerice et Ducharme, 1997). Dans la suite l'indice « 0 » dénote un opérateur probabiliste calculé sous $H_0 : X \sim f(\cdot)$.

Théorème 2.2. *Soit* $\mu_k = \mathbb{E}_0(X^k), k \geq 0$. *Soit aussi la matrice* $\mathbf{M}_k = [\mu_{i+i'}]_{i,i'=0,...,k-1}$, *le vecteur* $\boldsymbol{\mu}_k = (\mu_k, \mu_{k+1}, ...\mu_{2k-1})^T$ *et la constante* $c_k = \mu_{2k} - \boldsymbol{\mu}_k^T \mathbf{M}_k^{-1} \boldsymbol{\mu}_k$. *Alors les polynômes*

$$h_k(x) = c_k^{-1/2}(x^k - (1, x, x^2, ..., x^{k-1})\mathbf{M}_k^{-1}\boldsymbol{\mu}_k)$$

satisfont la condition du Théorème 2.1.

Les moments de la LNB ont des expressions explicites complexes. Il en va de même des coefficients des polynômes $h_k(\cdot)$ dont les expressions exactes sont très longues dès lors que $k > 2$. C'est pourquoi il est préférable de les exprimer sous une forme approximative. Mais ces calculs doivent être faits avec soin, car si les approximations numériques sont effectuées au niveau des éléments du Théorème 2.2, il en découle des erreurs d'arrondis qui détruisent l'orthonormalité. Ainsi, les coefficients des $h_k(\cdot)$ doivent être calculés en valeurs exactes, puis convertis en approximations numériques. En utilisant le logiciel MATHEMATICA, on obtient ainsi :

$$\begin{aligned}
h_1(x) &= -1.3979 + 0.4063x \\
h_2(x) &= 2.2836 - 1.6128x + 0.18247x^2 \\
h_3(x) &= 4.0815 + 4.5719x - 1.2053x^2 + 0.0862x^3 \\
h_4(x) &= 8.0795 - 12.0946x + 5.1951x^2 - 0.8249x^3 + 0.0431x^4 \\
h_5(x) &= -18.1064 + 33.1385x - 19.7207x^2 + 5.0168x^3 - 0.5665x^4 + 0.0233x^5
\end{aligned}$$

Soit $D_1, ..., D_n$ un échantillon aléatoire de *PCS*. Pour effectuer un test lisse de l'hypothèse nulle $H_0 : D \sim LNB$ au niveau asymptotique α, il suffit de calculer la statistique T_K avec les D_i en lieu et place des X_i et de confronter la valeur observée de cette statistique de test au quantile d'ordre $1 - \alpha$ de la loi χ^2_K. Le package de R BENFORDSMOOTHTEST, disponible sur le site GitHub, permet d'effectuer ce test lisse pour K allant jusqu'à 7.

Remarque 2.3. *Dans le Théorème 2.1, l'approximation* χ^2_K *est basée sur une convergence quand* $n \to \infty$. *Si* n *est petit,* $x^2_{K,1-\alpha}$ *peut donner une mauvaise approximation du quantile exact de la loi de* T_K. *S'il est nécessaire d'assurer un contrôle précis de l'erreur de Type I, on peut approcher la valeur exacte de ce quantile par la méthode de Monte-Carlo. C'est ce qui est fait dans le Package* BENFORDSMOOTHTEST, *dès lors que* n*< 100.*

Remarque 2.4. *Soit* $g(x)$ *une alternative fixée à la LNB. La puissance du test basé sur* T_K *peut être assez bien approximée par l'expression suivante qui se trouve dans Inglot et al. (1994, Théorème 2.1) et dont les conditions d'applicabilité sont rencontrées par virtuellement tous les* $g(x)$ *raisonnables dans le présent contexte. Soit* $\nu_k = \sqrt{n} \sum_{x=1}^{9} h_k(x)g(x)$ *regroupés dans* $\boldsymbol{\nu} = (\nu_1, ..., \nu_K)^T$ *et* $\Sigma = [\sigma_{i,j}]_{i,j=1,...,K}$, *où* $\sigma_{i,j} = \sum_{x=1}^{9} h_i(x)h_j(x)g(x) - \nu_i\nu_j/n$. *Posons la décomposition spectrale* $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$

où $\mathbf{\Lambda} = Diag\{\lambda_1, \ldots, \lambda_K\}$ est la matrice des valeurs propres de Σ et $\mathbf{P}$ et la matrice de ses vecteurs propres normalisés. Alors, uniformément en $t > 0$,

$$\mathbb{P}_g[T_K > t] = \mathbb{P}\left[\sum_{k=1}^{K} \lambda_k \chi_1^2(\delta_k^2) > t\right] + O(n^{-1/2}), \tag{2}$$

où δ_k sont les composantes de $\Lambda^{-1/2}\mathbf{P}\boldsymbol{\nu}$. Par ailleurs, pour approcher la loi de la somme pondérée des $\chi_1^2(\delta_k^2)$, on peut utiliser une approximation de Liu et al. (2009) commodément basée sur des moments. Adaptée au présent problème, cette approximation s'écrit :

$$\mathbb{P}\left[\sum_{k=1}^{K} \lambda_k \chi_1^2(\delta_k^2) > x_{K,1-\alpha}^2\right] \simeq \mathbb{P}\left[\chi_\ell^2(d^2) > \frac{a(x_{K,1-\alpha}^2 - c_1)}{\sqrt{c_2}} + (\ell + d)\right],$$

où $\ell = a^2 - 2d^2$, $d^2 = max[s_1 a^3 - a^2, 0]$, $a = 1/(s_1 - \sqrt{s_1^2 - s_2})$ si $s_1^2 > s_2$ et $1/s_1$ sinon, avec $s_1 = c_3/c_2^{3/2}$ et $s_2 = c_4/c_2^2$ avec $c_i = tr(\Sigma^i) + i\nu^T\Sigma^{(i-1)}\nu$, $i = 1, ...,4$. Cette approximation donne en général de bons résultats dans le contexte de tests d'adéquation (Duchesne et Lafaye De Micheaux, 2010). L'existence d'une expression explicite approchant la puissance du test lisse permet de mieux comprendre comment évolue cette puissance en fonction de $g(x)$. En effet, pour plusieurs des tests concernant la LNB (voir Section 3), la seule façon de calculer la puissance est par simulations, ce qui ne permet pas une bonne étude de la sensibilité de cette puissance en regard des écarts entre $f(x)$ et $g(x)$.

Remarque 2.5. *Le choix de l'hyperparamètre K est un élément important de la stratégie d'un test lisse. De la Remarque 2.4, si K est trop petit, certains couples (λ_k, δ_k^2) prenant de grandes valeurs n'apparaîtront pas dans la somme de l'équation (2) et le test perdra de la puissance car il ne tiendra pas compte de ces importants écarts à H_0 ; si K est trop grand, cette puissance est diluée par la considération de termes λ_k ou δ_k^2 nuls ou négligeables dans la statistique de test. Pour choisir judicieusement cet hyperparamètre, Ledwina (1994) a proposé une stratégie « data - driven » qui consiste à calculer, pour une valeur à choisir K_{max},*

$$\hat{K} = \underset{1 \le k \le K_{max}}{\arg\max} \left\{T_k - k\, log(n)\right\},$$

et à utiliser la statistique de test $T_{\hat{K}}$ qui $\xrightarrow{L}$ χ_1^2 sous H_0.

Ledwina (1994) explique que cette expression pour $\hat{K}$ découle de l'application de la procédure de sélection de modèles de Schwarz à une famille de densités emboîtant la loi postulée sous H_0. Une version locale du test du rapport de vraisemblances maximales mène ensuite à $T_{\hat{K}}$. La convergence vers la loi χ_1^2 vient du fait que (Ledwina, 1994), sous H_0, $\hat{K} \xrightarrow{Pr} 1$. Ainsi le test basé sur $T_{\hat{K}}$ possède le niveau asymptotique visé et des p-values asymptotiques peuvent être obtenues. Quand H_0 ne tient pas, $\hat{K}$ converge vers une valeur de $k \le K_{max}$ telle que $\delta_k^2 > 0$. Or les δ_k^2 non nuls sont de l'ordre $O(n)$ selon la Remarque 2.4 et comme $\log(n)/n \longrightarrow 0$, un résultat de Ducharme (1997, Sec. 3) assure que la probabilité à la droite de l'équation (2) converge vers 1. Il en découle que le test basé sur $T_{\hat{K}}$ est consistant dès lors que les δ_k^2 dans (2) ne sont pas nuls.

Le terme de pénalité $log(n)$ vient de l'application du critère bayésien de Schwarz. Quelques auteurs ont suggéré d'autres pénalités issues de critères alternatifs. Il a cependant été montré (Kallenberg, 2002) que le choix de n'importe quelle pénalité entre 2 et $log(n)$ préserve le niveau et la consistance du test. Ainsi, la théorie ne permet pas un choix judicieux de cette pénalité. Par contre, de nombreuses simulations par

Ledwina et ses coauteurs ont montré que la pénalité $log(n)$ est en général adéquate et que le test « data - driven » basé sur $T_{\hat{K}}$ est un bon compromis dans la famille de tests lisses $\{T_k, k = 1, ..., K_{max}\}$.

Enfin, on peut choisir K_{max} en exploitant de l'information contextuelle au problème (ce qu'on appelle le cas « horizon fini »), soit le laisser tendre vers ∞ (« horizon infini ») à une vitesse qui dépend de n. Ledwina et ses coauteurs ont beaucoup travaillé sur cette vitesse et les résultats théoriques obtenus sont impressionnants, mais d'une utilité pratique limitée, car ils s'expriment sous la forme $K_{max} = o(n^j)$ pour un j dépendant du contexte du problème. Heureusement la puissance du test basé sur $T_{\hat{K}}$ en fonction de K_{max} plafonne rapidement, de sorte que le cas « horizon fini » avec $5 \leq K_{max} \leq 7$ donne en général de bons résultats.

Remarque 2.6. *Dans la famille des tests lisses $\{T_k, k = 1, ...K_{max}\}$, on retrouve souvent des statistiques déjà proposées dans la littérature sur la base de considérations intuitives. C'est le cas ici, où $T_1/5.102674 = (a^*)^2$, la statistique introduite par Judge et Schechter (2009) et que l'on retrouve dans le package R* BENFORDTEST *(fonction* MEANDIGIT.BENFTEST*). Cette statistique est aussi reliée à la statistique du « Distortion Factor (DF) model » de Nigrini (1996).*

3. Les compétiteurs aux tests lisses

Comme signalé dans l'introduction, le nombre de tests d'adéquation à la LNB est plutôt limité au-delà du test du χ^2 de Pearson (1) et ceci facilite la tâche des fraudeurs. Soit $D_1, ..., D_n$ un échantillon aléatoire de *PCS*. Pour $d = 1, .., 9$, dénotons par $\hat{p}_d = n_d/n$, la proportion de d dans l'échantillon et π_d les probabilités de la LNB. Soit aussi $S_d = \sum_{j=1}^{d} \hat{p}_j$ et $S_d^* = \sum_{j=1}^{d} \pi_j$. Posons $Z_d = S_d - S_d^*$ et $t_d = (\pi_d + \pi_{d+1})/2$, pour $d = 1, .., 8$ et $t_9 = (\pi_9 + \pi_1)/2$. Lesperance et al. (2016) considèrent les versions discrètes des tests suivants basés sur les écarts entre la distribution cumulative empirique et la distribution cumulative théorique de la LNB :

$$W_n^2 = n \sum_{i=1}^{n} Z_d^2 t_d \qquad \text{(Cramer-von Mises)},$$

$$U_n^2 = n \sum_{i=1}^{n} (Z_d - \bar{Z})^2 t_d, \qquad \text{(Watson)},$$

dont une variante, dite de Freedman, se retrouve dans Joenssen (2014) et

$$A^2 = n \sum_{i=1}^{8} \frac{Z_d^2 t_d}{T_d(1 - T_d)} \qquad \text{(Anderson-Darling)}.$$

Morrow (2014) considère la statistique de Kolmogorov $K_n = \sqrt{n} \max_{1 \leq d \leq 9} |S_d - S_d^*|$. Enfin, un certain nombre d'auteurs considèrent des tests basés sur les écarts entre les probabilités π_d de la LNB et les $\hat{p}_d$. Leemis et al. (2000) proposent la statistique $m = \max_{1 \leq d \leq 9} |\hat{p}_d - \pi_d|$ alors que Cho et Gaines (2007) suggèrent $d = \sqrt{\sum_{d=1}^{9} (\hat{p}_d - \pi_d)^2}$ et Drake et Nigrini (2000) introduisent la « Mean Average Deviation » : $MAD = \frac{1}{9} \sum_{d=1}^{9} |\hat{p}_d - \pi_d|$. Enfin, dans un autre ordre d'idée, Judge et Schechter (2009) proposent la statistique $a^* = (\bar{D} - 3.44027)/(5.55973)$ qui est liée à la statistique T_1 du test lisse (voir Remarque 2.6). Au meilleur de nos connaissances, cette nomenclature couvre pratiquement tous les tests existants à

ce jour pour la LNB, sauf la statistique J_p^2 de Joenssen (2013b) inspirée du test de Shapiro-Wilks, qui ne sera pas considéré plus avant en raison de son mauvais comportement, un test adaptant la statistique de Hotelling qui semble peu intéressant car certains hyperparamètres doivent être choisis (sans critère précis à ce jour), et un test bayésien (Geyer et Williamson, 2004) qui ne permet pas le contrôle fréquentiste des erreurs de Type I et II, élément important pour les auditeurs.

Le package R BENFORDTESTS (version 1.2.0, 2015) maintenu par Joenssen (2013a) permet d'effectuer certains de ces tests via les fonctions CHISQ.BENFTEST (test du χ^2), KS.BENFTEST (test de Kolmogorov), MDIST.BENFTEST (test basé sur *m*) , EDIST.BENFTEST (test basé sur *d*), USQ.BENFTEST (variante de Freedman du test de Watson), JPSQ.BENFTEST (test basé sur J_p^2) et MEANDIGIT.BENFTEST (test basé sur a^*), auquels s'ajoute le test adaptant la statistique de Hotelling.

Signalons que Lesperance et al. (2016) et Joenssen (2014) recommandent le test de Watson ou sa variante de Freedman qui donnent de bons résultats dans leurs simulations dont nous reprenons certains éléments aux Sections 4 et 5, en ceci que leur puissance se range parmi les plus élevées pour les quelques alternatives qu'ils considèrent. Dans la suite de ce travail, nous comparons la puissance des présents tests lisses à celles de certains des tests plus haut sur une plage plus large d'alternatives. Plus précisément, comme représentant de la famille des tests lisses $\{T_k, k = 1, ...K_{max}\}$, nous retenons le test T_2 et sa version data-driven $T_{\hat{K}}$ (avec $K_{max} = 5$) ; comme représentant des tests basés sur les écarts entre distributions cumulatives, nous choisissons la statistique U_n^2 de Watson. Nous considérons le test MAD de Drake et Nigrini (2000) comme représentant des tests basés sur des écarts entre les π_d de la LNB et les $\hat{p}_d$, auxquel nous ajoutons le test classique du χ^2 de Pearson (1).

4. Les alternatives

Pour étudier la puissance des tests lisses de la Section 2 et la comparer à celle des tests retenus à la Section 3, nous devons préciser des alternatives pour lesquelles cette puissance sera calculée. Pour ce faire, nous considérons des familles d'alternatives indexées par un paramètre, génériquement noté β, et emboîtant la LNB. Notre première famille d'alternatives est celle de Rodriguez (2004) donnée par :

$$\mathbb{P}[D = d] = p_d^{(Rod)}(\beta) = \begin{cases} \frac{1}{9}(1 + \frac{10}{9}\ln(10) + x\ln(x) - (x+1)\ln(x+1)) & \text{si } \beta = 0 \\ \log_{10}[1 + 1/x] & \text{si } \beta = -1 \\ \frac{\beta+1}{9\beta} - ((x+1)^{(\beta+1)} - x^{(\beta+1)})/(\beta(10^{(\beta+1)} - 1)) & \text{sinon} \end{cases}, \quad (3)$$

pour $\beta \in \mathbb{R}$, où on peut remarquer que quand $\beta = 0$, on retrouve la loi de Stigler Lee et al. (2010), $\beta = -1$ donne la LNB et quand $\beta \to \pm\infty$, on a la loi uniforme discrète sur $\{1, 2, ..., 9\}$, notée $UD[\{1, .., 9\}]$. La deuxième famille d'alternatives est celle de la LNB généralisée de Pietronero et al. (2001) donnée par :

$$\mathbb{P}[D = d] = p_d^{(Piet)}(\beta) = \begin{cases} \frac{(d+1)^{(1-\beta)} - d^{(1-\beta)}}{10^{(1-\beta)} - 1} & \text{si } \beta \neq 1 \\ \log_{10}[1 + 1/x] & \text{si } \beta = 1 \end{cases},$$

où $\beta \in \mathbb{R}$ avec la LNB correspondant au cas $\beta = 1$. La troisième famille de lois alternatives est celle de Hürlimann (2009) où

$$\mathbb{P}[D = d] = p_d^{(Hurl)}(\beta) = \frac{1}{2}(\log_{10}[1 + x]^\beta - \log_{10}[x]^\beta - (1 - \log_{10}[1 + x])^\beta + (1 - \log_{10}[x])^\beta),$$

avec $\beta \in \mathbb{R}$. Notons que cette famille a la particularité de donner la LNB pour les deux valeurs $\beta = 1, 2$. La quatrième famille est celle d'un mélange de lois LNB et UD de la forme $(1-\beta) \times LNB + \beta \times UD[\{0, 1, ..., 9\}]$ avec $\beta \in [0, 1]$, considérée par Lesperance et al. (2016). La cinquième famille de lois, aussi considérée par Lesperance et al. (2016), est celle d'une loi LNB contaminée où

$$\mathbb{P}[D = d] = p_d^{(conta-1)}(\beta) = \begin{cases} \pi_d/(1+2\beta) & \text{si } d \neq 1, 9 \\ (\pi_d + \beta)/(1+2\beta) & \text{si } d = 1, 9 \end{cases},$$

où $\beta \geq 0$ et la LNB correspond au cas $\beta = 0$. Enfin la dernière famille de lois est nouvelle; elle est obtenue en contaminant plus lourdement que la famille précédente l'aile de droite de la LNB :

$$\mathbb{P}[D = d] = p_d^{(conta-2)}(\beta) = \begin{cases} \pi_d/(1+10\beta) & \text{si } d = 1, ..., 5 \\ (\pi_d + (d-5)\beta)/(1+10\beta) & \text{si } d = 6, ..., 9 \end{cases},$$

où $\beta \geq 0$ et la LNB correspond au cas $\beta = 0$. Ensemble, ces 6 familles couvrent de nombreuses alternatives à la LNB. Elles ont aussi été, pour les cinq premières, reconnues comme étant des alternatives plausibles, pouvant être rencontrées dans les applications. L'étude de la puissance des différents tests d'adéquation sur ces familles devrait donner une bonne idée de leur comportement en pratique. Terminons cette section en signalant que les cinq premières familles sont, à notre connaissance, les seules familles de lois existantes dans la littérature emboîtant la LNB.

5. Simulations

Dans cette section, nous étudions la puissance de différents tests d'adéquation à la LNB pour les alternatives de la Section 4. Notre but est de ranger les tests retenus à la Section 3 du meilleur au moins puissant, et ce pour chacune des six familles d'alternatives. Ceci permet de déterminer si les tests T_2 et $T_{\hat{K}}$ se rangent fréquemment parmi les meilleurs, justifiant ainsi leur introduction par le présent travail et leur utilisation en pratique. Pour chacune de ces familles d'alternatives, nous avons approché par simulation, selon la méthode exposée au paragraphe suivant, la courbe de puissance, en fonction de leur paramètre β, des tests retenus à la Section 3 pour des tailles d'échantillon variant entre n = 25 et n = 5000 et pour une large gamme de valeurs de β dans le domaine des valeurs autorisées. En regardant attentivement ces courbes de puissances, nous avons constaté que dans chaque famille, l'ordre des fonctions de puissance des différents tests ne varie pas selon les tailles d'échantillon considérées. Ceci permet une réduction importante de l'analyse des résultats de la simulation : pour obtenir le rang des différents tests selon leur courbe de puissance, il suffit de sélectionner une seule taille d'échantillon par famille, laquelle servira de représentant, et d'exhiber les courbes de puissances des tests à cette taille. Nous avons choisi cette taille d'échantillon représentative de façon à ce que l'examen visuel soit aisé. La plage de valeurs de β représentée a aussi été choisie pour une bonne lisibilité des graphiques et de façon à éviter les résultats triviaux (courbe de puissance trop proche de $\alpha = 0.05$ et 1).

Ainsi pour la famille de Rodriguez, la taille d'échantillon représentative est $n = 100$ avec $\beta \in [-4, 2]$; pour celle de Pietronero, $n = 25$ avec $\beta \in [-0.25, 2.5]$; pour la famille de Hurlimann, $n = 750$ avec $\beta \in [0.5, 2.75]$; pour la famille de mélange de LNB et UD , $n = 250$ avec $\beta \in [0.0, 0.45]$. Enfin, pour les deux LNB contaminées, nous prenons $n = 500$ avec $\beta \in [0, 0.06]$ et $\beta \in [0.0, 0.009]$ respectivement. Les tests sont effectués au niveau 5% et en regard de la Remarque 2.3, les quantiles de référence de tous les

tests sont approximés par Monte Carlo (50 000 réplications) afin de permettre une juste comparaison des puissances. Les puissances des tests pour chaque triplet (famille, n, β) sont approximées par le nombre de rejets parmi 10 000 réplications. Notons que pour le test T_2, les formules de la Remarque 2.4 sont aussi calculées et produisent des résultats remarquablement proches de ceux de la simulation.

Les graphiques des courbes de puissance des cinq tests retenus aux tailles d'échantillon représentatives apparaissent à la Figure 1, un panneau par famille d'alternatives. Comme les puissances de cette figure sont approximées de 10 000 réplications, elles contiennent un bruit statistique que l'on peut évaluer à environ ± 0.01 (au niveau de confiance 95%) lorsque la puissance est autour de 0.5. Ce point est pris en considération dans les conclusions qui suivent.

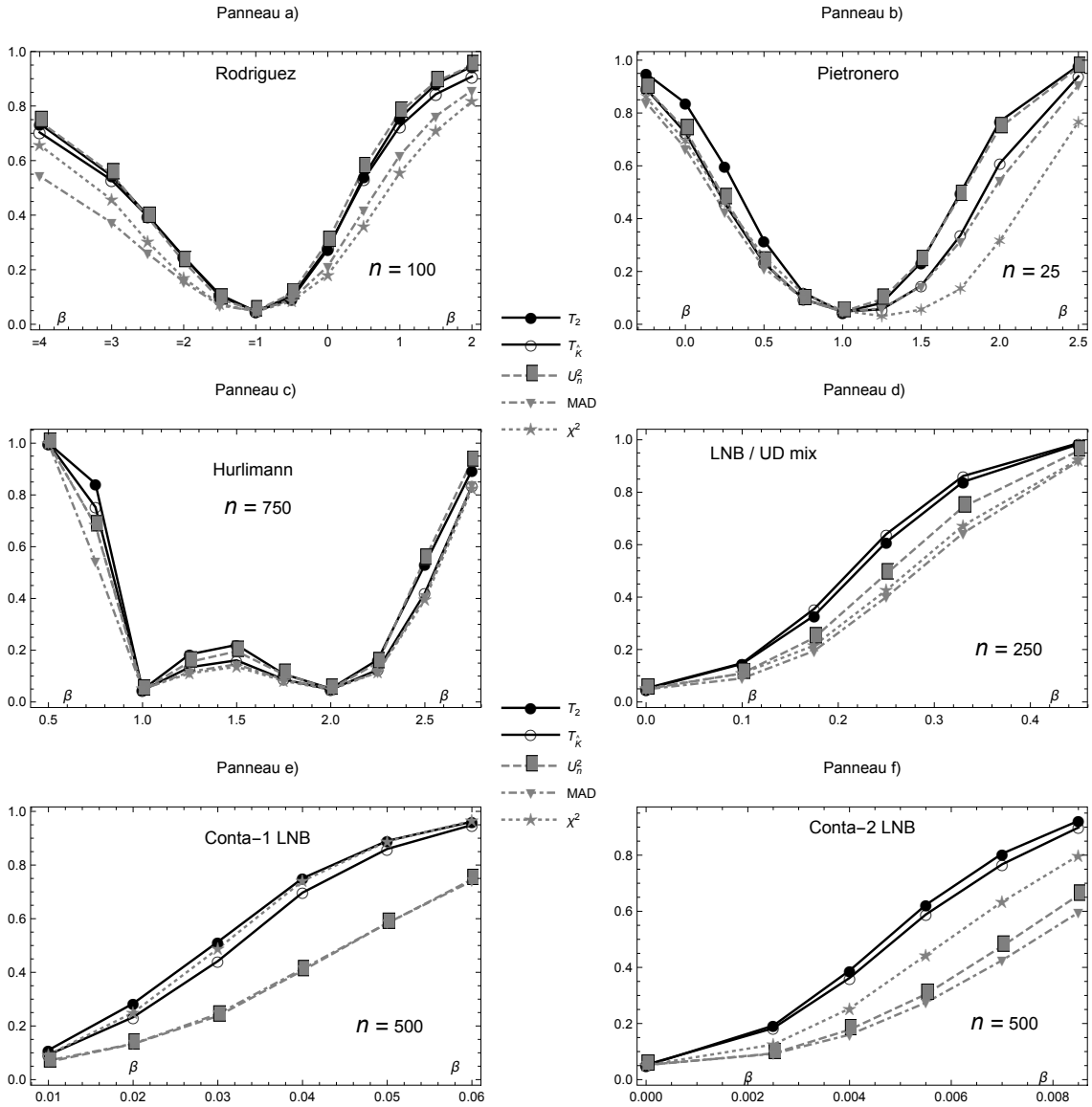

Figure 1: *Courbes de puissance, en fonction du paramètre β, de divers tests (basés sur 10 000 réplications) au niveau 5% pour l'hypothèse nulle de loi LNB. Les familles d'alternatives sont décrites à la Section 4. Les tests considérés sont T_2, (trait plein avec cercle plein), $T_{\hat{K}}$ (trait plein avec cercle vide), U_n^2 (tiret avec rectangle), MAD (tiret-pointillé avec triangle) et le test χ^2 de Pearson (pointillé avec étoile), dont les expressions se trouvent à la Section 3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 50 000 répétitions.*

Pour les lois de Rodriguez du Panneau a), le test de Watson basé sur U_n^2 est comparable à T_2 et suivi de très près par $T_{\hat{K}}$. Les tests basés sur MAD et χ^2 sont moins puissants (avec MAD meilleur que χ^2 si $\beta > -1$ et inversement sinon). Pour les lois de Pietronero du Panneau b), T_2 est le test le plus puissant sur toute la plage de valeurs de β, suivi de $T_{\hat{K}}$ et de U_n^2 quand $\beta < 1$ et inversement quand $\beta > 1$.

Encore là, les tests MAD et χ^2 ferment la marche, comparables quand $\beta < 1$ mais avec χ^2 nettement moins puissant que MAD à droite. Pour la famille de Hurlimann du Panneau c), quand $\beta < 1, T_2$ domine, suivi de $T_{\hat{K}}$ puis de U_n^2 alors que pour $\beta > 1$, T_2 et U_n^2 sont comparables mais $T_{\hat{K}}$ est légèrement moins puissant. Encore là, MAD et χ^2 ferment la marche. Pour la famille de mélange LNB et UD du Panneau d), $T_{\hat{K}}$ domine T_2, MAD et χ^2 sont les moins puissants et U_n^2 a un comportement intermédiaire. Pour la première famille de lois contaminées du Panneau e), T_2 et χ^2 dominent, suivi de près par $T_{\hat{K}}$ alors que U_n^2 et MAD ont une puissance nettement plus faible. Enfin pour la famille de lois contaminées du Panneau f), les deux tests lisses T_2 et $T_{\hat{K}}$ dominent nettement les autres, MAD et U_n^2 étant de loin inférieurs alors que χ^2 a un comportement intermédiaire.

En conclusion, on retire de cette expérience de simulation que le test T_2 est toujours parmi les deux meilleurs tests en termes de puissance, disputant souvent la tête avec le test basé sur $T_{\hat{K}}$. Le test MAD est toujours parmi les deux plus mauvais et ne devrait pas être utilisé. Les tests du χ^2 et U_n^2 ont des comportements plus variables. En regard de notre définition de « bon » test d'adéquation évoquée dans l'introduction, nous pouvons donc recommander l'un ou l'autre des tests T_2 ou $T_{\hat{K}}$. Leur introduction est ainsi justifiée et le fait qu'il s'agisse de tests nouveaux, augmentant l'arsenal des auditeurs, ajoute à leur utilité. Nous confirmons Morrow (2014) que le test du χ^2 de Pearson est généralement inférieur . Enfin, nos résultats atténuent les conclusions de Lesperance et al. (2016) et Joenssen (2014) qui recommandent le test U_n^2 de Watson.

Bibliographie

R. Abdullahi and N. Mansor. Fraud triangle theory and fraud diamond theory. understanding the convergent and divergent for future research. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 5(4) : 38–45, 10 2015. doi: 10.6007/ijarafms/v5-i4/1823.

L. Arshadi and A. H. Jahangir. Benford's law behavior of internet traffic. *Journal of Network and Computer Applications*, 40 :194–205, 4 2014. ISSN 1084-8045. doi: 10.1016/j.jnca.2013.09.007.

M. Ausloos, R. Cerqueti, and T. A. Mir. Data science for assessing possible tax income manipulation : The case of italy. *Chaos, Solitons & Fractals*, 104 :238–256, 11 2017. ISSN 0960-0779. doi: 10.1016/j.chaos.2017.08.012.

F. Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4) :551–572, 1938.

A. Berger, T. Hill, and E. Rogers. Benford online bibliography, 05 2015. URL `http://www.benfordonline.net`.

B. Boulerice and G. R. Ducharme. Smooth tests of goodness-of-fit for directional and axial data. *Journal of Multivariate Analysis*, 60(1) :154–175, 01 1997. doi: 10.1006/jmva.1996.1650.

A. Cerioli, L. Barabesi, A. Cerasa, M. Menegatti, and D. Perrotta. Newcomb–benford law and the detection of frauds in international trade. *Proceedings of the National Academy of Sciences*, 116(1) :106–115, dec 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806617115.

W. T. Cho and B. Gaines. Breaking the (benford) law. *The American Statistician*, 61(3) :218–223, 8 2007. doi: 10.1198/000313007x223496.

P. D. Drake and M. J. Nigrini. Computer assisted analytical procedures using benford's law. *Journal of Accounting Education*, 18(2) :127–146, 3 2000. ISSN 0748-5751. doi: 10.1016/s0748-5751(00)00008-7.

G. R. Ducharme. Consistent selection of the actual model in regression analysis. *Journal of Applied Statistics*, 24(5) : 549–558, oct 1997. doi: 10.1080/02664769723530.

P. Duchesne and P. Lafaye De Micheaux. Computing the distribution of quadratic forms : Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4) :858–862, 4 2010. doi: 10.1016/j.csda.2009.11.025.

C. Durtschi, H. William, and P. Carl. The effective use of benford's law to assist in detecting fraud in accounting data. *Journal Forensic Account*, 5 :17–34, 01 2004. ISSN 1524-5586.

N. Gauvrit, J.-C. Houillon, and J.-P. Delahaye. Generalized benford's law as a lie detector. *Advances in Cognitive Psychology*, 13(2) :121–127, jun 2017. doi: 10.5709/acp-0212-x.

A. Geyer and J. Martí. Applying benford's law to volcanology. *Geology*, 40(4) :327–330, 04 2012. ISSN 0091-7613. doi: 10.1130/G32787.1.

C. L. Geyer and P. P. Williamson. Detecting fraud in data sets using benford's law. *Communications in Statistics - Simulation and Computation*, 33(1) :229–246, jan 2004. doi: 10.1081/sac-120028442.

J. Hein, R. Zobrist, C. Konrad, and G. Schuepfer. Scientific fraud in 20 falsified anesthesia papers. *Der Anaesthesist*, 61 (6) :543–549, jun 2012. doi: 10.1007/s00101-012-2029-x.

T. P. Hill. The first digit phenomenon : A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data. *American Scientist*, 86(4) :358–363, 1998. ISSN 00030996.

W. Hürlimann. Generalizing benford's law using power laws : Application to integer sequences. *International Journal of Mathematics and Mathematical Sciences*, 2009 :1–10, 07 2009. doi: 10.1155/2009/970284.

T. Inglot, W. Kallenberg, and T. Ledwina. Power approximations to and power comparison of smooth goodness-of-fit tests. *Scandinavian journal of statistics*, (21) :131–145, 1994. ISSN 0303-6898.

D. W. Joenssen. Statistical tests for evaluating conformity to benford's law. *The Comprehensive R Archive Network*, 2013a. URL `https://cran.r-project.org/package=BenfordTests`.

D. W. Joenssen. A new test for benford's distribution. *Abstract-Proceedings of the 3rdJoint Statistical Meeting DAGStat*, 03 2013b.

D. W. Joenssen. Testing for benford's law : A monte carlo comparison of methods. *SSRN Electronic Journal*, 11 2014. doi: 10.2139/ssrn.2545243.

G. Judge and L. Schechter. Detecting problems in survey data using benford's law. *Journal of Human Resources*, 44(1) : 1–24, 2009. doi: 10.3368/jhr.44.1.1.

W. Kallenberg. The penalty in data driven neyman's tests. *Mathematical methods of statistics*, 11 :323–340, 2002. ISSN 1066-5307.

W. C. Kallenberg and L. Teresa. Data driven smooth tests for composite hypotheses comparison of powers. *Journal of Statistical Computation and Simulation*, 59(2) :101–121, oct 1997. doi: 10.1080/00949659708811850.

M.-P. Lecoutre. Cognitive models and problem spaces in ?purely random ? situations. *Educational Studies in Mathematics*, 23(6) :557–568, dec 1992. doi: 10.1007/bf00540060.

T. Ledwina. Data-driven version of neyman's smooth test of fit. *Journal of the American Statistical Association*, 89(427) : 1000–1005, sep 1994. doi: 10.1080/01621459.1994.10476834.

J. Lee, W. K. T. Cho, and G. G. Judge. Stigler's approach to recovering the distribution of first significant digits in natural data sets. *Statistics & Probability Letters*, 80(2) :82–88, jan 2010. doi: 10.1016/j.spl.2009.09.015.

L. M. Leemis, B. W. Schmeiser, and D. L. Evans. Survival distributions satisfying benford's law. *The American Statistician*, 54(4) :236–241, nov 2000. doi: 10.1080/00031305.2000.10474554.

M. Lesperance, W. J. Reed, M. A. Stephens, C. Tsao, and B. Wilton. Assessing conformance with benford's law : Goodness-of-fit tests and simultaneous confidence intervals. *PLOS ONE*, 11(3) :e0151235, mar 2016. doi: 10.1371/journal.pone.0151235.

H. Liu, Y. Tang, and H. H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4) :853–856, feb 2009. ISSN 0167-9473. doi: 10.1016/j.csda.2008.11.025.

J. Morrow. Benford's Law, Families of Distributions and a Test Basis. *Centre for Economic Performance, LSE*, (dp1291), Aug. 2014.

S. Newcomb. Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*, 4 :39–40, Jan. 1881. ISSN 00029327.

J. Neyman. »smooth test» for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4) :149–199, jul 1937. doi: 10.1080/03461238.1937.10404821.

M. J. Nigrini. A taxpayer compliance application of benford's law. *The Journal of the American Taxation Association*, 18 (1) :72–91, 04 1996. ISSN 0198-9073.

M. J. Nigrini and S. J. Miller. Benford's law applied to hydrology data—results and relevance to other geophysical data. *Mathematical Geology*, 39(5) :469–490, aug 2007. doi: 10.1007/s11004-007-9109-5.

L. Pietronero, E. Tosatti, V. Tosatti, and A. Vespignani. Explaining the uneven distribution of numbers in nature : the laws of benford and zipf. *Physica A : Statistical Mechanics and its Applications*, 293(1-2) :297–304, apr 2001. ISSN 0378-4371. doi: 10.1016/s0378-4371(00)00633-6.

P. N. Posch. A survey on sequences and distribution functions satisfying the first-digit-law. *Journal of Statistics and Management Systems*, 11(1) :1–19, jan 2008. doi: 10.1080/09720510.2008.10701294.

J. C. W. Rayner and D. J. Best. Smooth tests of goodness of fit : An overview. *International Statistical Review / Revue Internationale de Statistique*, 58(1) :9, apr 1990. ISSN 03067734, 17515823. doi: 10.2307/1403470.

R. J. Rodriguez. First significant digit patterns from mixtures of uniform distributions. *The American Statistician*, 58(1) : 64–71, feb 2004. doi: 10.1198/0003130042782.

M. Sambridge, H. Tkalcic, and P. Arroucau. Benford's law of first digits : From mathematical curiosity to change detector. *Asia Pacific Mathematics Newsletter(APMN)*, 1 :1–5, 01 2011.

O. Thas. *Comparing Distributions*. Springer New York, 1 edition, 2010. ISBN 978-0-387-92710-7. doi: 10.1007/978-0-387-92710-7.

Chapitre 3

Le jeu du chat et de la souris

3.1 Loi de Newcomb-Benford multivariée

Rappelons, comme vu dans la section 1.4, qu'en 1881 Newcomb, après avoir observé une détérioration plus marquée des tables de logarithmes pour les pages des nombres commençant par 1 que pour les pages commençant par 9, formula l'hypothèse : « pour toute liste de nombres prise dans un ensemble de données arbitraires, davantage de nombres tendent à avoir leur premier chiffre égal à un » (Newcomb, 1881). Il proposa alors les probabilités du premier chiffre significatif représentées dans la Figure 3.1.1, dont nous avons fait usage dans le chapitre précédent, à l'instar de Lesperance *et al.* 2016; Nigrini 1993; Nigrini et Mittermaier 1997; Drake et Nigrini 2000, dans le but de détecter des fraudes.

En revanche, il est moins connu que Newcomb (1881) avait aussi calculé les probabilités d'apparition du second chiffre significatif (SCS) $d_2 \in \{0, \ldots, 9\}$, et reporté ces probabilités dans son article (voir Figure 3.1.1).

Dig.	First Digit.	Second Digit.
0	. . .	0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

Figure 3.1.1 – Loi de probabilité du premier et second chiffre significatif. Newcomb (1881)

Disposant de la loi pour le premier et le second chiffre significatif, Hill (1995) (voir

aussi Cerioli *et al.*, 2018) a proposé une généralisation de la loi de Newcomb-Benford.

Soit X une variable aléatoire continue positive, p un nombre entier positif et $\mathscr{D}_1(X)$, $\mathscr{D}_2(X)$, ..., $\mathscr{D}_p(X)$ respectivement le premier, second, ..., $p-$ième chiffre significatif de X. Cette généralisation postule que la distribution conjointe de ces p chiffres significatifs prend la forme :

$$\mathbb{P}\left[\mathscr{D}_1(X) = d_1, \mathscr{D}_2(X) = d_2, \ldots, \mathscr{D}_p(X) = d_p\right] = \log_{10}\left(1 + \frac{1}{\sum_{l=1}^{p} 10^{p-l} d_l}\right) \tag{3.1.1}$$

avec $d_1 \in \{1, \ldots, 9\}$ et $d_l \in \{0, \ldots, 9\}, \forall\, l = 2, \ldots, p$.

Dans le cas particulier où $p = 2$, nous obtenons la distribution bivariée de la loi de Newcomb-Benford.

$$\mathbb{P}\left[\mathscr{D}_1(X) = d_1, \mathscr{D}_2(X) = d_2\right] = \log_{10}\left(1 + \frac{1}{(10d_1 + d_2)}\right) \tag{3.1.2}$$

avec $d_1 \in \{1, \ldots, 9\}$ et $d_2 \in \{0, \ldots, 9\}$.

Il en résulte que lorsqu'une série de nombres $\{X_i, i \geq 1\}$ satisfait à la loi de Newcomb-Benford bivariée, la probabilité de trouver un second chiffre significatif dépend du premier chiffre d_1. En particulier, le coefficient de corrélation de Spearman $\rho_S\left(\mathscr{D}_1(X), \mathscr{D}_2(X)\right) = 0.062$.

Les probabilités en pourcentage (%) de cette loi bivariée se trouvent dans la table 3.1.

Premier chiffre significatif \ *Second chiffre significatif* $d_1 \mid d_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	4.14	3.78	3.48	3.22	3.00	2.80	2.63	2.48	2.35	2.23
2	2.12	2.02	1.93	1.85	1.77	1.70	1.64	1.58	1.52	1.47
3	1.42	1.38	1.34	1.30	1.26	1.22	1.19	1.16	1.13	1.10
4	1.07	1.05	1.02	1.00	0.98	0.95	0.93	0.91	0.90	0.88
5	0.86	0.84	0.83	0.81	0.80	0.78	0.77	0.76	0.74	0.73
6	0.72	0.71	0.69	0.68	0.67	0.66	0.65	0.64	0.63	0.62
7	0.62	0.61	0.60	0.59	0.58	0.58	.057	0.56	.055	0.55
8	0.54	0.53	0.53	0.52	0.51	0.51	0.50	0.50	0.49	0.49
9	0.48	0.47	0.47	0.46	0.46	0.45	0.45	0.45	0.44	0.44

TABLE 3.1 – Distribution conjointe $\mathbb{P}(\mathscr{D}_1(X) = d_1, \mathscr{D}_2(X) = d_2)$ du premier et deuxième chiffre significatif, en pourcentage (%).

De la formule 3.1.1, nous obtenons aisément la formule permettant de calculer la loi marginale de chacun des chiffres significatifs,

$$\mathbb{P}\left[\mathscr{D}_p(X) = d_p\right] = \sum_{d_1=1}^{9} \sum_{d_2=0}^{9} \cdots \sum_{d_{p-1}=0}^{9} \mathbb{P}\left[\mathscr{D}_1(X) = d_1, \ldots, \mathscr{D}_p(X) = d_p\right] \tag{3.1.3}$$

En particulier les lois marginales du premier et second chiffre significatif sont :

$$\mathbb{P}\left[\mathscr{D}_1\left(X\right)=d_1\right]=\log_{10}\left(1+\frac{1}{d_1}\right), \tag{3.1.4}$$

$$\begin{aligned}\mathbb{P}\left[\mathscr{D}_2\left(X\right)=d_2\right]&=\sum_{d_1=1}^{9}\mathbb{P}\left[\mathscr{D}_1\left(X\right)=d_1,\mathscr{D}_2\left(X\right)=d_2\right] \\ &=\sum_{d_1=1}^{9}\log_{10}\left(1+\frac{1}{10d_1+d_2}\right)\end{aligned} \tag{3.1.5}$$

La Table 3.2 montre les probabilités (en %) pour les quatre (4) premiers chiffres significatifs. En particulier, nous retrouvons la table de Newcomb (1881) de la Figure 3.1.1 dans les deux premières colonnes de la Table 3.2.

Notons que si une série de nombres $\{X_i, i \geq 1\}$ satisfait la loi de Newcomb-Benford multivariée, alors plus p est grand, plus $\mathbb{P}\left[\mathscr{D}_p\left(X\right)=d_p\right]$ semble se rapprocher de la loi uniforme discrète sur les entiers $\{0,1,\ldots,9\}$ (cf $\mathbb{P}\left[\mathscr{D}_3\left(X\right)=d_3\right]$ et $\mathbb{P}\left[\mathscr{D}_4\left(X\right)=d_4\right]$, colonne 3 et 4 de la Table 3.2).

d	$\mathbb{P}\left[\mathscr{D}_1\left(X\right)=d_1\right]$	$\mathbb{P}\left[\mathscr{D}_2\left(X\right)=d_2\right]$	$\mathbb{P}\left[\mathscr{D}_3\left(X\right)=d_3\right]$	$\mathbb{P}\left[\mathscr{D}_4\left(X\right)=d_4\right]$
0	$\cdots$	11.97	10.18	10.02
1	30.10	11.39	10.14	10.01
2	17.61	10.88	10.10	10.01
3	12.49	10.43	10.06	10.01
4	9.69	10.03	10.02	10.00
5	7.92	9.67	9.98	10.00
6	6.69	9.34	9.94	9.99
7	5.80	9.04	9.90	9.99
8	5.12	8.76	9.86	9.99
9	4.58	8.50	9.83	9.98

TABLE 3.2 – Probabilités marginales en pourcentage (%) du premier, second, troisième et quatrième chiffre significatif de la loi Newcomb-Benford multivariée.

En bon statisticien que nous sommes, énoncer la loi conjointe, les lois marginales et ne pas parler des lois conditionnelles pourrait être considéré comme une faute professionnelle. Aussi, soit X une variable aléatoire continue positive et $\mathscr{D}_1$, $\mathscr{D}_2$, respectivement, le premier et le second chiffre significatif.

La distribution conditionnelle de $\mathscr{D}_1(X)$ conditionnellement à la valeur de $\mathscr{D}_2(X)$, soit $\left(\mathscr{D}_1\left(X\right)|\mathscr{D}_2\left(X\right)\right)$, prend la forme

$$\mathbb{P}\left(\mathscr{D}_1\left(X\right)=d_1|\mathscr{D}_2\left(X\right)=d_2\right)=\frac{\mathbb{P}\left(\mathscr{D}_1\left(X\right)=d_1,\mathscr{D}_2\left(X\right)=d_2\right)}{\mathbb{P}\left(\mathscr{D}_2\left(X\right)=d_2\right)} \tag{3.1.6}$$
$$=\frac{\log_{10}\left(1+\frac{1}{(10d_1+d_2)}\right)}{\sum_{d_1=1}^{9}\log_{10}\left(1+\frac{1}{10d_1+d_2}\right)}$$

avec $d_1 \in \{1,\ldots,9\}$ et $d_2 \in \{0,\ldots,9\}$.

Les probabilités conditionnelles (en %) de $\left(\mathscr{D}_1\left(X\right)|\mathscr{D}_2\left(X\right)\right)$ sont données dans la Table 3.3.

PREMIER CHIFFRE SIGNIFICATIF \ SECOND CHIFFRE SIGNIFICATIF										
$d_1 \setminus d_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	34.59	33.18	31.94	30.85	29.87	28.99	28.20	27.47	26.81	26.21
2	17.71	17.14	17.74	17.72	17.67	17.62	17.55	17.48	17.40	17.32
3	11.90	12.11	12.28	12.43	12.55	12.65	12.74	12.82	12.88	12.94
4	8.96	9.19	9.39	9.57	9.73	9.87	10.00	10.12	10.23	10.32
5	7.19	7.40	7.60	7.78	7.94	8.09	8.23	8.36	8.48	8.59
6	6.00	6.20	6.39	6.56	6.71	6.86	6.99	7.12	7.24	7.35
7	5.15	5.33	5.50	5.66	5.81	5.95	6.08	6.2	6.32	6.43
8	4.51	4.68	4.84	4.99	5.12	5.25	5.38	5.49	5.60	5.71
9	4.01	4.17	4.31	4.45	4.58	4.7	4.82	4.93	5.03	5.14
$\sum$	100	100	100	100	100	100	100	100	100	100

TABLE 3.3 – Distribution conditionnelle $\mathbb{P}(\mathscr{D}_1(X) = d_1|\mathscr{D}_2 = d_2)$, en pourcentage (%) sous la loi bivariée de Newcomb-Benford.

Réciproquement la distribution conditionnelle de $\mathscr{D}_2(X)$ conditionnellement à la valeur de $\mathscr{D}_1(X)$, soit $\left(\mathscr{D}_2\left(X\right)|\mathscr{D}_1\left(X\right)\right)$ sous la loi Newcomb-Benford est :

$$\mathbb{P}\left(\mathscr{D}_2\left(X\right)=d_2|\mathscr{D}_1\left(X\right)=d_1\right)=\frac{\mathbb{P}\left(\mathscr{D}_1\left(X\right)=d_1,\mathscr{D}_2\left(X\right)=d_2\right)}{\mathbb{P}\left(\mathscr{D}_1\left(X\right)=d_1\right)} \tag{3.1.7}$$
$$=\frac{\log_{10}\left(1+\frac{1}{(10d_1+d_2)}\right)}{\log_{10}\left(1+\frac{1}{d_1}\right)}$$

avec $d_1 \in \{1,\ldots,9\}$ et $d_2 \in \{0,\ldots,9\}$.

Nous donnons les probabilités conditionnelles (en %) de $\left(\mathscr{D}_2\left(X\right)|\mathscr{D}_1\left(X\right)\right)$ dans la Table 3.4

PREMIER CHIFFRE SIGNIFICATIF \ SECOND CHIFFRE SIGNIFICATIF											
$d_1 \setminus d_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	$\sum$
1	13.75	12.55	11.55	10.69	9.95	9.31	8.75	8.25	7.80	7.40	100
2	12.03	11.47	10.96	10.50	10.07	9.67	9.31	8.97	8.65	8.36	100
3	11.40	11.04	10.70	10.38	10.08	9.79	9.52	9.27	9.03	8.80	100
4	11.07	10.80	10.55	10.30	10.07	9.85	9.64	9.43	9.24	9.05	100
5	10.86	10.65	10.45	10.25	10.06	9.88	9.71	9.54	9.38	9.22	100
6	10.72	10.55	10.38	10.22	10.06	9.90	9.76	9.61	9.47	9.33	100
7	10.62	10.47	10.33	10.19	10.05	9.92	9.79	9.66	9.54	9.42	100
8	10.55	10.42	10.29	10.17	10.05	9.93	9.82	9.70	9.59	9.49	100
9	10.49	10.37	10.26	10.15	10.04	9.94	9.84	9.73	9.64	9.54	100

TABLE 3.4 – Distribution conditionnelle $\mathbb{P}(\mathscr{D}_2(X) = d_2 | \mathscr{D}_1 = d_1)$, en pourcentage (%) sous la loi bivariée de Newcomb-Benford.

3.2 Pourquoi utiliser le second chiffre significatif pour la détection de fraudes ?

Les fraudeurs sont sournois et malins. Étant conscient que les « bons » ont à leur disposition un outil de détection de fraude, notamment le « smooth test » (test lisse) appliqué au premier chiffre significatif vu dans le chapitre précédent, ils pourraient tenter une falsification des données en adaptant les premiers chiffres significatifs fabriqués pour passer entre les mailles du filet.

Les études de Burgstahler et Dichev (1997) et Degeorge *et al.* (1999) ont montré que la politique de résultat des entreprises peut se résumer en trois objectifs : - atteindre des résultats positifs (aversion pour les pertes), - mettre en évidence les performances (résultat en progression) et - répondre aux attentes des analystes financiers. En effet, un seul faux pas peut être préjudiciable, par exemple la chute de sa valeur boursière. Le premier objectif, l'aversion pour les pertes, reconnu comme un biais cognitif et émotionnel en finance comportementale par les psychologues, génère le besoin d'embellir les résultats : par exemple, un bénéfice de 6 997 800 euros sera généralement arrondi à 7 millions, donc une modification de plusieurs chiffres significatifs. En effet l'impact psychologique de ce dernier montant sera plus important que le montant originel alors que la différence entre les deux, est infime. Carslaw (1988) fut le premier à constater, dans son étude sur des firmes néo-zélandaises, que le second chiffre des résultats est marqué par un excès de 0 et un déficit de 9. En effet, les managers ont tendance à arrondir les résultats.

Par ailleurs Diekmann (2007), dans une étude de psychologie expérimentale, a demandé à des étudiants en sciences sociales de créer des nombres comportant quatre chiffres. Diekmann a constaté que, dans ce cas, le premier chiffre significatif de ces nombres est à peu près conforme à la loi Newcomb-Benford de (3.1.4) mais pas les autres, en ce sens que leur loi marginale s'écarte significativement de celles de la Table 3.2. Il a conclu

que pour détecter des fraudes impliquant la falsification de plusieurs chiffres, on devrait non seulement considérer le premier chiffre, mais également regarder au-delà du premier chiffre. Dans une autre étude, Burns (2009) a demandé à des participants de deviner des quantités réelles, telles que la dette nationale brute des États-Unis ou le pic de consommation d'électricité en été à Melbourne. Il a constaté que, bien que les réponses des participants au premier chiffre significatif ne suivent pas parfaitement la loi (3.1.4), elles se conforment mieux à cette distribution qu'à la distribution uniforme mais encore là, pas les autres chiffres significatifs. On peut conclure de ces travaux que si l'on tente de frauder en modifiant plusieurs chiffres significatifs, le premier pourrait être proche de la loi de Newcomb-Benford univariée mais les autres vont s'éloigner de la loi marginale de la Table 3.2 attendue.

L'un des écueils principaux à la validation ou à la réfutation statistique d'hypothèses est le « manque de puissance », c'est à dire l'incapacité d'affirmer si les différences observées sont le fruit de l'action d'un facteur donné ou de simples variations aléatoires. Lorsque les données sont sensées se conformer à la loi de Benford, un rejet de l'hypothèse nulle H_0 suggère qu'une certaine forme de manipulation des données pourrait avoir eu lieu. Joenssen (2013c) montre un gain de puissance dans le rejet de ce H_0 dû à une utilisation accrue de l'information en utilisant les deux premiers chiffres significatifs, car, si H_0 est vraie, les distributions des premier et deuxième chiffres significatifs ne sont pas indépendantes. Il fournit une analyse succincte en comparant la puissance fournie par un test d'adéquation à un chiffre et un test d'adéquation à deux chiffres. Il en déduit que « n'utiliser qu'un seul chiffre est moins informatif que deux ». Nous retrouvons donc l'ADN de la statistique : «plus il y a de données, plus nous pouvons aller loin dans l'analyse ».

Dans la partie suivante, nous nous baserons sur les travaux de Joenssen (2013c) et Wong (2010) pour faire l'état de l'art des tests d'adéquation existant pour la loi conjointe $(\mathscr{B}_1, \mathscr{B}_2)$ du premier et second chiffre significatif et développer un « smooth test » appliqué à cette loi bivariée.

3.3 Tests d'adéquation à la loi bivariée Newcomb-Benford

3.3.1 État de l'art

Soit X une variable aléatoire continue positive et $\mathscr{D}_1, \mathscr{D}_2$, respectivement le premier, défini sur $\{1, \ldots, 9\}$, et le second, défini sur $\{0, \ldots, 9\}$, chiffre significatif. Considérons le problème de tester $H_0 : (\mathscr{D}_1, \mathscr{D}_2)$ obéit à la loi de Newcomb-Benford bivariée de l'équation (3.1.2) que nous notons $\mathscr{B}_{(1,2)}$. Le premier test d'adéquation adapté à la résolution de ce H_0 fut le bon vieux test du χ^2 de Pearson (1900). Transformons le couple $(\mathscr{D}_1, \mathscr{D}_2)$ en une valeur univariée $\mathscr{D}_{12} = (10 * \mathscr{D}_1 + \mathscr{D}_2)$, définie sur $\{10, \ldots, 99\}$. Notons $\pi_{d_{12}}$ la

distribution de $\mathscr{D}_{12}$ obtenue déduite de (3.1.2) :

$$\mathbb{P}\left[\mathscr{D}_{12}\left(X\right)=d_{12}\right]=\pi_{d_{12}}=\mathbb{P}\left[\mathscr{D}_{1}\left(X\right)=d_{1},\mathscr{D}_{2}\left(X\right)=d_{2}\right] \tag{3.3.1}$$

avec $d_{12}\in\{10,\ldots,99\}$, $d_1\in\{1,\ldots,9\}$ et $d_2\in\{0,\ldots,9\}$. Soit $n_{d_{12}}$ le nombre de fois que, dans un jeu de n données, on observe $\mathscr{D}_{12}=d_{12}$. La statistique du test du χ^2 de Pearson est définie par

$$\chi^2=\sum_{d_{12}=10}^{99}\frac{\left(n_{d_{12}}-n\pi_{d_{12}}\right)^2}{n\pi_{d_{12}}} \tag{3.3.2}$$

et si $H_0:(\mathscr{D}_1,\mathscr{D}_2)\sim\mathscr{B}_{(1,2)}$ est vraie $\chi^2\longrightarrow\chi^2_{89}$, la loi du χ^2 à 89 degrés de liberté.

Par la suite Wong (2010) développe d'autres stratégies de tests pour l'hypothèse nulle d'une loi de Newcomb-Benford bivariée.

Soit $\widehat{\pi_{d_{12}}}=\dfrac{n_{d_{12}}}{n}$, la proportion de d_{12} dans un jeu de données de taille n. Notons $S_{d_{12}}=\sum_{i=10}^{d_{12}}\widehat{\pi_i}$ et $T_{d_{12}}=\sum_{i=10}^{d_{12}}\pi_i$. Posons $Z_{d_{12}}=S_{d_{12}}-T_{d_{12}}$ et $t_{d_{12}}=\dfrac{\pi_{d_{12}}+\pi_{d_{12}+1}}{2}$, pour $d_{12}\in\{10,\ldots,98\}$ et $t_{99}=\dfrac{\pi_{99}+\pi_{10}}{2}$. Les tests suivants, basés sur les écarts entre la distribution cumulative empirique (les $S_{d_{12}}$) et la distribution cumulative théorique (les $T_{d_{12}}$) de la loi de $\mathscr{D}_{12}$ sous l'hypothèse de Newcomb-Benford $\mathscr{B}_{(1,2)}$ sont des extensions au cas discret de tests bien connues pour des données continues.

$$W^2=\frac{1}{n}\sum_{d_{12}=10}^{99}Z_{d_{12}}^2t_{d_{12}} \qquad \text{(Cramer-von Mises)} \tag{3.3.3}$$

$$U^2=\frac{1}{n}\sum_{d_{12}=10}^{99}\left(Z_{d_{12}}-\overline{Z}\right)^2t_{d_{12}} \qquad \text{(Freedman-Watson)} \tag{3.3.4}$$

où $\overline{Z}=\sum_{d_{12}=10}^{99}t_{d_{12}}Z_{d_{12}}$,

$$A^2=\frac{1}{n}\sum_{d_{12}=10}^{98}\frac{Z_{d_{12}}^2t_{d_{12}}}{T_{d_{12}}\left(1-T_{d_{12}}\right)}, \qquad \text{(Anderson-Darling)} \tag{3.3.5}$$

On rejette H_0 si l'une ou l'autre de ces statistiques de tests dépasse la quantité appropriée de la loi de $\mathscr{D}_{12}$ sous H_0.

Nous avons approché par simulation pour les niveaux de significations $\alpha=10\%$, $\alpha=5\%$ et $\alpha=1\%$ via un million de réplications de Monte Carlo et pour sept (7) tailles d'échantillon, les seuils critiques sous H_0 pour chacun des quatre tests(cf Tables 3.5, 3.6 et 3.7).

	50	100	250	500	1000	3000	5000
W^2	0,351	0,350	0,348	0,348	0,348	0,347	0,348
U^2	0,155	0,154	0,153	0,152	0,152	0,152	0,152
A^2	2,031	1,972	1,929	1,922	1,916	1,909	1,913
χ^2	109,375	108,051	107,107	106,767	106,615	106,497	106,504

TABLE 3.5 – Quantiles au seuil critique $\alpha = 0.1$ en utilisant 1000000 d'échantillons aléatoires de taille n issus de la loi $\mathscr{D}_{12}$ sous l'hypothèse d'une loi de Newcomb-Benford bivariée $\mathscr{B}_{(1,2)}$

	50	100	250	500	1000	3000	5000
W^2	0,464	0,462	0,461	0,467	0,462	0,462	0,460
U^2	0,190	0,189	0,188	0,187	0,187	0,187	0,187
A^2	2,644	2,557	2,501	2,489	2,480	2,471	2,468
χ^2	116,973	114,730	113,160	112,632	112,278	112,167	112,139

TABLE 3.6 – Quantiles au seuil critique $\alpha = 0.05$ en utilisant 1000000 d'échantillons aléatoires de taille n issus de la loi $\mathscr{D}_{12}$ sous l'hypothèse d'une loi de Newcomb-Benford bivariée $\mathscr{B}_{(1,2)}$

	50	100	250	500	1000	3000	5000
W^2	0,741	0,749	0,746	0,746	0,739	0,744	0,746
U^2	0,270	0,270	0,270	0,270	0,269	0,269	0,269
A^2	4,200	4,021	3,928	3,897	3,853	3,860	3,865
χ^2	133,182	128,854	125,489	124,344	123,620	123,159	123,030

TABLE 3.7 – Quantiles au seuil critique $\alpha = 0.01$ en utilisant 1000000 d'échantillons aléatoires de taille n issus de la loi $\mathscr{D}_{12}$ sous l'hypothèse d'une loi de Newcomb-Benford bivariée $\mathscr{B}_{(1,2)}$

3.3.2 Fonction connectrice ou de lien (Connector function)

Soit X une variable aléatoire discrète, $f_X(\cdot)$ sa densité de probabilité (ou fonction de masse) définie sur $\mathcal{X} = \{x_1, ..., x_R\}$ et $F_X(\cdot)$ sa fonction de distribution cumulative (*cdf*) ou de répartition. Considérons $f_0(\cdot)$ une densité de probabilité fixée définie sur $\mathcal{X} = \{x_1, ..., x_R\}$ et $F_0(\cdot)$ sa fonction de répartition. On souhaite construire un test d'ajustement (goodness-of-fit test « GoF Test »), pour tester l'hypothèse $H_0 : f_X(\cdot) = f_0(\cdot)$. La « fonction connectrice ou de lien (connector function) » entre $F_X(\cdot)$ et $F_0(\cdot)$ est la fonction en escalier :

$$c(x; F_X, F_0) = \frac{f_X(x)}{f_0(x)} \qquad \text{pour } x_{r-1} \leq x \leq x_r,\ r = 1, ..., R, \tag{3.3.6}$$

avec $x_0 = -\infty$. Cette fonction connectrice est utile car elle permet de réexprimer H_0 sous la forme alternative : $c(x; F_X, F_0) \equiv 1$. Il existe une multitude d'expressions sources de

fonctions connectrices, dont celle de Neyman (1937) :

$$c(x; F_X, F_0) = \exp\left\{\sum_{m=1}^{\infty} \theta_m h_m(x) - K(\theta)\right\}, \tag{3.3.7}$$

où $K(\theta)$ est une constante de normalisation et $\{h_1(\cdot), h_2(\cdot), ...\}$ une suite de fonctions orthonormées relativement à $f_0(\cdot)$, c'est-à-dire satisfaisant

$$\sum_{r=1}^{R} h_m(x_r) h_{m'}(x_r) f_0(x_r) = \delta_{m,m'} \qquad \forall m, m' \geq 1,$$

le delta de Kronecker. Les fonctions de liens dépendent de $f_0(\cdot)$ et donc doivent être recalculées dans chaque cas. Algeri et Zhang (2020) proposent d'autres exemples résumés à la Table 3.8.

Méthodes	$c(x; F_X, F_0)$
Barton (1956)	$1 + \sum_{j=1}^{\infty} \tau_j h_j(x)$
Devroye et Györfi (1985)	$\max\left\{0, 1 + \sum_{j=1}^{\infty} \tau_j h_j(x)\right\} / K_{\boldsymbol{\tau}}$
Gajek (1986)	$\max\left\{0, 1 + \sum_{j=1}^{\infty} \tau_j h_j(x) - K_{\boldsymbol{\tau}}\right\}$

TABLE 3.8 – Formes populaires de $c(x; F_X, F_0)$. L'ensemble des fonctions $h_j(\cdot)$ forme une base orthonormée relativement à $f_0(\cdot)$.

En pratique, ces fonctions connectrices sont exprimées à l'ordre $M < \infty$, c'est-à-dire que les sommes infinies dans la formule 3.3.7 et la Table 3.8 sont tronquées à M termes. $c(x; F_X, F_0)$ s'exprime alors sous la forme $c_M(x)$ et dans le cas discret $M \leq R-1$ car $c_{R-1}(x; F_X, F_0) = c(x; F_X, F_0)$. En particulier Ducharme *et al.* (2020) ont calculé la fonction de lien pour la loi de Newcomb-Benford $\mathscr{D}_1$ pour le premier chiffre significatif.

Le choix de M n'est pas neutre car il affecte la puissance du test que nous produisons plus loin. Choisir un bon M dépendra des valeurs du « coefficient de Fourier » $\theta_m = \sum_{r=1}^{R} c(x_r; F_X, F_0) h_m(x_r) f_0(x_r)$: ceux qui sont nuls ne doivent pas être pris en compte dans la stratégie de test car ils correspondent à une surparamétrisation de la fonction connectrice et donc une perte de puissance.

En utilisant la nouvelle formulation $H_0 : c(x; F_X, F_0) \equiv 1$ et en supposant que l'approximation $c_M(x)$ soit bornée, le problème se réduit à celui de tester $H_0 : c_M \equiv 1$, ce qui se ramène à tester l'hypothèse nulle

$$H_0 : \theta_1 = ... = \theta_M = 0 \tag{3.3.8}$$

Pour ce faire, remarquons que la dérivée de la log-vraisemblance (basée sur la fonction connectrice de Neyman de (3.3.7))

$\log\left(f_M(\cdot)\right) = \log\left(\left(1 + \sum_{m=1}^{M} \theta_m h_m(x)\right) f_0(\cdot)\right)$, par rapport à θ donne les « scores » $U_m = \sum_{i=1}^{n} h_m(X_i)$. Il découle du théorème central limite et de l'orthogonalité des

fonctions $\{h_j(\cdot)\}$ que sous H_0, $n^{-1/2}(U_1, .., U_M)$ suit asymptotiquement la multinormale $\mathscr{N}_M(0, I_M)$, où I_M est la matrice identité d'ordre M.

Ainsi, à l'instar de ce qui a été fait dans le chapitre précédent, la statistique de test lisse (smooth test)

$$\mathcal{S}_M = n^{-1} \sum_{i=1}^{M} U_m^2 \tag{3.3.9}$$

suit asymptotiquement la loi χ^2_M sous l'hypothèse nulle $H_0 : f_X(\cdot) = f_0(\cdot)$. On retrouve donc le test lisse de la section 2 du chapitre précédent et on rejette H_0 si $\mathcal{S}_M > x_{M,1-\alpha}$, avec $x_{M,1-\alpha}$ le $1-\alpha$-ième quantile de la loi χ^2_M.

3.3.3 Smooth test appliqué à la loi de $\mathscr{D}_{12}$ sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$

En utilisant les mêmes méthodes que dans le chapitre précédent et notamment en exploitant le théorème de Boulerice et Ducharme (1997), on obtient les expressions suivantes pour les polynômes $h_m, m \in \{1, \ldots, 5\}$ à l'aide du logiciel MATHEMATICA, en considérant $f_0(d_{12}) = \pi_{d_{12}}$

$$\begin{aligned}
h_1(d_{12}) &= -1.5475 + 0.0401\, d_{12}, \\
h_2(d_{12}) &= 2.7958 - 0.1674\, d_{12} + 0.001\, d_{12}^2, \\
h_3(d_{12}) &= -5.1878 + 0.4869\, d_{12} - 0.0115\, d_{12}^2 + 0.00007\, d_{12}^3, \\
h_4(d_{12}) &= 9.7252 - 1.2375\, d_{12} + 0.0476\, d_{12}^2 - 0.0006\, d_{12}^3 + 3.3718 \times 10^{-6} d_{12}^4, \\
h_5(d_{12}) &= -18.3299 + 2.9372\, d_{12} - 0.1586\, d_{12}^2 + 0.0037\, d_{12}^3 \\
&\quad - 0.00003\, d_{12}^4 + 1.4943 \times 10^{-7}\, d_{12}^5.
\end{aligned}$$

On a déjà mentionné qu'une question importante dans l'application des tests lisses est le choix du meilleur M. En effet, choisir M trop grand surparamétrise la fonction connectrice et entraîne un effet de dilution de la puissance de test. Par contre, un M trop petit conduit à un test avec une bonne puissance sur certaines alternatives. mais médiocre pour d'autres.

Pour résoudre ce problème, Ledwina (1994) a suggéré d'incorporer le choix de M dans la procédure de test (« data driven $\mathcal{S}_{DD}$ ») en utilisant le critère d'information bayésien (BIC, Schwarz (1978)) qui offre un compromis entre la qualité de l'ajustement et la complexité du modèle. Nous avons utilisé cette méthode de sélection du M dans le chapitre précédent et suggéré de prendre

$$\hat{m} = \arg \max_{1 \leq m \leq 5} \{\mathcal{S}_m - m \log(n)\}.$$

En s'appuyant sur les travaux de Akaike (1973) qui propose le critère AIC, Inglot et Ledwina (2006) vont proposer une autre manière de choisir M :

> « use BIC only when an alternative is very distant from the null distribution, otherwise use AIC ».

Cette combinaison des 2 critères demande d'abord de définir une valeur maximale $d(n)$ de M. Les critères BIC ($S1$) et AIC ($A1$) pour la statistique de test (3.3.9), sont

$$S1 = min\{1 \leq m \leq d(n) : \mathcal{S}_m - k \log m \geq \mathcal{S}_j - j \log n, \quad j = 1, \ldots, d(n)\} \tag{3.3.10}$$

$$A1 = min\{1 \leq m \leq d(n) : \mathcal{S}_m - 2m \geq \mathcal{S}_j - 2j, \quad j = 1, \ldots, d(n)\} \tag{3.3.11}$$

Pour utiliser l'approche de Inglot et Ledwina (2006), il faut maintenant décider si on est proche du modèle de l'hypothèse nulle ou non. Malheureusement, on ne connaît pas à priori le vrai modèle. Pour résoudre ce problème, Inglot et Ledwina (2006) proposent d'utiliser la règle suivante. Soit

$$I_n(c) = \mathbb{1}\left(\max_{1 \leq m \leq d(n)} \mid \sqrt{n}\widehat{h_m} \mid \leq \sqrt{c \log n}\right) \tag{3.3.12}$$

où c *UNE CONSTANTE POSITIVE À DÉTERMINER* et $\widehat{h_m} = \frac{1}{n}\sum_{i=1}^{n} h_m(\cdot)$.

Considérons la nouvelle pénalité :

$$\pi(j, n) = \{j \log n\}\{I_n(c)\} + \{2j\}\{1 - I_n(c)\} \tag{3.3.13}$$

qui permet d'obtenir une nouvelle statistique de test que nous noterons $\mathcal{S}_{NDD}$ (new data driven smooth test) où,

$$m^* = \min\{1 \leq m \leq d(n) : \mathcal{S}_m - \pi(m, n) \geq \mathcal{S}_j - \pi(j, n), \, j = 1, \ldots, d(n)\}$$
$$\mathcal{S}_{NDD} = \mathcal{S}_{m^*}$$

Remarque 5. Quand $c \to +\infty$, $\mathcal{S}_{NDD}$ converge vers $\mathcal{S}_{DD}$. Le choix d'un bon c, pour chaque $f_0(\cdot)$ se fait par une procédure empirique de calibrage utilisant des simulations et menant à un « bon compromis ». Dans le cadre de la loi de Newcomb-Benford, cette procédure a mené à la valeur $c = 1.3$ pour $n > 100$. Le lecteur verra à la Section 3.3.5 le travail de calibrage menant à ce choix. Notons que dans leur article, Inglot et Ledwina (2006) choisissent $c = 2.4$.

Dans la suite $\mathcal{S}_{NDD}$, correspond à la statistique de test « new data driven smooth » pour $c = 1.3$.

Pour compléter la stratégie du test de $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$ basé sur $\mathscr{D}_{12}$ nous avons approchés les seuils critiques sous H_0 pour chacun des tests $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, $\mathcal{S}_{DD}$, $\mathcal{S}_{NDD}$ par simulation de Monte Carlo pour les niveaux de significations $\alpha = 10\%$, $\alpha = 5\%$ et $\alpha = 1\%$ via un million de réplications et pour les sept (7) tailles d'échantillon considérées (cf Tables 3.9, 3.10 et 3.11).

	50	100	250	500	1000	3000	5000
$\mathcal{S}_1$	2,714	2,703	2,704	2,704	2,702	2,702	2,706
$\mathcal{S}_2$	4,565	4,589	4,591	4,612	4,611	4,604	4,608
$\mathcal{S}_3$	6,194	6,223	6,229	6,243	6,261	6,238	6,249
$\mathcal{S}_4$	7,720	7,748	7,757	7,771	7,779	7,763	7,774
$\mathcal{S}_5$	9,191	9,217	9,201	9,228	9,239	9,215	9,224
$\mathcal{S}_{DD}$	4,020	3,388	3,050	2,927	2,844	2,771	2,757
$\mathcal{S}_{NDD}$	5,752	4,602	3,461	3,140	2,967	2,826	2,797

TABLE 3.9 – Seuils critiques de la loi des statistiques $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, $\mathcal{S}_{DD}$, $\mathcal{S}_{NDD}$ sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$ au seuil $\alpha = 0.1$ en utilisant 1000000 échantillons aléatoires

	50	100	250	500	1000	3000	5000
S_1	3,817	3,830	3,829	3,848	3,840	3,831	3,843
S_2	5,947	5,968	5,971	5,997	5,987	5,976	5,990
S_3	7,793	7,818	7,818	7,807	7,814	7,812	7,820
S_4	9,511	9,511	9,496	9,499	9,481	9,482	9,500
S_5	11,153	11,126	11,079	11,071	11,066	11,073	11,073
S_{DD}	6,044	5,484	4,718	4,344	4,177	4,005	3,960
S_{NDD}	9,297	8,213	6,207	5,033	4,499	4,136	4,056

TABLE 3.10 – Seuils critiques de la loi des statistiques $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, $\mathcal{S}_{DD}$, $\mathcal{S}_{NDD}$ sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$ au seuil $\alpha = 0.05$ en utilisant 1000000 échantillons aléatoires

	50	100	250	500	1000	3000	5000
S_1	6,586	6,608	6,634	6,659	6,600	6,645	6,661
S_2	9,260	9,238	9,259	9,233	9,189	9,180	9,198
S_3	11,564	11,442	11,413	11,377	11,330	11,323	11,364
S_4	13,723	13,520	13,396	13,3111	13,250	13,274	13,264
S_5	15,796	15,423	15,255	15,174	15,140	15,088	15,092
S_{DD}	12,194	10,433	9,148	8,726	8,377	7,830	7,484
S_{NDD}	15,025	14,415	13,495	12,582	11,224	9,0222	8,334

TABLE 3.11 – Seuils critiques de la loi des statistiques $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, $\mathcal{S}_{DD}$, $\mathcal{S}_{NDD}$ sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$ au seuil $\alpha = 0.01$ en utilisant 1000000 échantillons aléatoires

3.3.4 Alternatives

Dans cette section, nous construisons des familles d'alternatives pour mener une étude de la puissance de différents tests. Idéalement, ces familles devraient intégrer la loi $\mathscr{B}_{(1,2)}$ de $(\mathscr{D}_1, \mathscr{D}_2)$ (cf 3.1.2) sous H_0, au moins approximativement, de sorte que les courbes de puissance puissent être à peu près collées sur le niveau choisi du test. Pour le cas de la loi $\mathscr{B}_1$ de $\mathscr{D}_1$ sous H_0, il existe quelques familles d'alternatives répertoriées dans le chapitre précédent. Certaines de ces familles sont facilement étendues au cas bivarié. En particulier, la fonction de densité de la famille bivariée de Rodriguez (2004) notée $\mathcal{R}_{(1,2)}(\gamma)$ prend la forme

$$\mathbb{P}\left[\mathscr{D}_1 = d_1, \mathscr{D}_2 = d_2\right] = \begin{cases} \dfrac{1}{810}(9 + 19\ln(10) + 9(10d_1 + d_2)\ln(10d_1 + d_2) - \\ 9(10d_1 + d_2 + 1) + 19\ln(10d_1 + d_2 + 1)) & si\,\gamma = 0 \\ \log_{10}[1 + \dfrac{1}{10d_1 + d_2}] & si\,\gamma = -1 \\ \frac{\gamma+1}{90\gamma} - \frac{(10d_1+d_2+1)^{\gamma+1} - (10d_1+d_2)^{\gamma+1}}{\gamma(100^{\gamma+1} - 10^{\gamma+1})} & sinon \end{cases} \tag{3.3.14}$$

, avec $\gamma \in \mathbb{R}$ et $(d_1, d_2) \in \{1, \ldots, 9\} \times \{0, \ldots, 9\}$. Le cas $\gamma = -1$ correspond à la loi Newcomb-Benford bivariée $\mathscr{B}_{(1,2)}$ (3.1.2), tandis que le cas $\gamma = 0$ donne la version bivariée de la distribution de Stigler $S_{(1,2)}$ (Lee *et al.*, 2010). Bien entendu, la somme de ces probabilités sur d_2 donne la distribution univariée, $\mathcal{R}_1(\gamma)$ de Rodriguez (2004) et la somme sur d_1 donne la distribution marginale de $\mathcal{R}_2(\gamma)$. À partir de là, les densités conditionnelles $\mathcal{R}_{(1|2)}(\gamma)$ et $\mathcal{R}_{(2|1)}(\gamma)$ sont facilement obtenues. Également, la famille Newcomb-Benford Généralisée définie dans Pietronero *et al.* (2001) est étendue en bivariée et notée $\mathcal{GB}_{(1,2)}(\gamma)$. Sa fonction de densité est donnée par

$$\mathbb{P}\left[\mathscr{D}_1 = d_1, \mathscr{D}_2 = d_2\right] = \begin{cases} \log_{10}[1 + 1/(10d_1 + d_2)] & \text{si } \gamma = 0 \\ \frac{(10d_1+d_2)^{-\gamma} - (10d_1+d_2+1)^{-\gamma}}{(10^{-\gamma} - 100^{-\gamma})} & \text{sinon,} \end{cases}$$

avec $\gamma \in \mathbb{R}$ et $(d_1, d_2) \in \{1, \ldots, 9\} \times \{0, \ldots, 9\}$. Bien entendu également, la somme de ces probabilités sur d_2 donne la distribution univariée, $\mathcal{GB}_1(\gamma)$ de Pietronero *et al.* (2001) et la somme sur d_1 donne la distribution marginale de $\mathcal{GB}_2(\gamma)$. À partir de là, les densités conditionnelles $\mathcal{GB}_{(1|2)}(\gamma)$ et $\mathcal{GB}_{(2|1)}(\gamma)$ sont facilement obtenues avec les formules usuelles.

Nous pouvons donc construire les distributions bivariées où les variables aléatoires $(\mathscr{D}_1, \mathscr{D}_2)$ suivent les couples de lois $\left(\mathscr{B}_1, \mathcal{R}_{(2|1)}(\gamma)\right)$, $\left(\mathcal{R}_1, \mathscr{B}_{(2|1)}(\gamma)\right)$, $\left(\mathcal{R}_1(\gamma), \mathcal{R}_{(2|1)}(\gamma)\right)$,

$\Big(\mathscr{B}_1, \mathcal{GB}_{(2|1)}(\gamma)\Big)$, $\Big(\mathcal{GB}_1(\gamma), \mathscr{B}_{(2|1)}\Big)$ et $\Big(\mathcal{GB}_1(\gamma), \mathcal{GB}_{(2|1)}(\gamma)\Big)$.

Wong (2010) a proposé une famille d'alternative obtenue en faisant une $\gamma-$mixture, $\gamma \in [0,1]$ entre la loi de $\mathscr{D}_{12}$ sous H_0 et d'autres distributions définies sur $\{10,\ldots,99\}$. En particulier, il considère la mixture $\gamma\mathscr{B}_{(1,2)} + (1-\gamma)U_{[10,99]}$, et une autre mixture avec la distribution empirique bivariée de Hill (1988), notée $\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\mathcal{H}_{(1,2)}$. À ceux-ci nous avons ajouté

- une $\gamma-$mixture entre la loi $\mathscr{B}_{(1,2)}$ sous H_0 et la distribution de Stigler $S_{(1,2)}$ notée $\gamma\mathscr{B}_{(1,2)} + (1-\gamma)S_{(1,2)}$
- une $\gamma-$mixture entre la loi $\mathscr{B}_{(1,2)}$ sous H_0 et le produit extérieur de la distribution bivariée de Stigler du premier chiffre significatif S_1 et l'uniforme sur $\{0,1,\ldots,9\}$ notée $\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,\mathcal{S}_1 \otimes U_{[0,9]}$.
- une $\gamma-$mixture entre la loi $\mathscr{B}_{(1,2)}$ sous H_0 et le produit extérieur de l'uniforme sur $\{1,\ldots,9\}$ et la distribution de Stigler du deuxième chiffre significatif S_2 notée $\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,U_{[1,9]} \otimes S_2$.
- une $\gamma-$mixture entre la loi $\mathscr{B}_{(1,2)}$ sous H_0 et le produit extérieur de la distribution de Hill du premier chiffre significatif $\mathcal{H}_1$ et l'uniforme sur $\{0,1,\ldots,9\}$ notée $\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,\mathcal{H}_1 \otimes U_{[0,9]}$.
- une $\gamma-$mixture entre la loi $\mathscr{B}_{(1,2)}$ sous H_0 et le produit extérieur de l'uniforme sur $\{1,\ldots,9\}$ et la distribution de Hill du deuxième chiffre significatif $\mathcal{H}_2$ notée $\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,U_{[1,9]} \otimes \mathcal{H}_2$.

Notons que dans tous ces cas, si $\gamma = 1$, la $\gamma-$mixture est la loi $\mathscr{B}_{(1,2)}$ sous H_0.

Une autre approche pour créer des alternatives intéressantes dans le contexte de la détection de fraude est de considérer le cas où le fraudeur ne modifie que le second chiffre significatif $\mathscr{D}_2$, laissant le premier chiffre significatif intact, ou modifié pour être concordant avec la loi de Newcomb-Benford du premier chiffre significatif. Ce faisant, il aura tendance à négliger la corrélation entre le premier et le second chiffre significatif, ce qui conduit à une distribution bivariée où $\mathscr{D}_1 \sim \mathscr{B}_1$ tandis que $\mathscr{D}_2$ est indépendant de $\mathscr{D}_1$ et suit une distribution différente de la loi de Newcomb-Benford pour $\mathscr{D}_2$. À cet égard, nous avons considéré le cas où $\mathscr{D}_2 \sim \mathcal{R}_2(\gamma)$. Nous avons également examiné le cas où le fraudeur aurait légèrement modifié les deux chiffres indépendamment pour répondre à ses besoins, en utilisant une approche similaire. Cela conduit à considérer $(\mathscr{D}_1, \mathscr{D}_2) \sim (\mathcal{R}_1(\gamma) \perp \mathcal{R}_2(\gamma))$ où le symbole $\perp$ dénote l'indépendance. Nous nommons cette approche, la famille Indépendance. En se basant sur cela, nous définissons aussi $(\mathcal{B}_1 \perp \mathcal{GB}_2(\gamma))$ et $(\mathcal{GB}_1(\gamma) \perp \mathcal{GB}_2(\gamma))$.

Enfin, une autre famille d'alternatives englobant la famille indépendance peut être obtenue via des copules, en particulier la copule de Farlie-Gumbel-Morgenstein(FGM). Soit $F(\cdot)$, $G(\cdot)$ respectivement les fonctions de répartitions de $\mathscr{D}_1$ et $\mathscr{D}_2$. La fonction de répartition bivariée de la copule FGM est définie par :

$$C(d_1, d_2; \gamma, F, G) = F(d_1) \times G(d_2) \times [1 + \gamma\,(1 - F(d_1))\,(1 - G(d_2))]\,, \tag{3.3.15}$$

$(d_1, d_2) \in \{1, \ldots, 9\} \times \{0, \ldots, 9\}$. Dans le cas $\gamma = 0$, la copule correspond à la famille indépendance. Nous considérons ici les copules $C(\gamma, \mathscr{B}_1, \mathscr{B}_2)$, $C(\gamma, \mathscr{B}_1, S_2)$, $C(\gamma, \mathscr{B}_1, \mathcal{H}_2)$ et $C(\gamma, \mathcal{S}_1, \mathscr{B}_2)$.

Nous subdivisons l'ensemble ainsi des 21 alternatives en deux groupes. Un sous ensemble de 8 « `Training` » et un autre 13 de « testing ». Pour une meilleure clarté nous n'affichons dans la suite de ce document que les travaux liés aux alternatives de « `Training` ». Les graphiques liés aux alternatives de « `Testing` » se trouvent dans les annexes. La Table 3.12 donne un récapitulatif de toutes les alternatives que nous avons construits.

Familles	Alternatives	γ	Relation avec $\mathscr{B}_{12}$	Échantillon
Mixture	$\gamma\mathscr{B}_{(1,2)} + (1-\gamma)U_{[10,99]}$	$[0,1]$	$\gamma = 1 \Rightarrow \mathcal{B}_{(1,2)}$	`Training`
Mixture	$\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\mathcal{H}_{(1,2)}$	$[0,1]$	$\gamma = 1 \Rightarrow \mathcal{B}_{(1,2)}$	`Testing`
Mixture	$\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,\mathcal{S}_1 \otimes U_{[0,9]}$	$[0,1]$	$\gamma = 1 \Rightarrow \mathcal{B}_{(1,2)}$	`Training`
Mixture	$\gamma\mathscr{B}_{(1,2)} + (1-\gamma)S_{(1,2)}$	$[0,1]$	$\gamma = 1 \Rightarrow \mathcal{B}_{(1,2)}$	`Testing`
Mixture	$\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,U_{[1,9]} \otimes S_2$	$[0,1]$	$\gamma = 1 \Rightarrow \mathcal{B}_{(1,2)}$	`Testing`
Mixture	$\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,\mathcal{H}_1 \otimes U_{[0,9]}$	$[0,1]$	$\gamma = 1 \Rightarrow \mathcal{B}_{(1,2)}$	`Testing`
Mixture	$\gamma\mathscr{B}_{(1,2)} + (1-\gamma)\,U_{[1,9]} \otimes \mathcal{H}_2$	$[0,1]$	$\gamma = 1 \Rightarrow \mathcal{B}_{(1,2)}$	`Testing`
$\left(F_1, F_{2\|1}\right)$	$\left(\mathscr{B}_1, \mathcal{R}_{(2\|1)}(\gamma)\right)$	$\mathbb{R}$	$\gamma = -1 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Training`
$\left(F_1, F_{2\|1}\right)$	$\left(\mathcal{R}_1(\gamma), \mathscr{B}_{(2\|1)}(\gamma)\right)$	$\mathbb{R}$	$\gamma = -1 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Training`
$\left(F_1, F_{2\|1}\right)$	$\left(\mathcal{R}_1(\gamma), \mathcal{R}_{(2\|1)}(\gamma)\right)$	$\mathbb{R}$	$\gamma = -1 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Testing`
$\left(F_1, F_{2\|1}\right)$	$\left(\mathscr{B}_1, \mathcal{GB}_{(2\|1)}(\gamma)\right)$	$\mathbb{R}$	$\gamma = 0 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Testing`
$\left(F_1, F_{2\|1}\right)$	$\left(\mathcal{GB}_1(\gamma), \mathscr{B}_{(2\|1)}\right)$	$\mathbb{R}$	$\gamma = 0 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Testing`
$\left(F_1, F_{2\|1}\right)$	$\left(\mathcal{GB}_1(\gamma), \mathcal{GB}_{(2\|1)}(\gamma)\right)$	$\mathbb{R}$	$\gamma = 0 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Testing`
Copule	$C(\gamma, \mathscr{B}_1, \mathscr{B}_2)$	$[-1,1]$	$\gamma \simeq 0.135 \Rightarrow \mathcal{B}_{(1,2)}$	`Training`
Copule	$C(\gamma, \mathscr{B}_1, S_2)$	$[-1,1]$	$\gamma \simeq 0.135 \Rightarrow \mathcal{B}_{(1,2)}$	`Testing`
Copule	$C(\gamma, \mathscr{B}_1, \mathcal{H}_2)$	$[-1,1]$	$\gamma \simeq 0.135 \Rightarrow \mathcal{B}_{(1,2)}$	`Training`
Copule	$C(\gamma, \mathcal{S}_1, \mathscr{B}_2)$	$[-1,1]$	$\gamma \simeq 0.135 \Rightarrow \mathcal{B}_{(1,2)}$	`Testing`
Indépendance	$(\mathcal{R}_1(\gamma) \perp \mathcal{R}_2(\gamma))$	$\mathbb{R}$	$\gamma = -1 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Training`
Indépendance	$(\mathcal{B}_1 \perp \mathcal{R}_2(\gamma))$	$\mathbb{R}$	$\gamma = -1 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Training`
Indépendance	$(\mathcal{B}_1 \perp \mathcal{GB}_2(\gamma))$	$\mathbb{R}$	$\gamma = -1 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Testing`
Indépendance	$(\mathcal{GB}_1(\gamma) \perp \mathcal{GB}_2(\gamma))$	$\mathbb{R}$	$\gamma = -1 \Rightarrow \mathcal{B}_1 \perp \mathcal{B}_2$	`Testing`

TABLE 3.12 – Liste des alternatives

3.3.5 Calibrage du c

Il est utopique d'espérer l'existence d'un c (cf (3.3.13)) qui donnerait le meilleur test en terme de puissance pour toutes les alternatives. Nous chercherons alors un c faisant un bon compromis.

Pour parvenir à ce résultat, nous faisons varier c entre 0 et 3 avec un pas de 0.1 (c'est-à-dire $c \in \{0, 0.1, 0.2, 0.3, \ldots, 3\}$). Les tests $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, $\mathcal{S}_{DD}$ et $\mathcal{S}_{NDD}$ sont effectués au niveau 5% et leur puissance est approximée par Monte Carlo (10000 réplications) afin de permettre une juste comparaison des résultats. Les puissances des tests pour chaque triplet $(\text{famille}, n, \gamma)$ sont approximées par le nombre de rejets parmi les 10000 réplications. Pour donner une idée des résultats obtenus, les graphiques des courbes de puissance de six tests, soit

— le meilleur et le pire test parmi $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5\}$,
— $\mathcal{S}_{DD}$,
— $\{\mathcal{S}_{NDD}, c \in \{1.1, 1.3, 1.5\}\}$

apparaissent (un panneau par famille d'alternatives et taille d'échantillon) dans les Figures 3.3.1 à 3.3.8. Nous attirons l'attention du lecteur sur le fait que les tests sur certaines alternatives de « `Training` » étant semblables, nous affichons uniquement les alternatives donnant des résultats différents. Comme les puissances de ces figures sont approximées de 10000 réplications, elles contiennent un bruit statistique que l'on peut évaluer à environ ± 0.01 (au niveau de confiance 95%) lorsque la puissance est autour de 0.5.

Ne perdons pas de vue que l'objectif est de trouver un c dont le test $\mathcal{S}_{NDD}$ associé soit meilleur que le test $\mathcal{S}_{DD}$. Pour ce faire, nous nous concentrons sur l'enveloppe des courbes de puissance du meilleur et du pire test des $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5\}$. Ensuite on choisit le c dont le test associé est si possible $\geq \mathcal{S}_{DD}$ et le plus proche du meilleur test parmi $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5\}$, de façon à peu près uniforme en terme de n, de la famille d'alternatives et de la valeur de γ. De toutes ces figures, il ressort que la valeur de c faisant compromis est **1.3**.

Famille des mixtures: Benford Uniforme

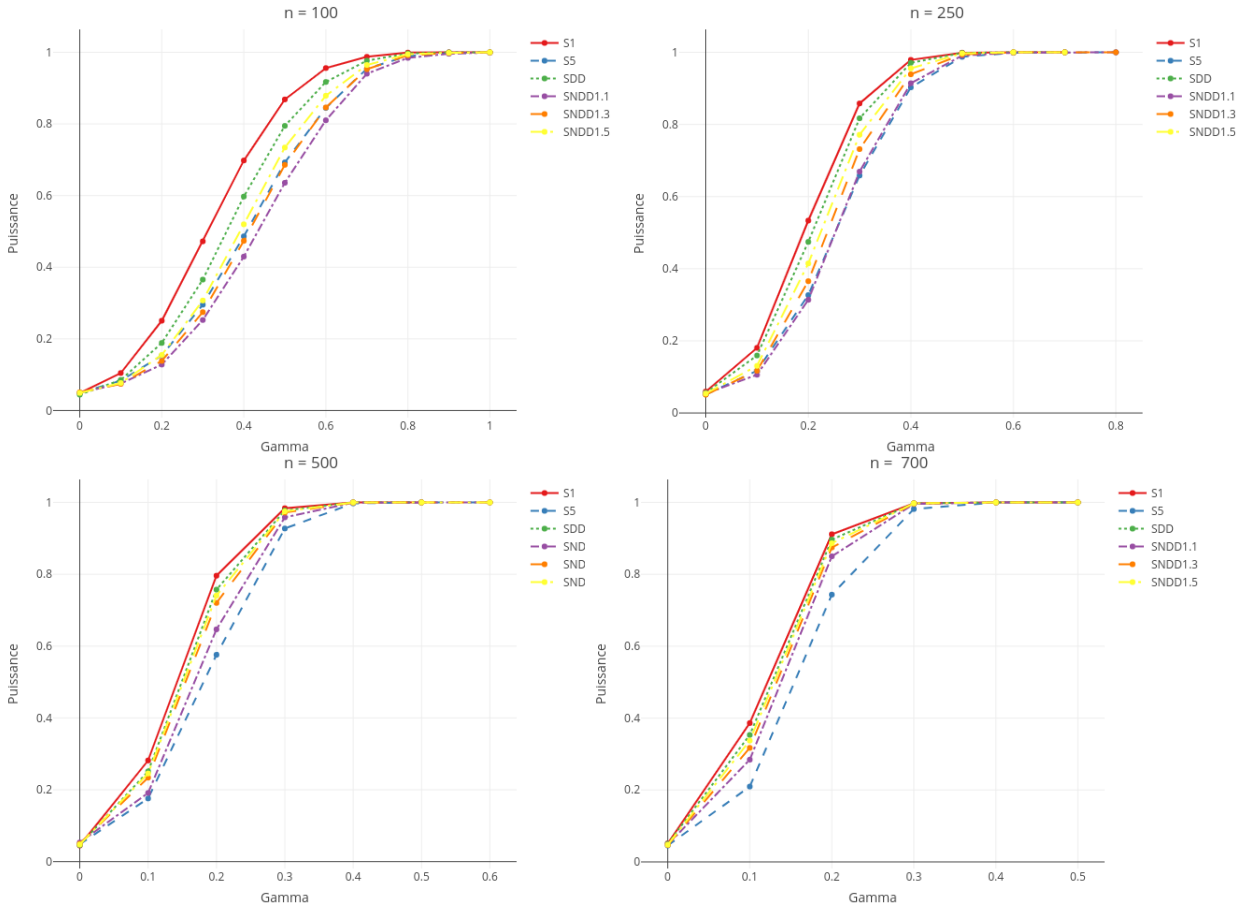

FIGURE 3.3.1 – Mixture Benford Uniforme $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma U_{[10,99]}$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des mixtures: Benford Stigler Uniforme

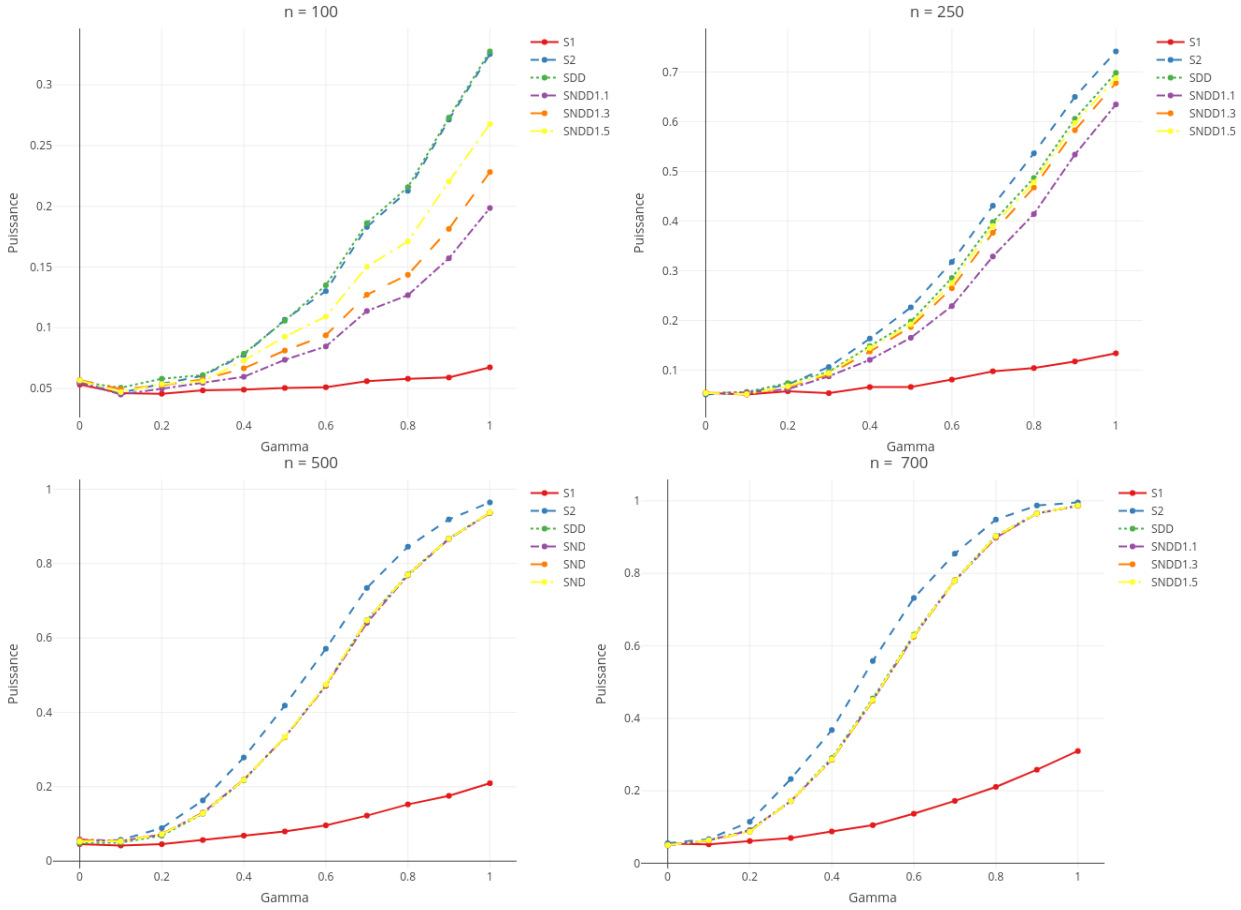

FIGURE 3.3.2 – Mixture Benford Stigler Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, \mathcal{S}_1 \otimes U_{[0,9]}$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille Indépendance: Benford Rodriguez

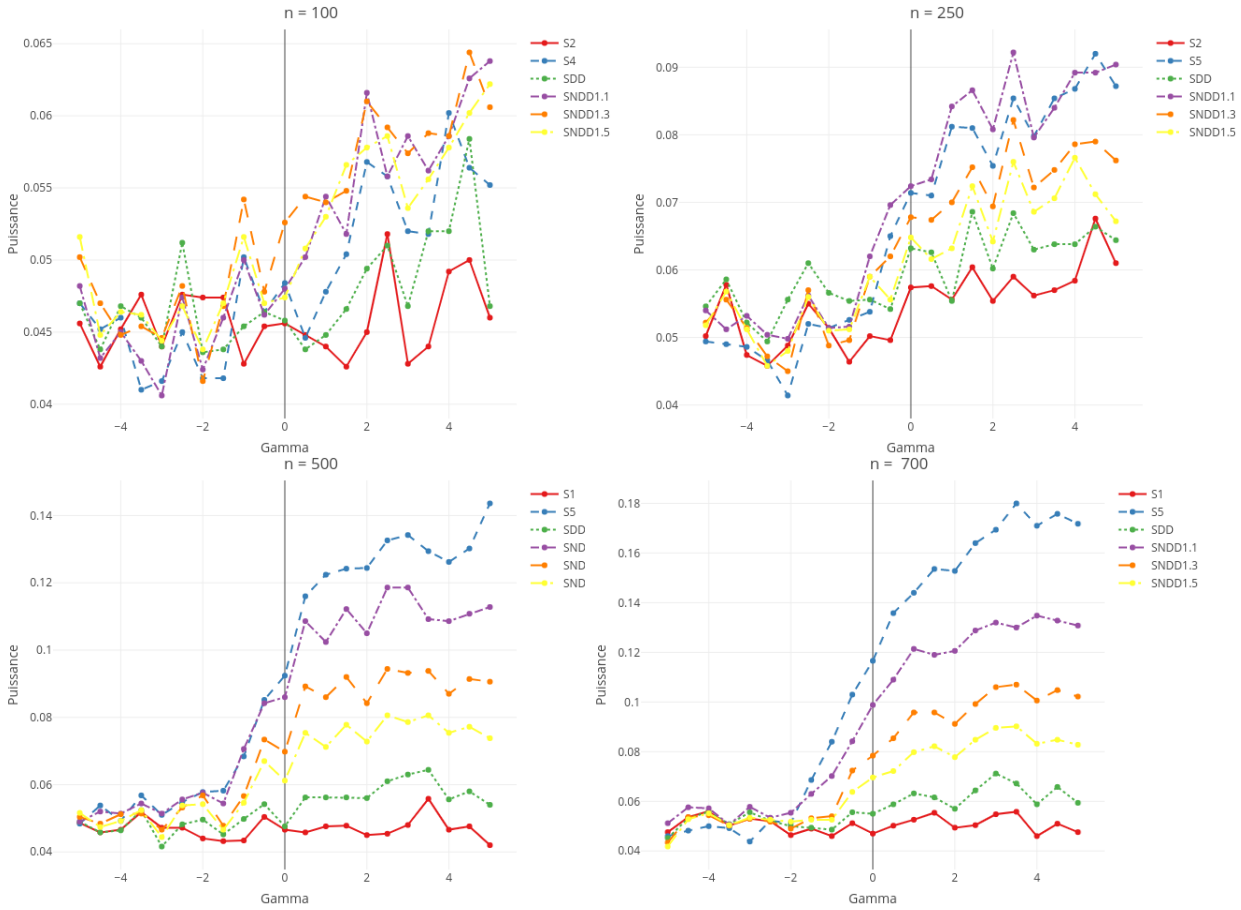

FIGURE 3.3.3 – Indépendance Benford Rodriguez ($\mathcal{B}_1 \perp \mathcal{R}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille Indépendance: Rodriguez Rodriguez

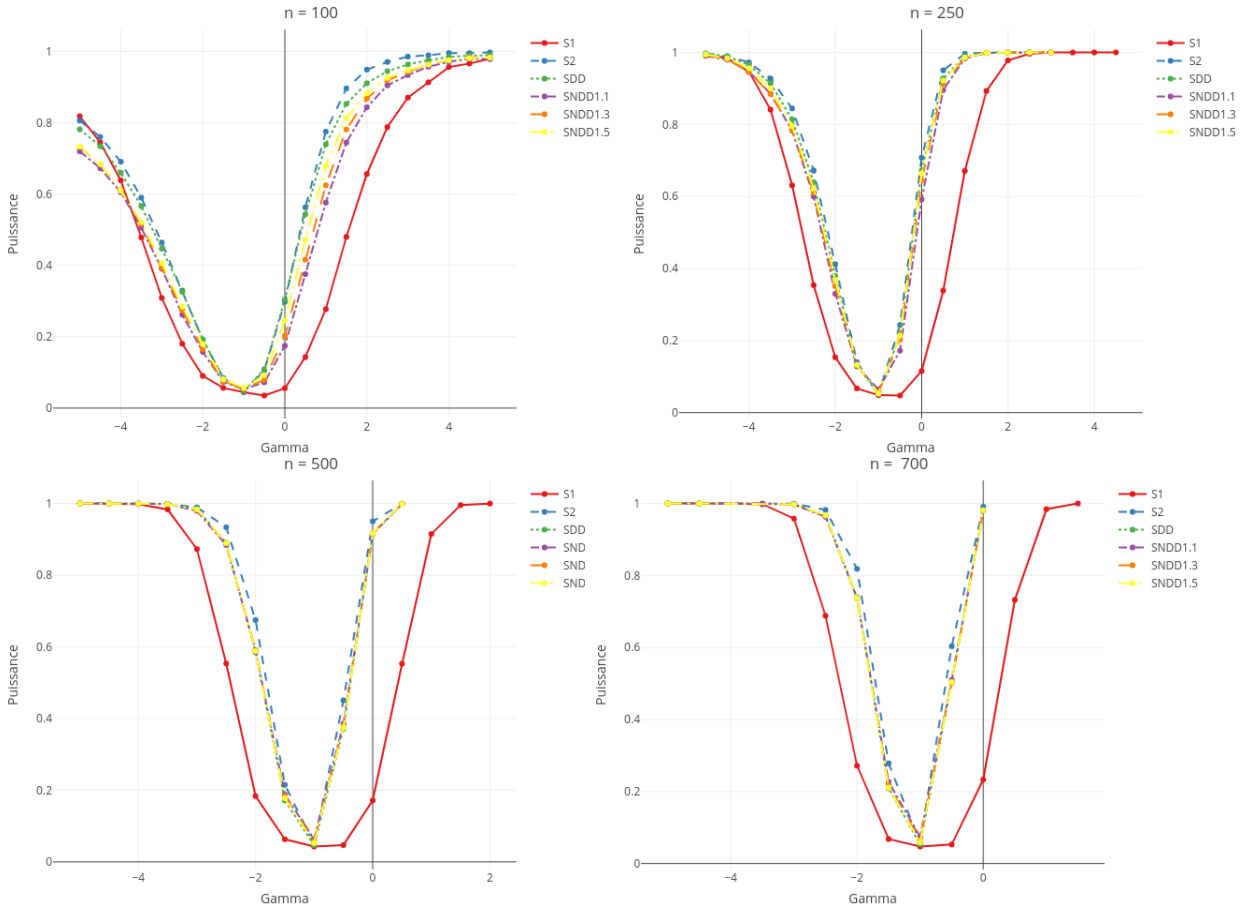

FIGURE 3.3.4 – Indépendance Rodriguez Rodriguez $(\mathcal{R}_1(\gamma) \perp \mathcal{R}_2(\gamma))$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille de Copules: Benford Hill

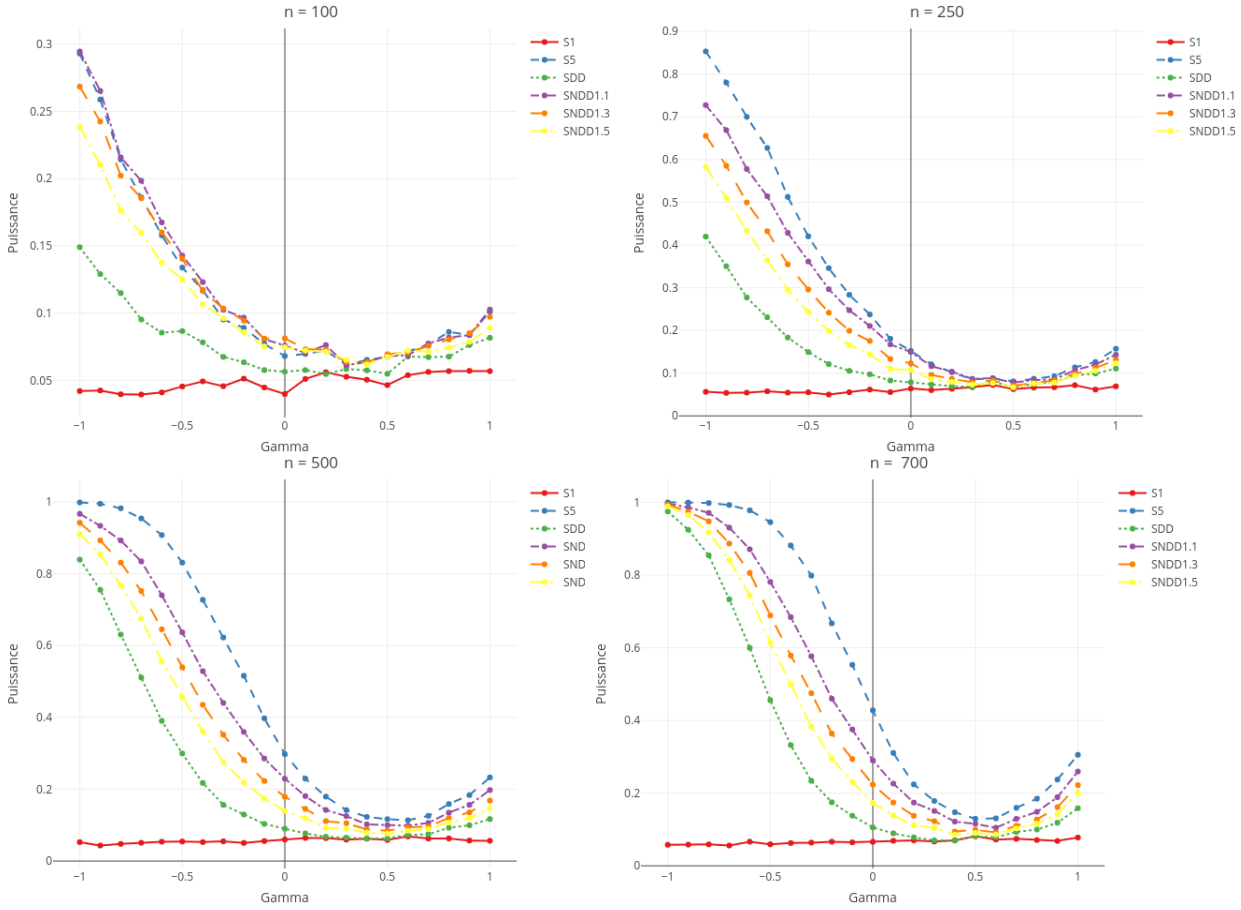

FIGURE 3.3.5 – Copule Benford Hill $C(\gamma, \mathscr{B}_1, H_2)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille de Copules: Benford Benford

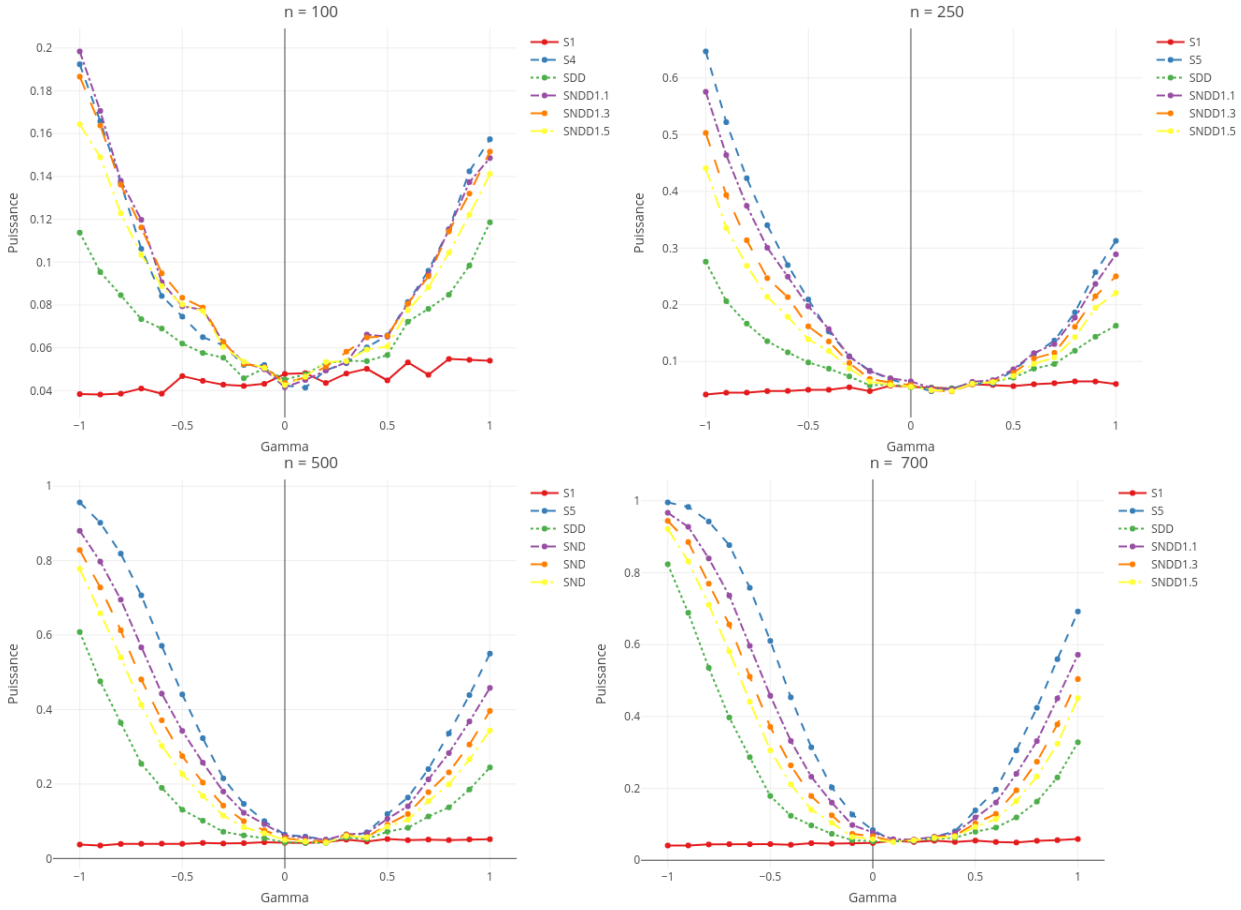

FIGURE 3.3.6 – Copule Benford Benford $C(\gamma, \mathscr{B}_1, \mathscr{B}_2)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des conditionnelles: Rodriguez sachant Benford

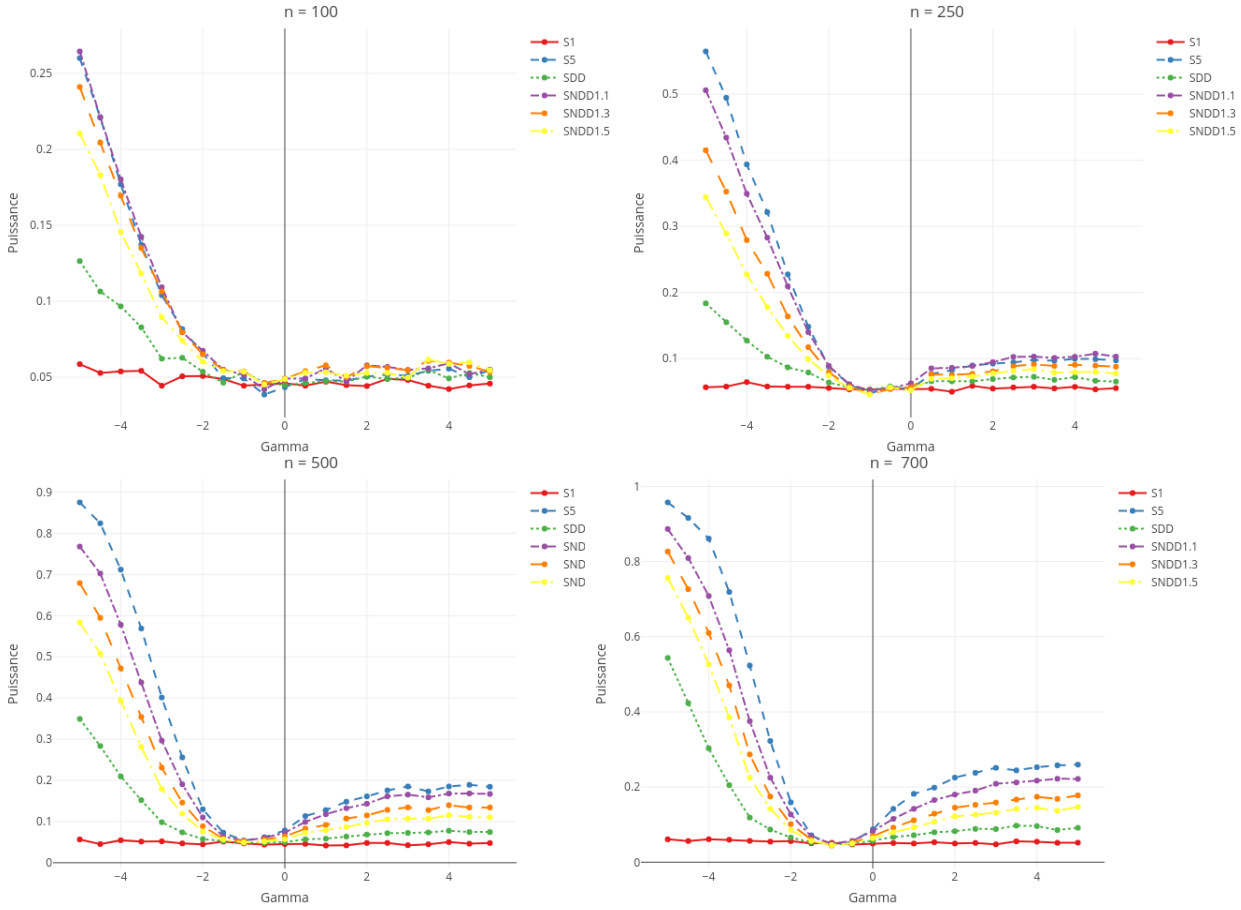

FIGURE 3.3.7 – Rodriguez sachant Benford $\left(\mathscr{B}_1, \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des conditionnelles: Benford sachant Rodriguez

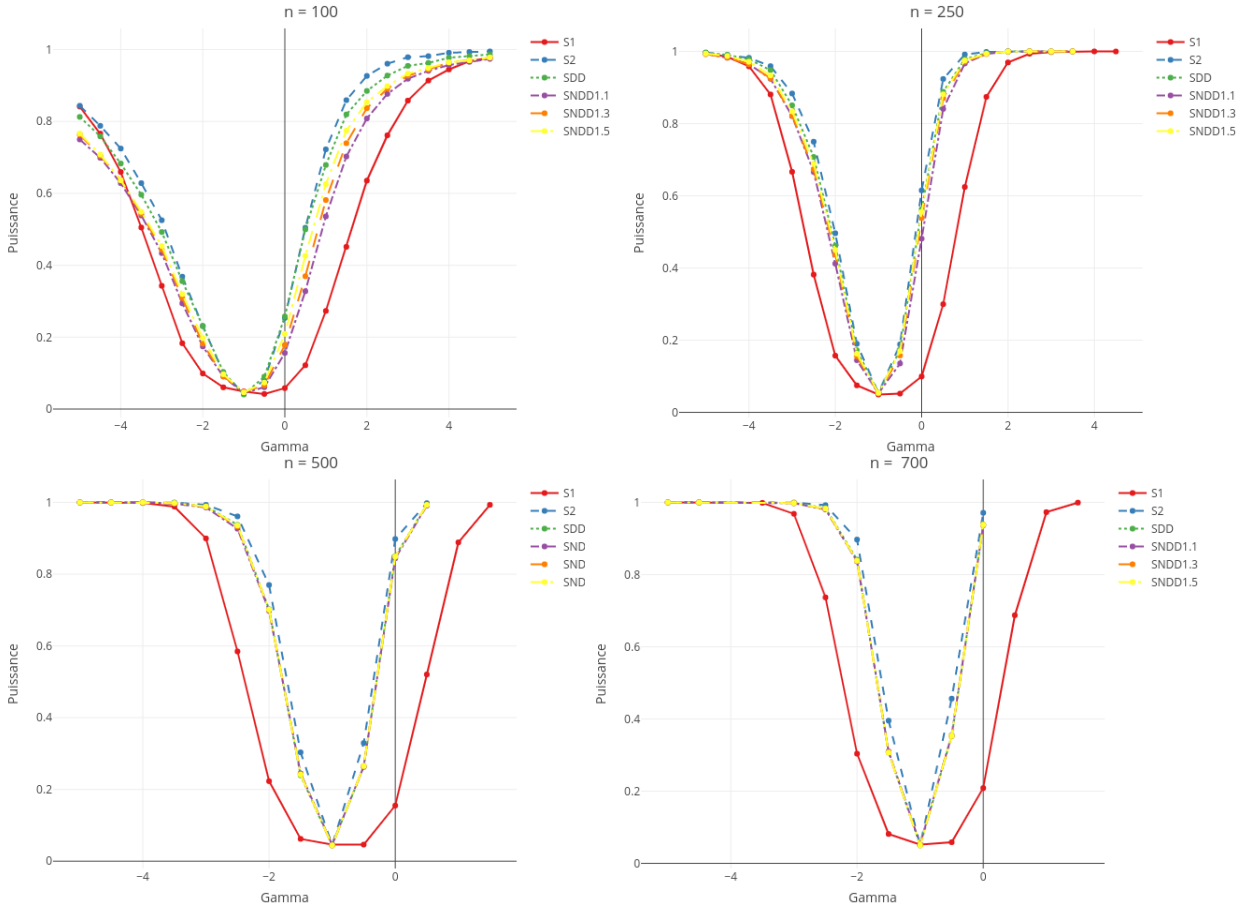

FIGURE 3.3.8 – Benford sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathscr{B}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

3.3.6 Comparaison du test « new data driven smooth test $\mathcal{S}_{NDD}$ » et des tests classiques

Nous utiliserons dans ce contexte les alternatives de « `Training` » de la section 3.3.4, et rajoutons des alternatives proposées dans Wong (2010). Wong imagine la situation où des transactions sont traités plusieurs fois et donc propose une contamination de la loi de Newcomb-Benford par addition et multiplication. Concrètement il s'agit respectivement de rajouter/ multiplier une valeur γ aux probabilités de la loi de Newcomb-Benford univariée définie sur $\{10, \ldots, 99\}$.

Dans la suite $\mathcal{S}_{NDD}$, correspond à la statistique de test « new data driven smooth »

pour $c = 1.3$.

Les tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 sont effectués au niveau 5% et leur puissance est approximée par Monte Carlo (10000 réplications) pour chaque triplet (famille, n, γ) . Les courbes de puissance résultantes apparaissent aux figures 3.3.9 à 3.3.18. De ces figures, nous remarquons qu'aucun des tests ne sort gagnant tout le temps mais nous pouvons regrouper leur comportement en deux groupes.

- **Cas où W^2 ou U^2 sont meilleurs** (cf Figures 3.3.9, 3.3.10, 3.3.12, 3.3.15, 3.3.16, 3.3.17, et 3.3.18) : La courbe de puissance de la statistique de test $\{\mathcal{S}_{NDD}, c = 1.3\}$, se situe généralement entre celle du W^2 et du U^2. Quand la taille de l'échantillon n devient grand, sa courbe de puissance se rapproche de la « meilleure courbe ». Nous pouvons déduire que dans ces situations $\mathcal{S}_{NDD}$ est un bon compromis dans la mesure où l'on ne sait pas lequel du W^2ou U^2 est le plus puissant pour une alternative donnée.

- **Cas où χ^2 est le meilleur test** (cf Figures 3.3.11, 3.3.13 et 3.3.14) : La puissance de la statistique de test $\mathcal{S}_{NDD}$, est très faible.

En conclusion, si W^2 ou U^2 sont plus puissants que le test du χ^2, $\{\mathcal{S}_{NDD}, c = 1.3\}$ offre un bon compromis. Par contre si le test du χ^2 est le plus puissant des tests classiques considérés ici, alors le test $\mathcal{S}_{NDD}$ ne devrait pas être utilisé. Le problème c'est qu'on ne peut savoir à l'avance pour une alternative donnée, si le test du χ^2 est préférable aux autres tests.

Famille des mixtures: Benford Uniforme

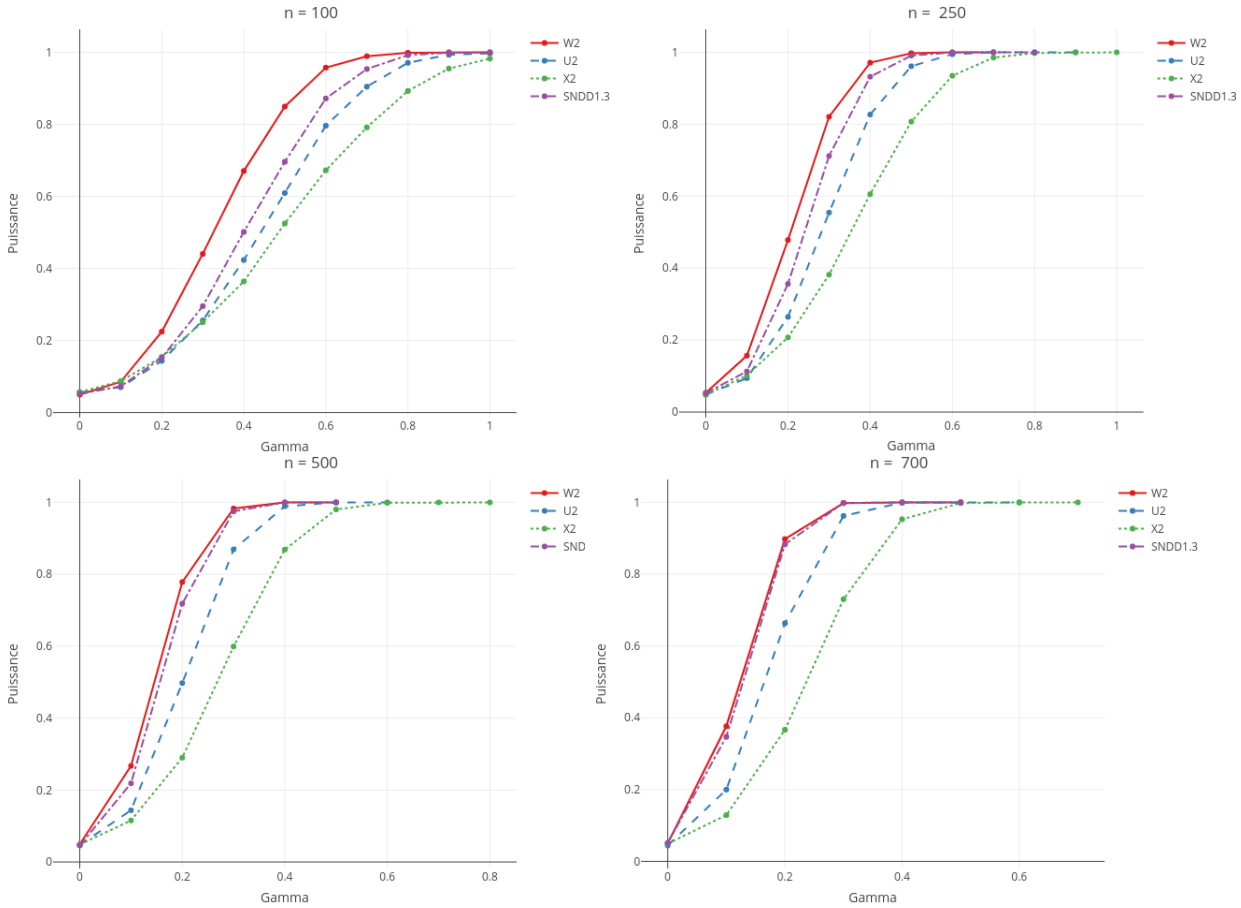

FIGURE 3.3.9 – Mixture Benford Uniforme $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma U_{[10,99]}$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Stigler Uniforme

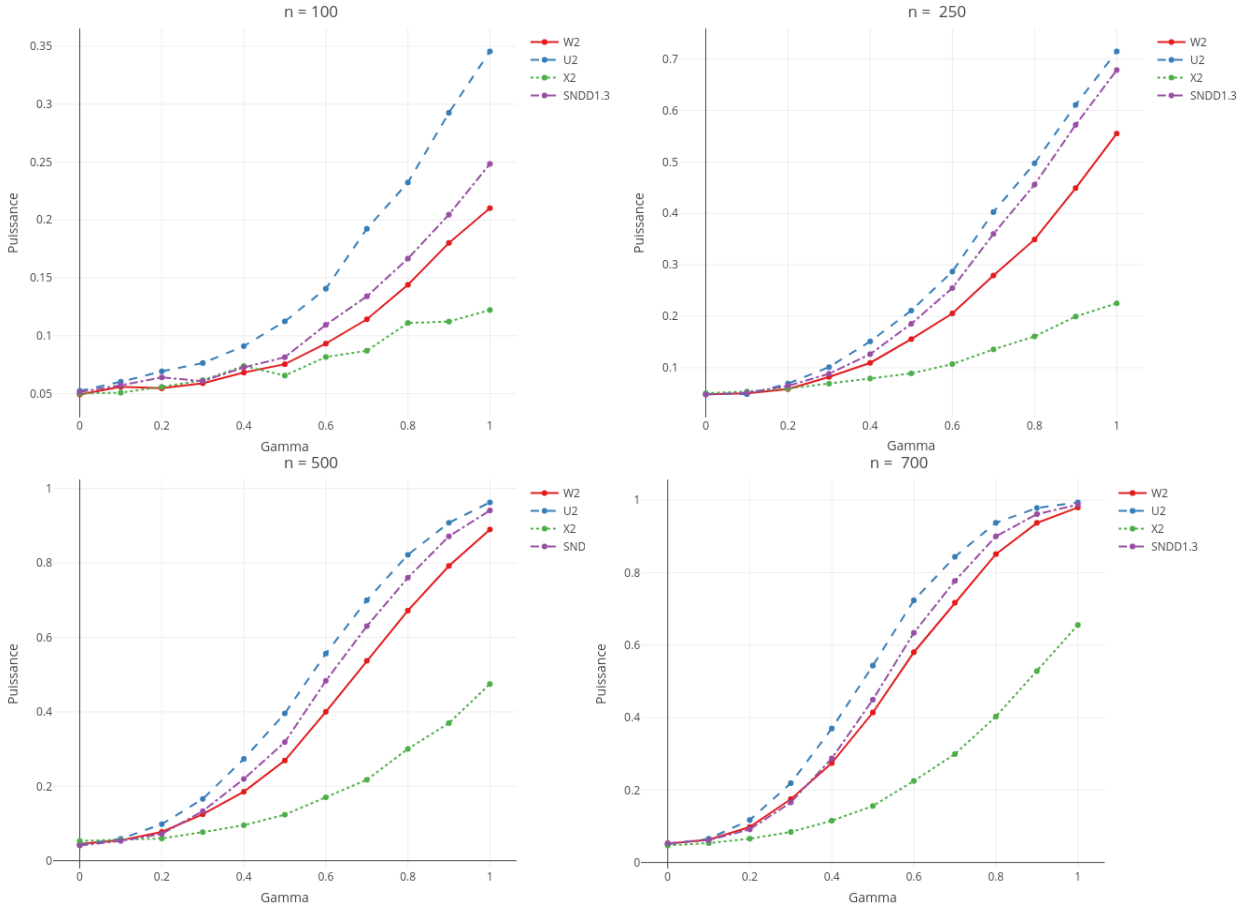

FIGURE 3.3.10 – Famille des mixtures : Mixture Benford Stigler Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, \mathcal{S}_1 \otimes U_{[10,99]}$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Benford Rodriguez

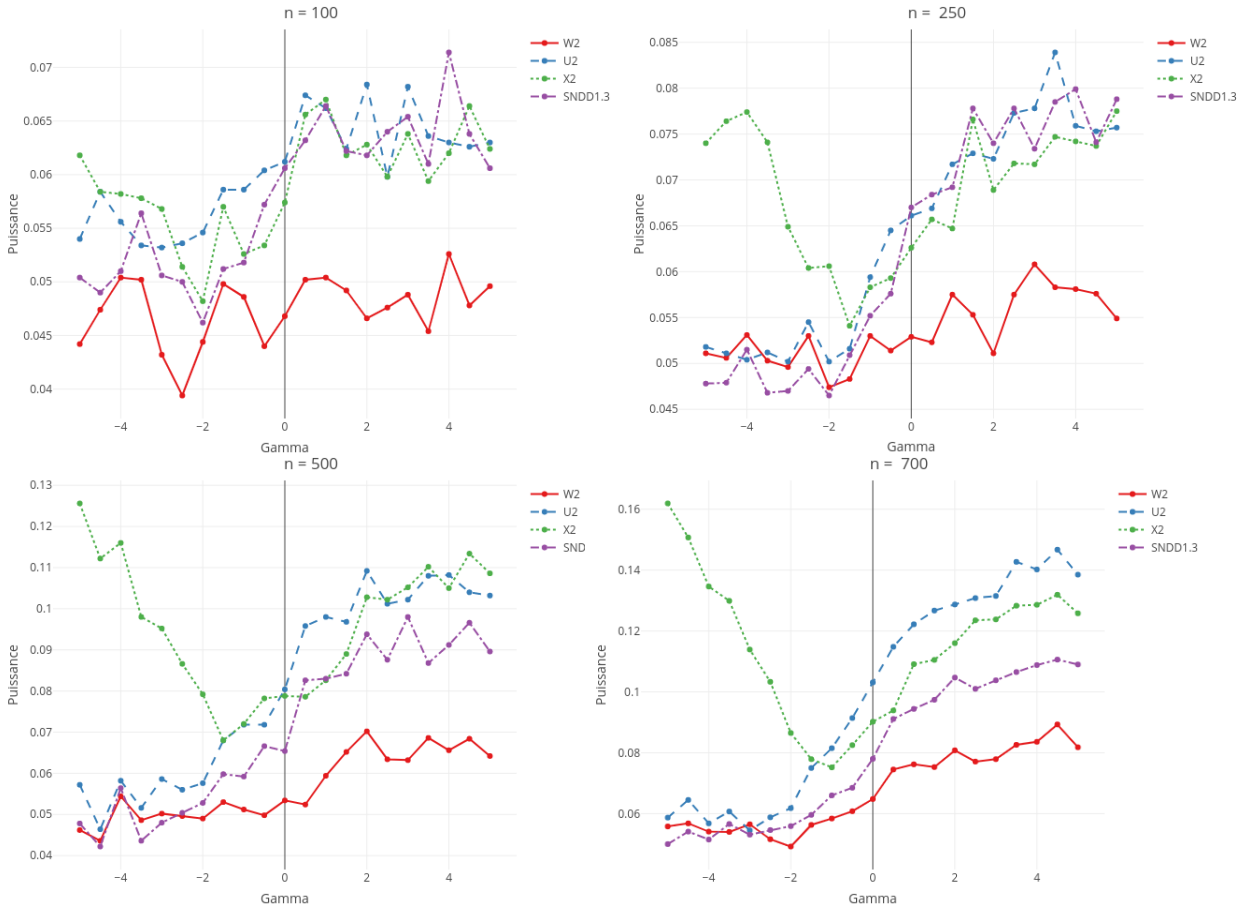

FIGURE 3.3.11 – Famille des indépendances : Indépendance Benford Rodriguez ($\mathscr{B}_1 \perp \mathcal{R}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Rodriguez Rodriguez

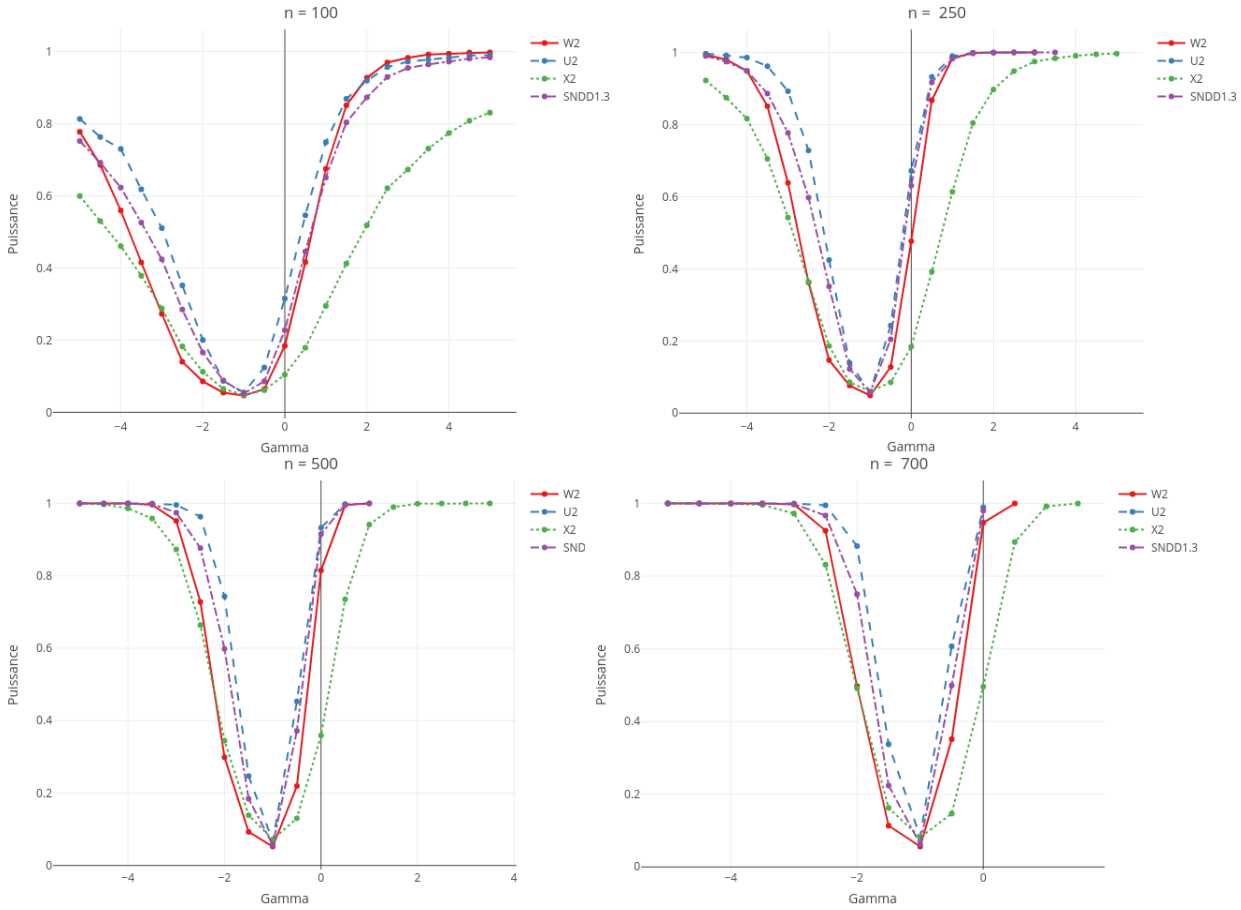

FIGURE 3.3.12 – Famille des indépendances : Indépendance Rodriguez Rodriguez $(\mathcal{R}_1(\gamma) \perp \mathcal{R}_2(\gamma))$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille de Copules: Benford Hill

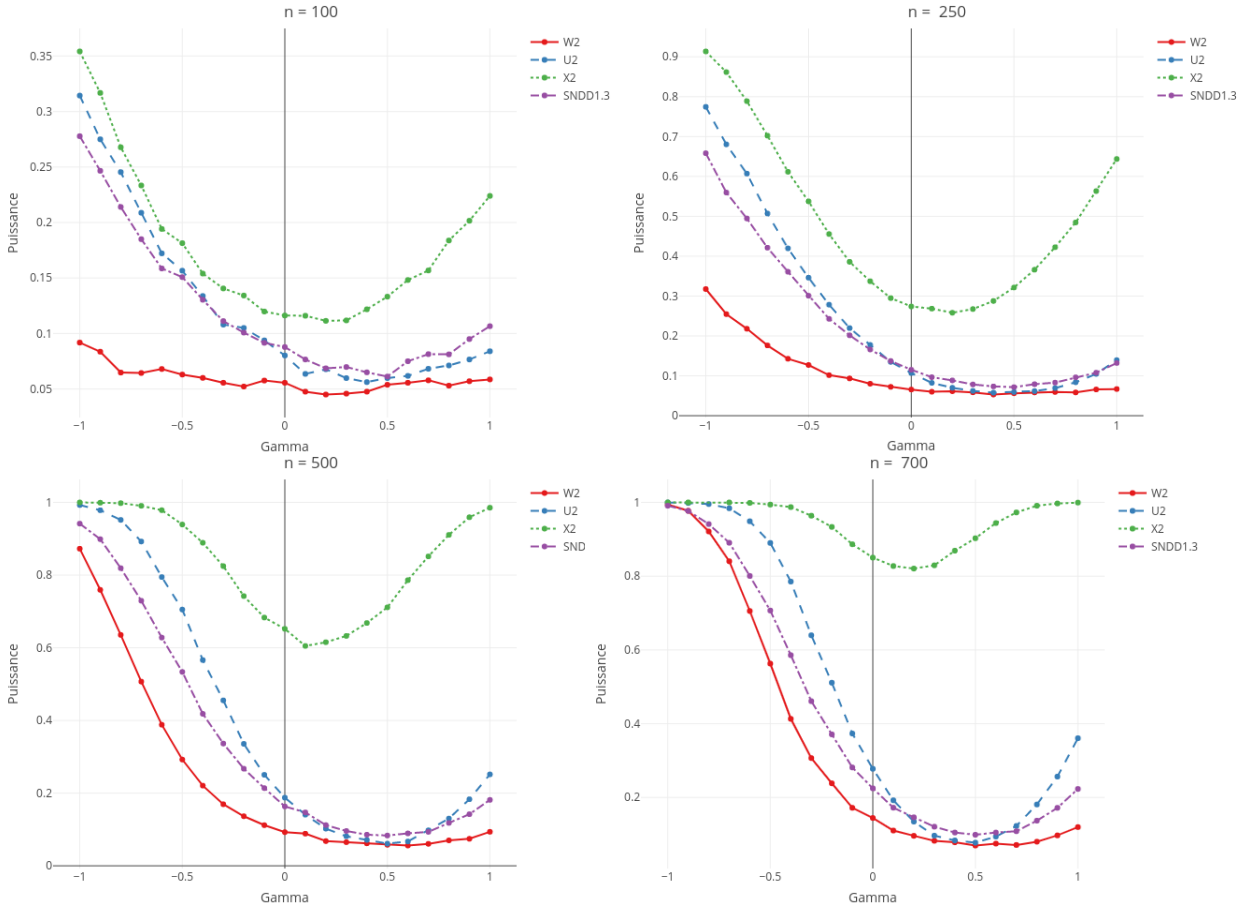

FIGURE 3.3.13 – Famille des copules : Copule Benford Hill $C(\gamma, \mathscr{B}_1, \mathcal{H}_2)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille de Copules: Benford Benford

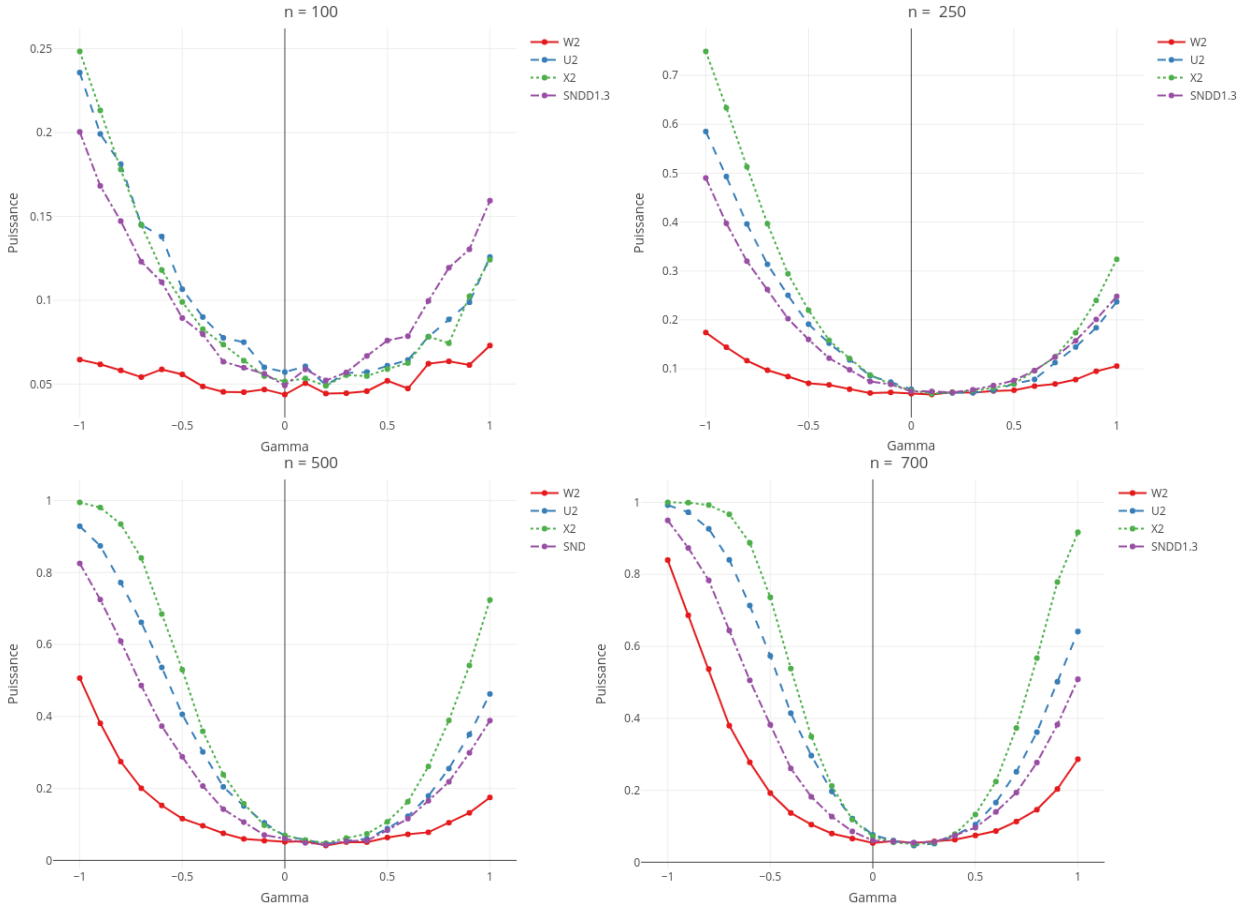

FIGURE 3.3.14 – Famille des copules : Copule Benford Benford $C(\gamma, \mathscr{B}_1, \mathscr{B}_2)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Rodriguez sachant Benford

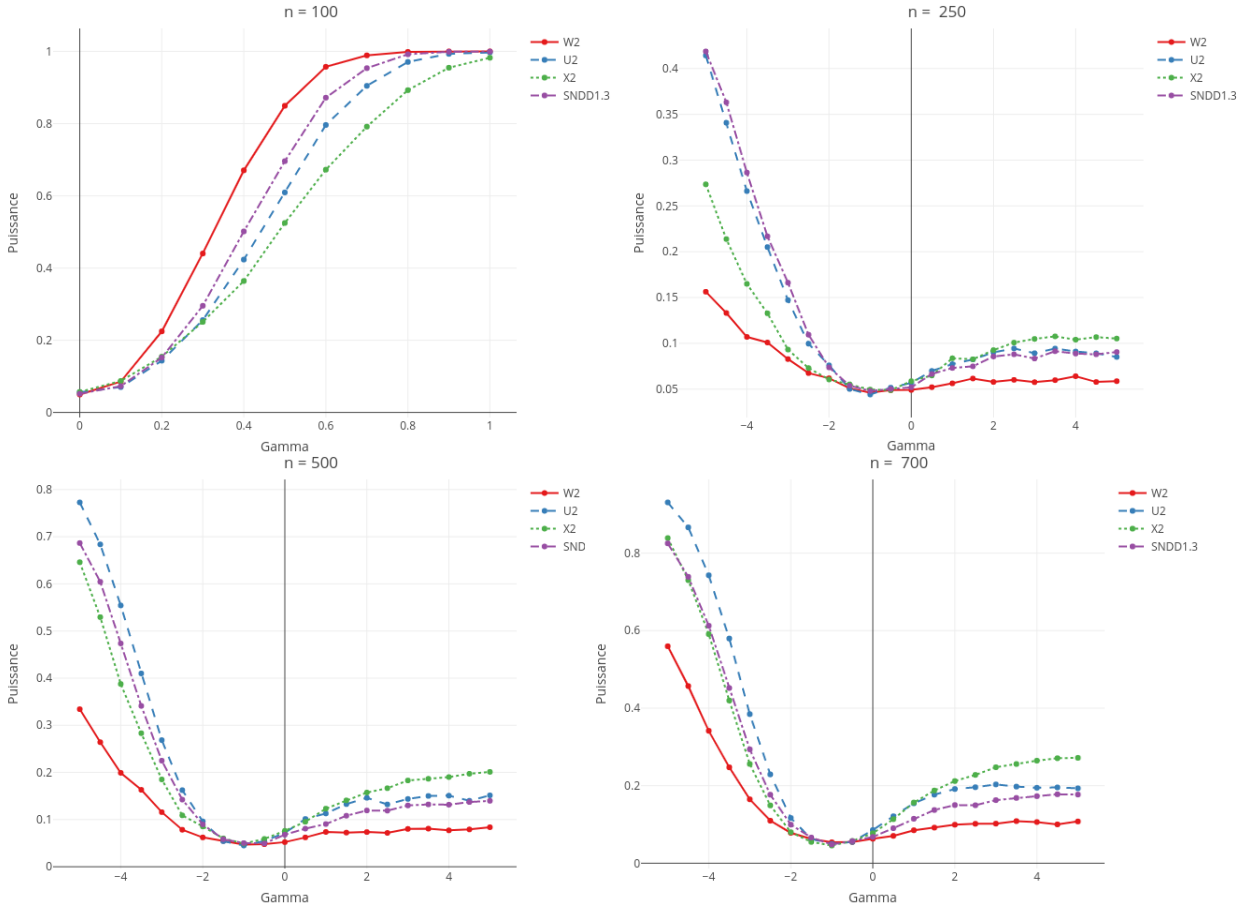

FIGURE 3.3.15 – Famille des conditionnelles : Rodriguez sachant Benford $\left(\mathscr{B}_1, \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Benford sachant Rodriguez

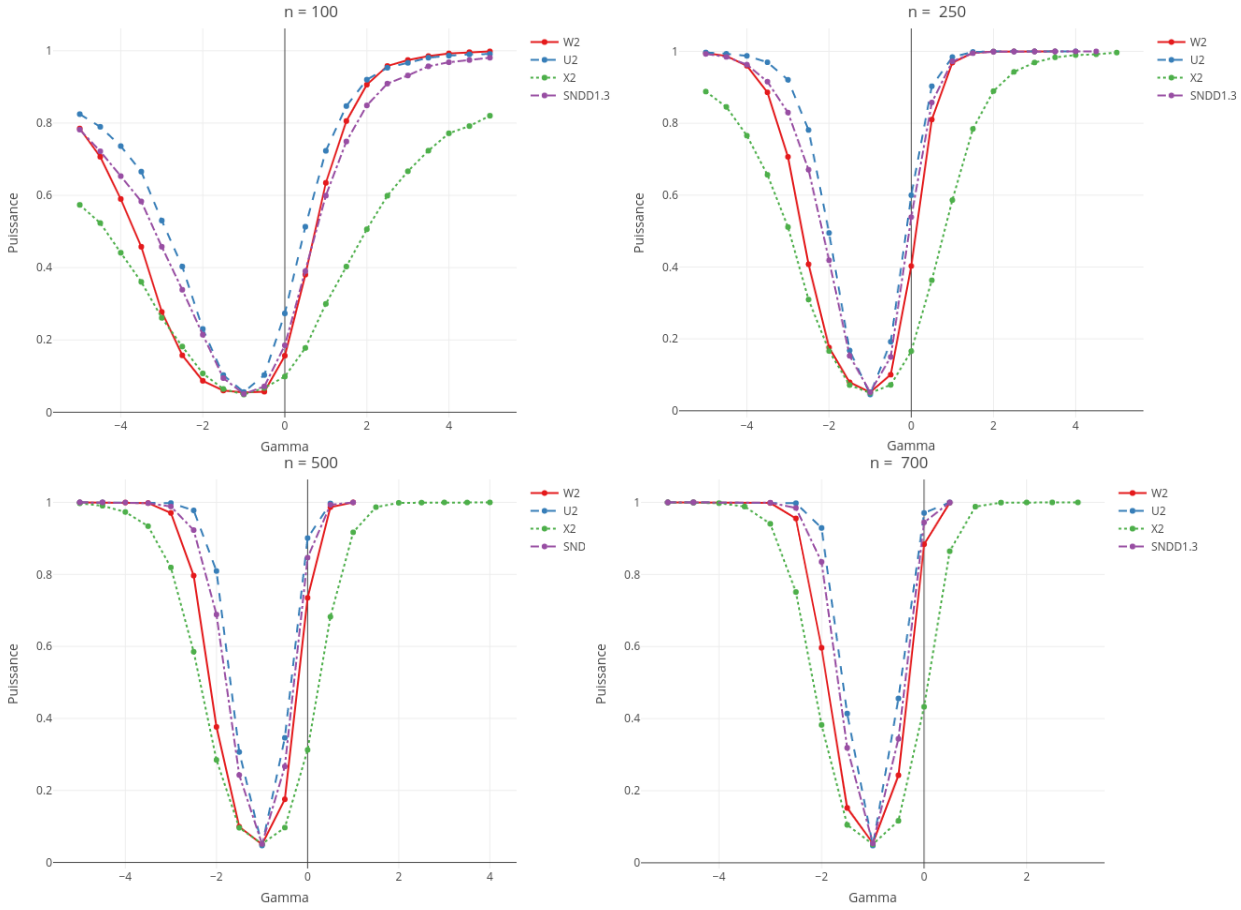

FIGURE 3.3.16 – Famille des conditionnelles : Benford sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathscr{B}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Alternative de Wong: Modification additive de Benford

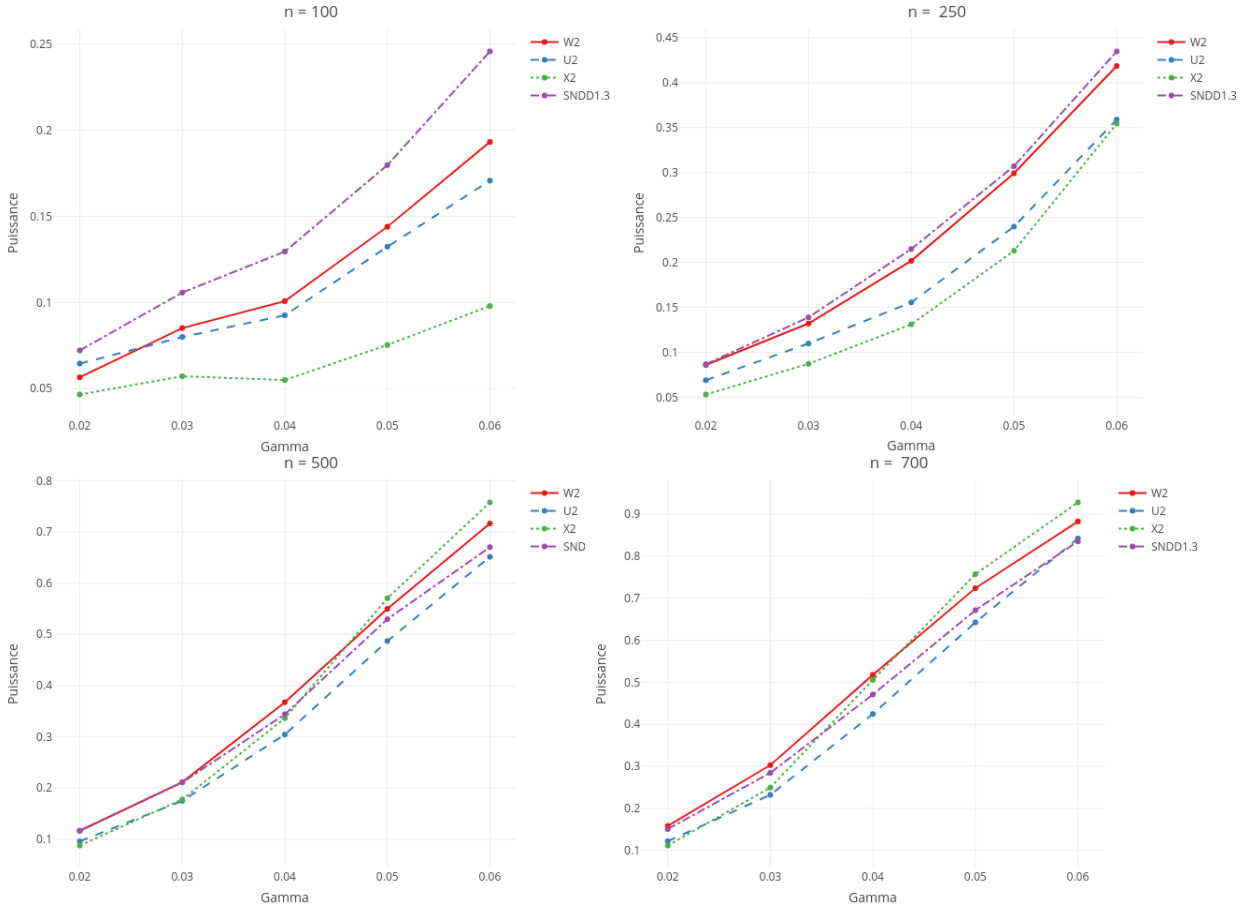

FIGURE 3.3.17 – Famille de Wong additive : Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Alternative de Wong: Modification multiplicative de Benford

FIGURE 3.3.18 – Famille de Wong multiplicative : Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Dans la section suivante, nous allons explorer le comportement des alternatives pour essayer de comprendre pourquoi le $\mathcal{S}_{NDD}$ n'offre pas le meilleur compromis dans toutes les situations.

3.3.7 Analyse des alternatives

Pour essayer de comprendre pourquoi les tests classiques U^2 et W^2 donnent parfois des puissances nettement inférieures au test du χ^2 pour certaines alternatives étudiées, nous avons analysé les fréquences d'apparitions des chiffres significatifs et les fonctions de répartition desdites alternatives. Les figures 3.3.19 à 3.3.28 représentent les résultats de cette analyse.

Nous remarquons deux types de comportements des courbes de fonctions de réparti-

tion. Pour les alternatives (Copule Benford Benford, Copule Benford Hill et Indépendance Benford Rodriguez) où le test de χ^2 est meilleur (cf Figures 3.3.14, 3.3.13 et 3.3.3) , les courbes respectives de leur fonction de répartition (à gauche $\gamma = 0$, à droite $\gamma = 1$) se positionnent à peu près au même niveau que la fonction de répartition de la loi de Newcomb-Benford (cf Figures 3.3.24, 3.3.23, et 3.3.21). Or les expressions de U^2 (3.3.4) et W^2 (3.3.3) mesurent les écarts entre la fonction de répartition empirique et celle attendue sous H_0. De même, dans le cas où W^2 ou U^2 sont plus puissants que le test du χ^2 (cf Figures 3.3.9, 3.3.10, 3.3.12, 3.3.15, 3.3.16, 3.3.17, et 3.3.18) , les courbes des fonctions de répartition des alternatives concernées sont notablement différentes de la courbe de la fonction de répartition de la loi de Newcomb-Benford (cf Figures 3.3.19, 3.3.20, 3.3.22, 3.3.26). En pratique, il est difficile de prévoir à l'avance quels tests classiques (U^2, W^2) et χ^2 vont le mieux se comporter puisque l'on ne connaît pas en général l'alternative réelle. Ce point est important car le test $\mathcal{S}_{NDD}$ à une puissance intermédiaire entre ces trois tests. Si l'écart entre leur puissance est grand, $\mathcal{S}_{NDD}$ n'offrira pas un un bon compromis, se situant en terme de puissance entre ces tests, souvent plus proche de U^2 et W^2

L'objectif de la prochaine section est de mettre en place un test lisse pouvant faire un meilleur compromis dans le cas où le test du χ^2 est meilleur.

`Famille des mixtures: Benford Uniforme`

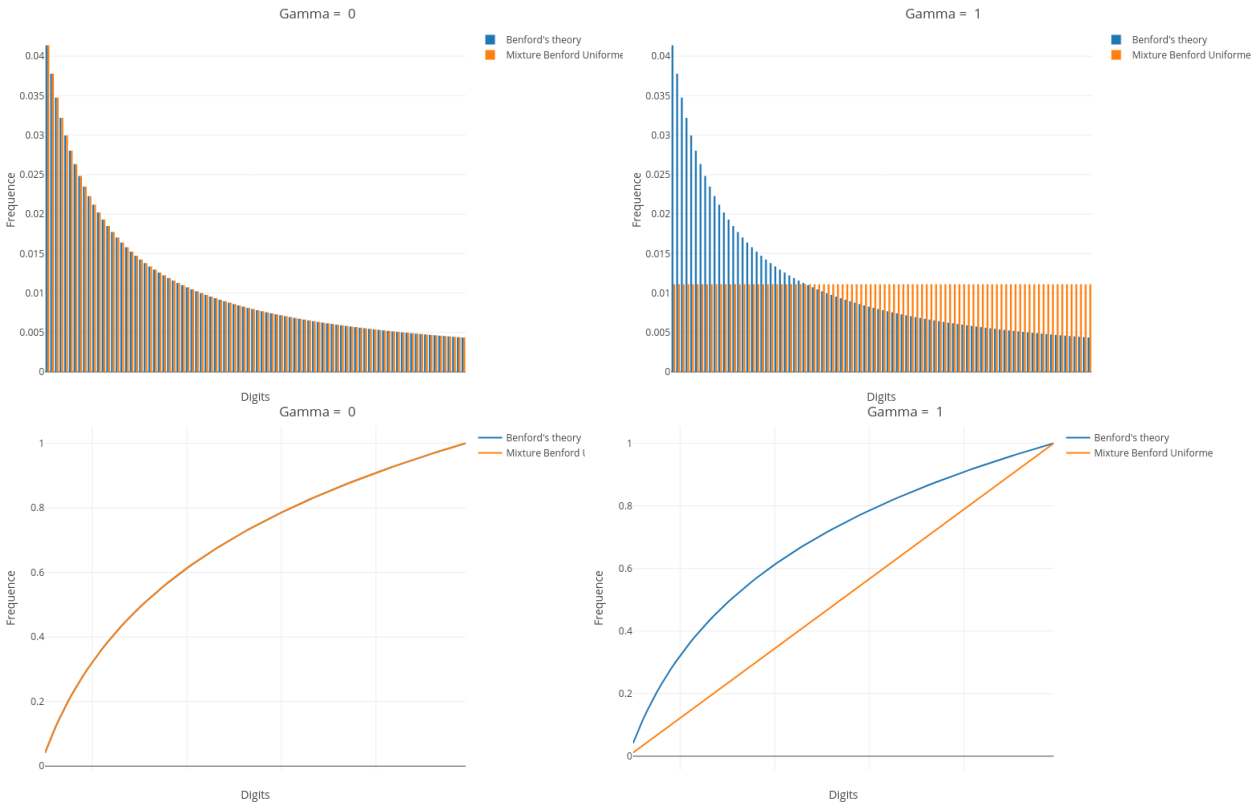

FIGURE 3.3.19 – Mixture Benford Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma U_{[10,99]}$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille des mixtures: Benford Stigler Uniforme`

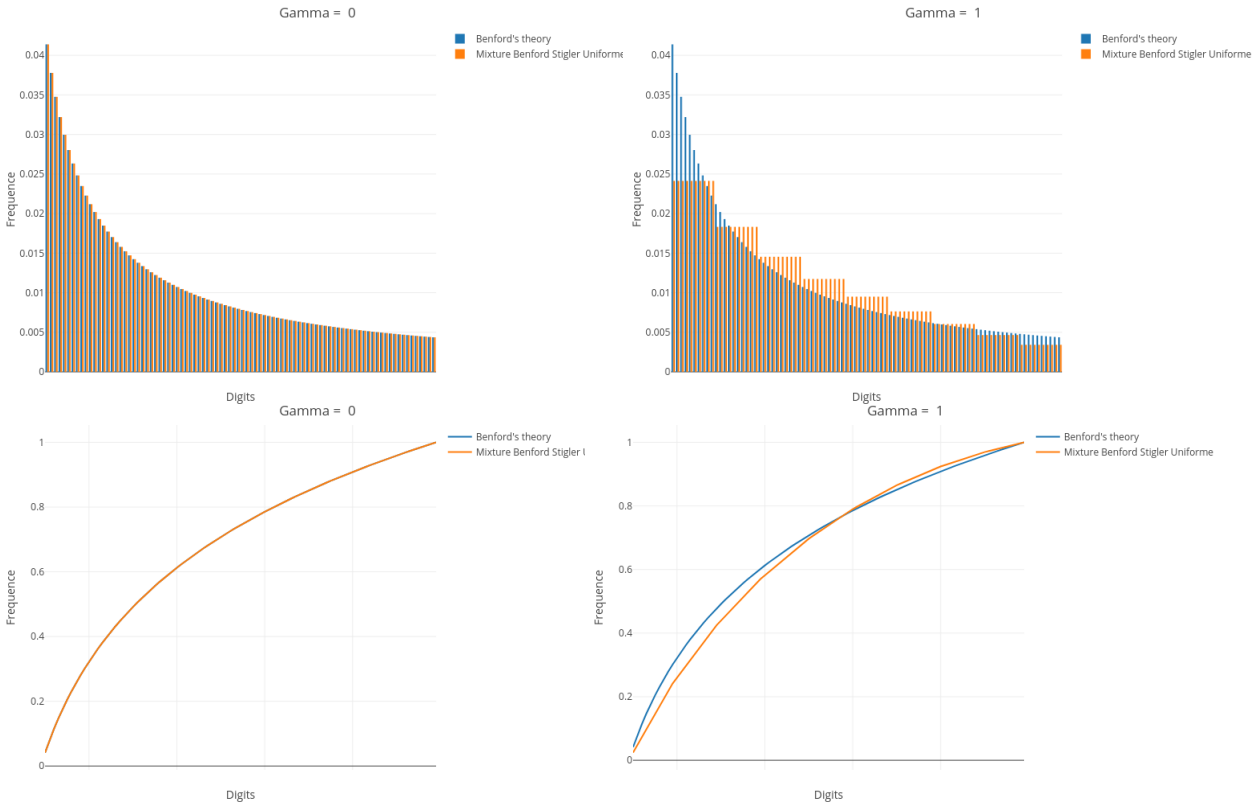

FIGURE 3.3.20 – Mixture Benford Stigler Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, \mathcal{S}_1 \otimes U_{[10,99]}$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille Indépendance: Benford Rodriguez`

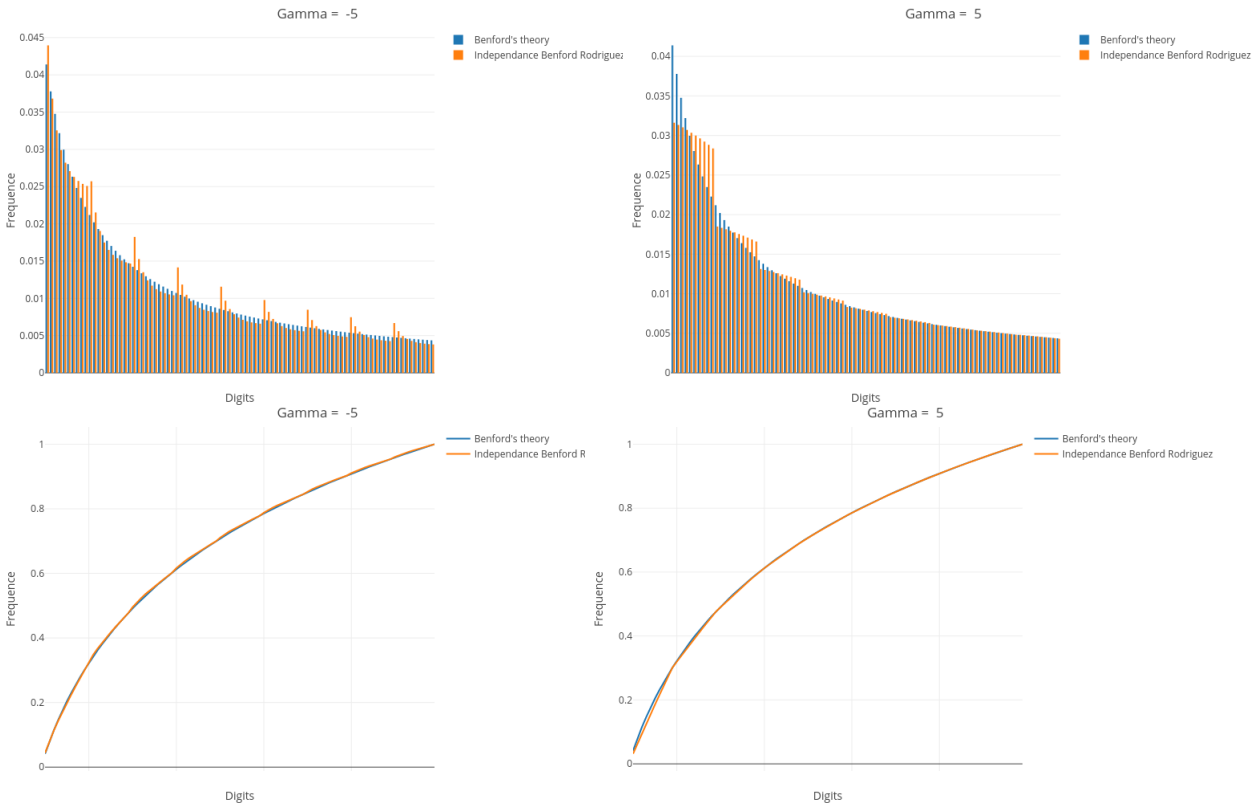

FIGURE 3.3.21 – Famille des indépendances : Indépendance Benford Rodriguez $(\mathscr{B}_1 \perp \mathcal{R}_2(\gamma))$: :
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille Indépendance: Rodriguez Rodriguez`

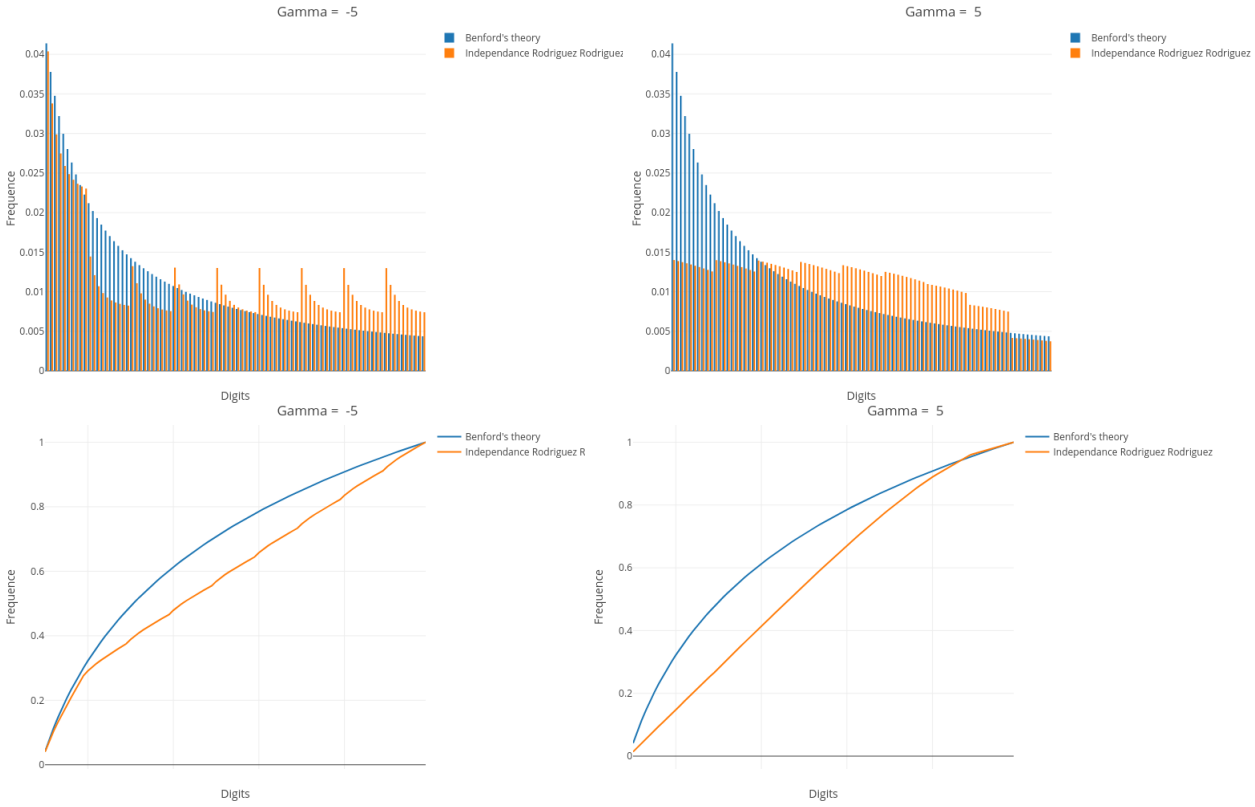

FIGURE 3.3.22 – Famille des indépendances : Indépendance Rodriguez Rodriguez $(\mathcal{R}_1(\gamma) \perp \mathcal{R}_2(\gamma))$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille de Copules: Benford Hill`

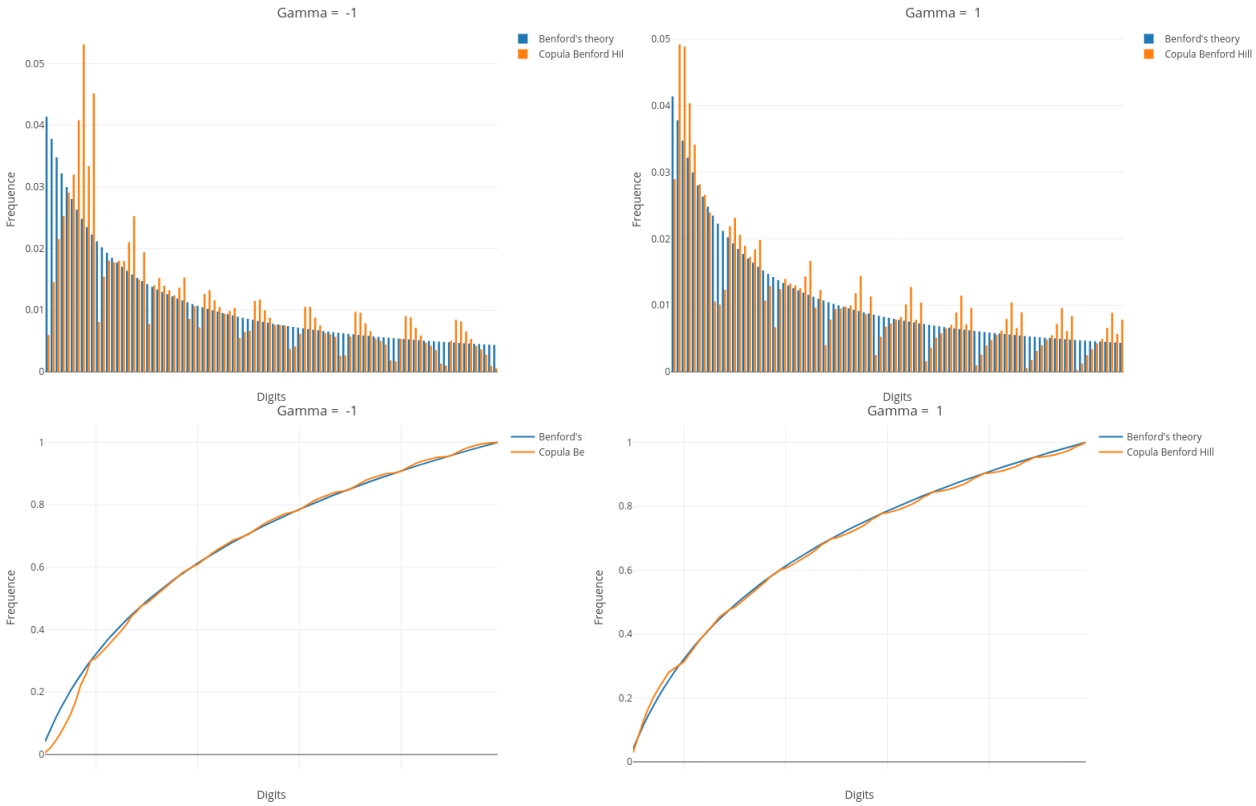

FIGURE 3.3.23 – Famille des copules : Copule Benford Hill $C(\gamma, \mathscr{B}_1, \mathcal{H}_2)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille de Copules: Benford Benford`

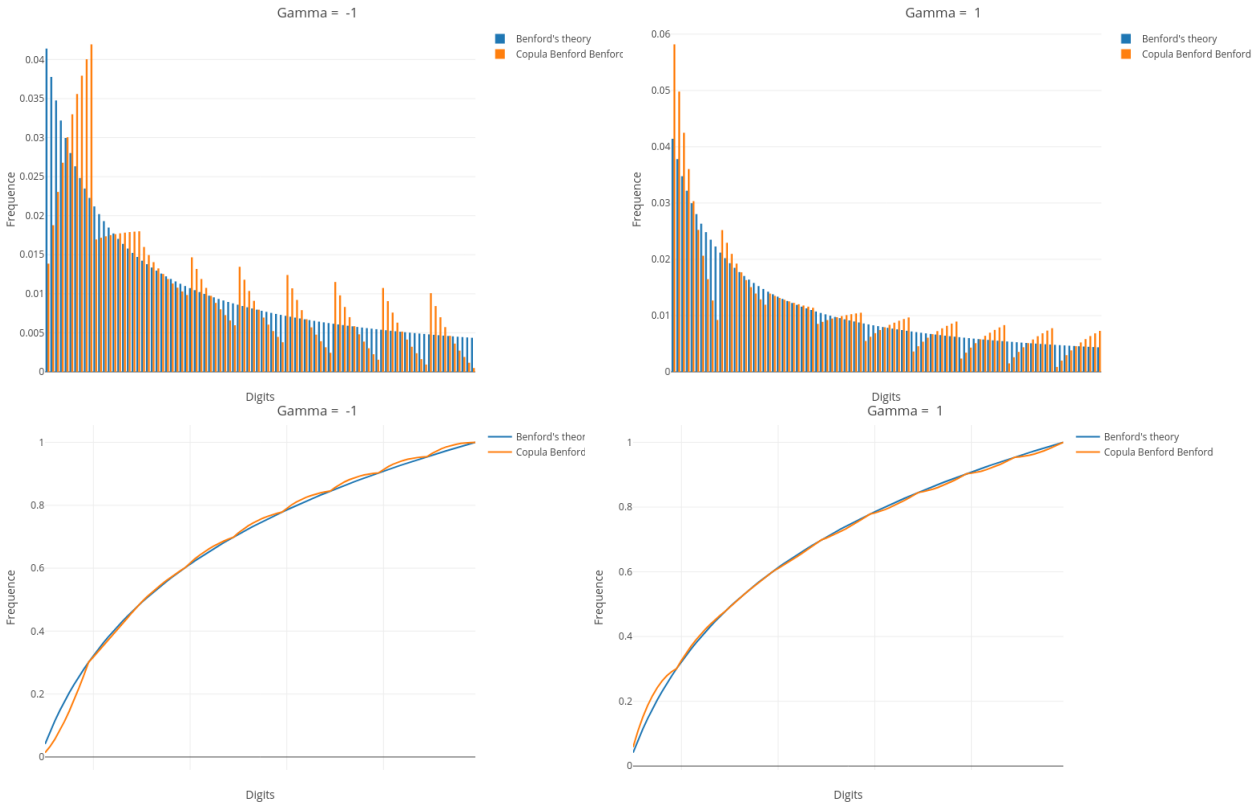

FIGURE 3.3.24 – Famille des copules : Copule Benford Benford $C(\gamma, \mathscr{B}_1, \mathscr{B}_2)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille des conditionnelles: Rodriguez sachant Benford`

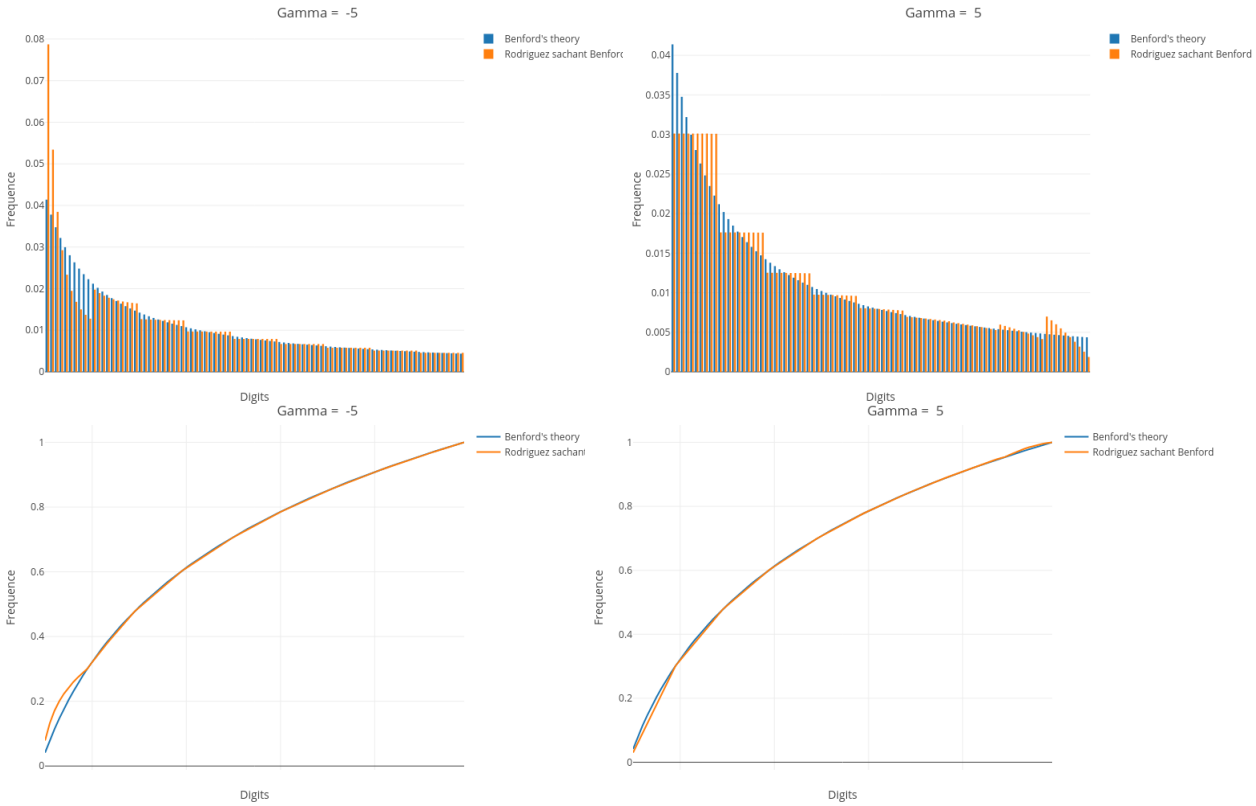

FIGURE 3.3.25 – Famille des conditionnelles : Rodriguez sachant Benford $\left(\mathscr{B}_1, \mathcal{R}_{(2|1)}(\gamma)\right)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille des conditionnelles: Benford sachant Rodriguez`

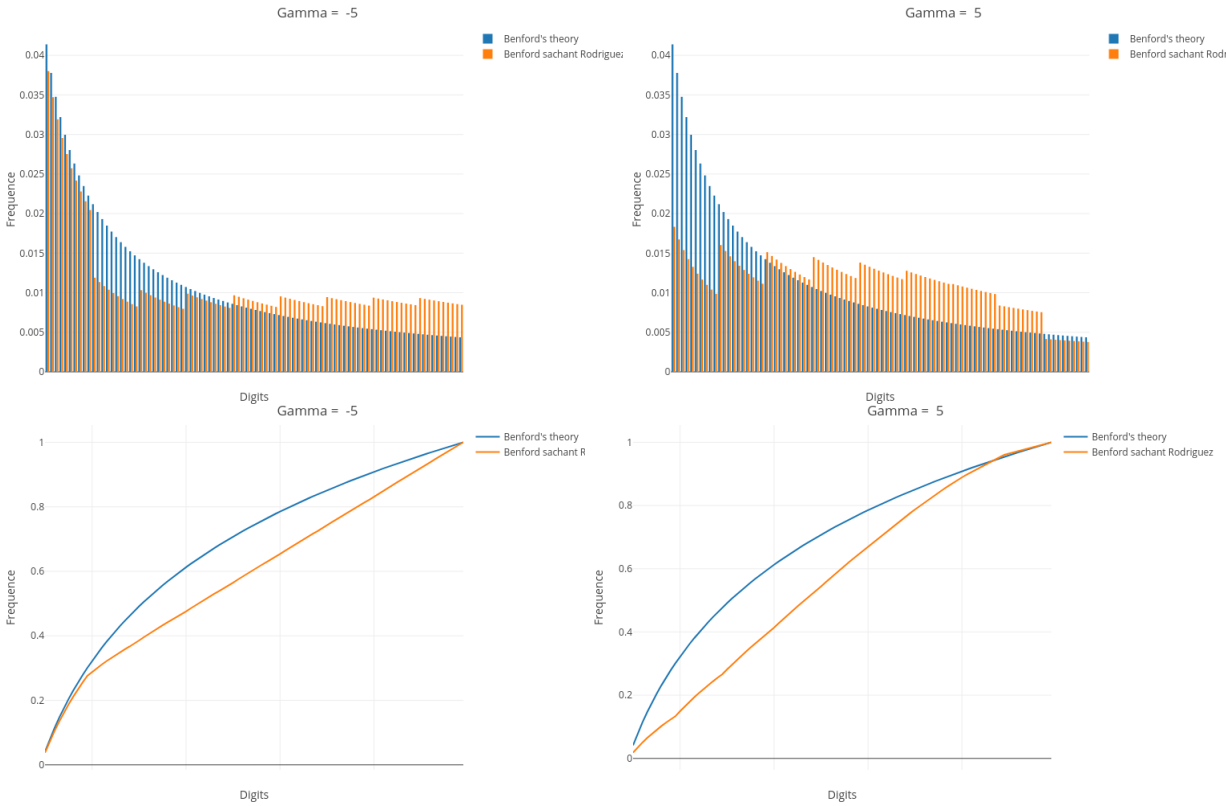

FIGURE 3.3.26 – Famille des conditionnelles : Benford sachant Rodriguez $\Big(\mathcal{R}_1(\gamma), \mathscr{B}_{(2|1)}(\gamma)\Big)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Alternative de Wong: Modification additive de Benford`

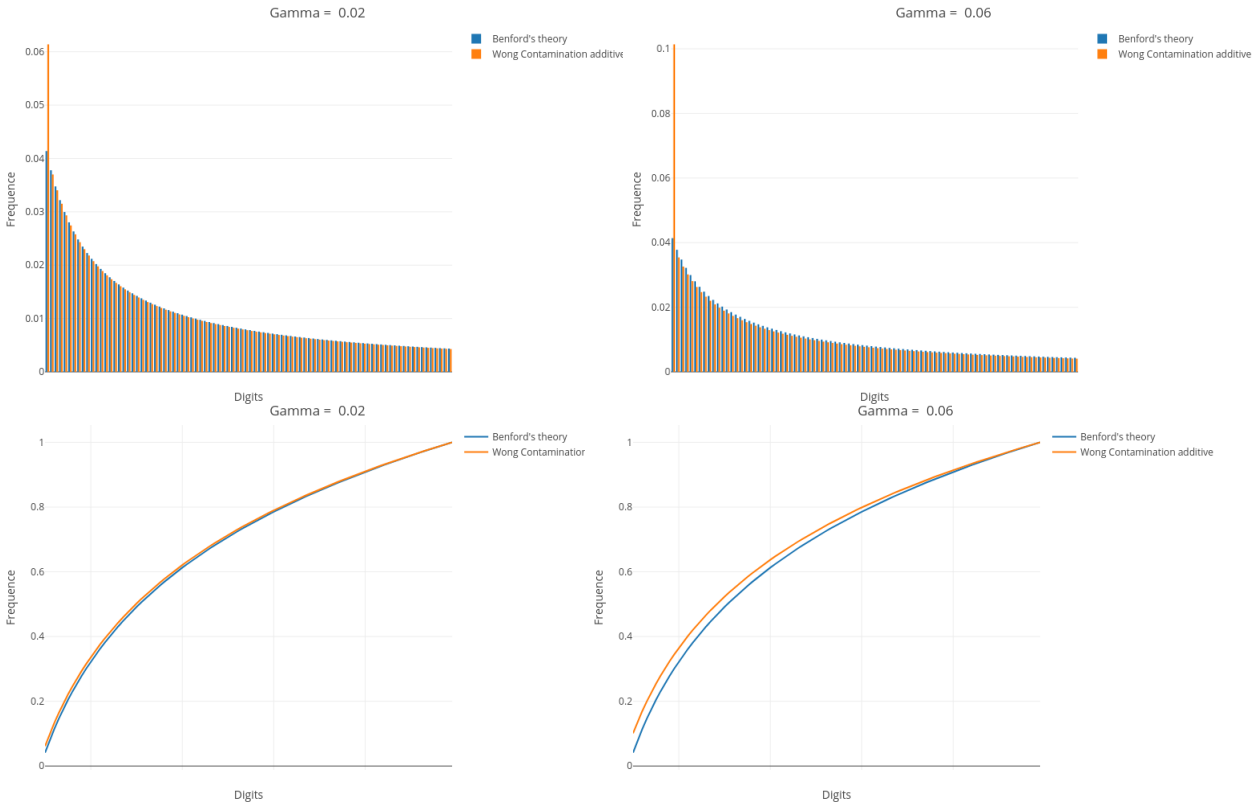

FIGURE 3.3.27 – Famille de Wong additive :
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Alternative de Wong: Modification multiplicative de Benford

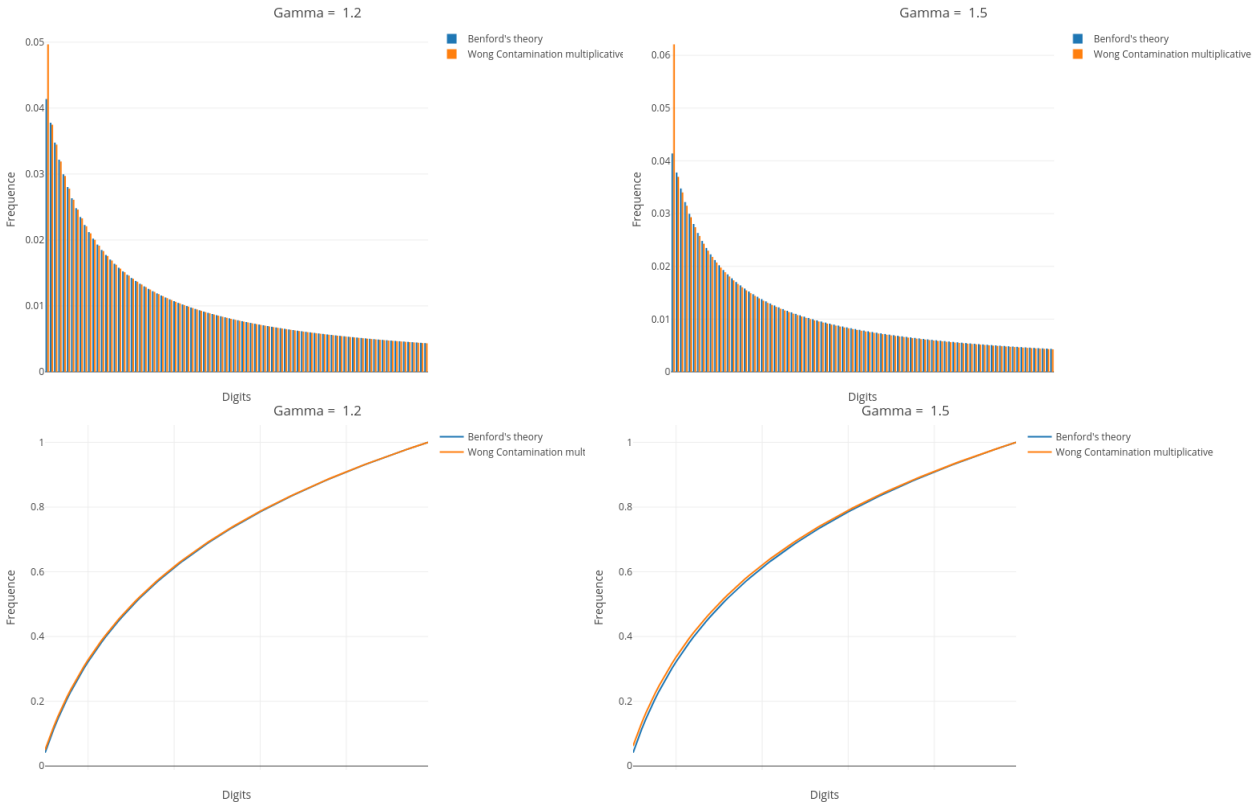

FIGURE 3.3.28 – Famille de Wong multiplicative :
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

3.3.8 Extension de la fonction connectrice ou de lien (Connector function) au cas bivarié

Nous avons défini dans la section 3.3.2, la notion de fonction connectrice dans le cas univarié. Dans cette section, nous allons étendre cette définition au cas bivarié.

Soit (X_1, X_2) un couple de variables aléatoires discrètes, $f_X(x_1, x_2) = f_X(x_1) f_X(x_2|x_1)$ la densité de probabilité (ou fonction de masse) associée et $F_X(x_1, x_2)$ la fonction de distribution cumulative (*cdf*) ou de répartition. Considérons $f_0(x_1, x_2) = f_0(x_1) f_0(x_2|x_1)$ une densité de probabilité de loi bivariée fixée et $F_0(x_1, x_2)$ la fonction de répartition (*cdf*) associée. On veut tester l'hypothèse nulle $H_0 : f_X(x_1, x_2) \sim f_0(x_1, x_2)$. La fonction connectrice de Neyman (1937) donnée à l'équation 3.3.7 devient :

$$c(x_1, x_2; F_X, F_0) = \exp\left\{\sum_{m=1}^{M} \theta_m h_m(x_1) - K(\theta) + \sum_{m'=1}^{M'} \lambda_{m'}(x_1) h_{m'}(x_2|x_1) - K^*(\lambda(x_1))\right\},$$

où $K(\theta)$, $K^*(\lambda(x_1))$ sont des constantes de normalisation , $\{h_1(\cdot), h_2(\cdot), ...\}$ une suite de fonctions orthonormées relativement à $f_0(x_1)$ et $\{h_1(x_2|x_1), h_2(x_2|x_1), ...\}$ une suite de fonctions orthonormées relativement à $f_0(x_2|x_1)$. Dans le cas de la loi bivariée de Newcomb-Benford, les fonctions $\{h_1(x_2|x_1), h_2(x_2|x_1), ...\}$ sont données dans la section 3.3.9.

Sachant qu'un bon travail statistique préliminaire devrait faire en sorte que $f_0(\cdot)$ soit proche de $f_X(\cdot)$ on aura que θ_m et $\lambda_{m'}$ qui seront proche de 0, ce qui implique que $K(\theta) \simeq K^*(\lambda(x_1)) \simeq 0$. Nous obtenons donc

$$c_{M,M'}(x_1, x_2) \simeq 1 + \sum_{m=1}^{M} \theta_m h_m(x_1) + \sum_{m'=1}^{M'} \lambda_{m'}(x_1) h_{m'}(x_2|x_1).$$

Par conséquence, l'hypothèse nulle H_0 implique que $\theta_1 = ... = \theta_M = 0$ et $\lambda_1(x_1) = ... = \lambda_{M'}(x_1) = 0$

La dérivée de la log-vraisemblance de $c_{M,M'}(x_1, x_2)$ par rapport à θ donne le score $U_m = \sum_{i=1}^{n} h_m(X_{1_i})$ et, $U^*_{m'} = \sum_{j=1}^{n} h_{m'}\left(X_{2_j} \mid X_{1_j}\right)$ quand on le dérive par rapport aux $\lambda_{m'}$

En utilisant ce qui a été fait dans Ducharme *et al.* (2020) , la statistique de test lisse est donnée par

$$S_{Cond_{M,M'}} = n^{-1} \sum_{i=1}^{M} U_m^2 + n^{-1} \sum_{i=1}^{M'} U_{m'}^{*^2} \tag{3.3.16}$$

et suit asymptotiquement la loi $\chi^2_{M+M'}$ sous l'hypothèse nulle $H_0 : f_X(\cdot, \cdot) = f_0(\cdot, \cdot)$. Pour des raisons de simplicité, dans la suite nous posons $M = M'$.

3.3.9 Smooth test conditionnel appliqué à loi de Newcomb-Benford Bivariée

Comme vu dans la section 3.3.6, le test $\mathcal{S}_{NDD}$ ne donne pas tout le temps un bon compromis. Nous optons alors pour une mise en place d'un nouveau test lisse basé également sur la théorie développée dans la section 3.3.8.

Nous ne transformons plus la loi bivariée $(\mathscr{D}_1, \mathscr{D}_2)$ en univariée $(\mathscr{D}_{12})$. Nous nous appuyons sur l'héritage de Bayes et nous considérons

$$(\mathscr{D}_1, \mathscr{D}_2) = (\mathscr{D}_1, \mathscr{D}_2|\mathscr{D}_1)$$

La fonction de masse du couple $(\mathscr{D}_2|\mathscr{D}_1)$ est le suivant ;

$$\begin{aligned} f_0(\mathscr{D}_2|\mathscr{D}_1) &= \mathbb{P}\left[\mathscr{D}_2 = d_2|\mathscr{D}_1 = d_1\right] \\ &= \frac{\mathbb{P}\left[(\mathscr{D}_2 = d_2) \cap (\mathscr{D}_1 = d_1)\right]}{\mathbb{P}\left[(\mathscr{D}_1 = d_1)\right]} \\ &= \frac{\log_{10}\left(1 + \dfrac{1}{10d_1 + d_2}\right)}{\log_{10}\left(1 + \dfrac{1}{d_1}\right)}. \end{aligned}$$

Cette distribution conditionnelle est donnée dans la Table 3.4. Et comme nouveau test lisse, exploitant cette loi conditionnelle, nous l'appelons « Smooth test conditionnel », notée $\mathcal{S}_{Cond}$.

En exploitant le théorème de Boulerice et Ducharme (1997) et en utilisant les $h_m(d_1)$, $m \in \{1, \ldots, 5\}$, $d_1 \in \{1, \ldots, 9\}$ calculés dans Ducharme *et al.* (2020), nous avons calculé les polynômes $h_m(d_2|d_1)$, $d_1 = 1, .., 9$, $d_2 \in \{0, \ldots, 9\}$ à l'aide du logiciel MATHEMATICA. On obtient les expressions suivantes pour les polynômes :

— Cas $d_1 = 1$

$$\begin{aligned} h_1(d_2|d_1 = 1) &= -1.3753 + 0.3497\, d_2 \\ h_2(d_2|d_1 = 1) &= 1.3877 - 1.1850\, d_2 + 0.1384\, d_2^2 \\ h_3(d_2|d_1 = 1) &= -1.1689 + 2.4352\, d_2 - 0.7506\, d_2^2 + 0.0573\, d_2^3 \\ h_4(d_2|d_1 = 1) &= 0.8457 - 4.0558\, d_2 + 2.3875\, d_2^2 - 0.4382\, d_2^3 \\ &\quad + 0.0248\, d_2^4 \\ h_5(d_2|d_1 = 1) &= -0.5272 + 6.2839\, d_2 - 6.0619\, d_2^2 + 1.9544\, d_2^3 \\ &\quad - 0.2536\, d_2^4 + 0.012\, d_2^5 \end{aligned}$$

— Cas $d_1 = 2$

$$\begin{aligned} h_1(d_2|d_1 = 2) &= -1.4528 + 0.3486\, d_2 \\ h_2(d_2|d_1 = 2) &= 1.4943 - 1.2054\, d_2 + 0.1378\, d_2^2 \\ h_3(d_2|d_1 = 2) &= -1.2757 + 2.5082\, d_2 - 0.7565\, d_2^2 + 0.0570\, d_2^3 \\ h_4(d_2|d_1 = 2) &= 0.9335 - 4.2027\, d_2 + 2.4239\, d_2^2 - 0.4394\, d_2^3 \\ &\quad + 0.0247\, d_2^4 \\ h_5(d_2|d_1 = 2) &= -0.5881 + 6.5106\, d_2 - 6.1724\, d_2^2 + 1.9684\, d_2^3 \\ &\quad - 0.2535\, d_2^4 + 0.0114\, d_2^5 \end{aligned}$$

— Cas $d_1 = 3$

$$
\begin{aligned}
h_1(d_2|d_1=3) &= -1.4853 + 0.3484\,d_2 \\
h_2(d_2|d_1=3) &= 1.5392 - 1.2145\,d_2 + 0.1377\,d_2^2 \\
h_3(d_2|d_1=3) &= -1.3208 + 2.5401\,d_2 - 0.7595\,d_2^2 + 0.0567\,d_2^3 \\
h_4(d_2|d_1=3) &= 0.9707 - 4.2673\,d_2 + 2.4411\,d_2^2 - 0.4403\,d_2^3 \\
&\quad + 0.0246\,d_2^4 \\
h_5(d_2|d_1=3) &= -0.6140 + 6.6115\,d_2 - 6.2247\,d_2^2 + 1.9761\,d_2^3 \\
&\quad - 0.2536\,d_2^4 + 0.0113\,d_2^5
\end{aligned}
$$

— Cas $d_1 = 4$

$$
\begin{aligned}
h_1(d_2|d_1=4) &= -1.5034 + 0.3483\,d_2 \\
h_2(d_2|d_1=4) &= 1.5640 - 1.2197\,d_2 + 0.1377\,d_2 \\
h_3(d_2|d_1=4) &= -1.3458 + 2.5582\,d_2 - 0.76128\,d_2^2 + 0.0569\,d_2^3 \\
h_4(d_2|d_1=4) &= 0.9915 - 4.3037\,d_2 + 2.4511\,d_2^2 - 0.4409\,d_2^3 \\
&\quad + 0.0246\,d_2^4 \\
h_5(d_2|d_1=4) &= -0.6285 + 6.6689\,d_2 - 6.2551\,d_2^2 + 1.9809\,d_2^3 \\
&\quad - 0.2538\,d_2^4 + 0.0113\,d_2^5
\end{aligned}
$$

— Cas $d_1 = 5$

$$
\begin{aligned}
h_1(d_2|d_1=5) &= -1.5148 + 0.3482\,d_2 \\
h_2(d_2|d_1=5) &= 1.5799 - 1.2230\,d_2 + 0.1377\,d_2^2 \\
h_3(d_2|d_1=5) &= -1.3618 + 2.5697\,d_2 - 0.7624\,d_2^2 + 0.0569\,d_2^3 \\
h_4(d_2|d_1=5) &= 1.005 - 4.3271\,d_2 + 2.4576\,d_2^2 - 0.4414\,d_2^3 \\
&\quad + 0.0246\,d_2^4 \\
h_5(d_2|d_1=5) &= -0.6376 + 6.7057\,d_2 - 6.2749\,d_2^2 + 1.9842\,d_2^3 \\
&\quad - 0.2539\,d_2^3 + 0.0113\,d_2^5
\end{aligned}
$$

— Cas $d_1 = 6$

$$
\begin{aligned}
h_1(d_2|d_1=6) &= -1.5228 + 0.3482\,d_2 \\
h_2(d_2|d_1=6) &= 1.5908 - 1.2254\,d_2 + 0.1377\,d_2^2 \\
h_3(d_2|d_1=6) &= -1.3728 + 2.5778\,d_2 - 0.7632\,d_2^2 + 0.0569\,d_2^3 \\
h_4(d_2|d_1=6) &= 1.0138 - 4.3434\,d_2 + 2.4621\,d_2^2 - 0.4416\,d_2^3 \\
&\quad + 0.02465\,d_2^4 \\
h_5(d_2|d_1=6) &= -0.6440 + 6.7316\,d_2 - 6.2889\,d_2^2 + 1.9865\,d_2^3 \\
&\quad - 0.2541\,d_2^4 + 0.0113\,d_2^5
\end{aligned}
$$

— Cas $d_1 = 7$

$$
\begin{aligned}
h_1(d_2|d_1=7) &= -1.5286 + 0.3482\,d_2 \\
h_2(d_2|d_1=7) &= 1.5989 - 1.2271\,d_2 + 0.1376\,d_2^2 \\
h_3(d_2|d_1=7) &= -1.3810 + 2.5837\,d_2 - 0.7639\,d_2^2 + 0.0569\,d_2^3 \\
h_4(d_2|d_1=7) &= 1.0205 - 4.3554\,d_2 + 2.4655\,d_2^2 - 0.4419\,d_2^3 \\
&\quad + 0.0246\,d_2 \\
h_5(d_2|d_1=7) &= -0.6487 + 6.7503\,d_2 - 6.2990\,d_2^2 + 1.988\,d_2^3 \\
&\quad - 0.2541\,d_2 + 0.0113\,d_2
\end{aligned}
$$

— Cas $d_1 = 8$

$$
\begin{aligned}
h_1(d_2|d_1=8) &= -1.5330 + 0.3482\,d_2 \\
h_2(d_2|d_1=8) &= 1.6050 - 1.2284\,d_2 + 0.1376\,d_2^2 \\
h_3(d_2|d_1=8) &= -1.3871 + 2.5882\,d_2 - 0.7643\,d_2^2 + 0.0569\,d_2^3 \\
h_4(d_2|d_1=8) &= 1.0256 - 4.3646\,d_2 + 2.4681\,d_2^2 - 0.4420\,d_2^3 \\
&\quad + 0.0246\,d_2^4 \\
h_5(d_2|d_1=8) &= -0.6523 + 6.7651\,d_2 - 6.3072\,d_2^2 + 1.9896\,d_2^3 \\
&\quad - 0.2542\,d_2^4 + 0.0113\,d_2^5
\end{aligned}
$$

— Cas $d_1 = 9$

$$
\begin{aligned}
h_1(d_2|d_1=9) =& -1.5366 + 0.3482\, d_2 \\
h_2(d_2|d_1=9) =& 1.6099 - 1.2295\, d_2 + 0.1376\, d_2^2 \\
h_3(d_2|d_1=9) =& -1.3920\, d_2 + 2.5919\, d_2 - 0.7647\, d_2^2 + 0.0569\, d_2^3 \\
h_4(d_2|d_1=9) =& 1.0297\, d_2 - 4.3719\, d_2 + 2.4702\, d_2^2 - 0.4422\, d_2^3 \\
& + 0.0246\, d_2^4 \\
h_5(d_2|d_1=9) =& -0.6551\, d_2 + 6.7767\, d_2 - 6.3135\, d_2^2 + 1.9906\, d_2^3 \\
& - 0.2542\, d_2^4 + 0.0113\, d_2^5
\end{aligned}
$$

Soit X une variable aléatoire continue positive et $\mathscr{D}_1, \mathscr{D}_2$, respectivement le premier, défini sur $\{1, \ldots, 9\}$, et le second, défini sur $\{0, \ldots, 9\}$, chiffre significatif.

En adaptant l'équation 3.3.9, on obtient

$$
\mathcal{S}_{temp_{M,d_1}} = n^{-1} \sum_{i=1}^{M} \left(\sum_{j=1}^{n} h_i \left(\mathscr{D}_2 \left(X_j \right) \mid \mathscr{D}_1 \left(X_j \right) = d_1 \right) \right)^2, \; d_1 \in \{1, \ldots, 9\}
$$

$$
\mathcal{S}_{temp_M} = \sum_{d_1=1}^{9} \mathcal{S}_{temp_{M,d_1}}
$$

$\mathcal{S}_{temp_M}$ représente l'information provenant de $\mathscr{D}_2 \left(X_j \right) \mid \mathscr{D}_1 \left(X_j \right)$. Nous allons maintenant ajouter l'information provenant de $\mathscr{D}_1 \left(X_j \right)$. En utilisant les acquis de Ducharme *et al.* (2020) sur le test lisse pour le premier chiffre significatif, que nous notons $\mathcal{S}_{PCS}$, nous obtenons

$$
\begin{aligned}
\mathcal{S}_{Cond_M} &= \mathcal{S}_{temp_M} + \mathcal{S}_{PCS} \\
\mathcal{S}_{Cond_M} &= \left[\sum_{d_1=1}^{9} n^{-1} \sum_{i=1}^{M} \left(\sum_{j=1}^{n} h_i \left(\mathscr{D}_2 \left(X_j \right) \mid \mathscr{D}_1 \left(X_j \right) = d_1 \right) \right)^2 \right] + \left[n^{-1} \sum_{i=1}^{M} \left(\sum_{j=1}^{n} h_i \left(\mathscr{D}_1 \left(X_j \right) \right) \right)^2 \right]
\end{aligned}
\tag{3.3.17}
$$

Nous nous retrouvons confronté maintenant à la même question que précédemment dans la section 3.3.3 à savoir, le choix du meilleur M. Pour ce faire, nous avons utilisé la même approche pour déterminer un bon calibrage du c (cf section 3.3.5). De par nos simulations nous sommes arrivés à la conclusion que le c faisant compromis est le $c = +\infty$. Ce qui rejoint les travaux réalisés dans Ledwina (1994) dans le choix du M adéquat, qui a pour expression

$$
\hat{m} = \arg \max_{1 \leq m \leq 5} \left\{ \mathcal{S}_{Cond_m} - m \log(n) \right\}
$$

Ainsi, la nouvelle statistique de test servant de compromis $\mathcal{S}_{Cond,DD}$ est donné par $\mathcal{S}_{Cond_{\hat{m}}}$

Le seuil critique sous H_0 pour le $\mathcal{S}_{Cond,DD}$ est comme précédemment approché par simulation de Monte Carlo pour les niveaux de significations $\alpha = 10\%$, $\alpha = 5\%$ et $\alpha = 1\%$ via un million de réplications pour les sept (7) tailles d'échantillon considérées (cf Tables 3.13, 3.14 et 3.15).

	50	100	250	500	1000	3000	5000
$\mathcal{S}_{Cond,DD}$	21.364	20.588	19.293	18.520	17.890	17.125	16.945

TABLE 3.13 – Seuil critique de la loi de la statistique $\mathcal{S}_{Cond,DD}$ sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$ au seuil $\alpha = 0.1$ en utilisant 1000000 échantillons aléatoires

	50	100	250	500	1000	3000	5000
$\mathcal{S}_{Cond,DD}$	24.979	23.780	22.683	21.801	21.059	19.924	19.465

TABLE 3.14 – Seuil critique de la loi de la statistiques $\mathcal{S}_{Cond,DD}$ sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$ au seuil $\alpha = 0.05$ en utilisant 1000000 échantillons aléatoires

	50	100	250	500	1000	3000	5000
$\mathcal{S}_{Cond,DD}$	33.650	32.599	30.332	28.697	27.241	26.175	25.482

TABLE 3.15 – Seuil critique de la loi de la statistique $\mathcal{S}_{Cond,DD}$ sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$ au seuil $\alpha = 0.01$ en utilisant 1000000 échantillons aléatoires

3.3.10 Comparaison du test $\mathcal{S}_{NDD}$, $\mathcal{S}_{Cond,DD}$ et des tests classiques

Nous utiliserons dans ce contexte les alternatives utilisées dans la section 3.3.6.

Les tests $\mathcal{S}_{NDD}$, $\mathcal{S}_{Cond,DD}$, U^2, χ^2, W^2 sont effectués au niveau 5% et leur puissance est approximée par Monte Carlo (10000 réplications) pour chaque triplet $(\text{famille}, n, \gamma)$. Ces courbes de puissance apparaissent aux figures 3.3.29 à 3.3.38.

Nous pouvons faire les mêmes conclusions que dans la section 3.3.6 et nous attirons l'attention du lecteur sur les alternatives où χ^2 **est le meilleur test** (cf Figures 3.3.11 , 3.3.13 et 3.3.14) . Nous remarquons dans ces cas (cf Figures 3.3.31 , 3.3.33 et 3.3.34) que le test $\mathcal{S}_{Cond,DD}$ s'en sort aussi très bien. Nous arrivons à la conclusion :

— si W^2 ou U^2 sont plus puissants que le test du χ^2, $\mathcal{S}_{NDD}$ offre un bon compromis

— si χ^2 est plus puissant que les tests W^2 ou U^2, $\mathcal{S}_{Cond,DD}$ offre un bon compromis.

Ne disposant pas à priori d'information sur le test qui serait meilleur face à un jeu de données, nous ne pouvons recommander aucun test comme compromis à ce stade. Pour trouver un test compromis, dans la section suivante nous allons proposer une combinaison des tests $\mathcal{S}_{Cond,DD}$ et $\mathcal{S}_{NDD}$ basé sur les test du χ^2 et de W^2.

Famille des mixtures: Benford Uniforme

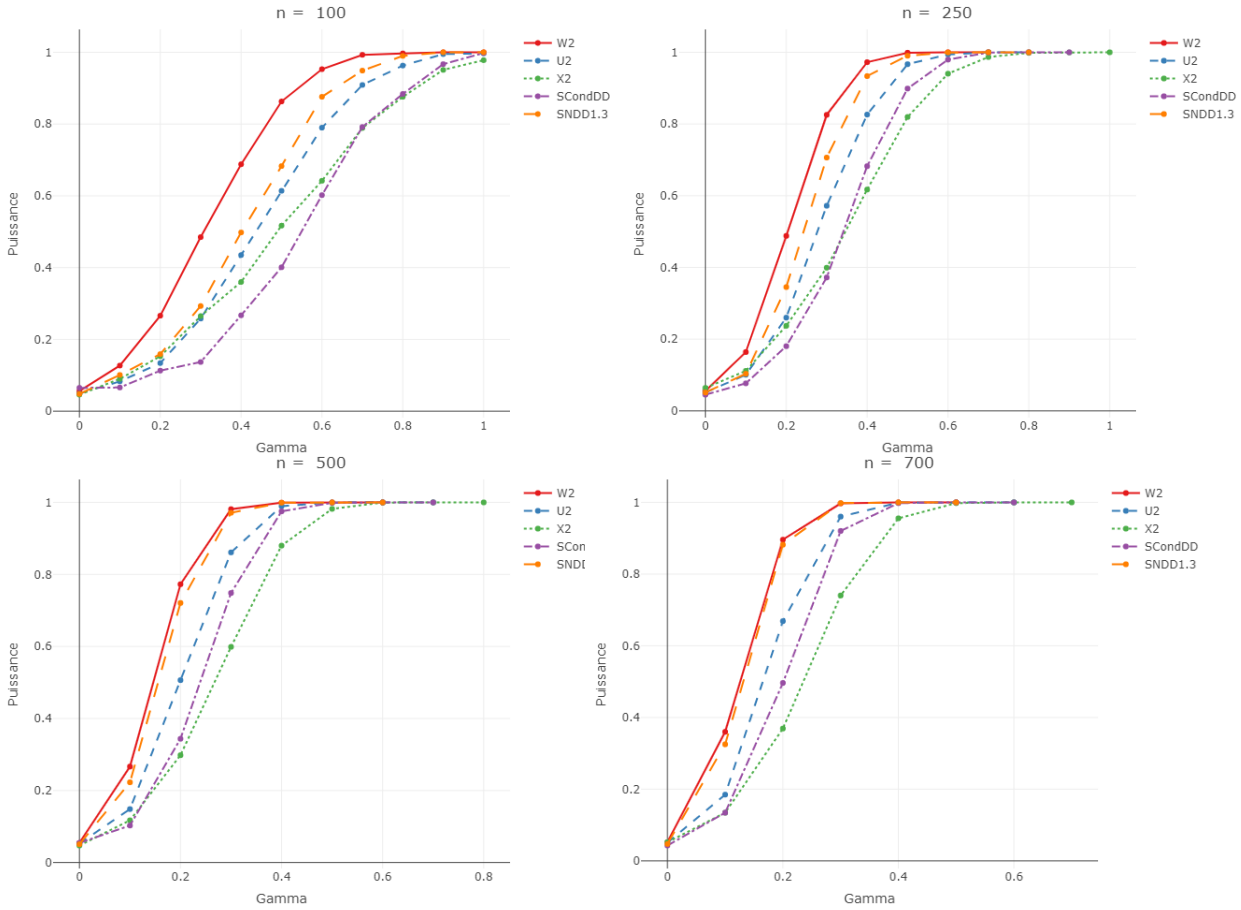

FIGURE 3.3.29 – Mixture Benford Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma U_{[10,99]}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Stigler Uniforme

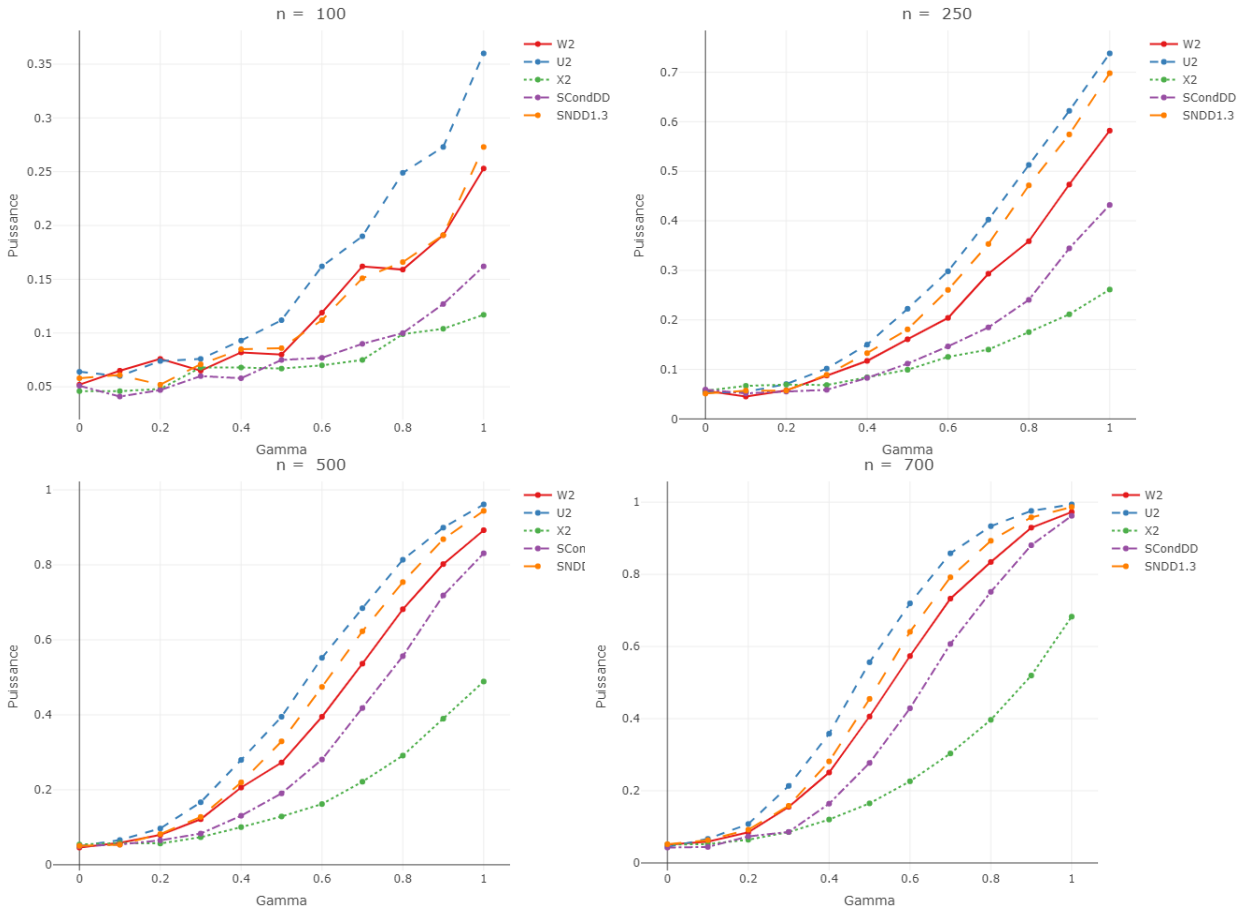

FIGURE 3.3.30 – Famille des mixtures : Mixture Benford Stigler Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\,\mathcal{S}_1 \otimes U_{[10,99]}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Benford Rodriguez

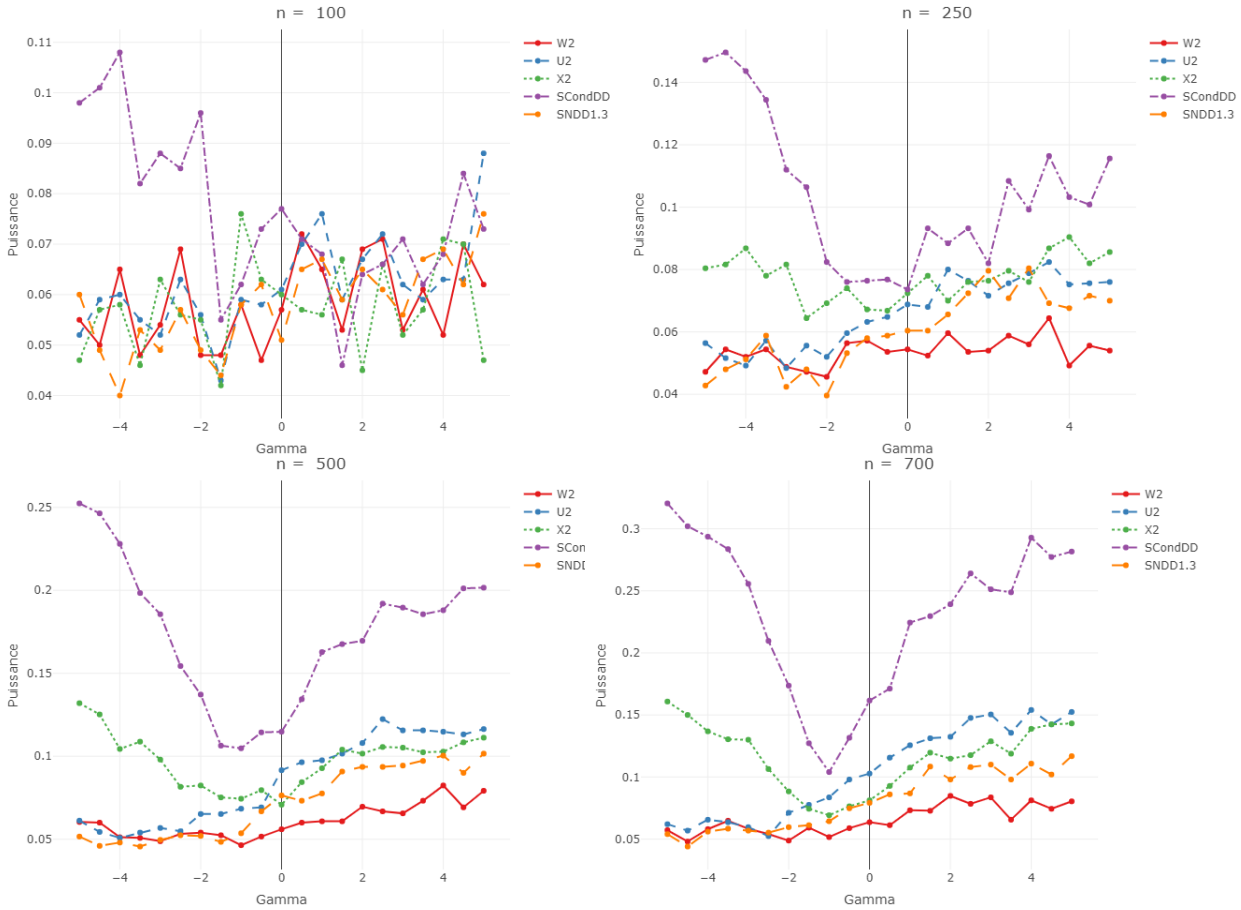

FIGURE 3.3.31 – Famille des indépendances : Indépendance Benford Rodriguez ($\mathscr{B}_1 \perp \mathcal{R}_2(\gamma)$) : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Rodriguez Rodriguez

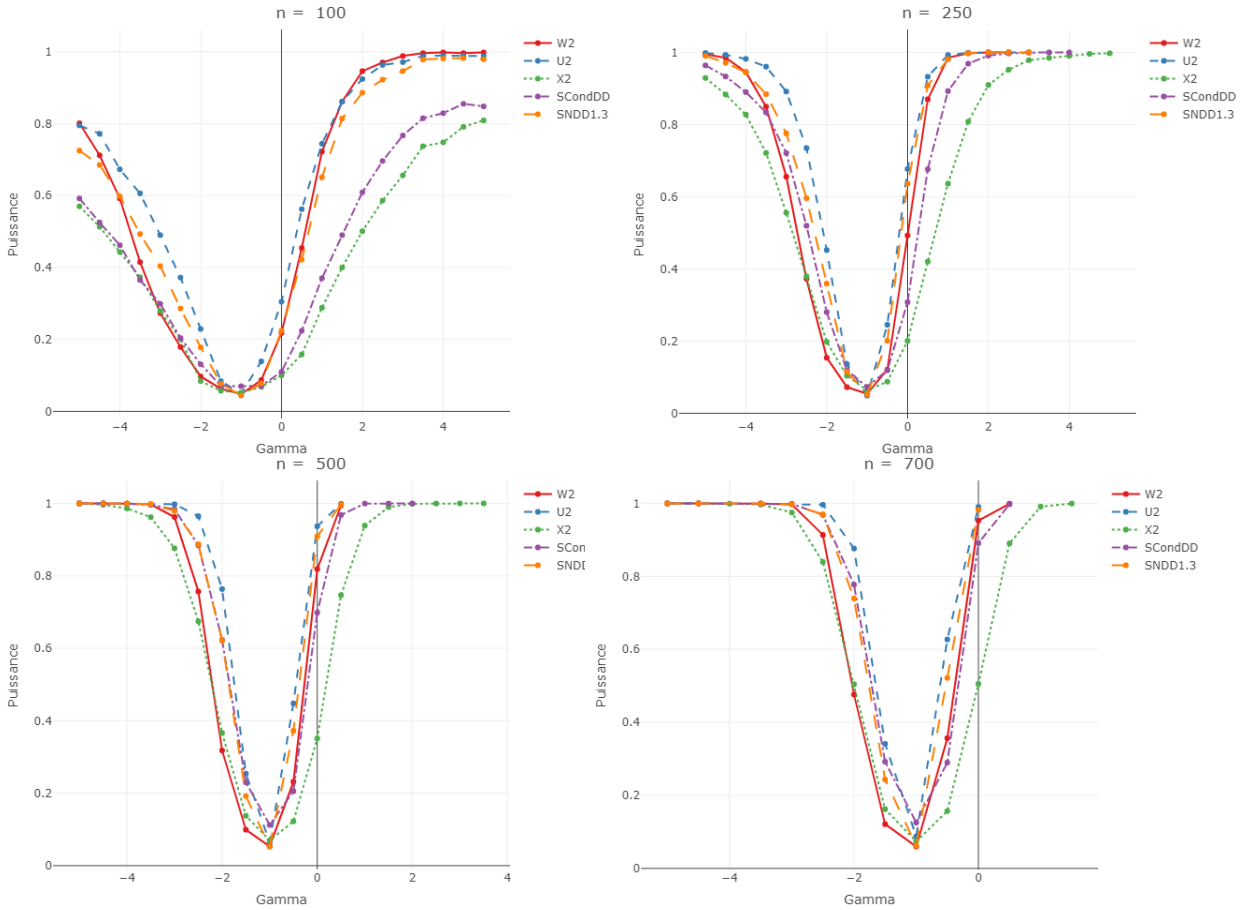

FIGURE 3.3.32 – Famille des indépendances : Indépendance Rodriguez Rodriguez ($\mathcal{R}_1(\gamma) \perp \mathcal{R}_2(\gamma)$) : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille de Copules: Benford Hill

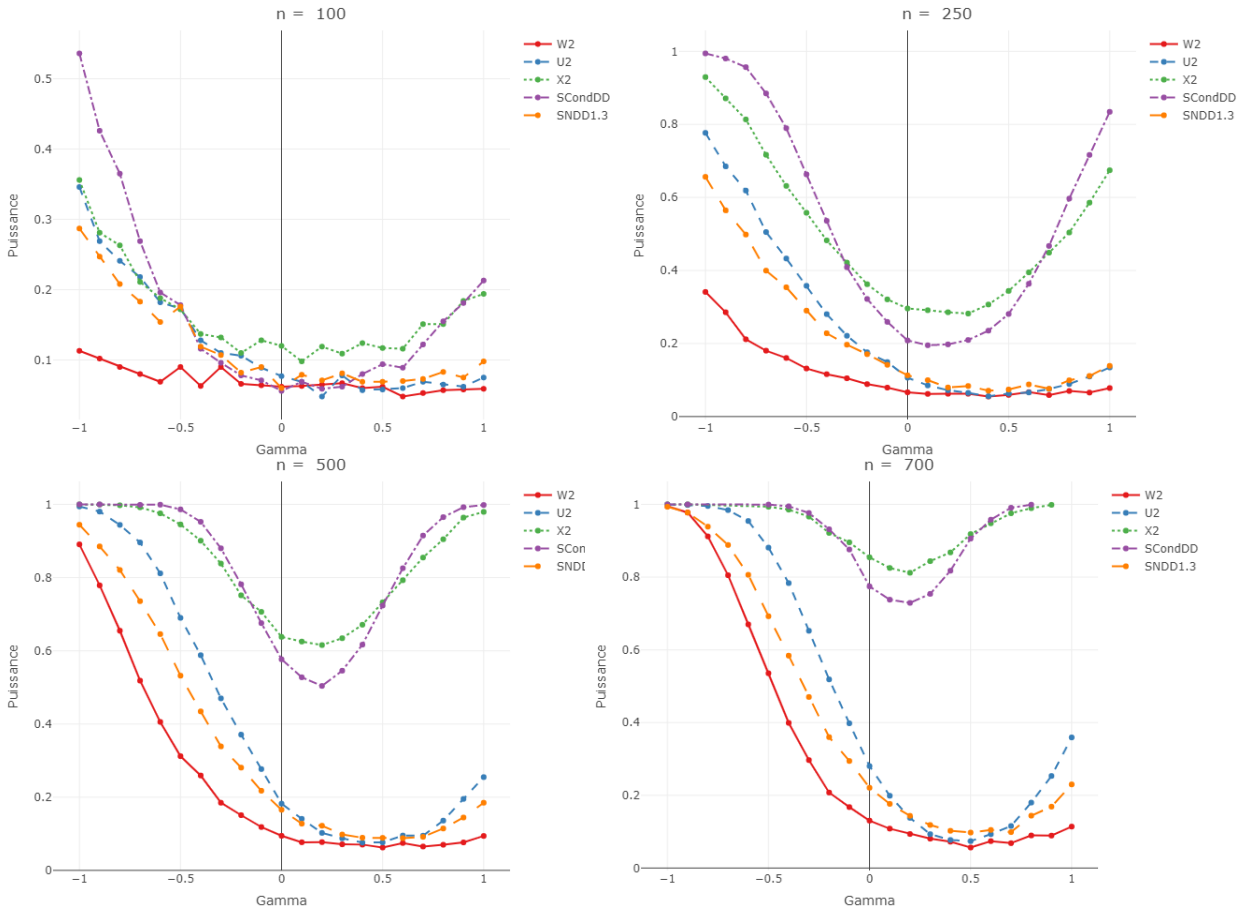

FIGURE 3.3.33 – Famille des copules : Copule Benford Hill $C(\gamma, \mathscr{B}_1, \mathcal{H}_2)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille de Copules: Benford Benford

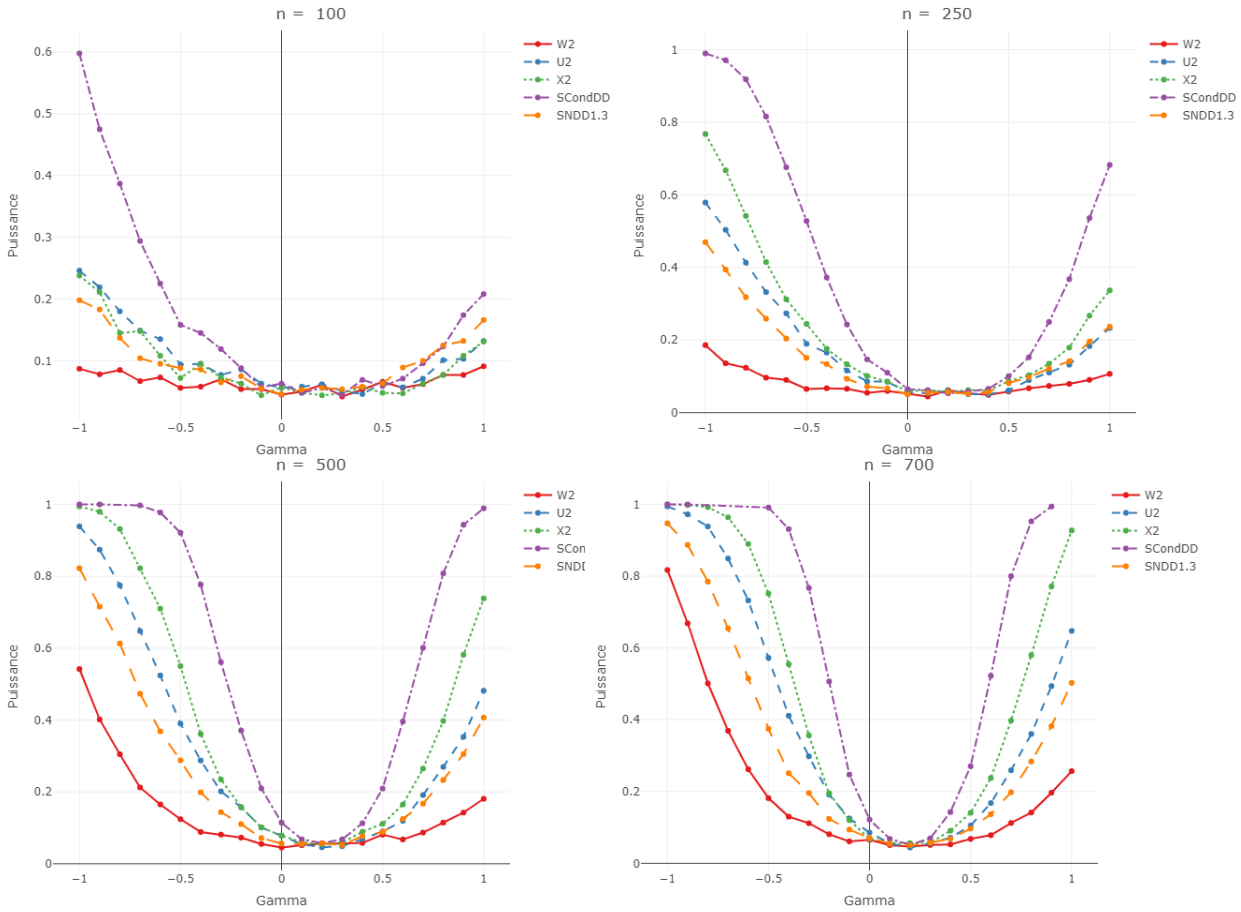

FIGURE 3.3.34 – Famille des copules : Copule Benford Benford $C(\gamma, \mathscr{B}_1, \mathscr{B}_2)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Rodriguez sachant Benford

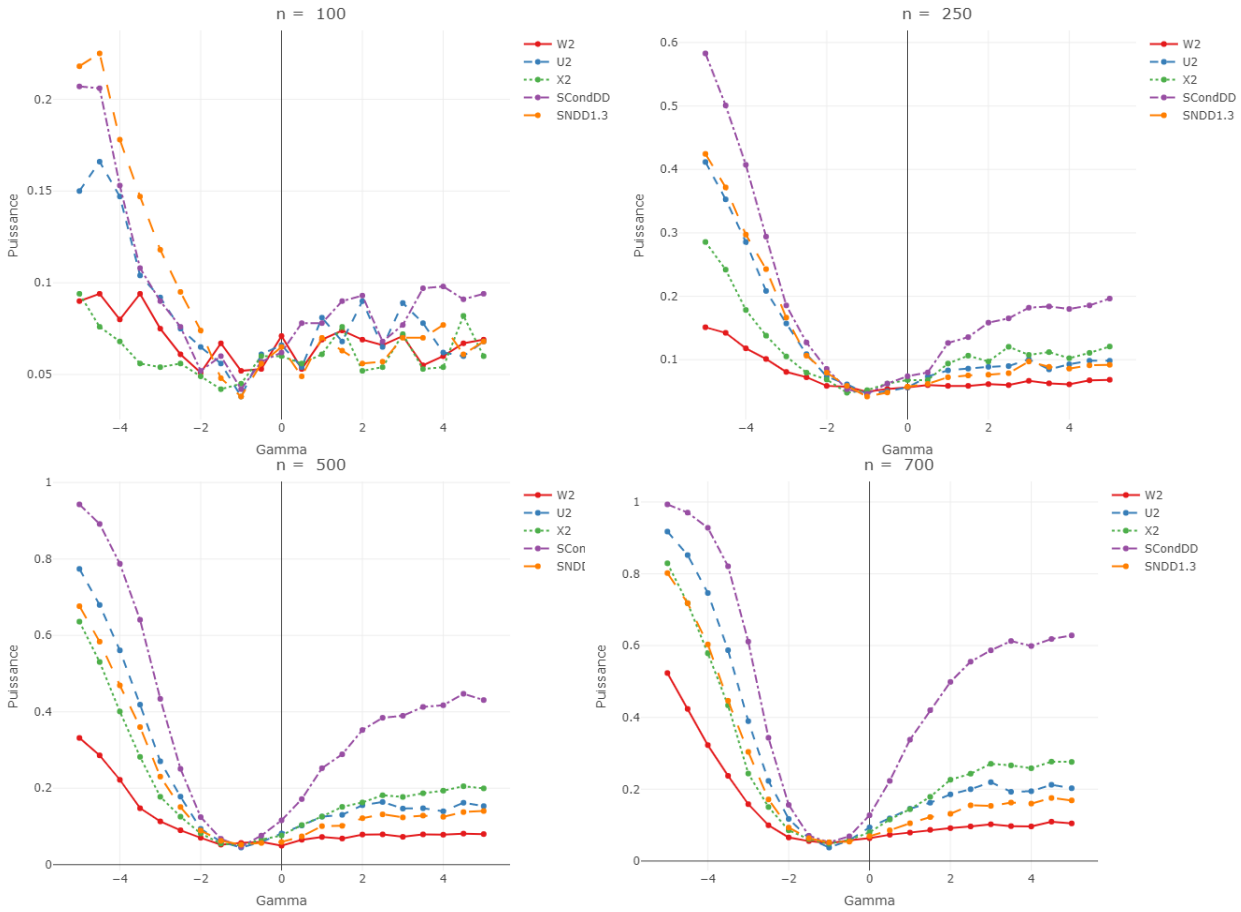

FIGURE 3.3.35 – Famille des conditionnelles : Rodriguez sachant Benford $\left(\mathscr{B}_1, \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Benford sachant Rodriguez

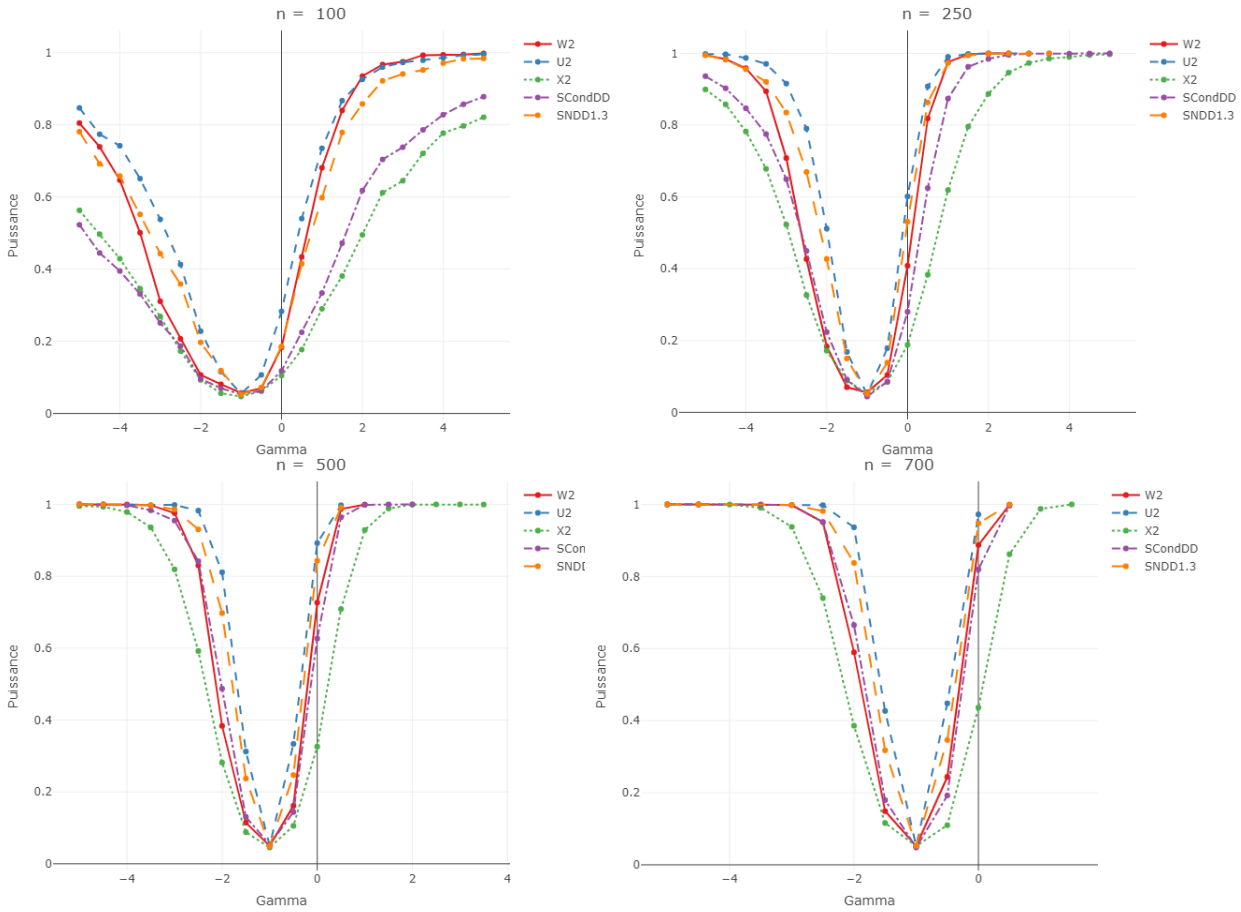

FIGURE 3.3.36 – Famille des conditionnelles : Benford sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathscr{B}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Alternative de Wong: Modification additive de Benford

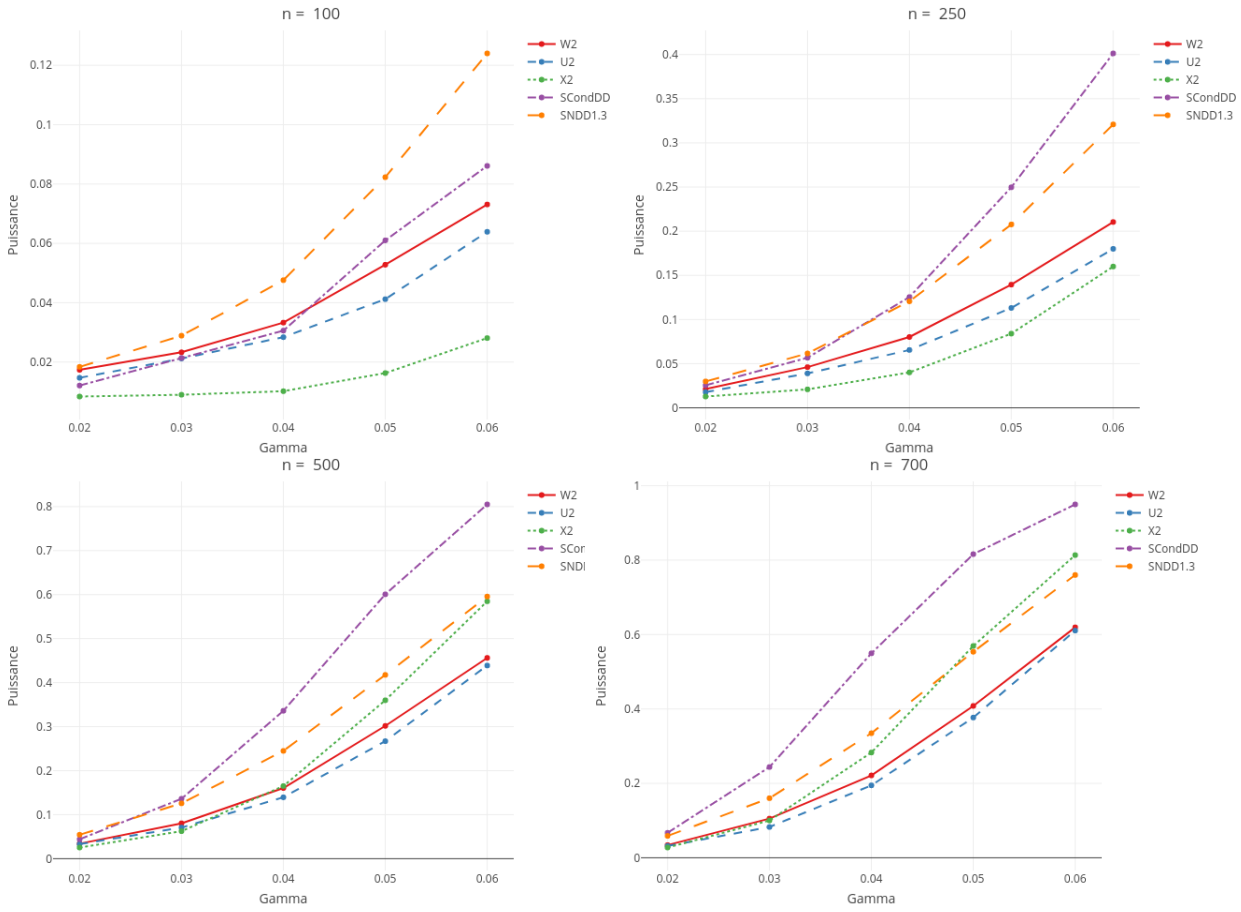

FIGURE 3.3.37 – Famille de Wong additive : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Alternative de Wong: Modification multiplicative de Benford

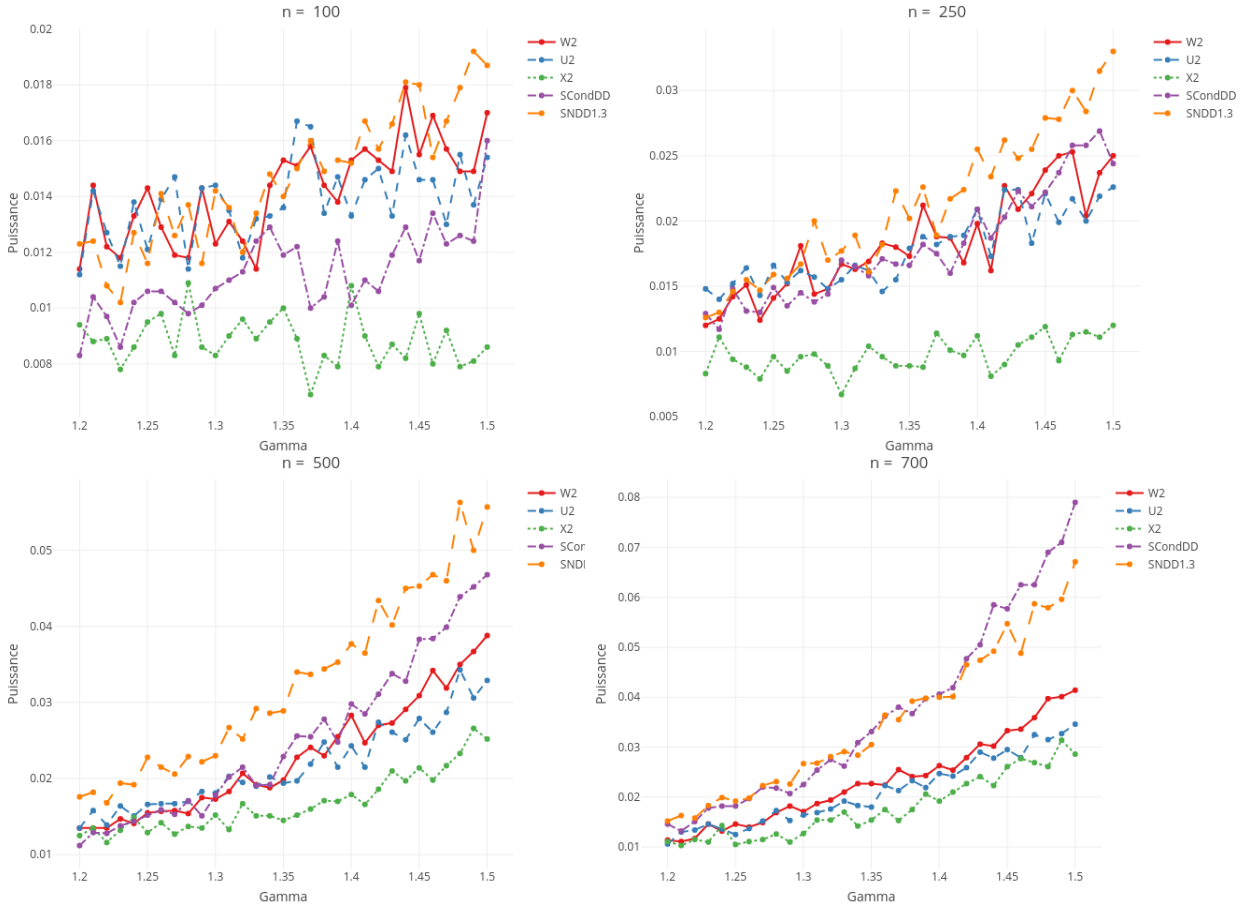

FIGURE 3.3.38 – Famille de Wong multiplicative : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

3.3.11 Test Oracle

Dans ce qui précède , nous avons remarqué que le test du χ^2 a souvent une assez mauvaise puissance, mais parfois, contre certaines alternatives (à l'instar du Copule Benford Hill, ou du Copule Benford Benford), il est meilleur que les tests classiques (U^2, W^2) sans toutefois battre le Smooth test conditionnel $\mathcal{S}_{Cond,DD}$ que l'on a proposé à la section 3.3.9. Ainsi, au vu de ce comportement erratique et ne disposant pas d'« oracle » sur l'alternative en cours, nous recommandons que le test du χ^2 ne devrait pas être utilisé en détection de fraudes.

Parallèlement, à la section 3.3.10, nous avons remarqué que le test lisse conditionnel $\mathcal{S}_{Cond,DD}$ était bien meilleur que le test $\mathcal{S}_{NDD}$ et parfois, le cas de figure inverse se produit. Or un examen des Figures 3.3.31 , 3.3.33 et 3.3.34 montre que le test $\mathcal{S}_{Cond,DD}$ est plus

puissant que $\mathcal{S}_{NDD}$ précisément dans les cas où le test du χ^2 est meilleur que les tests classiques et inversement.

Ceci suggère l'idée d'utiliser le test $\mathcal{S}_{Cond,DD}$ dans les cas où le test du χ^2 est meilleur que les tests classiques et autrement d'utiliser le test $\mathcal{S}_{NDD}$, qui lui est globalement comparable aux tests classiques.

Évidemment on ne sait pas en pratique contre une alternative inconnue si le test du χ^2 sera bon ou mauvais, mais on peut s'appuyer sur la p-value de ce test, laquelle est facilement calculable. Si la p-value du test du χ^2 est inférieure à celle d'un test classique, disons le W^2, on utiliserait alors le test $\mathcal{S}_{Cond,DD}$ et sinon le test lisse $\{\mathcal{S}_{NDD}, c = 1.3\}$. On aurait ainsi un autre test compromis « le smooth test oracle $\mathcal{STO}$ » entre $\mathcal{S}_{NDD}$ et le test $\mathcal{S}_{Cond,DD}$ qui, on peut l'espérer, sera comparable en termes de puissance à ces deux tests compromis.

Que le lecteur soit sans crainte nous n'inventons pas de tel procédé. En effet l'idée d'utiliser un test comme « oracle » pour choisir le test devant être utilisé pour un jeu de données remonte à Büning (2002) , du moins sous une forme « naïve ». Elle a été reprise par Inglot et Ledwina (2006) et plus récemment par Ledwina et Wyłupek (2014) sous une forme différente de celle que nous allons présenter.

En effet, ici, nous allons utiliser deux (2) « oracles » : le test du χ^2 et le test de W^2, ou plutôt leur p-values. Si la p-value du test du χ^2 est inférieure à celle du test W^2, on utilisera le test conditionnel et inversement dans l'autre cas.

$$\mathcal{STO} = \begin{cases} \mathcal{S}_{Cond,DD} & \text{Si } pvalue_{\chi^2} < pvalue_{W^2} \\ \mathcal{S}_{NDD} & \text{Si } pvalue_{\chi^2} > pvalue_{W^2} \end{cases} \tag{3.3.18}$$

Le rejet ou non de l'hypothèse H_0 est décrit dans l'algorithme 3.1

Pour mener l'étude de simulation, nous étudions les niveaux réels de puissance des tests afin de garantir une comparaison adéquate. En effet, le test $\mathcal{STO}$, étant basé sur l'utilisation des p-values du χ^2 et de W^2, nous avons calibré les quantiles des tests $\mathcal{S}_{NDD}$ et $\mathcal{S}_{Cond,DD}$ pour que le test $\mathcal{STO}$ soit bien de niveau 5%. On peut évaluer le bruit statistique après simulation à 0.009 (au niveau de confiance 95%) lorsque la puissance est autour de 0.5. Les seuils critiques sous H_0 pour chacun des tests $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, $\mathcal{S}_{DD}$, $\mathcal{S}_{NDD}$, $\mathcal{S}_{Cond_1}$, $\mathcal{S}_{Cond_2}$, $\mathcal{S}_{Cond_3}$, $\mathcal{S}_{Cond_4}$, $\mathcal{S}_{Cond_5}$, $\mathcal{S}_{Cond,DD}$, W^2, U^2, A^2 et χ^2 sont approchés par simulation de Monte Carlo pour le niveaux de significations $\alpha = 4.1\%$ via un million de réplications pour les sept (7) tailles d'échantillon considérées (cf Tables 3.16).

Algorithm 3.1 Algorithme du test $\mathcal{STO}$

Require: $X \leftarrow$ une variable aléatoire continue positive
Require: $p_{\chi^2} \leftarrow$ la pvalue de χ^2 sous H_0 associée à X
Require: $p_{W^2} \leftarrow$ la pvalue de W^2 sous H_0 associée à X
Require: $q_{\{\mathcal{S}_{Cond,DD},\alpha\}} \leftarrow$ la quantile d'ordre α sous H_0 du test $\{\mathcal{S}_{Cond,DD}\}$
Require: $v_{\{\mathcal{S}_{Cond,DD}\}} \leftarrow$ la valeur de la statistique du test $\{\mathcal{S}_{Cond,DD}\}$
Require: $q_{\{\mathcal{S}_{NDD},\alpha\}} \leftarrow$ la quantile d'ordre α sous H_0 du test $\mathcal{S}_{NDD}$
Require: $v_{\mathcal{S}_{NDD}} \leftarrow$ la valeur de la statistique du test $\mathcal{S}_{NDD}$
  if $p_{\chi^2} < p_{W^2}$ **then**
    if $v_{\{\mathcal{S}_{Cond,DD}\}} > q_{\{\mathcal{S}_{Cond,DD},\alpha\}}$ **then**
      return On rejette H_0
    else
      return On ne rejette pas H_0
    end if
  else
    if $v_{\mathcal{S}_{NDD}} > q_{\{\mathcal{S}_{NDD},\alpha\}}$ **then**
      return On rejette H_0
    else
      return On ne rejette pas H_0
    end if
  end if

	50	100	250	500	1000	3000	5000
$\mathcal{S}_{Cond_1}$	17,627	18.434	18.740	18.857	18,941	18,906	18,966
$\mathcal{S}_{Cond_2}$	31,128	31.983	32.111	32.150	32,230	32,195	32,208
$\mathcal{S}_{Cond_3}$	43,577	44.550	44.727	44.716	44,726	44,697	44,724
$\mathcal{S}_{Cond_4}$	55,076	56.485	56.749	56.825	56,838	56,797	56,784
$\mathcal{S}_{Cond_5}$	66,255	68.109	68.465	68.622	68,654	68,643	68,655
$\mathcal{S}_{Cond,DD}$	26,388	25.358	23.579	22.513	21,700	20,718	20,463
$\mathcal{S}_1$	4,157	4.166	4.173	4.152	4,178	4,178	4,183
$\mathcal{S}_2$	6,342	6.362	6.364	6.374	6,375	6,382	6,381
$\mathcal{S}_3$	8,248	8.249	8.238	8.243	8,252	8,247	8,240
$\mathcal{S}_4$	10,024	9.974	9.955	9.965	9,946	9,965	9,966
$\mathcal{S}_5$	11,702	11.646	11.599	11.595	11,580	11,574	11,587
$\mathcal{S}_{NDD,1.3}$	10,094	9.139	7.057	5.817	5,037	4,545	4,448
W^2	0,497	0.497	0.495	0.495	0,495	0,495	0,495
U^2	0,200	0.199	0.197	0.197	0,197	0,197	0,197
A^2	2,828	2.731	2.670	2.653	2,644	2,632	2,640
χ^2	119,093	116.554	114.694	114.197	113,840	113,593	113,555

TABLE 3.16 – Seuils critiques de la loi des statistiques $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, $\mathcal{S}_{NDD,1.3}$, $\mathcal{S}_{Cond_1}$, $\mathcal{S}_{Cond_2}$, $\mathcal{S}_{Cond_3}$, $\mathcal{S}_{Cond_4}$, $\mathcal{S}_{Cond_5}$, $\mathcal{S}_{Cond,DD}$, W^2, U^2, A^2 et χ^2 sous $H_0 : (\mathscr{D}_1, \mathscr{D}_2) \sim \mathscr{B}_{(1,2)}$au seuil $\alpha = 0.041$ en utilisant 1000000 échantillons aléatoires.

3.3.12 Comparaison du test $\mathcal{STO}$ et des tests classiques

Nous utiliserons dans ce contexte les alternatives utilisées dans la section (Alternatives).

Les tests $\mathcal{STO}$, U^2, χ^2, W^2 sont effectués au niveau 5% et leur puissance est approximée par Monte Carlo (10000 réplications) pour chaque triplet $(\text{famille}, n, \gamma)$. Ces courbes de puissance apparaissent aux figures 3.3.39 à 3.3.48.

Nous remarquons deux types de comportement de la courbe de puissance du test $\mathcal{STO}$.

— soit sa courbe de puissance se trouve dans la moitié supérieure de l'enveloppe formé par la courbe de puissance du « pire » et du « meilleur » test et tend proportionnellement à la taille de l' échantillon vers le « meilleur » test, cf par exemple les graphiques 3.3.39 et 3.3.40.

— soit sa courbe de puissance est meilleure que toutes les autres courbes de tests envisagés dans notre étude comme nous pouvons le remarquer sur les graphiques 3.3.44, 3.3.45, 3.3.47 et 3.3.48.

Dans tous les cas, le comportement de la courbe de puissance du test $\mathcal{STO}$ nous permet de conclure que le test $\mathcal{STO}$ est un « bon compromis » dans toutes les alternatives. Les résultats des simulations sur les alternatives « `Testing` » de la Table 3.12 se trouvent dans la section 3.3.13.

Famille des mixtures: Benford Uniforme

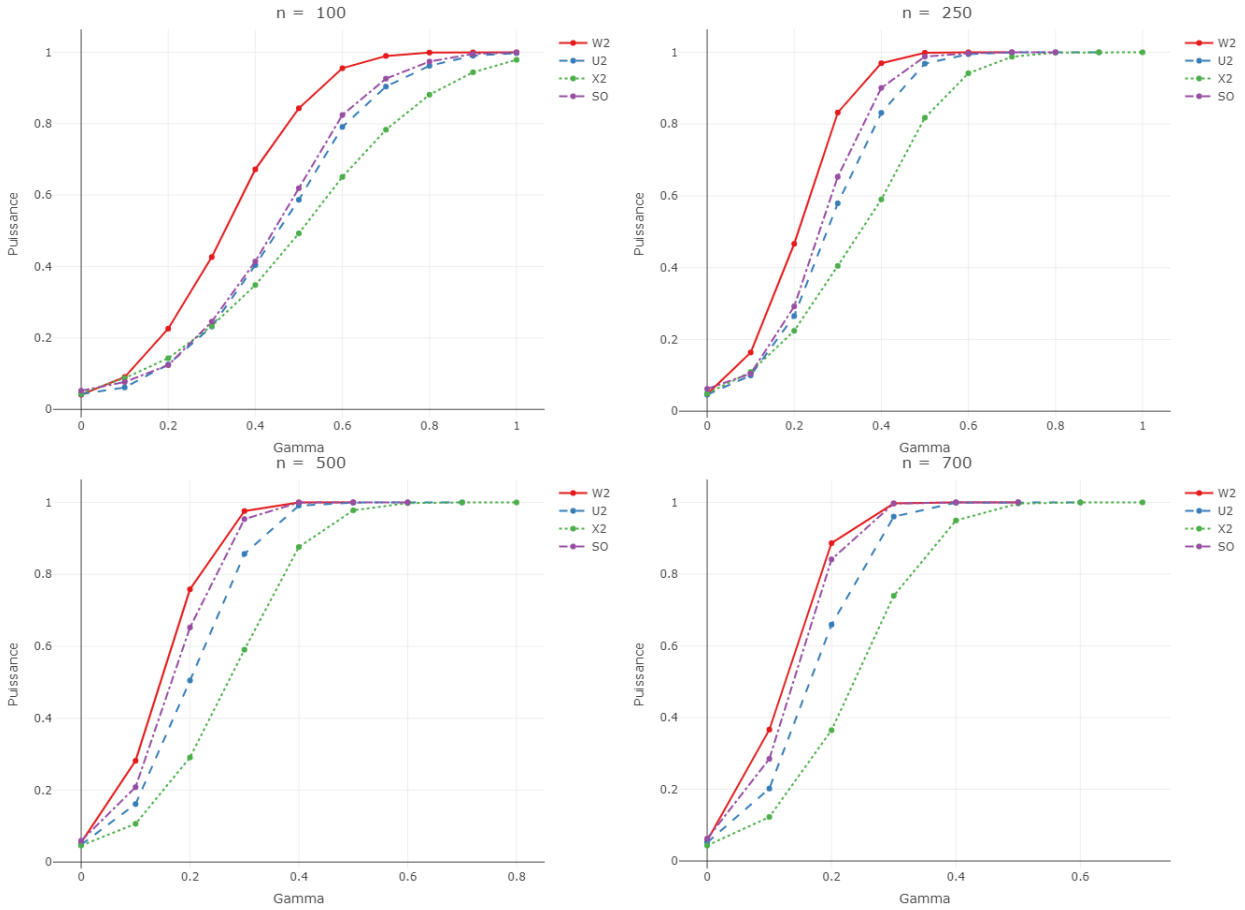

FIGURE 3.3.39 – Mixture Benford Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma U_{[10,99]}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Stigler Uniforme

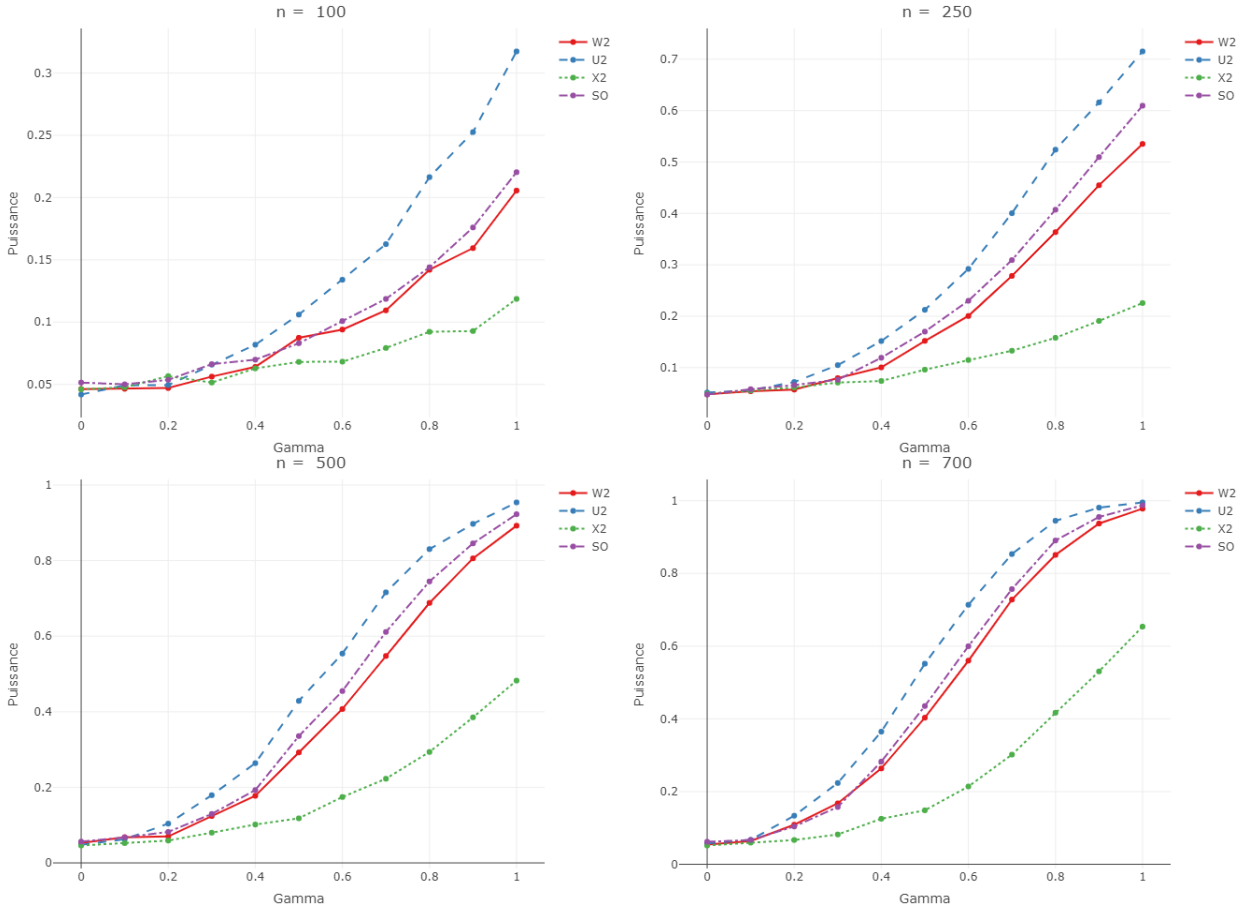

FIGURE 3.3.40 – Famille des mixtures : Mixture Benford Stigler Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, \mathcal{S}_1 \otimes U_{[10,99]}$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Benford Rodriguez

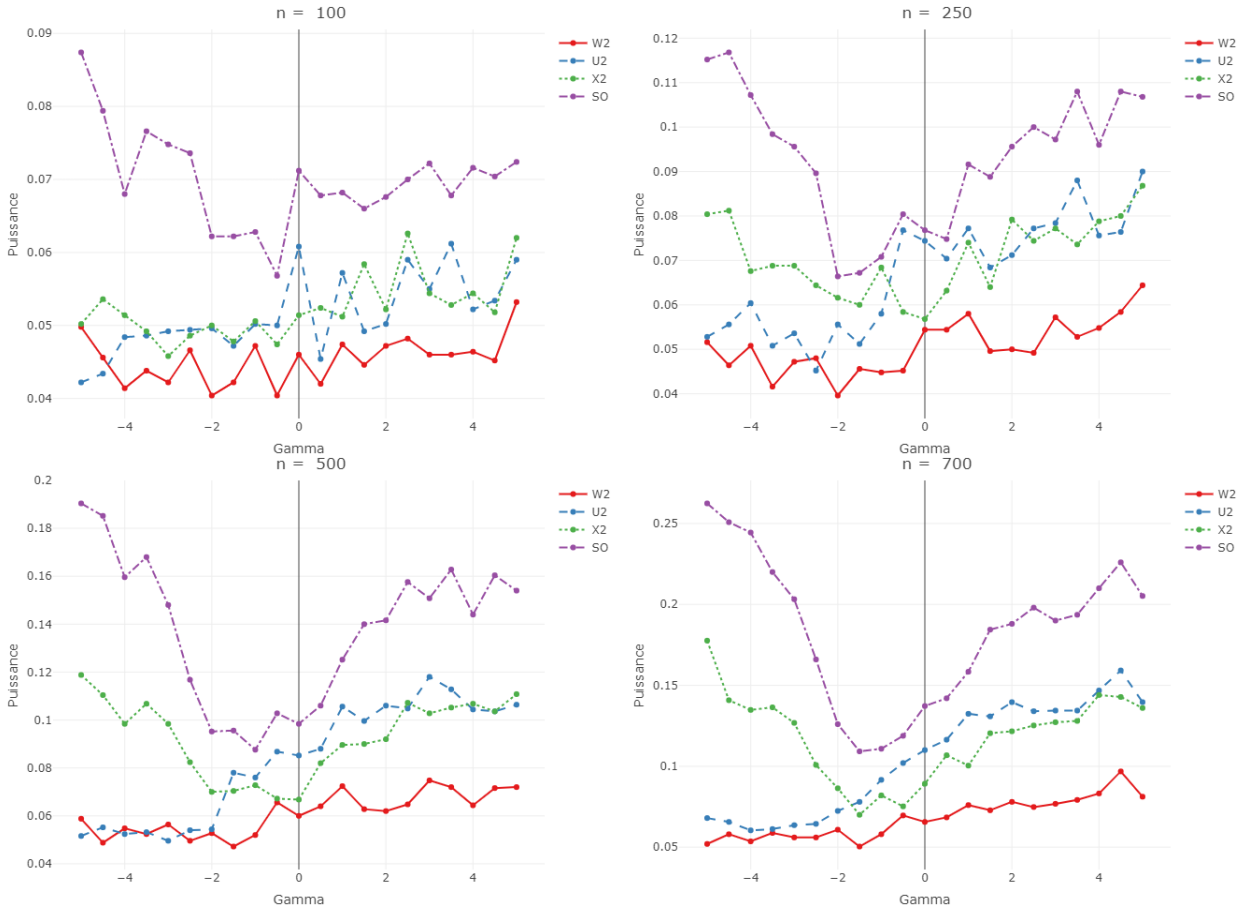

FIGURE 3.3.41 – Famille des indépendances : Indépendance Benford Rodriguez ($\mathscr{B}_1 \perp \mathcal{R}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Rodriguez Rodriguez

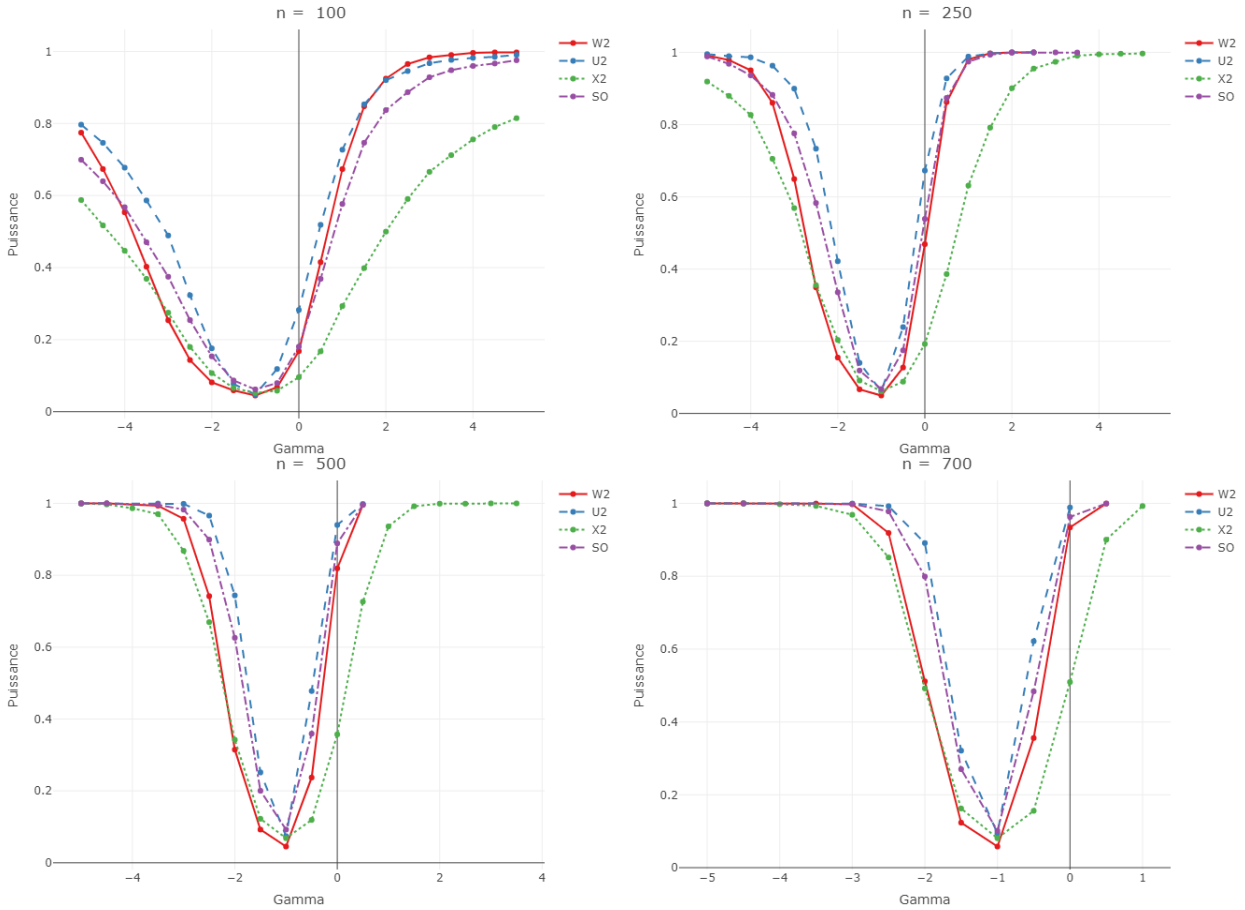

FIGURE 3.3.42 – Famille des indépendances : Indépendance Rodriguez Rodriguez ($\mathcal{R}_1(\gamma) \perp \mathcal{R}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille de Copules: Benford Hill

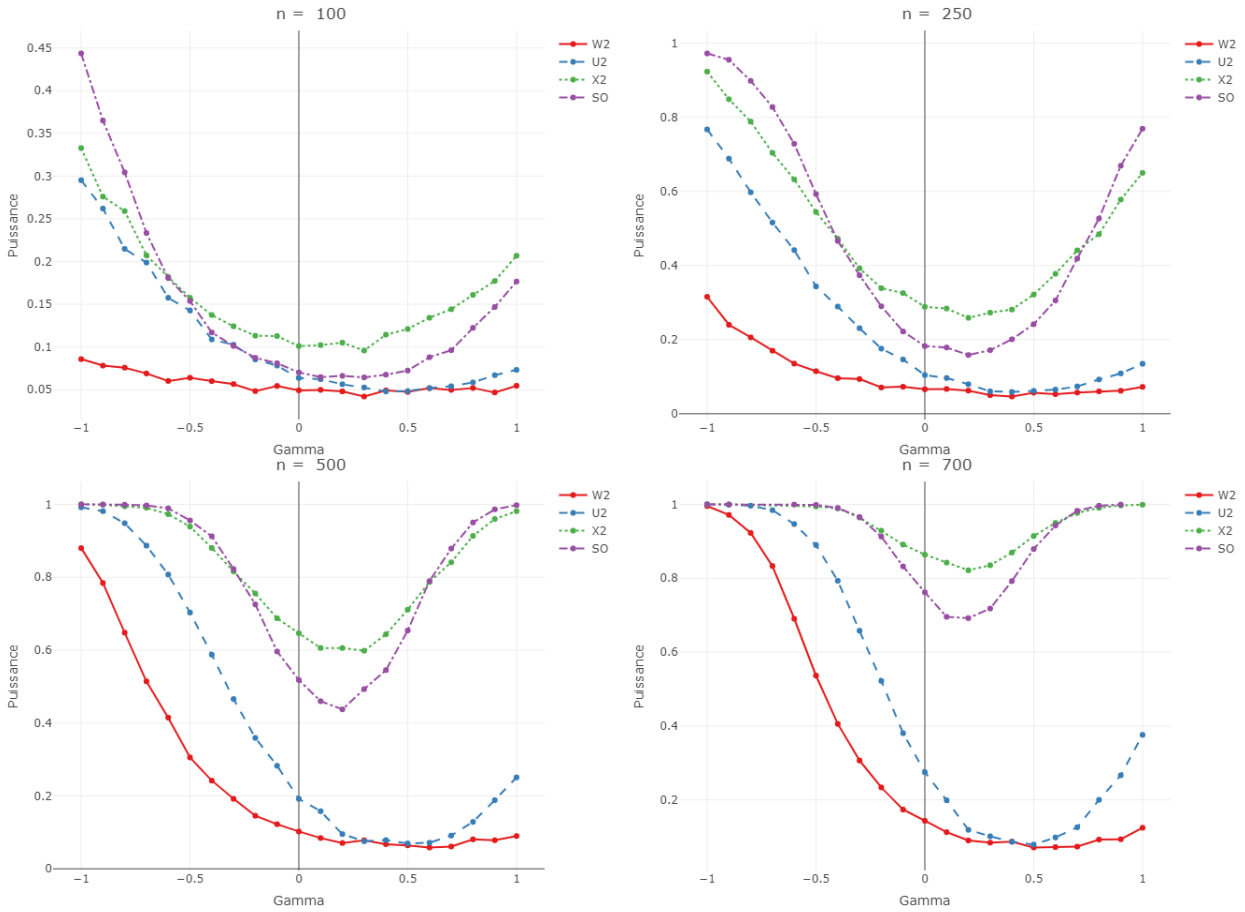

FIGURE 3.3.43 – Famille des copules : Copule Benford Hill $C(\gamma, \mathscr{B}_1, \mathcal{H}_2)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille de Copules: Benford Benford

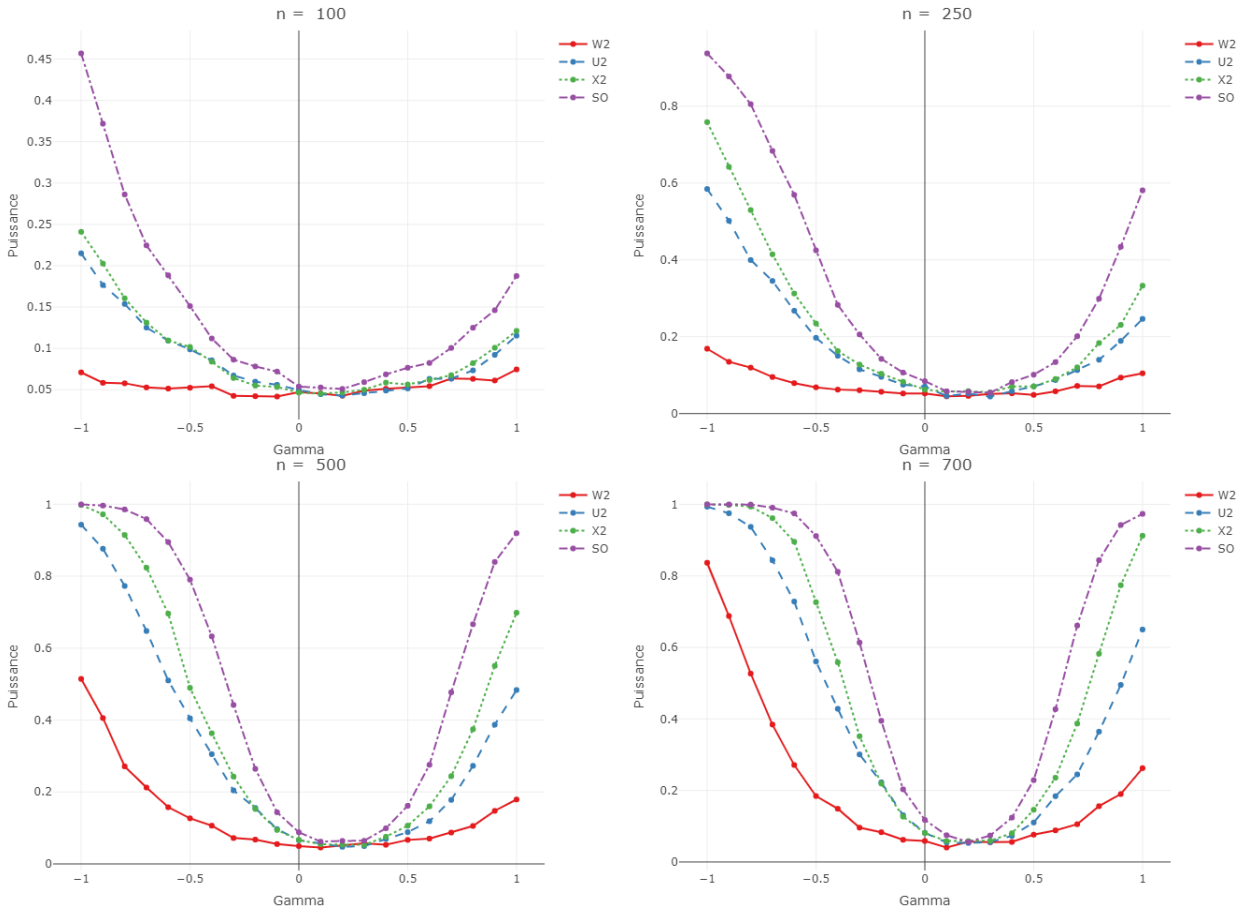

FIGURE 3.3.44 – Famille des copules : Copule Benford Benford $C(\gamma, \mathscr{B}_1, \mathscr{B}_2)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Rodriguez sachant Benford

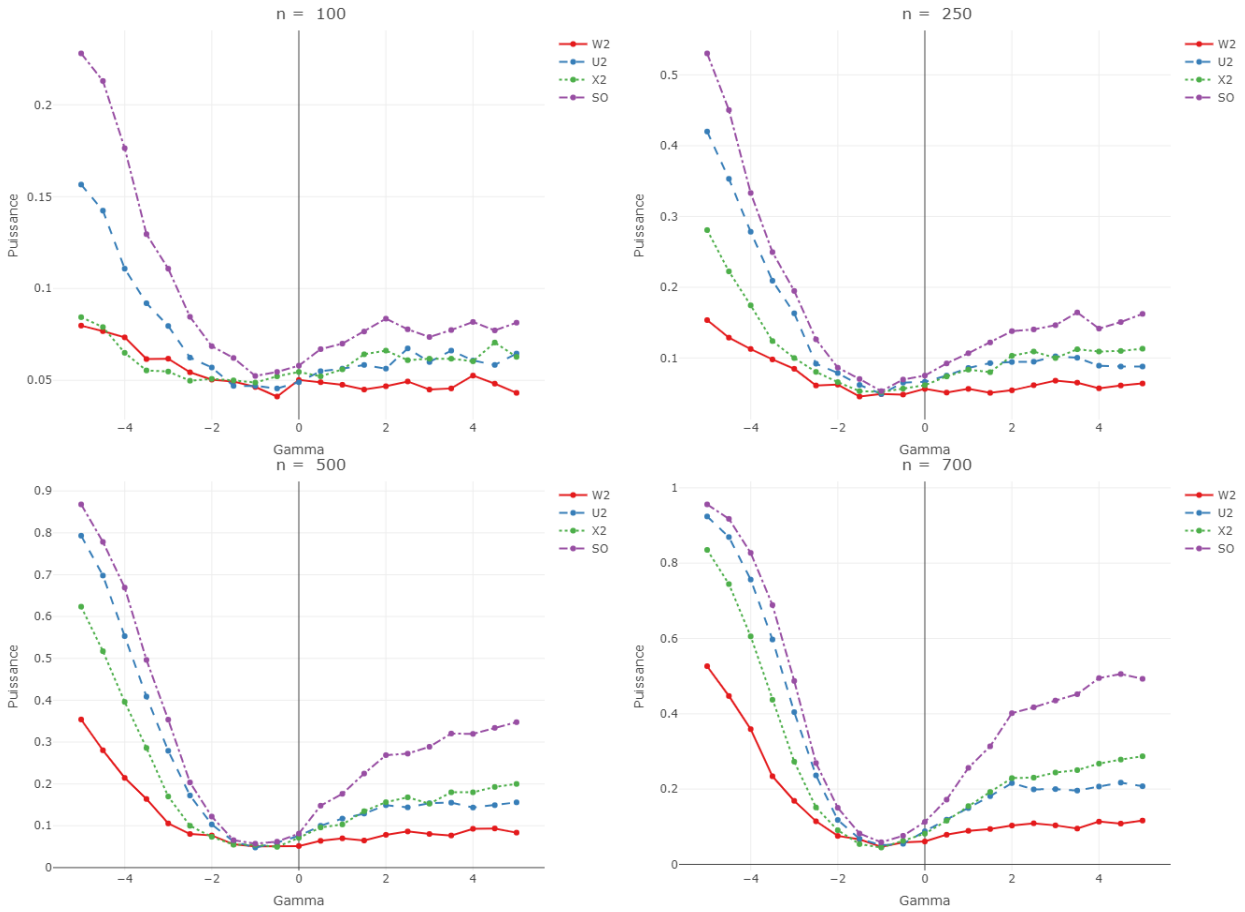

FIGURE 3.3.45 – Famille des conditionnelles : Rodriguez sachant Benford $\left(\mathscr{B}_1, \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Benford sachant Rodriguez

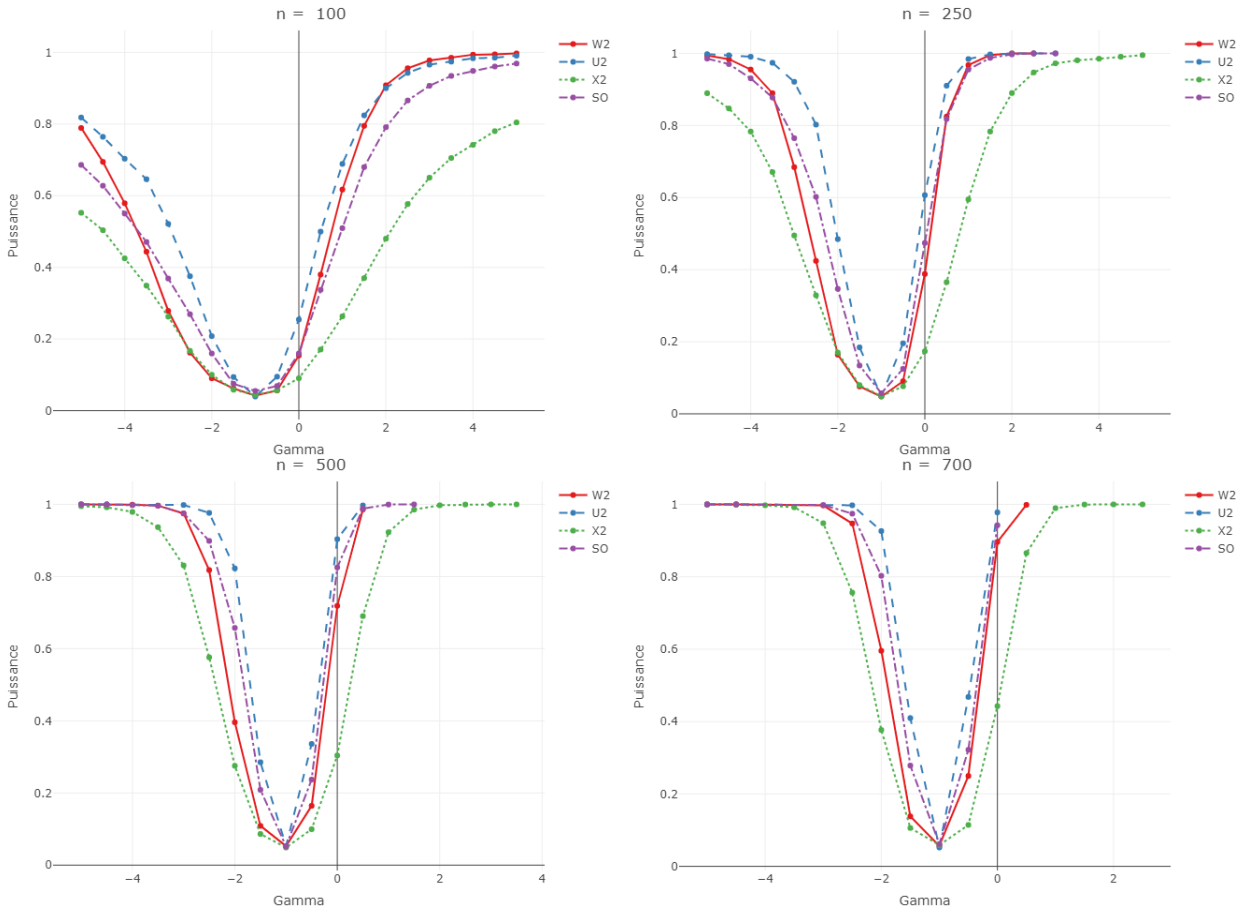

FIGURE 3.3.46 – Famille des conditionnelles : Benford sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathscr{B}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Alternative de Wong: Modification additive de Benford

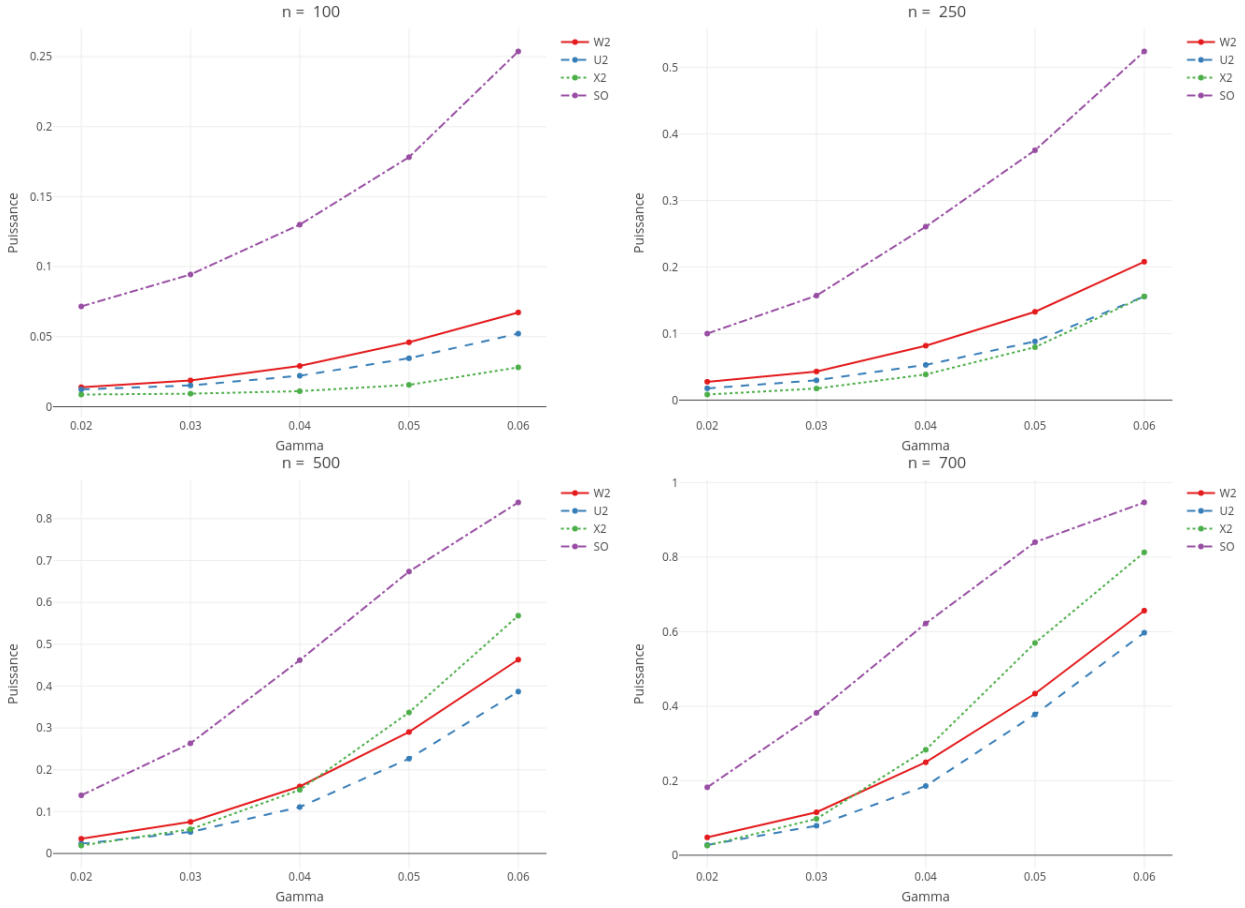

FIGURE 3.3.47 – Famille de Wong additive : Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Alternative de Wong: Modification multiplicative de Benford

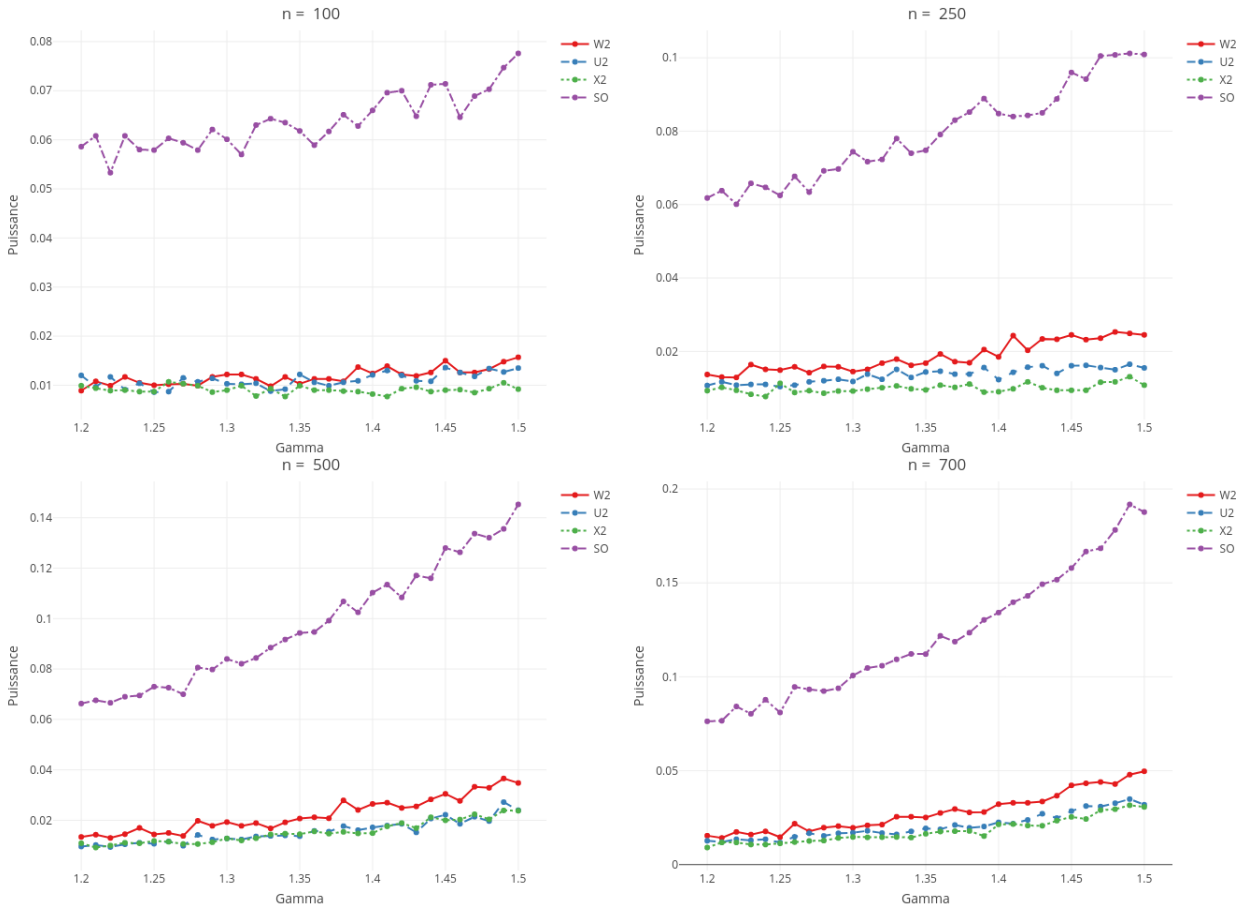

FIGURE 3.3.48 – Famille de Wong multiplicative : Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

3.3.13 Application du test $\mathcal{STO}$ sur les alternatives « Testing » de la section 3.3.4

Les termes « training » et « Testing » utilisés dans la section sur les alternatives est un clin d'œil à l'apprentissage automatique dont nous empruntons l'idée de vérifier notre test faisant « compromis » sur un échantillon n'ayant pas servi lors de l'apprentissage.

Nous rappelons au lecteur que les tests $\mathcal{STO}$, U^2, χ^2, W^2 sont effectués au niveau 5% et leur puissance est approximée par Monte Carlo (10000 réplications) pour chaque triplet (famille, n, γ) . Ces courbes de puissance apparaissent aux figures 3.3.49 à 3.3.61.

Pour toutes les figures de la famille des mixtures (Figures 3.3.49 à 3.3.53), les figures 3.3.55, 3.3.57, 3.3.58, 3.3.59, et 3.3.61 , les courbes des tests de U^2 et W^2 font partie du

top 3 en se partageant la première et la troisième position avec la seconde place, dans tout les cas, occupée par $\mathcal{STO}$ qui quand n croît se rapproche considérablement de la première position. Le test de χ^2 est le moins puissant.

Sur les figures 3.3.54, 3.3.56, et 3.3.60, nous remarquons que la courbe du test $\mathcal{STO}$, est la vedette en effet $\forall\, n$, γ elle reste la meilleure et de loin. Elle est suivie par la courbe du test de χ^2.

En conclusion, on retire de cette expérience de simulation que le test $\mathcal{STO}$ est toujours parmi les deux meilleurs tests en termes de puissance, disputant la tête suivant les cas avec l'un des tests classiques U^2, W^2 ou χ^2.

Comme il n'est pas possible de savoir a priori quel test est le meilleur classique pour l'alternative inconnue qui nous est fournie par la nature (ou le fraudeur !), il ressort de ce travail que le test $\mathcal{STO}$ pourrait être utilisé avec profit dans de très nombreux cas pour détecter des déviations de la loi de Newcomb-Benford bivariée. Notre test peut donc être recommandé dans un contexte de détection de fraudes et ajouté à l'arsenal des outils disponibles à cette fin.

Famille des mixtures: Benford Hill

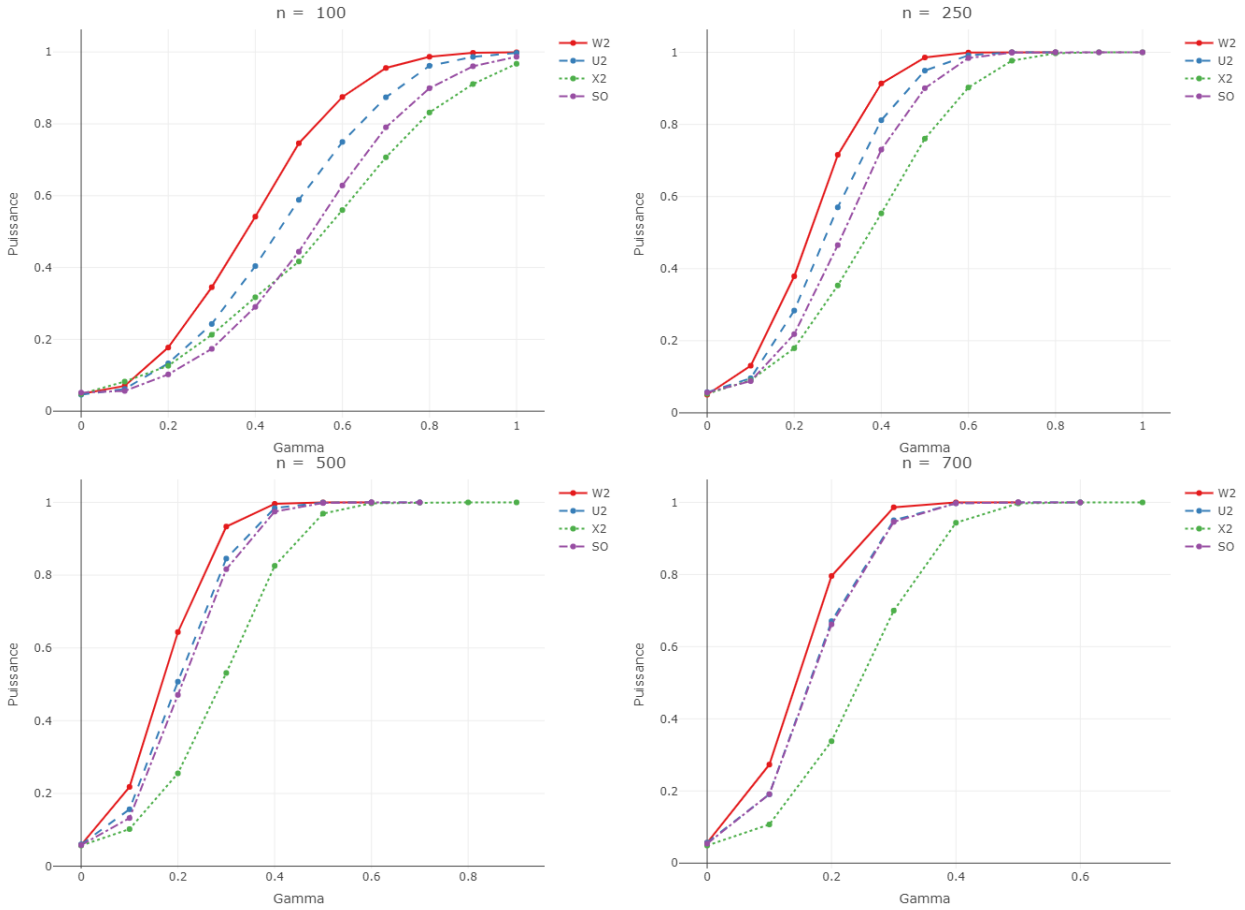

FIGURE 3.3.49 – Mixture Benford Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{H}_{(1,2)}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Stigler

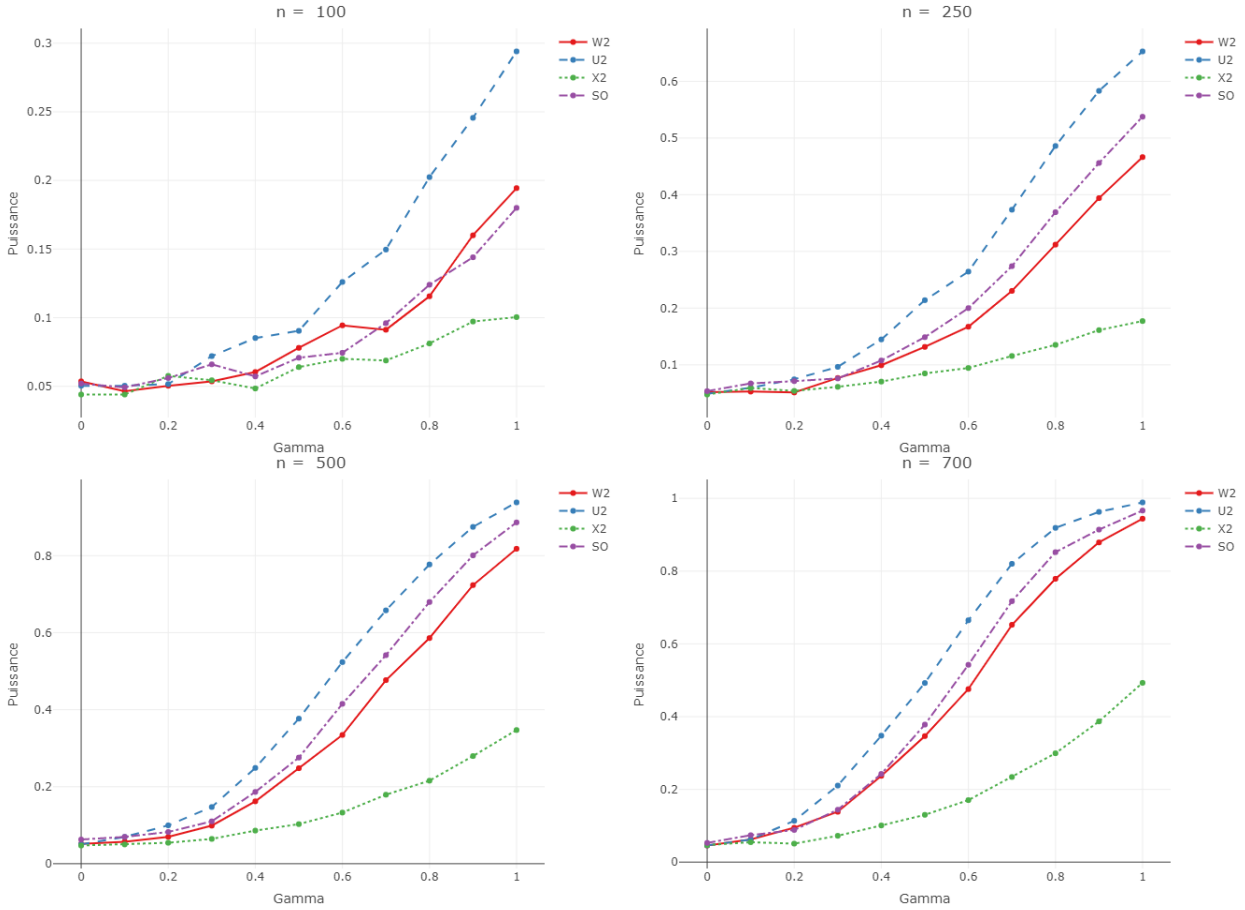

FIGURE 3.3.50 – Mixture Benford Stigler $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma\mathcal{S}_{(1,2)}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Uniforme Stigler

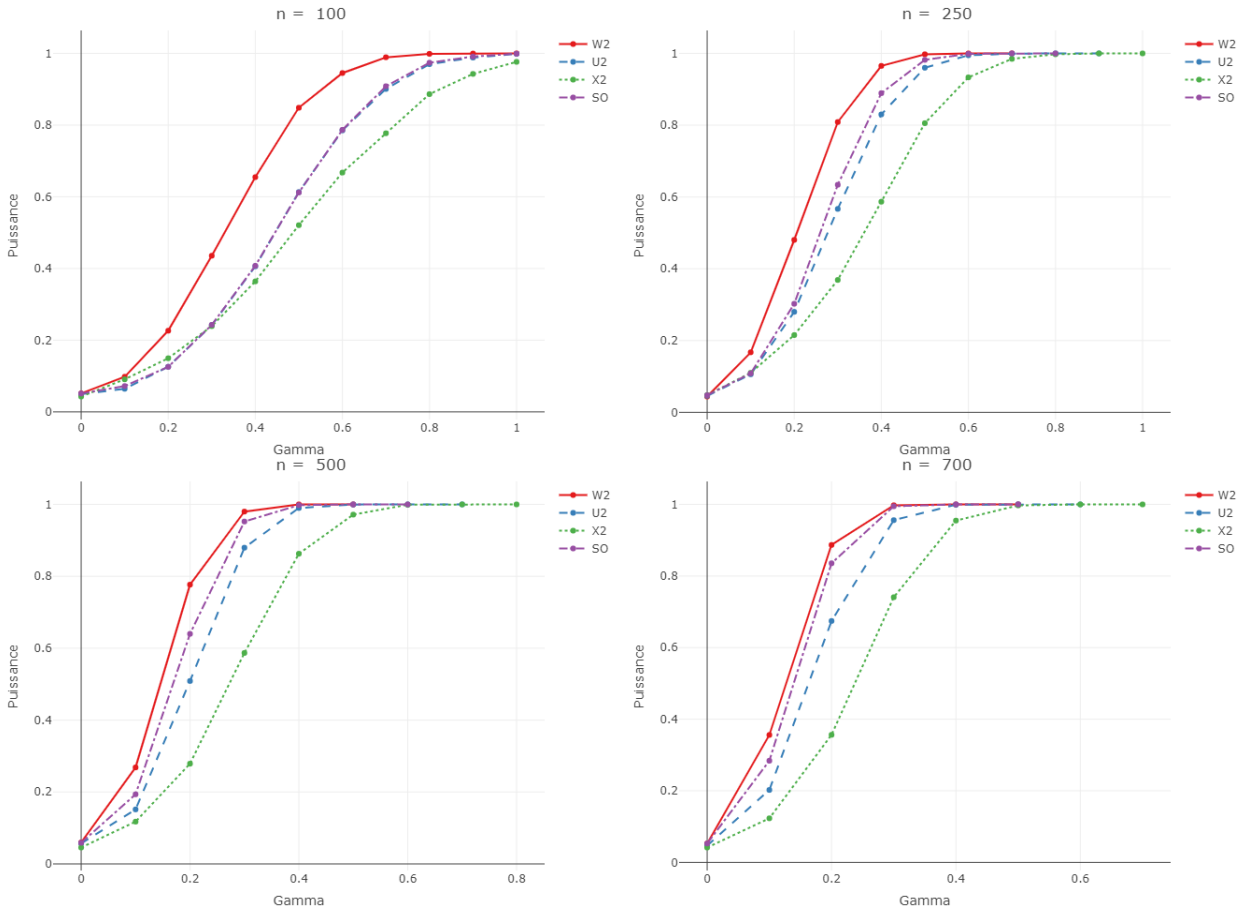

FIGURE 3.3.51 – Mixture Benford Uniforme Stigler $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma U_{[1,9]}\otimes S_2$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Hill Uniforme

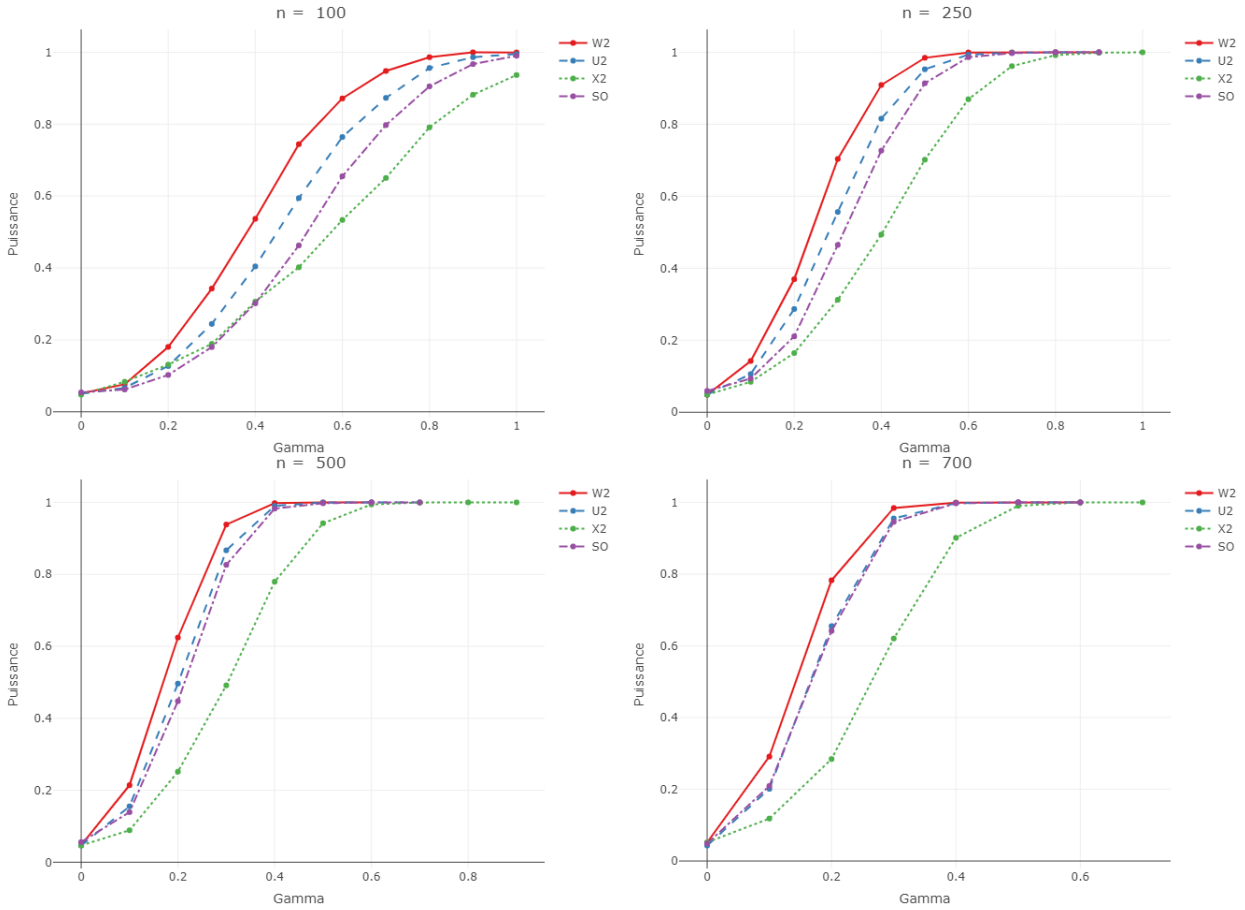

FIGURE 3.3.52 – Mixture Benford Hill Uniforme $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma\,\mathcal{H}_1\otimes U_{[0,9]}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Uniforme Hill

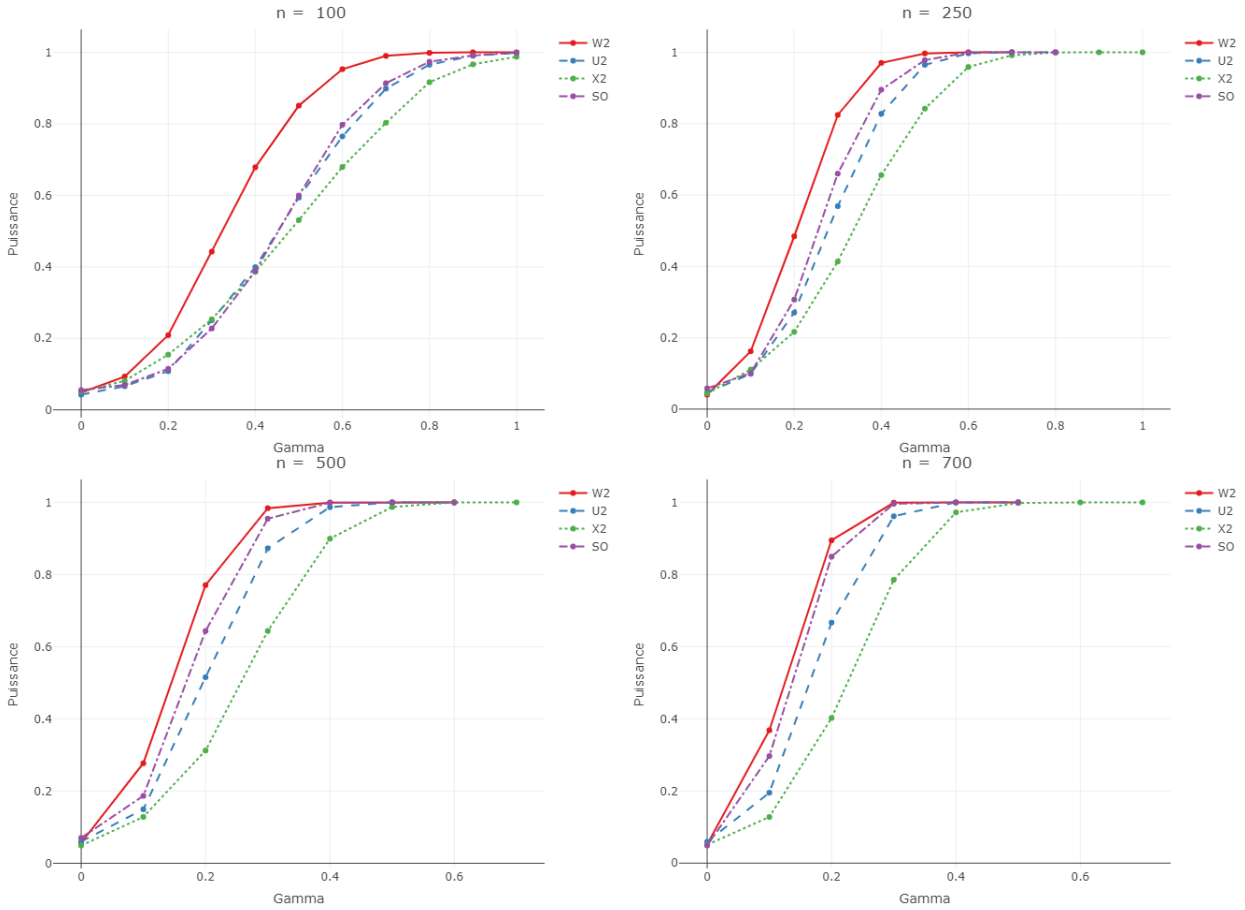

FIGURE 3.3.53 – Mixture Benford Uniforme Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma(1-\gamma)\, U_{[1,9]} \otimes \mathcal{H}_2$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des Copules: Benford Stigler

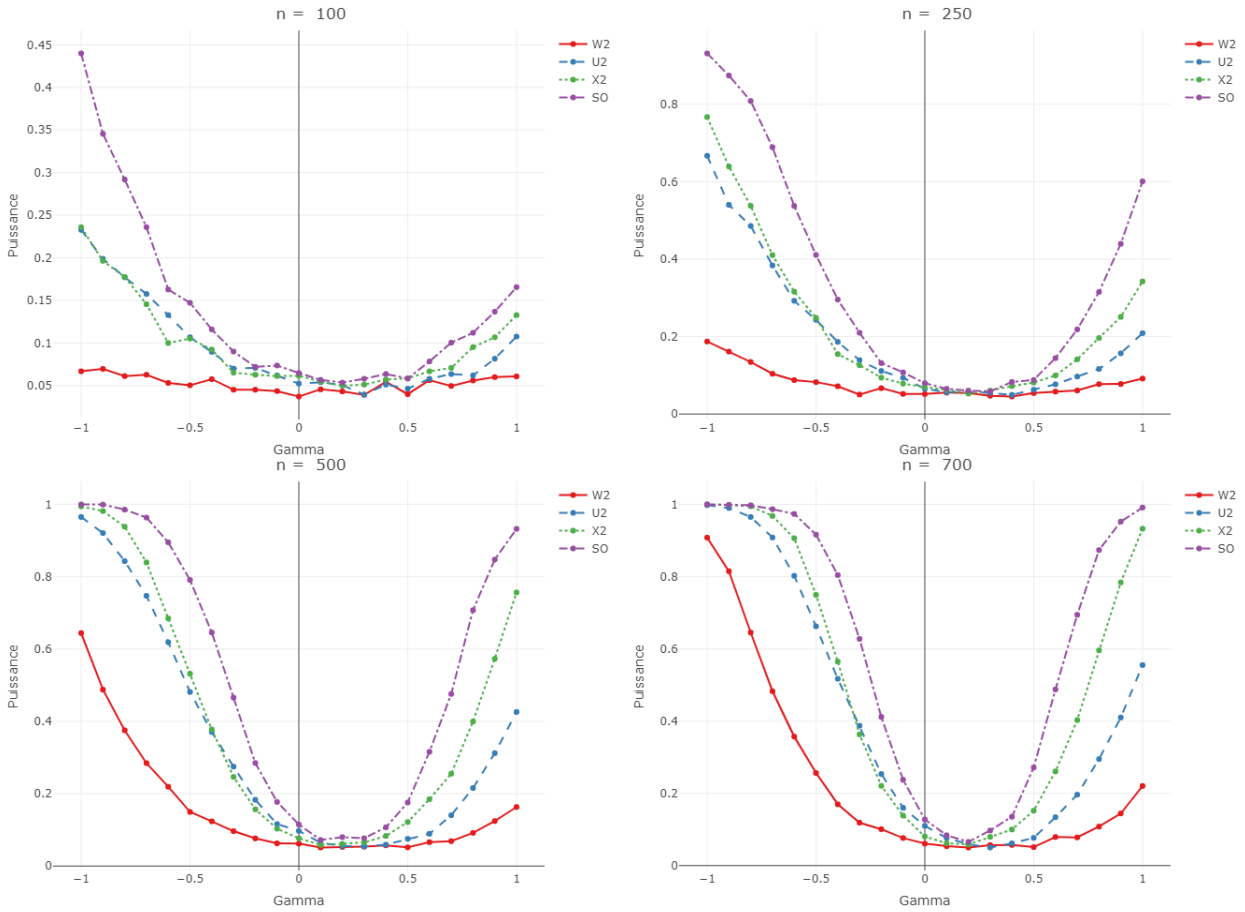

FIGURE 3.3.54 – Copule Benford Stigler $C(\gamma, \mathscr{B}_1, S_2)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des Copules: Stigler Benford

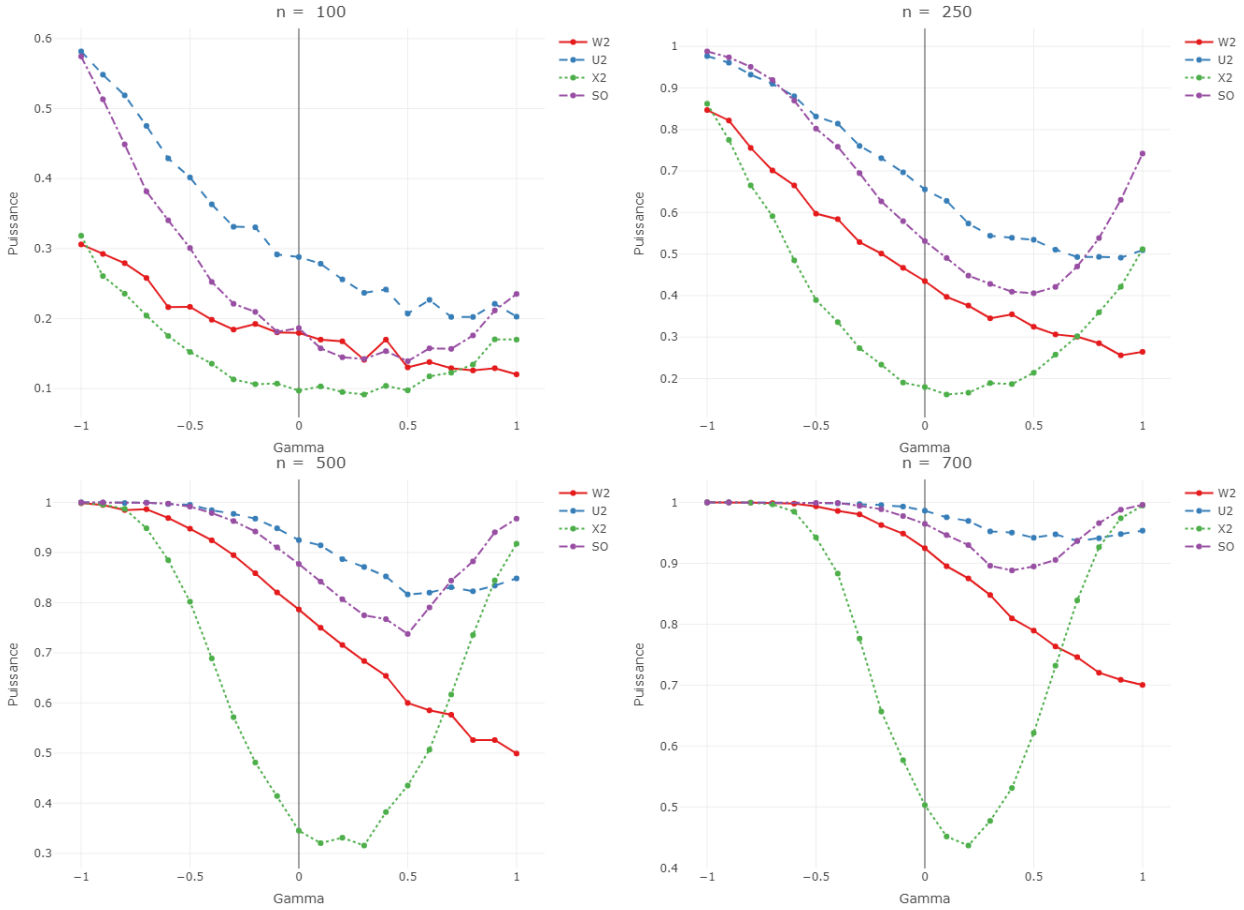

FIGURE 3.3.55 – Copule Stigler Benford $C(\gamma, \mathcal{S}_1, \mathscr{B}_2)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Benford Newcomb-Benford Généralisée

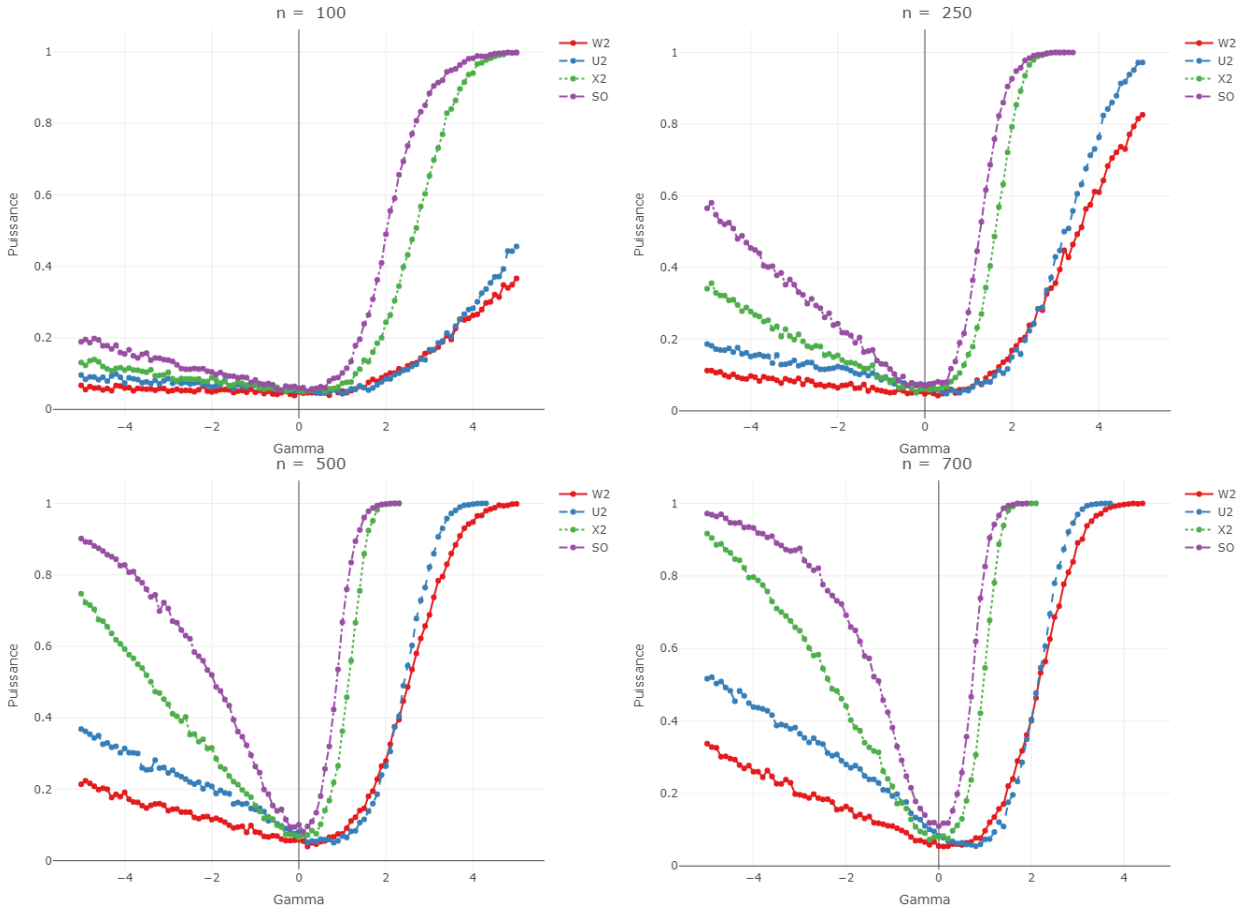

FIGURE 3.3.56 – Indépendance Benford Newcomb-Benford Généralisée ($\mathcal{B}_1 \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

`Famille Indépendance: Newcomb-Benford Généralisée Newcomb-Benford Généralisée`

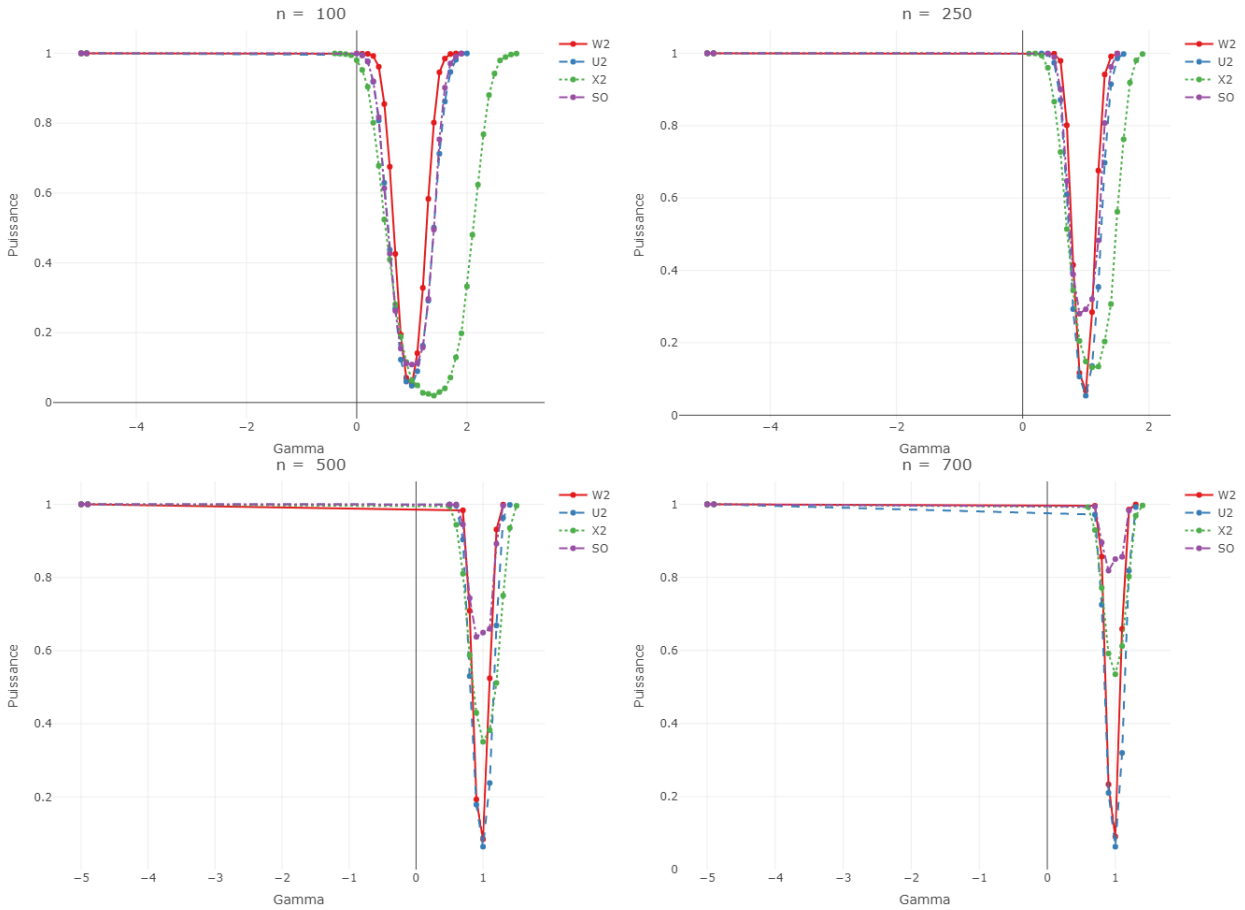

FIGURE 3.3.57 – Indépendance Newcomb-Benford Généralisée Newcomb-Benford Généralisée ($\mathcal{GB}_1(\gamma) \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Rodriguez sachant Rodriguez

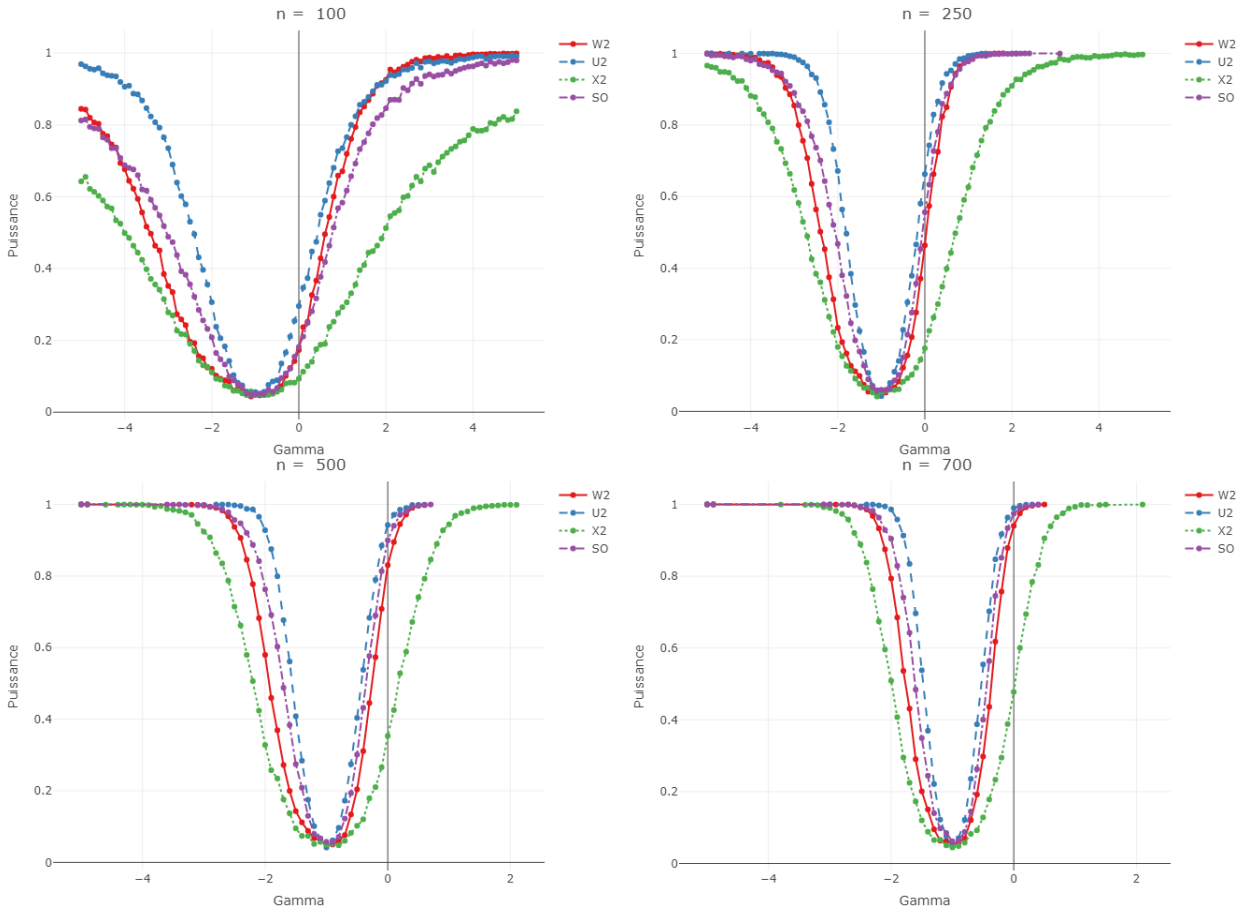

FIGURE 3.3.58 – Rodriguez sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Benford

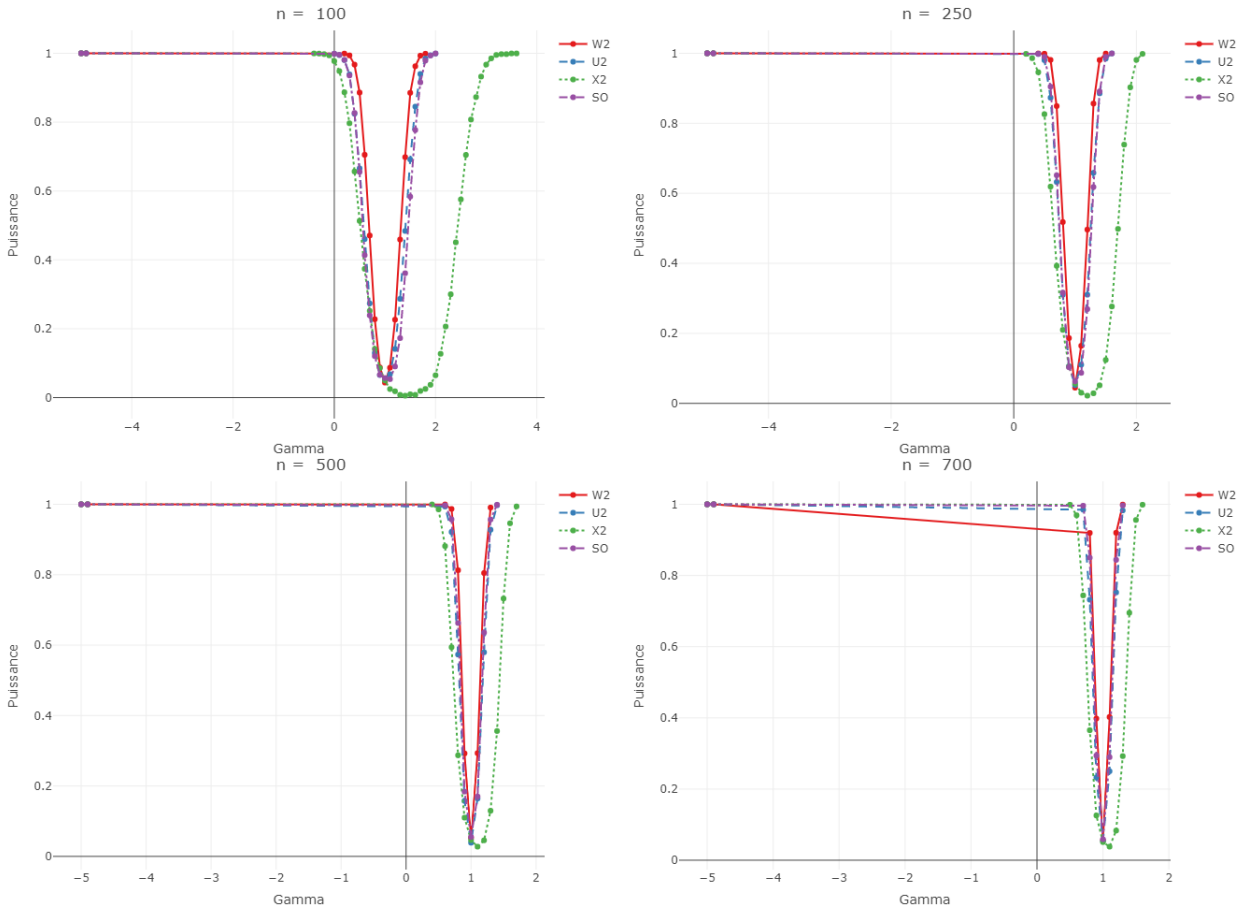

FIGURE 3.3.59 – Newcomb-Benford Généralisée sachant Benford $\left(\mathcal{GB}_1(\gamma), \mathscr{B}_{(2|1)}\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Benford sachant Newcomb-Benford Généralisée

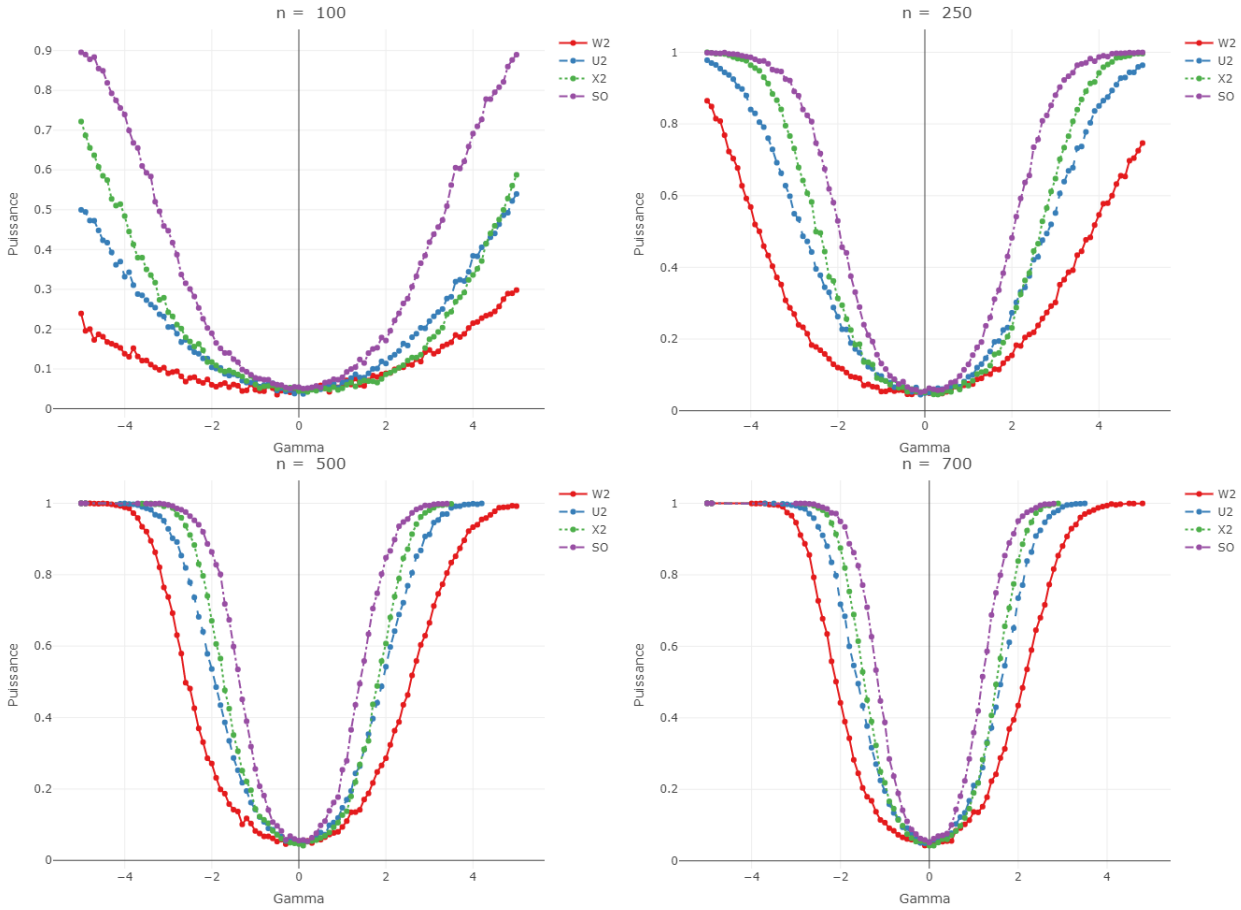

FIGURE 3.3.60 – Benford sachant Newcomb-Benford Généralisée $\left(\mathscr{B}_1, \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée

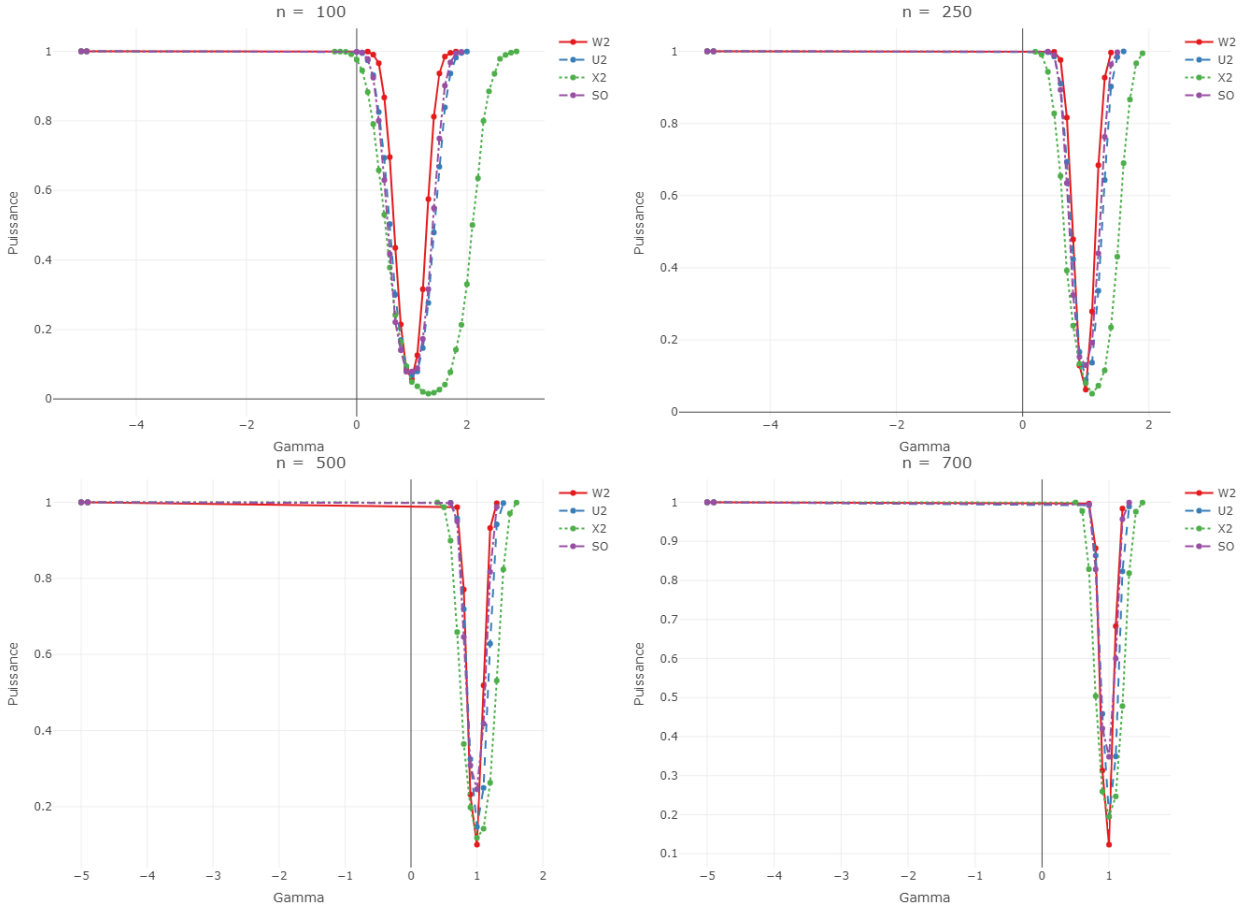

FIGURE 3.3.61 – Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée $\left(\mathcal{GB}_1(\gamma), \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi Newcomb-Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue) et $\mathcal{STO}$ (couleur violette) dont les expressions se trouvent respectivement aux Sections 3.3.3 et 3.3.18. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Conclusion générale : perspectives de la recherche

Nous avons présenté dans les précédents chapitres la démarche de recherche et les travaux menés pour détecter les fraudes en trouvant un test d'adéquation faisant compromis à la loi de Newcomb-Benford dans le cadre de notre projet de recherche. Comme nous l'avons dit dans le chapitre 3, nous considérons notre recherche comme « un jeu du chat et de la souris ». Aussi selon nous, cette recherche ouvre la voie à de nouveaux travaux sur le sujet. Dans ce chapitre nous synthétisons les contributions de notre recherche et approfondissons les perspectives de la recherche.

Notre objet de recherche comme dit dans la section (Introduction) est de proposer un test d'adéquation qui permet de tester l'hypothèse nulle $H_0 : X \sim \mathscr{B}$ en contrôlant le risque d'erreur de Type I et en tentant de minimiser l'erreur de Type II. Pour ce fait nous nous sommes intéressés en premier lieu au premier chiffre significatif (PCS) en y appliquant le smooth test de Neyman (1937). Nous avons comparé les résultats obtenus aux travaux de Lesperance *et al.* (2016) et Joenssen (2013c) au chapitre 2. Nous avons remarqué que notre test T_2 et le data driven $T_{\hat{K}}$ sont toujours parmi les « meilleurs » tests, d'où nous le recommandons dans le cas des premiers chiffres significatifs.

Les résultats étant prometteurs et sachant que les fraudeurs sont sournois et vicieux avec une capacité de vouloir toujours arriver à leurs fins avec seule limite leur imagination, nous nous sommes imaginés à leur place au chapitre 3. Nous avons donc considéré certaines possibilités de fraudes par exemple le cas où le fraudeur ne modifie que le second chiffre significatif, laissant le premier chiffre significatif intact, ou modifié pour être concordant avec la loi de Newcomb-Benford du premier chiffre significatif et inversement. Également, nous nous sommes mis dans la situation où le fraudeur modifie le premier et le second chiffre significatifs pour répondre à ses besoins. Le lecteur pourra consulter la section 3.3.4 pour plus de détails sur les différentes possibilités considérées. Nous avons également abordé la loi conjointe du premier et seconde chiffre significatifs en suivant deux écoles différentes :

- la transformation du couple $(\mathscr{D}_1, \mathscr{D}_2)$ en une valeur univariée $\mathscr{D}_{12} = (10 * \mathscr{D}_1 + \mathscr{D}_2)$, définie sur $\{10, \ldots, 99\}$. Le test compromis obtenu par l'application du Smooth test est le « new data driven smooth test $\mathcal{S}_{NDD}$ »
- en considérant le couple $(\mathscr{D}_1, \mathscr{D}_2)$. Le test compromis obtenu par l'application du Smooth test dans ce cas est le « Smooth test conditionnel $\mathcal{S}_{Cond,DD}$ »

Malheureusement, ces deux écoles ne nous ont pas fourni un test meilleur dans toutes

les alternatives considérées. Alors en se basant sur les travaux de Büning (2002); Inglot et Ledwina (2006) et Ledwina et Wyłupek (2014) nous avons proposé « le smooth test oracle $\mathcal{STO}$ » entre S_{NDD} et le test $\mathcal{S}_{Cond,DD}$.

En regard de notre définition de « bon » test d'adéquation évoquée dans l'introduction et des résultats de la section 3.3.12, nous pouvons donc recommander « le smooth test oracle $\mathcal{STO}$ » . La nécessité d'utiliser notre test se trouve justifiée par nos différentes simulations.

Dans nos travaux, nous nous sommes intéressés uniquement aux deux premiers chiffres significatifs. Ainsi nous pouvons nous projeter facilement à l'utilisation des n chiffres significatifs. Parallèlement, en se mettant à la place d'un fraudeur potentiel, nous pourrions considérer que pour p chiffres significatifs, il modifie les k premiers chiffres significatifs, $k < n$, de telle sorte qu'ils concordent avec la loi de Newcomb-Benford, $(\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_k) \sim \mathscr{B}_{(1,\ldots,k)}$ et laissant les $n - k$ derniers chiffres significatifs intacts ou non. Une application du smooth test tenant en compte ces différentes alternatives serait un outil plus que supplémentaire pour détecter les fraudes.

L'implémentation en R du package permettant l'utilisation du « smooth test oracle $\mathcal{STO}$ » est en cours de rédaction et sera incluse dans la prochaine version du package R BenfordSmoothTest qui a été développé dans le cadre de notre première contribution pour le premier chiffre significatif.

Bibliographie

Rabiâu Abdullahi et Noorhayati Mansor : Fraud Triangle Theory and Fraud Diamond Theory. Understanding the Convergent and Divergent For Future Research. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 5(4):38–45, octobre 2015.

Hirotogu Akaike : Information Theory and an Extension of the Maximum Likelihood Principle. *In Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281. 1973.

Sara Algeri et Xiangyu Zhang : Exhaustive goodness-of-fit via smoothed inference and graphics. juin 2020.

Laleh Arshadi et Amir Hossein Jahangir : Benford's law behavior of Internet traffic. *Journal of Network and Computer Applications*, 40:194–205, avril 2014. ISSN 1084-8045.

Marcel Ausloos, Roy Cerqueti et Tariq A. Mir : Data science for assessing possible tax income manipulation : The case of Italy. *Chaos, Solitons & Fractals*, 104:238–256, novembre 2017. ISSN 0960-0779.

Lucio Barabesi, Andrea Cerasa, Andrea Cerioli et Domenico Perrotta : Goodness-of-Fit Testing for the Newcomb-Benford Law With Application to the Detection of Customs Fraud. *Journal of Business & Economic Statistics*, 36(2):346–358, avril 2017.

Isabel Barbancho, Lorenzo J. Tardón, Ana M. Barbancho et Mateu Sbert : Benford'S Law For Music Analysis. 2015.

D. E. Barton : On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. *Scandinavian Actuarial Journal*, 1953(sup1):24–63, jan 1953.

D. E. Barton : A form of Neyman's uppsi 2 K test of goodness of fit applicable to grouped and discrete data. *Scandinavian Actuarial Journal*, 1955(1-2):1–16, jan 1955.

D. E. Barton : Neyman's test of goodness of fit when the null hypothesis is composite. *Scandinavian Actuarial Journal*, 1956(2):216–245, jul 1956.

Frank Benford : The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

Arno Berger, Ted Hill et Erika Rogers : Benford Online Bibliography, mai 2015. URL `http://www.benfordonline.net`.

Peter L. Bernstein : *Against the Gods : the remarkable story of risk*. Wiley John + Sons, décembre 1998. ISBN 978-0-471-29563-1.

Herbert Büning : An adaptive distribution-free test for the general two-sample problem. *Computational Statistics*, 17(2):297–313, jul 2002.

Bernard Boulerice et Gilles R. Ducharme : Smooth Tests of Goodness-of-Fit for Directional and Axial Data. *Journal of Multivariate Analysis*, 60(1):154–175, janvier 1997.

Jeff Boyle : An Application of Fourier Series to the Most Significant Digit Problem. *The American Mathematical Monthly*, 101(9):879–886, novembre 1994.

Krzysztof Burdzy : *Resonance, From Probability to Epistemology and Back*. IMPERIAL COLLEGE PRESS, octobre 2016.

David Burgstahler et Ilia Dichev : Earnings management to avoid earnings decreases and losses. *Journal of Accounting and Economics*, 24(1):99–126, dec 1997.

John Burke et Eric Kincanon : Benford's law and physical constants : The distribution of initial digits. *American Journal of Physics*, 59(10):952–952, octobre 1991.

Bruce D. Burns : Sensitivity to statistical regularities : People (largely) follow Benford's law. *In Proceedings of the 2009 CogSci Conference*, Amsterdam, The Netherlands, 2009. URL `http://csjarchive.cogsci.rpi.edu/Proceedings/2009/papers/637/paper637.pdf`.

Tom Van Caneghem : EARNINGS MANAGEMENT INDUCED BY COGNITIVE REFERENCE POINTS. *The British Accounting Review*, 34(2):167–178, juin 2002.

Tom Van Caneghem : The impact of audit quality on earnings rounding-up behaviour : some UK evidence. *European Accounting Review*, 13(4):771–786, décembre 2004.

Charles A. P. N. Carslaw : Anomalies in Income Numbers : Evidence of Goal Oriented Behavior. *The Accounting Review*, 63(2):321–327, 1988. URL `http://www.jstor.org/stable/248109`.

Andrea Cerioli, Lucio Barabesi, Andrea Cerasa, Mario Menegatti et Domenico Perrotta : Newcomb–Benford law and the detection of frauds in international trade. *Proceedings of the National Academy of Sciences*, 116(1):106–115, décembre 2018. ISSN 0027-8424.

Wendy Tam Cho et Brian Gaines : Breaking the (Benford) Law. *The American Statistician*, 61(3):218–223, août 2007.

Marco Corazza, Andrea Ellero et Alberto Zorzi : What Sequences obey Benford's Law ? *Department of Applied Mathematics, University of Venice, Working Papers*, janvier 2008.

Emanuele Crocetti et Giorgia Randi : Using the Benford's Law as a First Step to Assess the Quality of the Cancer Registry Data. *Frontiers in Public Health*, 4, octobre 2016.

Francois Degeorge, Jayendu Patel et Richard Zeckhauser : Earnings management to exceed thresholds. *The Journal of Business*, janvier 1999.

Jean-Paul Delahaye : L'étonnante loi de benford, novembre 1999. URL `https://www.pourlascience.fr/sr/logique-calcul/letonnante-loi-de-benford-1358.php`.

Luc Devroye et László Györfi : *Nonparametric density estimation : the L1 view*. John Wiley, New York, 1985. ISBN 0471816469.

Andreas Diekmann : Not the First Digit ! Using Benford's Law to Detect Fraudulent Scientif ic Data. *Journal of Applied Statistics*, 34(3):321–329, avril 2007.

Philip D. Drake et Mark J. Nigrini : Computer assisted analytical procedures using Benford's Law. *Journal of Accounting Education*, 18(2):127–146, mars 2000. ISSN 0748-5751.

Gilles R. Ducharme : Consistent selection of the actual model in regression analysis. *Journal of Applied Statistics*, 24(5):549–558, octobre 1997.

Gilles R. Ducharme, Samuel Kaci et Credo Vovor-Dassu : Tests d'adéquations lisses pour la loi de Newcomb-Benford. *Mathématiques appliquées et stochastiques*, 3(1), 2020.

Pierre Duchesne et Pierre Lafaye De Micheaux : Computing the distribution of quadratic forms : Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, avril 2010.

Jean-Marie Dufour, Abdeljelil Farhat, Lucien Gardiol et Lynda Khalaf : Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal*, 1(1):C154–C173, juin 1998.

Cindy Durtschi, Hillison William et Pacini Carl : The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. *Journal Forensic Account*, 5:17–34, janvier 2004. ISSN 1524-5586.

Patrick J. Farrell et Katrina Rogers-Stewart : Comprehensive study of tests for normality and symmetry : extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, 76(9):803–816, septembre 2006.

Leslaw Gajek : On Improving Density Estimators which are not Bona Fide Functions. *The Annals of Statistics*, 14(4):1612–1618, 1986. URL `http://www.jstor.org/stable/2241494`.

Nicolas Gauvrit, Jean-Charles Houillon et Jean-Paul Delahaye : Generalized Benford's Law as a Lie Detector. *Advances in Cognitive Psychology*, 13(2):121–127, juin 2017.

Vincent Genest et Christian Genest : La loi de Newcomb-Benford ou la loi du premier chiffre significatif. *Bulletin Association Mathématique du Québec*, L1(2):22–39, 2011a.

Vincent Genest et Christian Genest : La loi de newcomb-benford ou la loi du premier chiffre significatif. *Bulletin Association Mathématique du Québec*, L1(2):22–39, 2011b.

A. Geyer et J. Martí : Applying Benford's law to volcanology. *Geology*, 40(4):327–330, avril 2012. ISSN 0091-7613.

Christina Lynn Geyer et Patricia Pepple Williamson : Detecting Fraud in Data Sets Using Benford's Law. *Communications in Statistics - Simulation and Computation*, 33 (1):229–246, janvier 2004.

K. Gore et Kherdekar : A beautiful normal distribution and its model. *International Journal of Advanced Education and Research*, 2:32–34, juillet 2017. ISSN 2455-5746.

Liming Guan, Daoping He et David Yang : Auditing, integral approach to quarterly reporting, and cosmetic earnings management. *Managerial Auditing Journal*, 21(6): 569–581, juillet 2006.

Stefan Günnel et Karl-Heinz Tödter : Does Benford's Law hold in economic research and forecasting ? *Empirica*, 36(3):273–292, août 2008.

Mark Handcock et Morris Martina : *Relative distribution methods in the social sciences*. Springer, New York, 1999. ISBN 0387987789.

J. Hein, R. Zobrist, C. Konrad et G. Schuepfer : Scientific fraud in 20 falsified anesthesia papers. *Der Anaesthesist*, 61(6):543–549, juin 2012.

Theodore P. Hill : Random-number guessing and the first digit phenomenon. *Psychological Reports*, 62(3):967–971, jun 1988.

Theodore P. Hill : A statistical derivation of the significant-digit law. *Statist. Sci.*, 10 (4):354–363, 1995. ISSN 0883-4237.

Theodore P. Hill : The First Digit Phenomenon : A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data. *American Scientist*, 86(4):358–363, 1998.

Werner Hürlimann : Generalizing Benford's Law Using Power Laws : Application to Integer Sequences. *International Journal of Mathematics and Mathematical Sciences*, 2009:1–10, septembre 2009.

T. Inglot, Wilbert C. M. Kallenberg et Teresa Ledwina : Power approximations to and power comparison of smooth goodness-of-fit tests. *Scandinavian journal of statistics*, (21):131–145, 1994. ISSN 0303-6898.

Tadeusz Inglot et Teresa Ledwina : Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Applications*, 417(1):124–133, aug 2006.

Dieter William Joenssen : A New Test for Benford's Distribution. *Abstract-Proceedings of the 3rdJoint Statistical Meeting DAGStat*, mars 2013a.

Dieter William Joenssen : Statistical Tests for Evaluating Conformity to Benford's Law. *The Comprehensive R Archive Network*, 2013b. URL `https://cran.r-project.org/package=BenfordTests`.

Dieter William Joenssen : Two digit testing for Benford's Law. *In Proceedings of the ISI World Statistics Congress, 59th Session in Hong Kong*, décembre 2013c. URL `http://www.statistics.gov.hk/wsc/CPS021-P2-S.pdf`.

Dieter William Joenssen : Testing for Benford's Law : A Monte Carlo Comparison of Methods. *SSRN Electronic Journal*, novembre 2014.

George Judge et Laura Schechter : Detecting Problems in Survey Data Using Benford's Law. *Journal of Human Resources*, 44(1):1–24, 2009.

Wilbert C. M. Kallenberg : The penalty in data driven Neyman's tests. *Mathematical methods of statistics*, 11:323–340, 2002. ISSN 1066-5307.

Wilbert C. M. Kallenberg et Teresa Ledwina : Data driven smooth tests for composite hypotheses comparison of powers. *Journal of Statistical Computation and Simulation*, 59(2):101–121, octobre 1997.

Juha Kinnunen et Markku Koskela : Who Is Miss World in Cosmetic Earnings Management ? A Cross-National Comparison of Small Upward Rounding of Net Income Numbers among Eighteen Countries. *Journal of International Accounting Research*, 2 (1):39–68, janvier 2003.

Donald Knuth : *The art of computer programming*. Addison-Wesley Pub. Co, Reading, Mass, 1973. ISBN 0-201-89684-2.

Marie-Paule Lecoutre : Cognitive models and problem spaces in ?purely random ? situations. *Educational Studies in Mathematics*, 23(6):557–568, septembre 1992.

Teresa Ledwina : Data-Driven Version of Neyman's Smooth Test of Fit. *Journal of the American Statistical Association*, 89(427):1000–1005, septembre 1994.

Teresa Ledwina et Grzegorz Wyłupek : Detection of non-Gaussianity. *Journal of Statistical Computation and Simulation*, 85(17):3480–3497, nov 2014.

Joanne Lee, Wendy K. Tam Cho et George G. Judge : Stigler's approach to recovering the distribution of first significant digits in natural data sets. *Statistics & Probability Letters*, 80(2):82–88, janvier 2010.

Lawrence M. Leemis, Bruce W. Schmeiser et Diane L. Evans : Survival Distributions Satisfying Benford's Law. *The American Statistician*, 54(4):236–241, novembre 2000.

M. Lesperance, W. J. Reed, M. A. Stephens, C. Tsao et B. Wilton : Assessing Conformance with Benford's Law : Goodness-Of-Fit Tests and Simultaneous Confidence Intervals. *PLOS ONE*, 11(3):e0151235, mars 2016.

Eduardo Ley : On the Peculiar Distribution of the U.S. Stock Indexes' Digits. *The American Statistician*, 50(4):311, novembre 1996.

Huan Liu, Yongqiang Tang et Hao Helen Zhang : A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, février 2009. ISSN 0167-9473.

Daniel Manurung et Niki Hadian : Detection Fraud of Financial Statement with Fraud Triangle. 2013.

Fernando Marmolejo-Ramos et Jorge González-Burgos : A Power Comparison of Various Tests of Univariate Normality on Ex-Gaussian Distributions. *Methodology*, 9 (4):137–149, janvier 2013.

N. MEUSNIER : Dr Arbuthnot et Mr Hidden. Mathématiques, Providence Divine et Petite Vérole : sur le probable au début du XVIIIe siècle en Angleterre. *Centre d´analyse et de mathématiques sociales*, (162), janvier 1999.

Steven Miller : *Benford's Law*. Princeton University Press, mai 2015. ISBN 0691147612.

Tariq Ahmad Mir et Marcel Ausloos : Benford's law : A “sleeping beauty” sleeping in the dirty pages of logarithmic tables. *Journal of the Association for Information Science and Technology*, 69(3):349–358, septembre 2017.

John Morrow : Benford's Law, Families of Distributions and a Test Basis. *Centre for Economic Performance, LSE*, (dp1291), août 2014.

Simon Newcomb : Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*, 4:39–40, janvier 1881.

J. Neyman : «Smooth test» for goodness of fit. *Scandinavian Actuarial Journal*, 1937 (3-4):149–199, juillet 1937.

Mark J. Nigrini : *The detection of income tax evasion through an analysis of digital distributions*. phdthesis, University of Cincinnati, 1993.

Mark J. Nigrini : A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association*, 18(1):72–91, avril 1996. ISSN 0198-9073.

Mark J. Nigrini : An Assessment of the Change in the Incidence of Earnings Management Around the Enron-Andersen Episode. *Review of Accounting and Finance*, 4(1):92–110, janvier 2005.

Mark J. Nigrini et Steven J. Miller : Benford's Law Applied to Hydrology Data—Results and Relevance to Other Geophysical Data. *Mathematical Geology*, 39 (5):469–490, août 2007.

Mark J. Nigrini et Linda J. Mittermaier : The Use of Benford's Law as an Aid in Analytical Procedures. *Auditing-a Journal of Practice & Theory*, septembre 1997.

Jyrki Niskanen et Matti Keloharju : Earnings cosmetics in a tax-driven accounting environment : evidence from Finnish public firms. *European Accounting Review*, 9 (3):443–452, septembre 2000.

Hadi Alizadeh Noughabi et Naser Reza Arghami : Monte Carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation*, 81(8):965–972, août 2011.

Emanuel Parzen : Quantile Probability and Statistical Data Modeling. *Statistical Science*, 19(4):652–662, novembre 2004.

Karl Pearson : X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, juillet 1900.

Luis Pericchi et David Torres : Quick Anomaly Detection by the Newcomb–Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela. *Statistical Science*, 26(4):502–516, novembre 2011.

L. Pietronero, E. Tosatti, V. Tosatti et A. Vespignani : Explaining the uneven distribution of numbers in nature : the laws of Benford and Zipf. *Physica A : Statistical Mechanics and its Applications*, 293(1-2):297–304, avril 2001. ISSN 0378-4371.

H. Poincaré et A. Quiquet : *Calcul des probabilités*. Cours de la Faculté des Sciences de Paris : Cours de physique mathématique. Gauthier-Villars, 1912. URL `https://books.google.fr/books?id=SPJLAAAAMAAJ`.

Peter N. Posch : A survey on sequences and distribution functions satisfying the first-digit-law. *Journal of Statistics and Management Systems*, 11(1):1–19, janvier 2008.

J. C. W. Rayner et D. J. Best : Smooth Tests of Goodness of Fit : An Overview. *International Statistical Review / Revue Internationale de Statistique*, 58(1):9, avril 1990.

J. C. W. Rayner, O. Thas, D. J. Best, A. Walter, Shewhart et S. Samuel Wilks : *Smooth tests of goodness of fit using R*. John Wiley & Sons (Asia), Singapore Hoboken, NJ, 2009. ISBN 9780470824443.

Ricardo J. Rodriguez : First Significant Digit Patterns From Mixtures of Uniform Distributions. *The American Statistician*, 58(1):64–71, février 2004.

Malcolm Sambridge, Hrvoje Tkalcic et Pierre Arroucau : Benford's Law of First Digits : From Mathematical Curiosity to Change Detector. *Asia Pacific Mathematics Newsletter(APMN)*, 1:1–5, janvier 2011.

G. Schüpfer, J. Hein, M. Casutt, L. Steiner et C. Konrad : Vom Finanz- zum Wissenschaftsbetrug. *Der Anaesthesist*, 61(6):537–542, juin 2012.

Gideon Schwarz : Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2), mar 1978.

Paul Scott et Maria Fasli : Benford's law : An empirical investigation and a novel explanation. Csm technical report 349, Department of Computer Science, University of Essex, UK, 2001.

Pavithraa Seenivasan, Soumya Easwaran, Seshan Sridhar et Sitabhra Sinha : Using Skewness and the First-Digit Phenomenon to Identify Dynamical Transitions in Cardiac Models. *Frontiers in Physiology*, 6, janvier 2016.

Christopher J. Skousen, Liming Guan et T. Sterling Wetzel : Anomalies and Unusual Patterns in Reported Earnings : Japanese Managers Round Earnings. *Journal of International Financial Management and Accounting*, 15(3):212–234, octobre 2004.

Olivier Thas : *Comparing Distributions*. Springer New York, 2010. ISBN 978-0-387-92709-1.

Jacob K. Thomas : Unusual Patterns in Reported Earnings. *The Accounting Review*, 64(4):773–787, octobre 1989.

Karl-Heinz Tödter : Benford's Law as an Indicator of Fraud in Economics. *German Economic Review*, 10(3):339–351, août 2009.

H. Varian : *Benford's Law*, volume 26. 1972.

Wanda A. Wallace : Assessing the Quality of Data Used for Benchmarking and Decision-Making - Benford's Law, using Excel, provides an initial assessment of data quality. This "primer" using governmental data can benefit educators and practitioners alike., 2002. ISSN 0883-1483.

Christoph Watrin, Ralf Struffert et Robert Ullmann : Benford's law : an instrument for selecting tax audit targets ? *Review of Managerial Science*, 2(3):219–237, may 2008.

S. Wong : Testing BenfordĹaw with the first two significant digits. 2010.

B. W. Yap et C. H. Sim : Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155, décembre 2011.

Berna Yazici et Senay Yolacan : A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2):175–183, février 2007.

Annexe A

Les alternatives « Testing »

Tel que annoncé dans le dernier paragraphe de la section 3.3.4, on trouvera ici les graphiques des courbes de puissances liés aux simulations appliquées aux alternatives « Testing ».

A.1 Calibrage du c

Les graphiques des courbes de puissances sur les alternatives « Testing » dans le calibrage de c effectué à la section 3.3.5 sont présentés ci dessous.
Le lecteur remarquera que les conclusions tirées des alternatives « Training » à la section 3.3.5 dans le dernier paragraphe tiennent pour les alternatives « Testing » à savoir la valeur de c faisant compromis est **1.3**

Famille des mixtures: Benford Hill

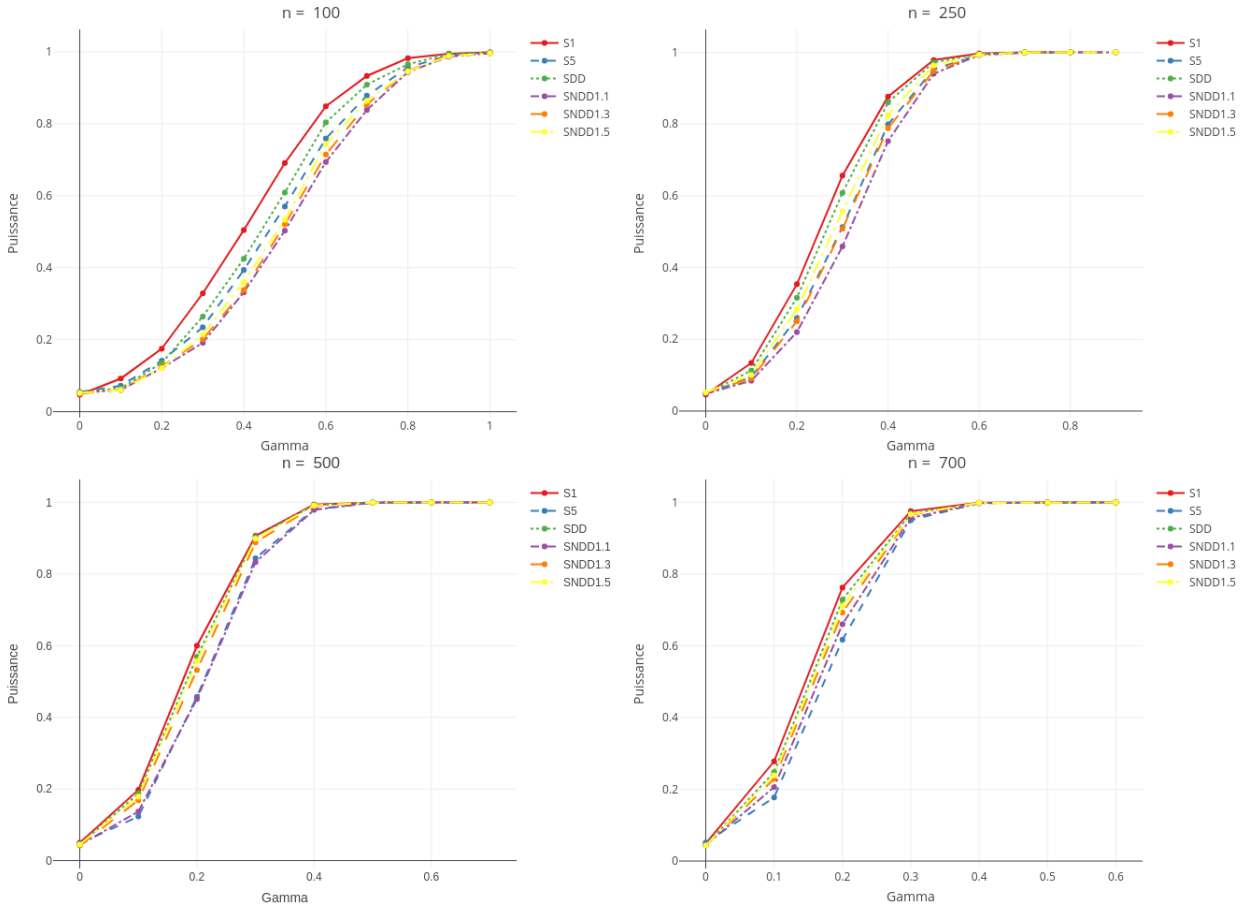

FIGURE A.1.1 – Mixture Benford Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{H}_{(1,2)}$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des mixtures: Benford Stigler

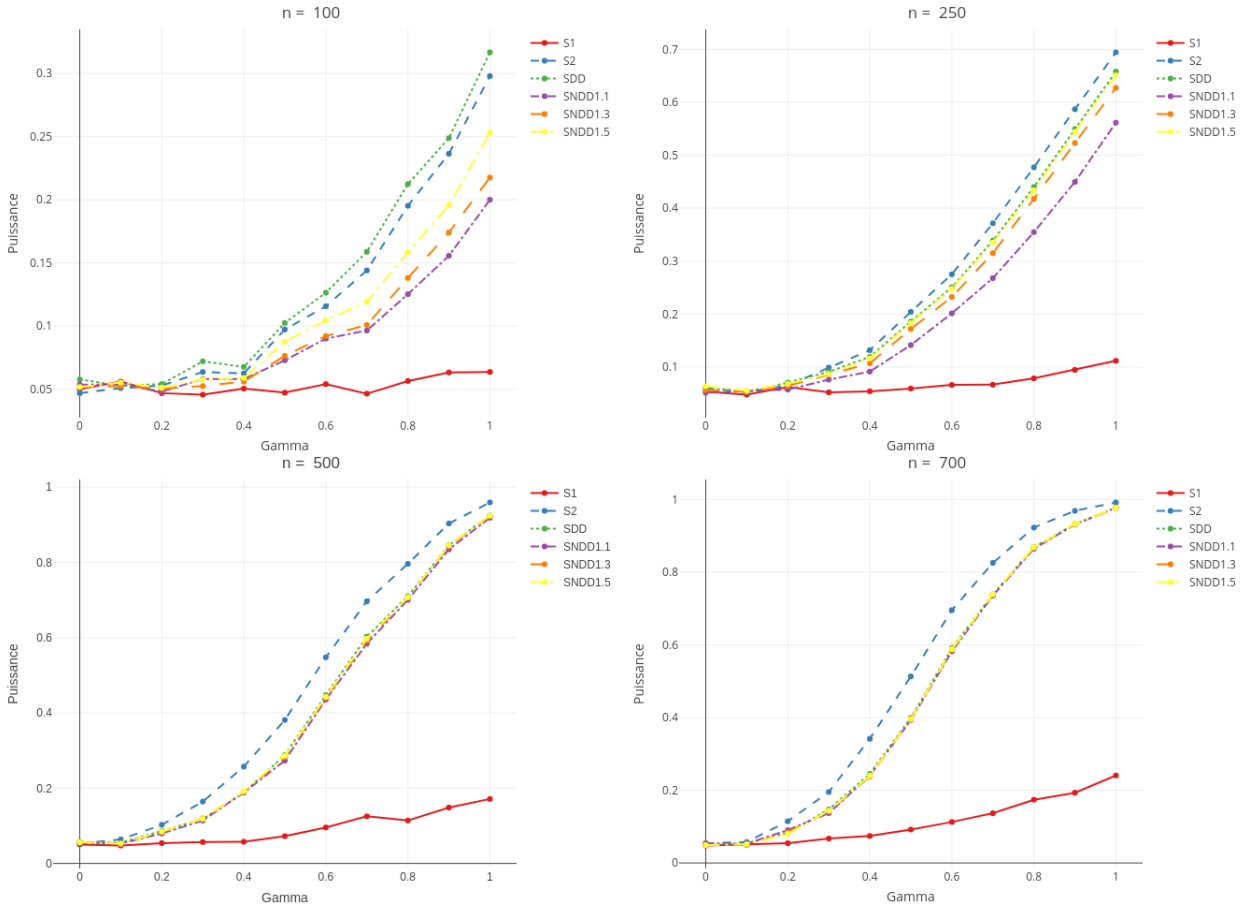

FIGURE A.1.2 – Mixture Benford Stigler $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma\mathcal{S}_{(1,2)}$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1,...,5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1,...,5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des mixtures: Benford Uniforme Stigler

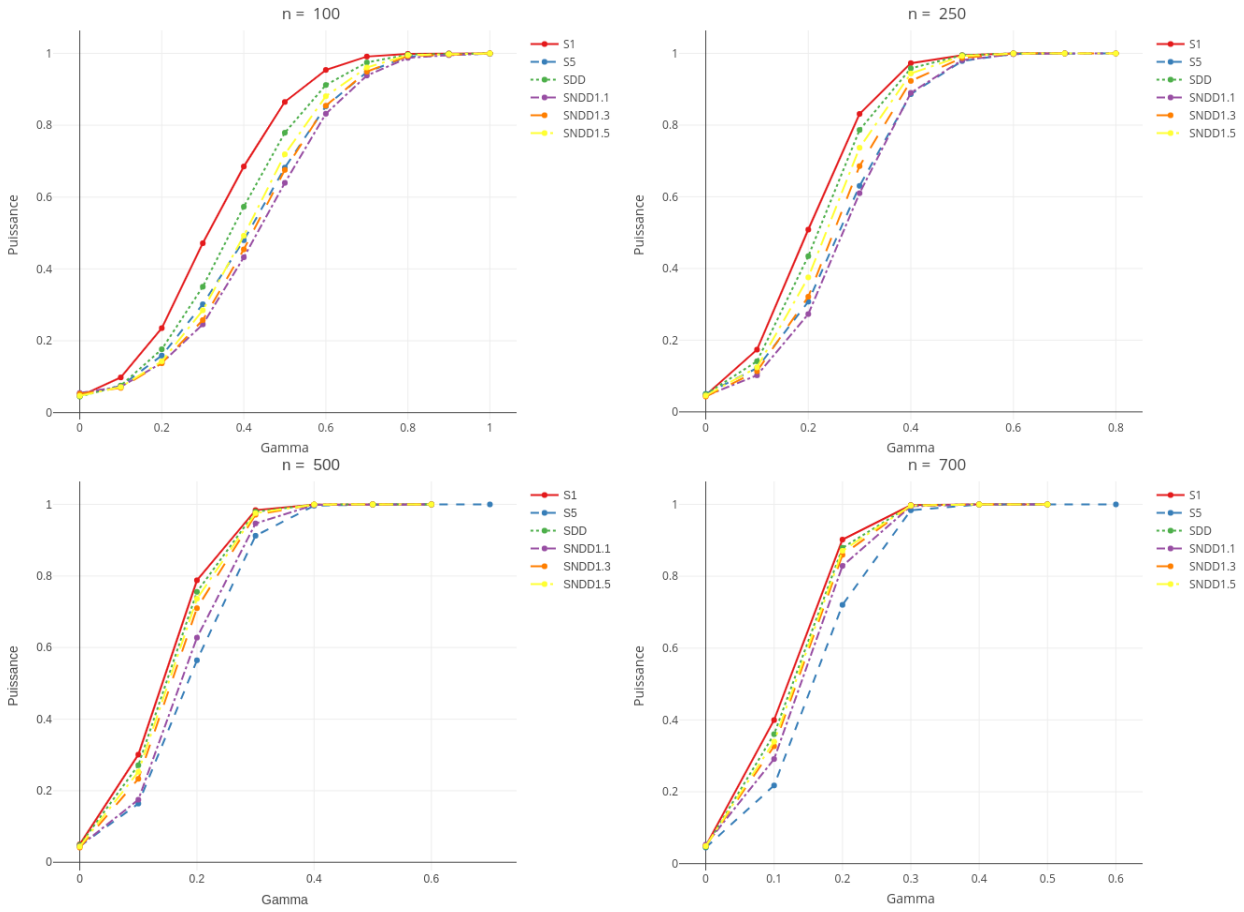

FIGURE A.1.3 – Mixture Benford Uniforme Stigler $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, U_{[1,9]} \otimes S_2$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des mixtures: Benford Hill Uniforme

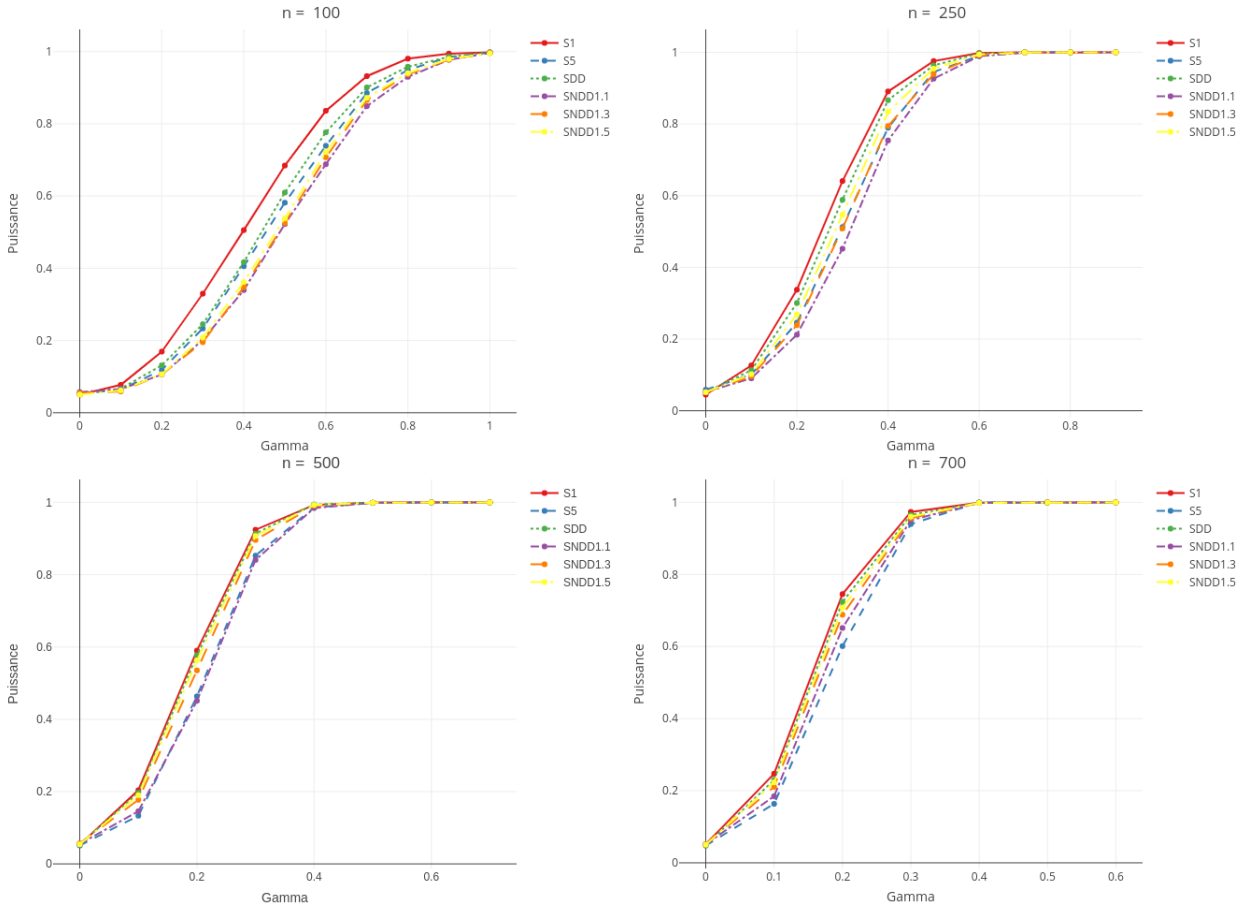

FIGURE A.1.4 – Mixture Benford Hill Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\,\mathcal{H}_1 \otimes U_{[0,9]}$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des mixtures: Benford Uniforme Hill

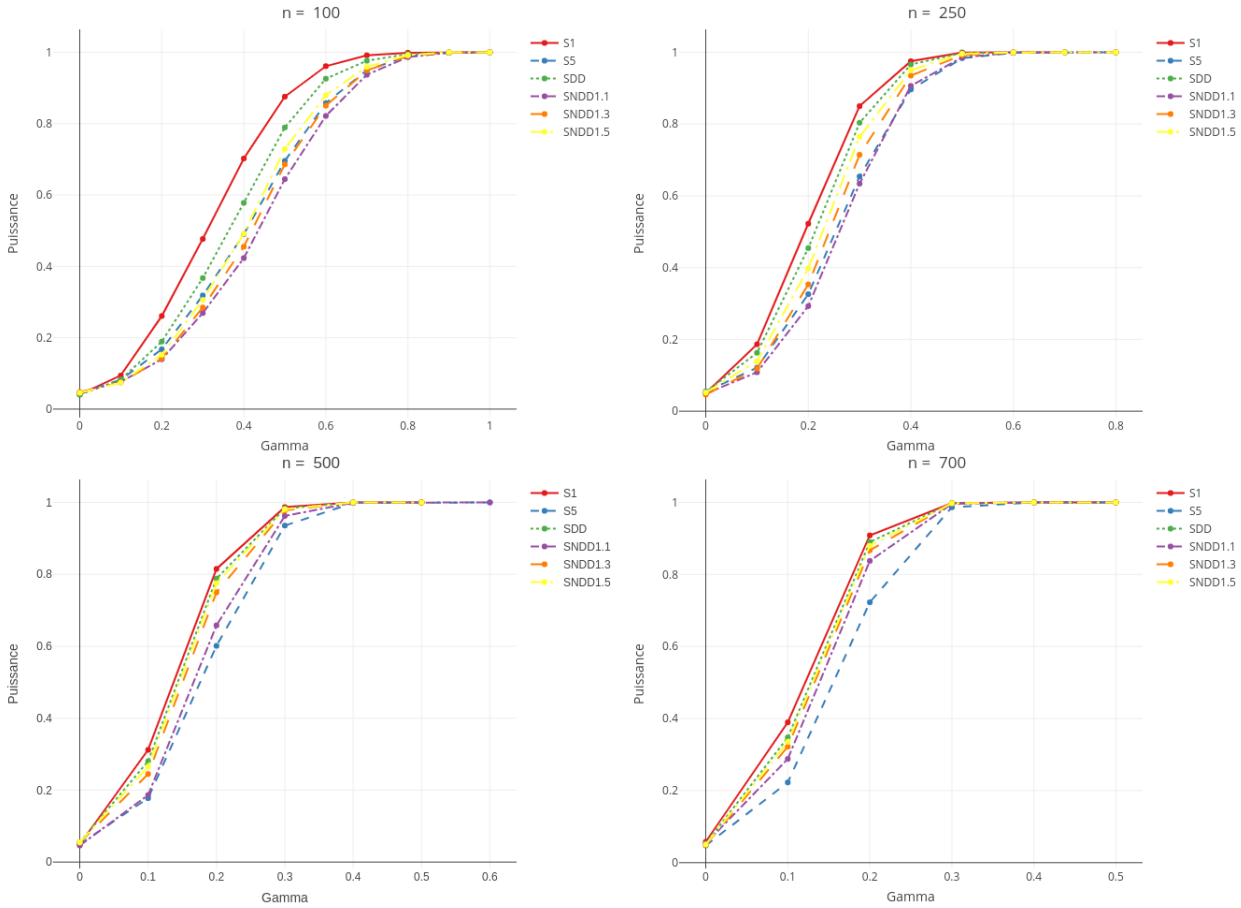

FIGURE A.1.5 – Mixture Benford Uniforme Hill $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma(1-\gamma)U_{[1,9]}\otimes\mathcal{H}_2$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des Copules: Benford Stigler

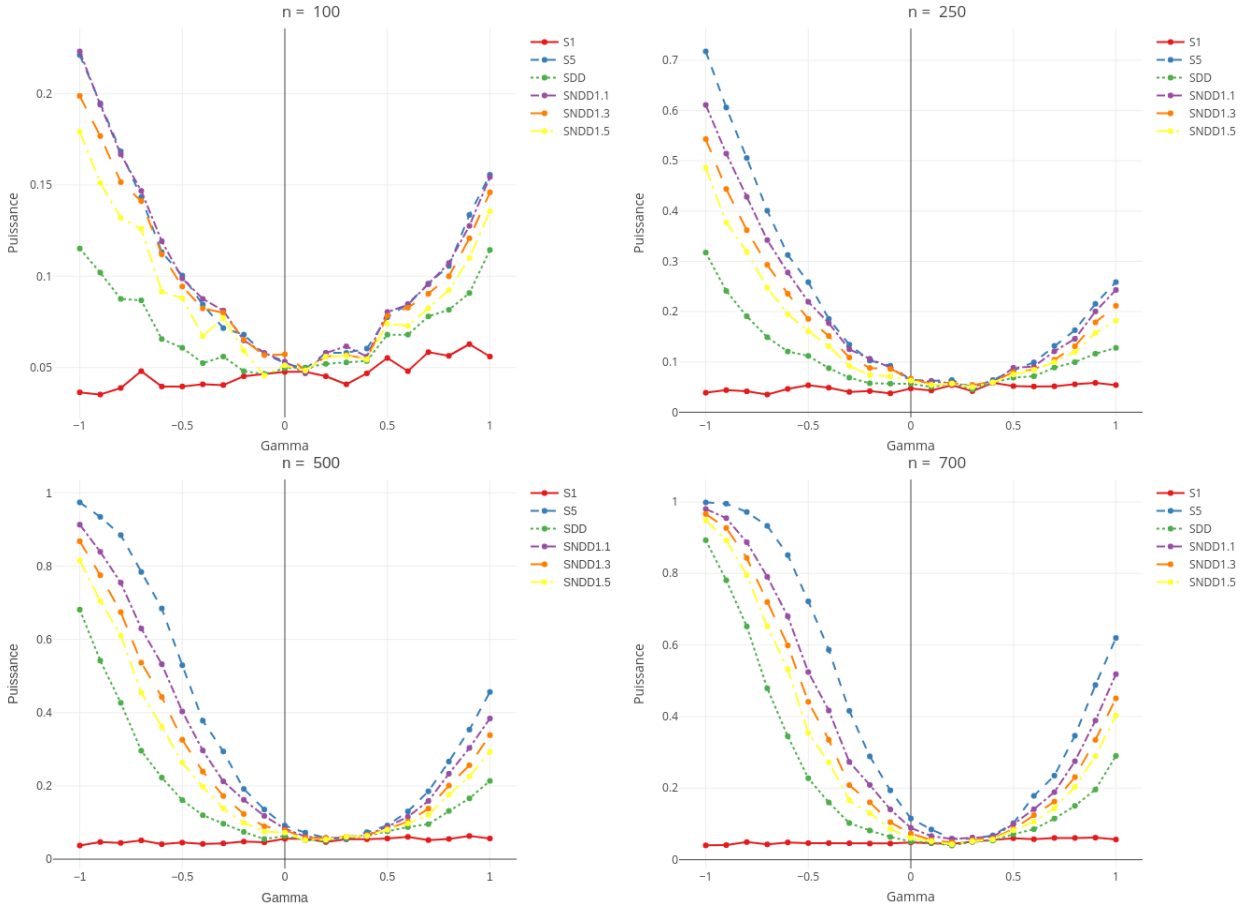

FIGURE A.1.6 – Copule Benford Stigler $C(\gamma, \mathscr{B}_1, S_2)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des Copules: Stigler Benford

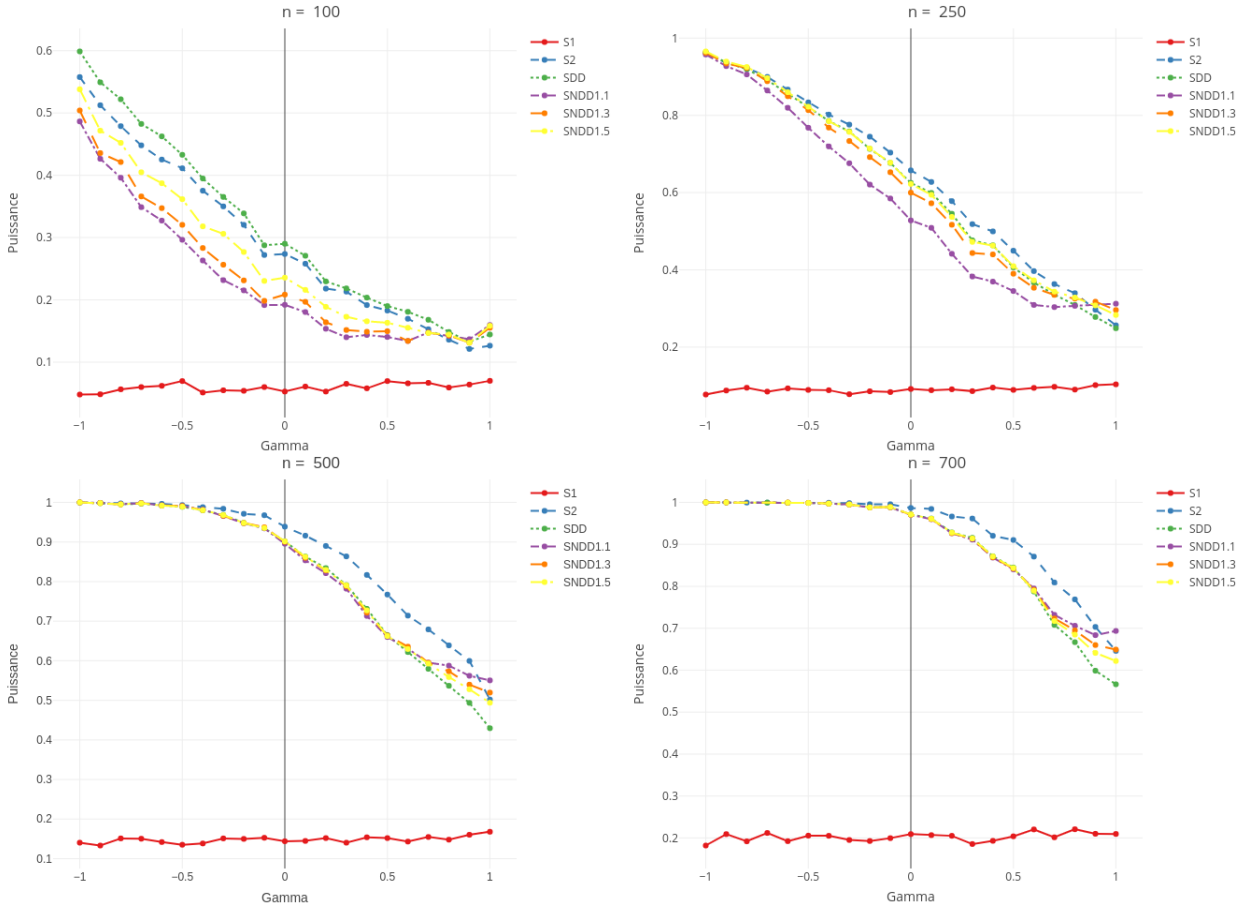

FIGURE A.1.7 – Copule Stigler Benford $C(\gamma, \mathcal{S}_1, \mathscr{B}_2)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille Indépendance: Benford Newcomb-Benford Généralisée

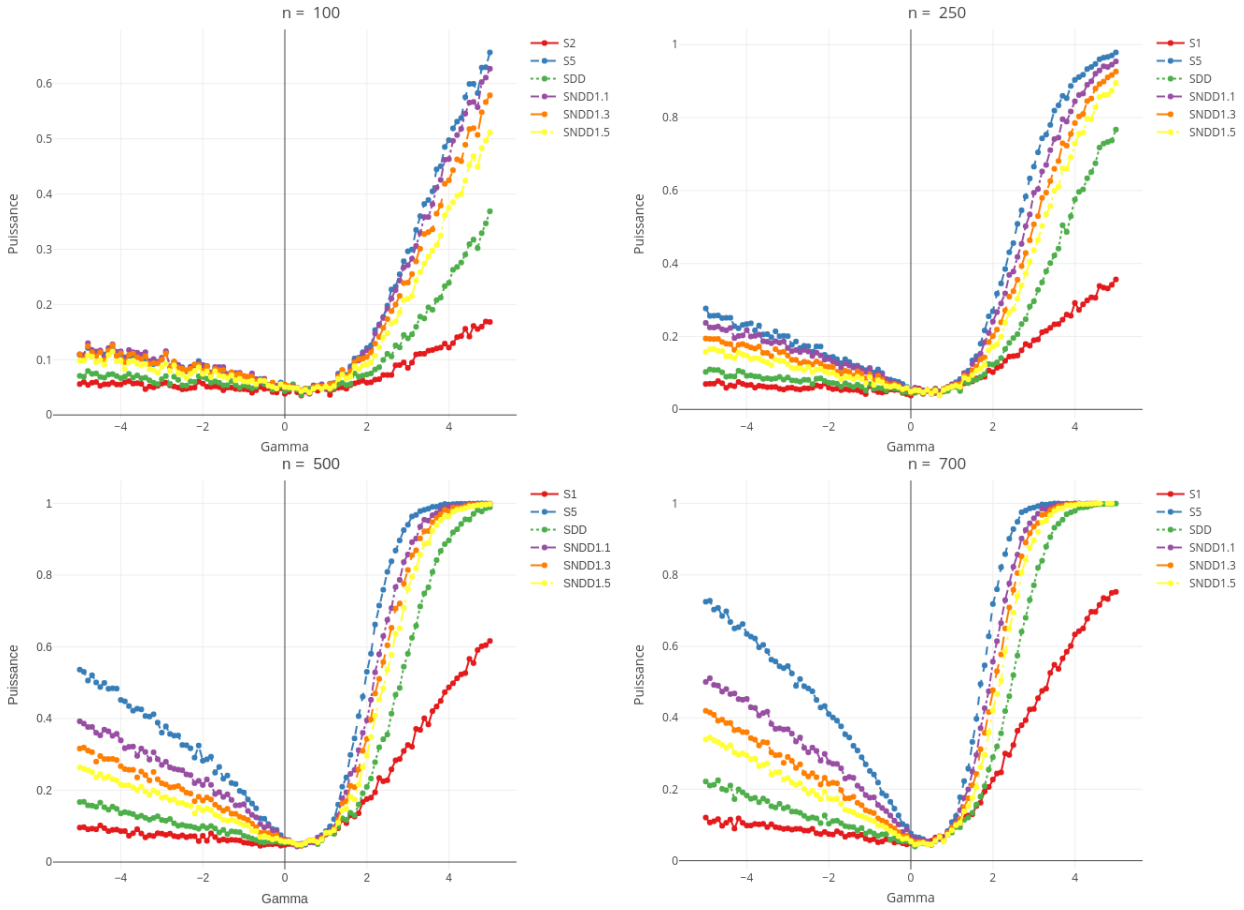

FIGURE A.1.8 – Indépendance Benford Newcomb-Benford Généralisée ($\mathcal{B}_1 \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille Indépendance: Newcomb-Benford Généralisée Newcomb-Benford Généralisée

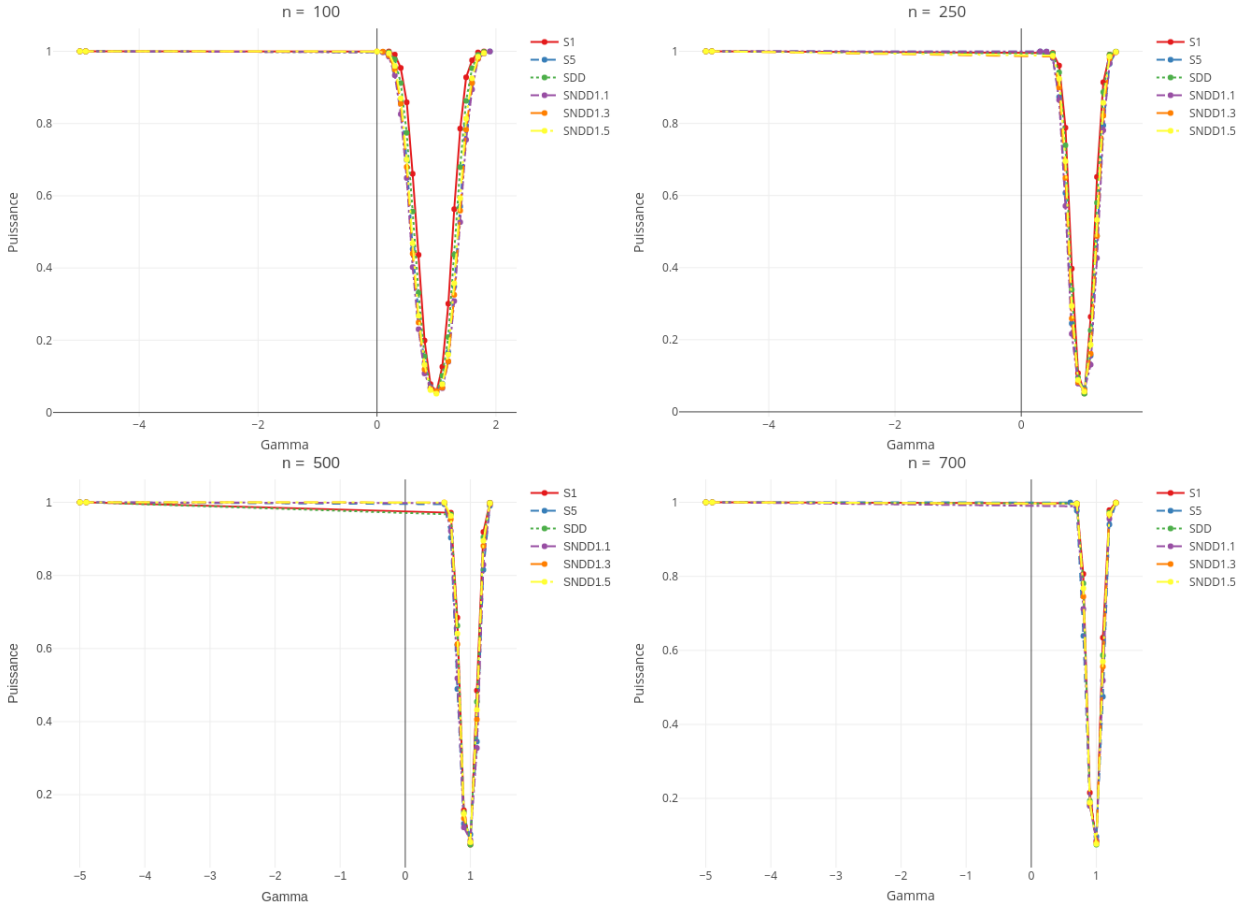

FIGURE A.1.9 – Indépendance Newcomb-Benford Généralisée Newcomb-Benford Généralisée ($\mathcal{GB}_1(\gamma) \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des conditionnelles: Rodriguez sachant Rodriguez

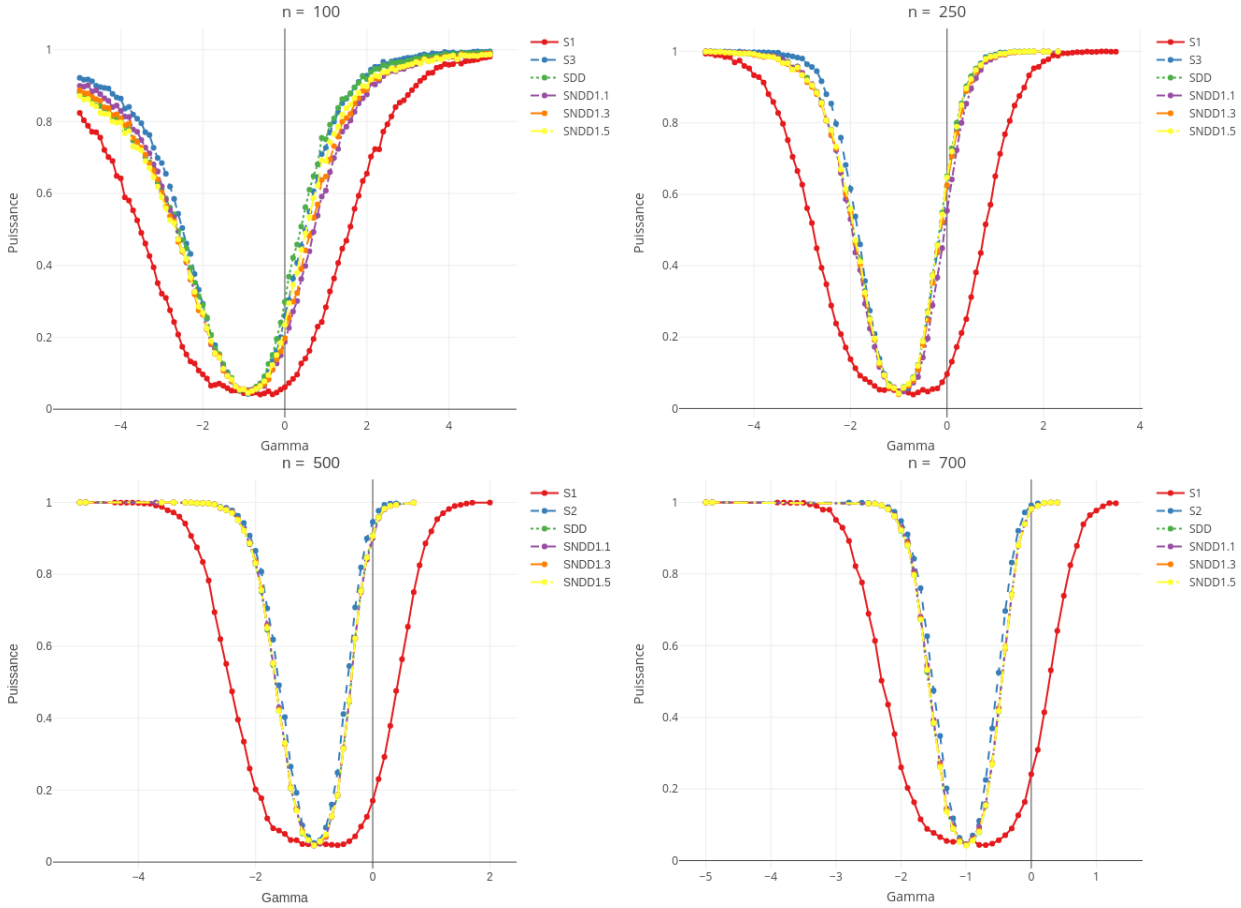

FIGURE A.1.10 – Rodriguez sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Benford

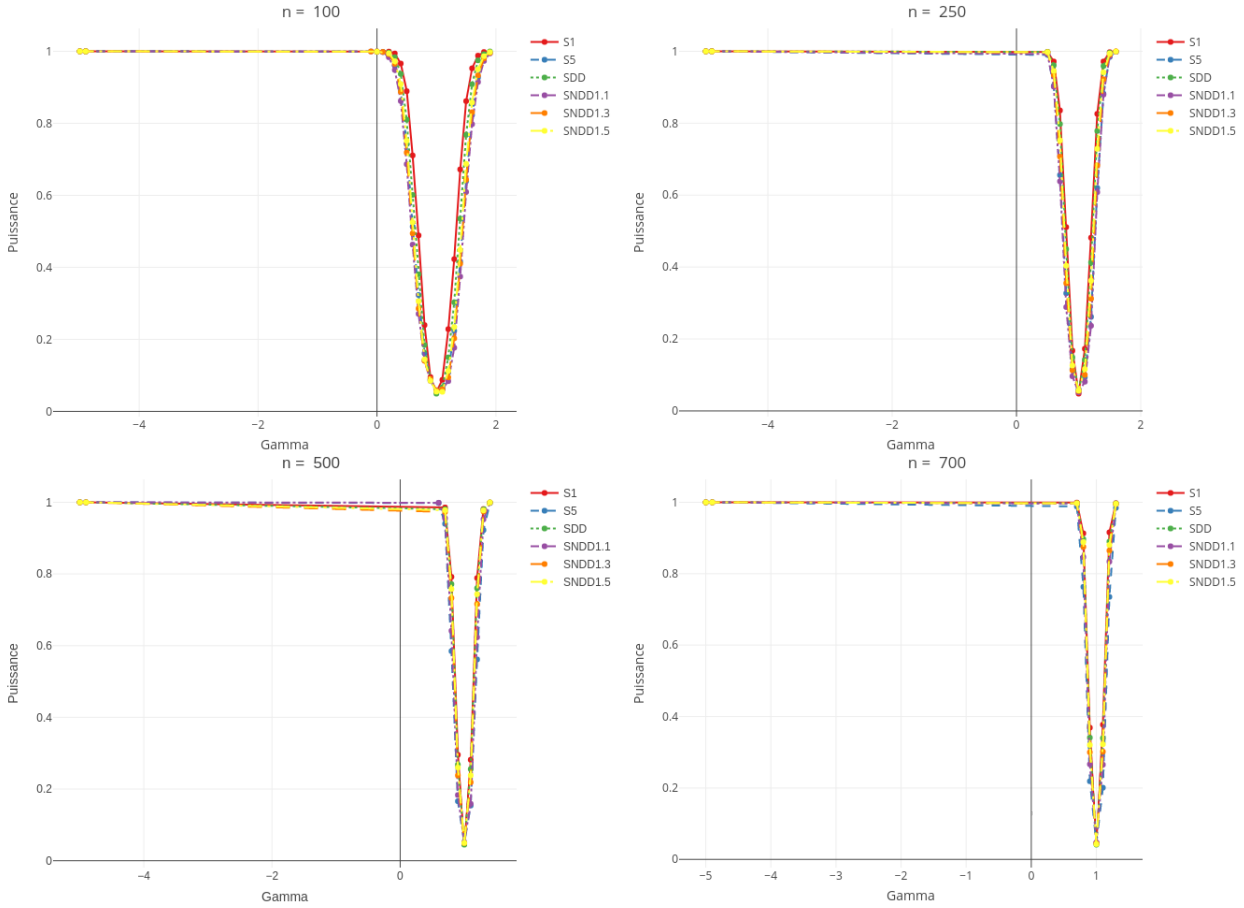

FIGURE A.1.11 – Newcomb-Benford Généralisée sachant Benford $\left(\mathcal{GB}_1(\gamma), \mathscr{B}_{(2|1)}\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des conditionnelles: Benford sachant Newcomb-Benford Généralisée

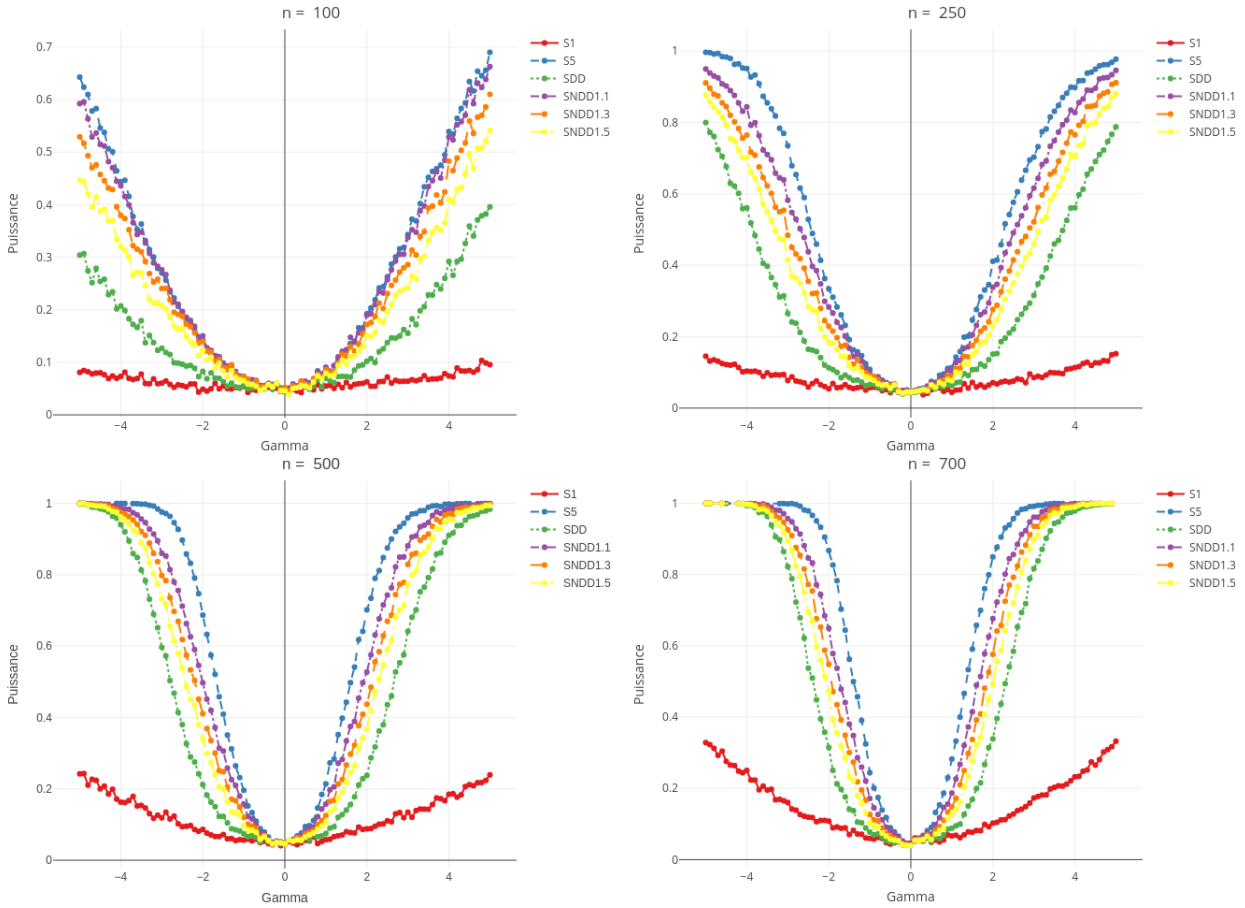

FIGURE A.1.12 – Benford sachant Newcomb-Benford Généralisée $\left(\mathscr{B}_1, \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée

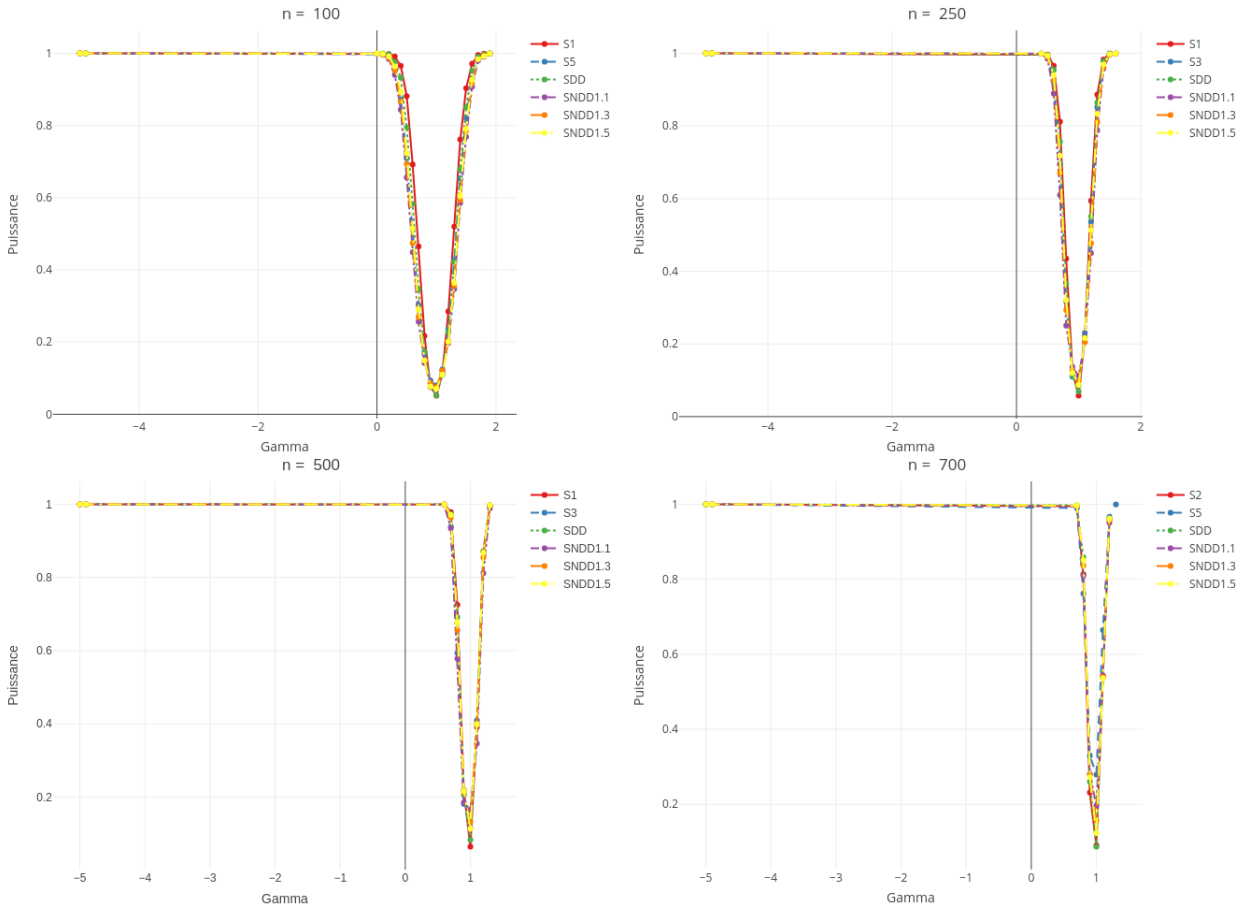

FIGURE A.1.13 – Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée $\left(\mathcal{GB}_1(\gamma), \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de loi Newcomb - Benford $\mathscr{B}_{(1,2)}$. Les tests représentés sont : le meilleur entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur rouge), $\mathcal{S}_{DD}$(couleur verte) , le pire entre $\mathcal{S}_i$, $i \in \{1, ..., 5\}$ (couleur bleue), $\mathcal{S}_{NDD1,1}$ avec $c = 1.1$ (couleur violette) , $\mathcal{S}_{NDD1,3}$ avec $c = 1.3$ (couleur orange) et $\mathcal{S}_{NDD1,5}$ avec $c = 1.5$ (couleur jaune), dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions. Lorsque $\gamma = 0$, on retombe sur la loi $\mathscr{B}_{(1,2)}$ sous l'hypothèse H_0.

A.2 Comparaison du test $\mathcal{S}_{NDD}$ et des tests classiques

Les graphiques des courbes de puissances sur les alternatives « Testing » dans le cadre de la comparaison du test $\mathcal{S}_{NDD}$ développé dans le cadre de nos travaux à la section 3.3.6 sont présentés ci dessous.

Le lecteur remarquera que les conclusions tirées des alternatives « Training » à la section 3.3.6 tiennent pour les alternatives « Testing » à savoir si W^2 ou U^2 sont plus puissants que le test du χ^2, $\{\mathcal{S}_{NDD}, c = 1.3\}$ offre un bon compromis. Par contre si le test du χ^2

est le plus puissant des tests classiques considérés ici, alors le test $\mathcal{S}_{NDD}$ ne devrait pas être utilisé.

Famille des mixtures: Benford Hill

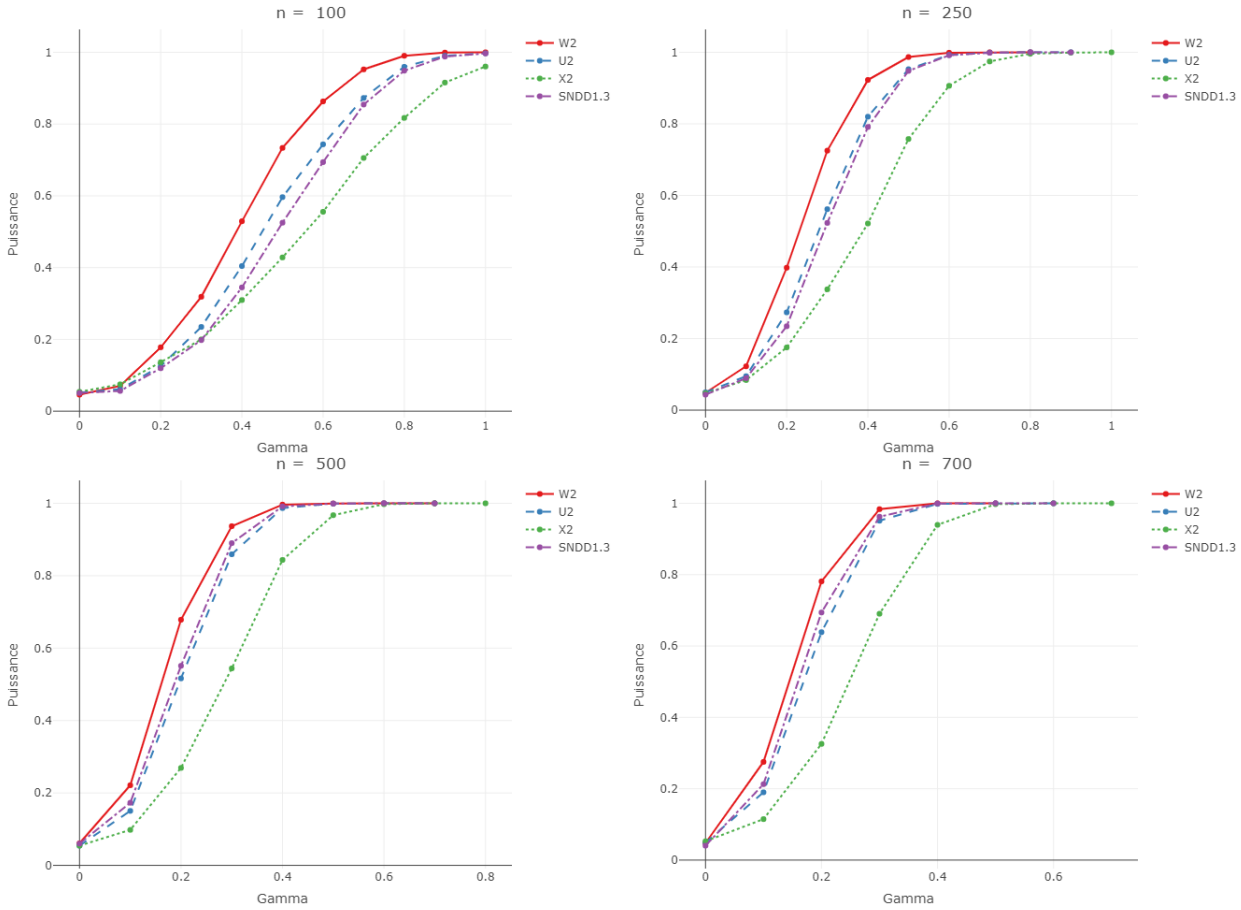

FIGURE A.2.1 – Mixture Benford Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{H}_{(1,2)}$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Stigler

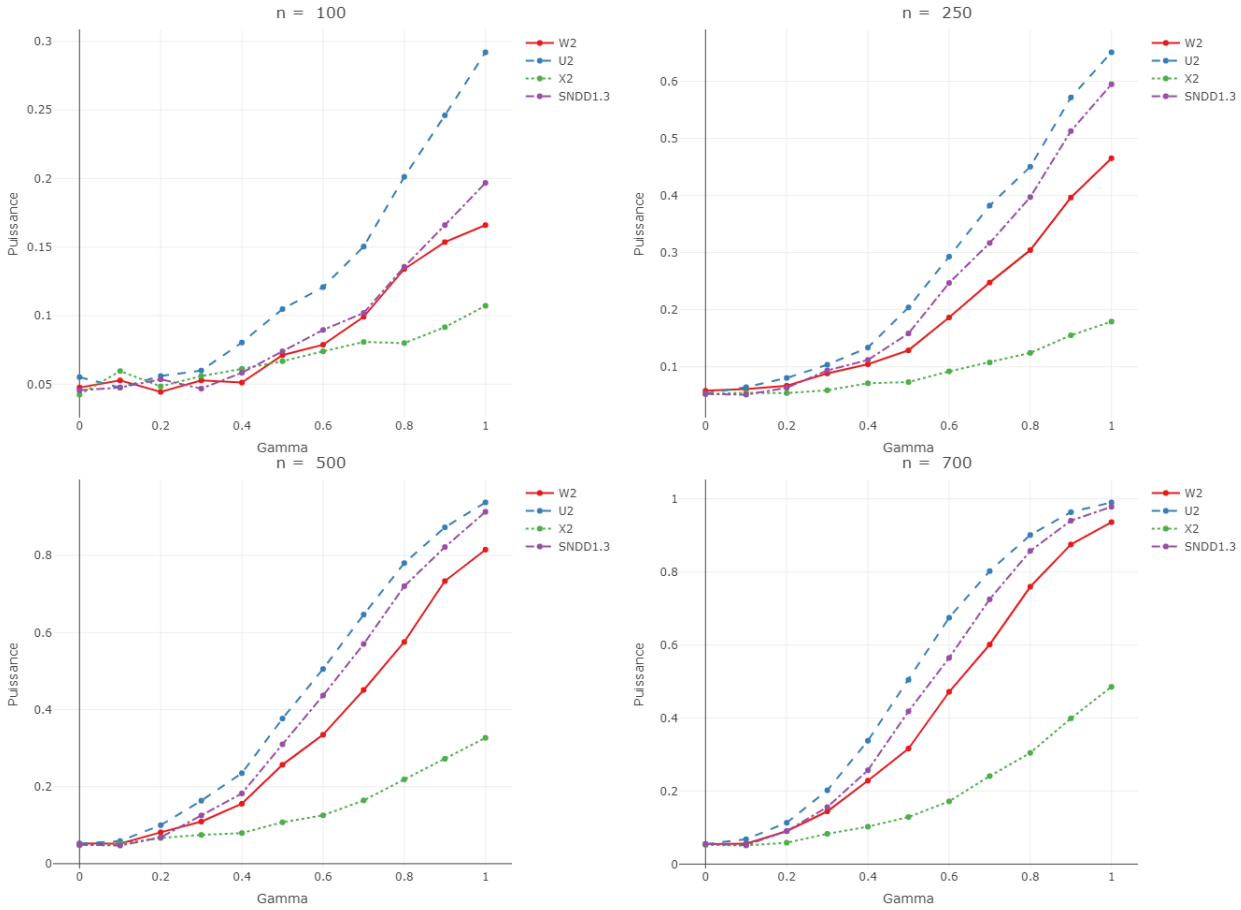

FIGURE A.2.2 – Mixture Benford Stigler $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{S}_{(1,2)}$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Uniforme Stigler

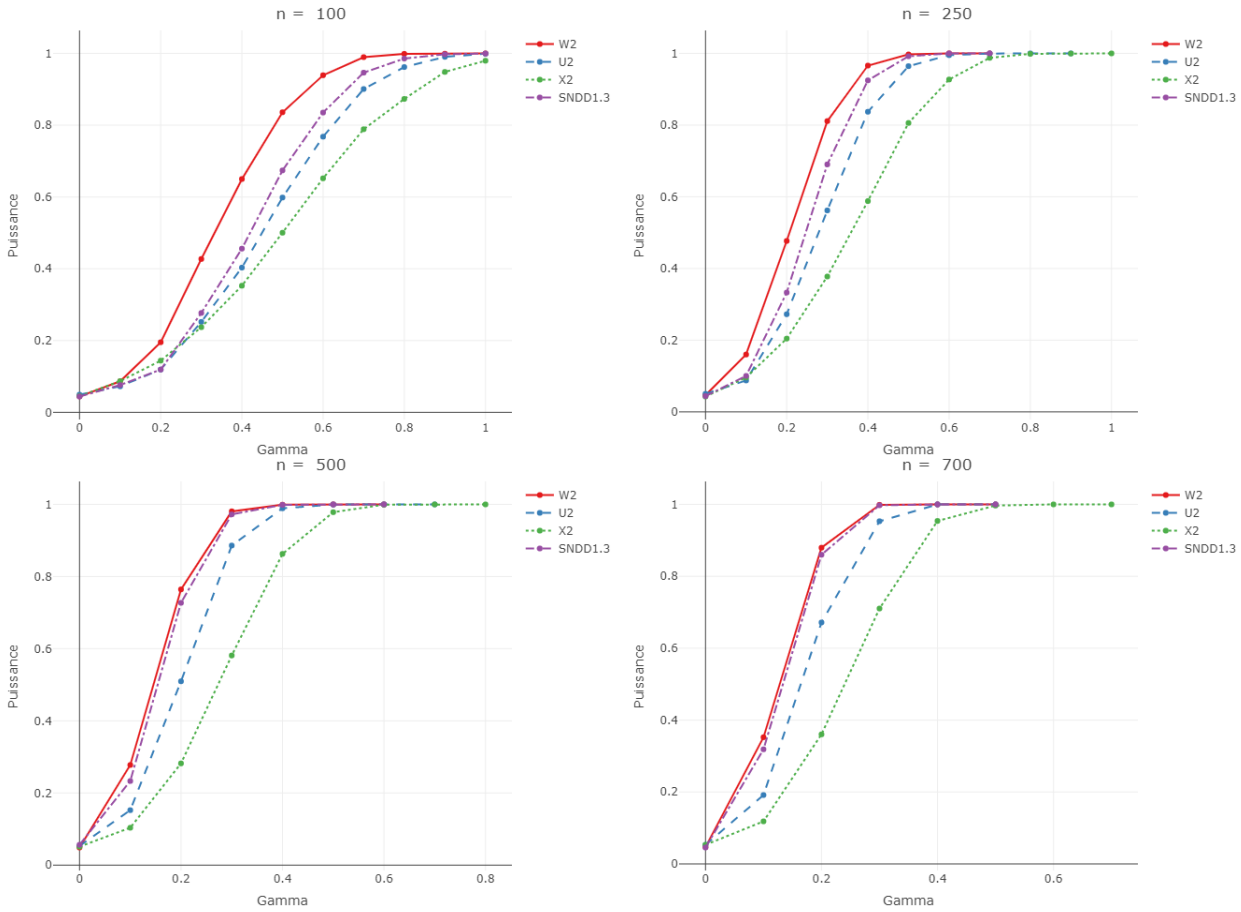

FIGURE A.2.3 – Mixture Benford Uniforme Stigler $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, U_{[1,9]} \otimes S_2$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Hill Uniforme

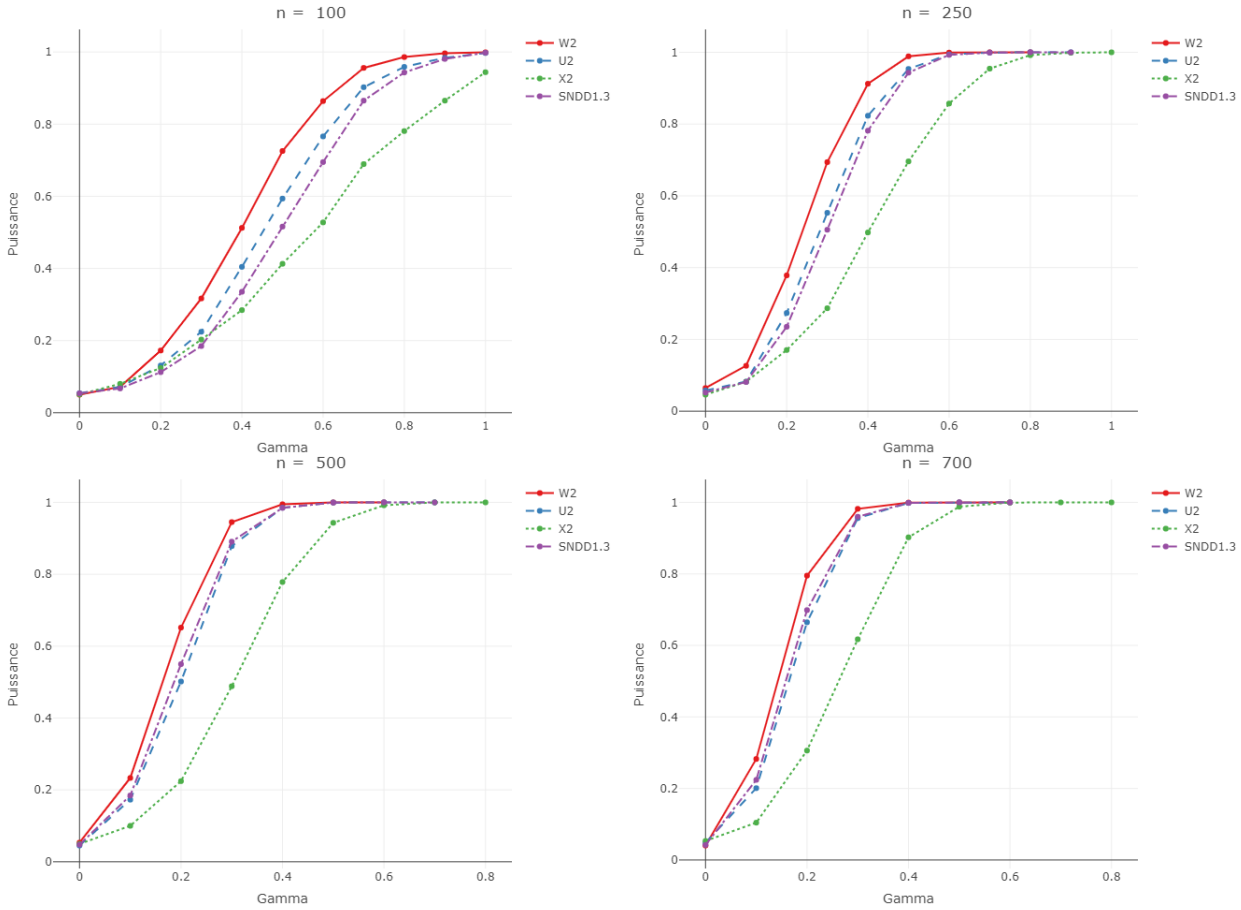

FIGURE A.2.4 – Mixture Benford Hill Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, \mathcal{H}_1 \otimes U_{[0,9]}$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Uniforme Hill

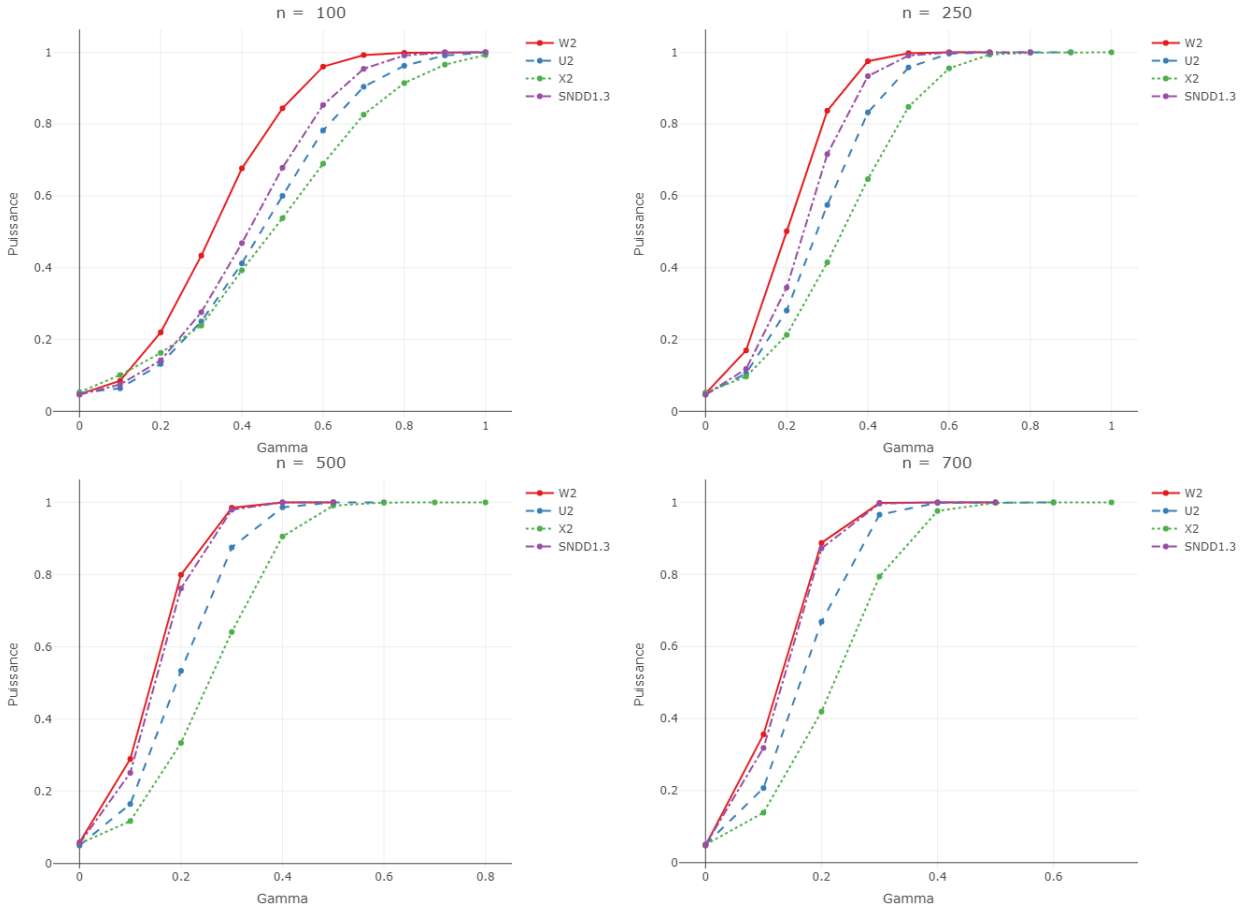

FIGURE A.2.5 – Mixture Benford Uniforme Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma(1-\gamma)\, U_{[1,9]} \otimes \mathcal{H}_2$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des Copules: Benford Stigler

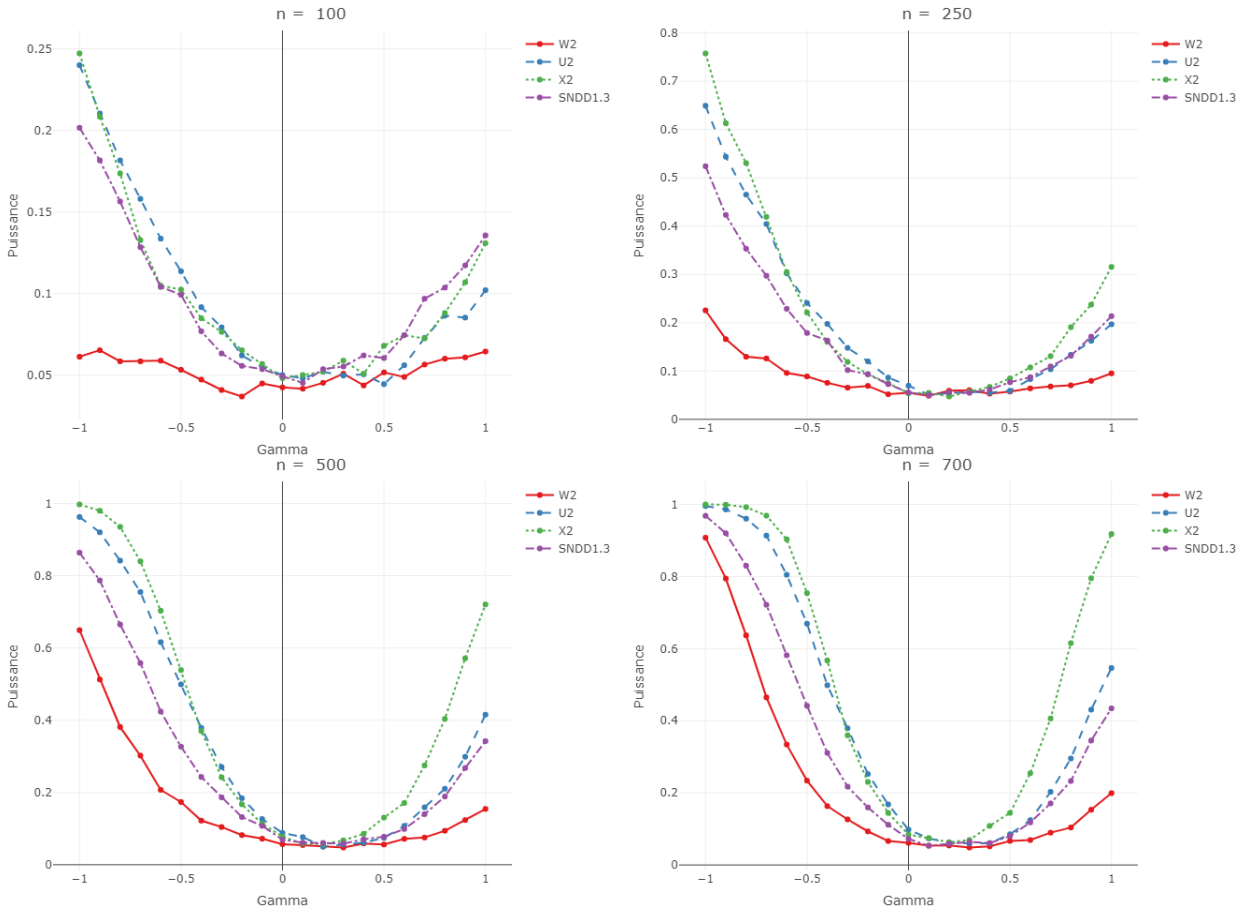

FIGURE A.2.6 – Copule Benford Stigler $C(\gamma, \mathscr{B}_1, S_2)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des Copules: Stigler Benford

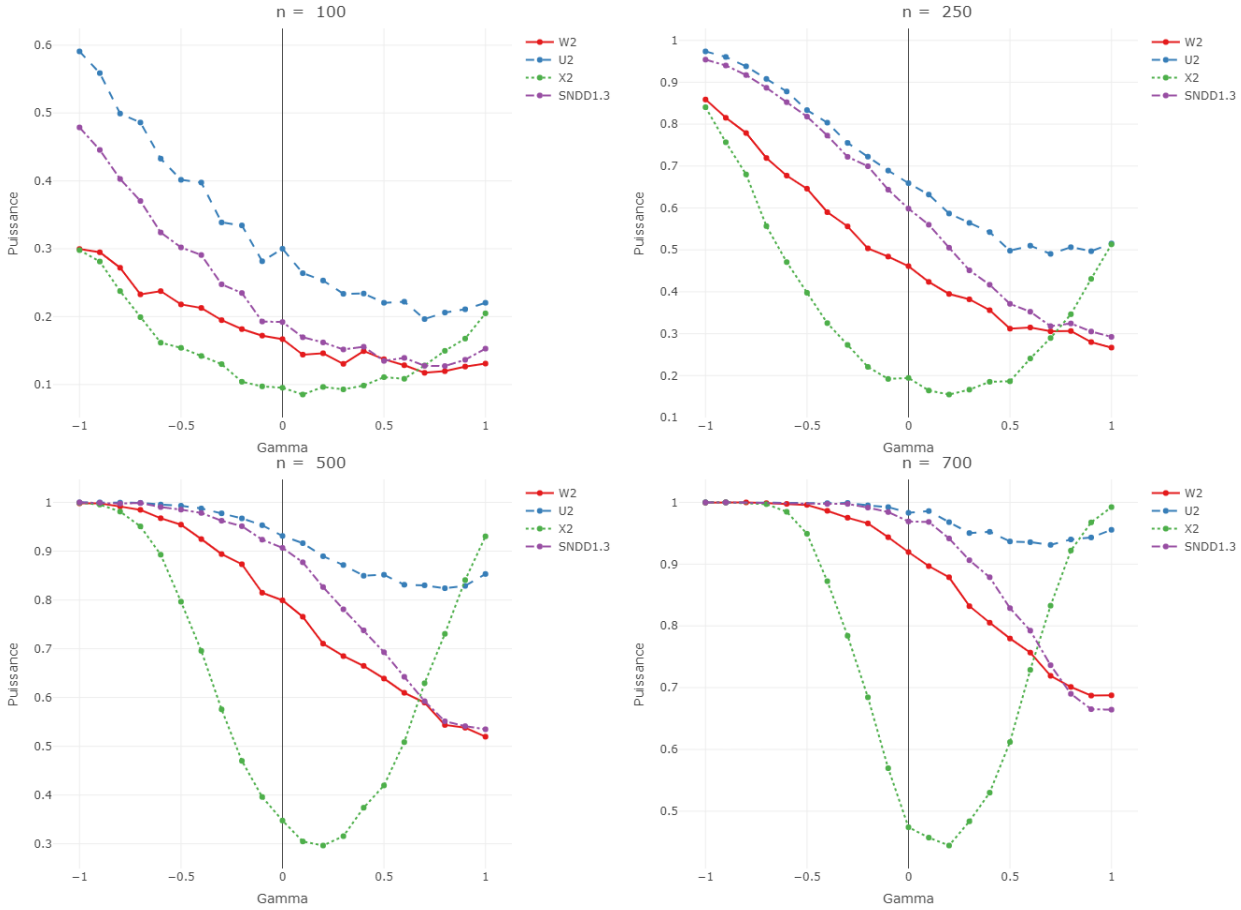

FIGURE A.2.7 – Copule Stigler Benford $C(\gamma, \mathcal{S}_1, \mathscr{B}_2)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Benford Newcomb-Benford Généralisée

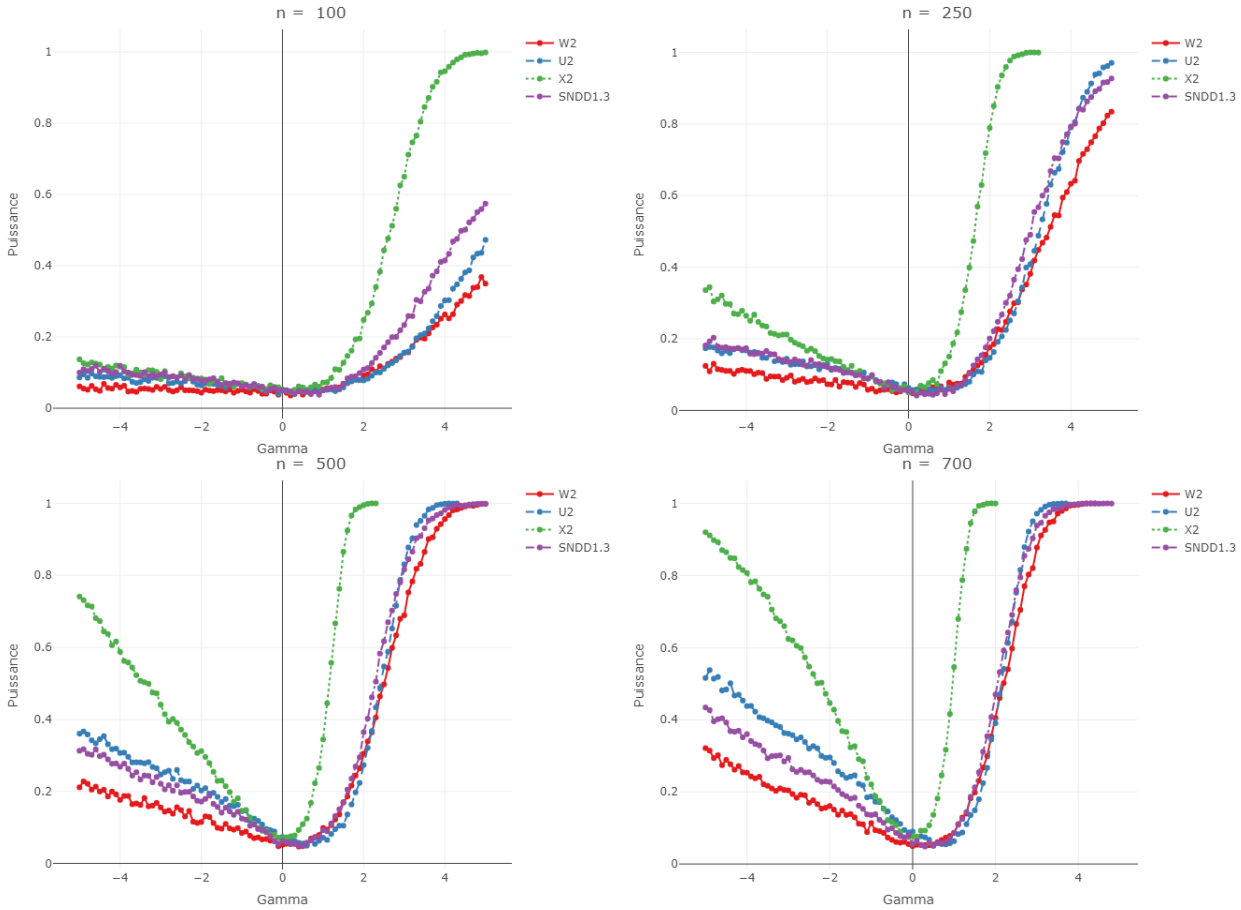

FIGURE A.2.8 – Indépendance Benford Newcomb-Benford Généralisée ($\mathcal{B}_1 \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Newcomb-Benford Généralisée Newcomb-Benford Généralisée

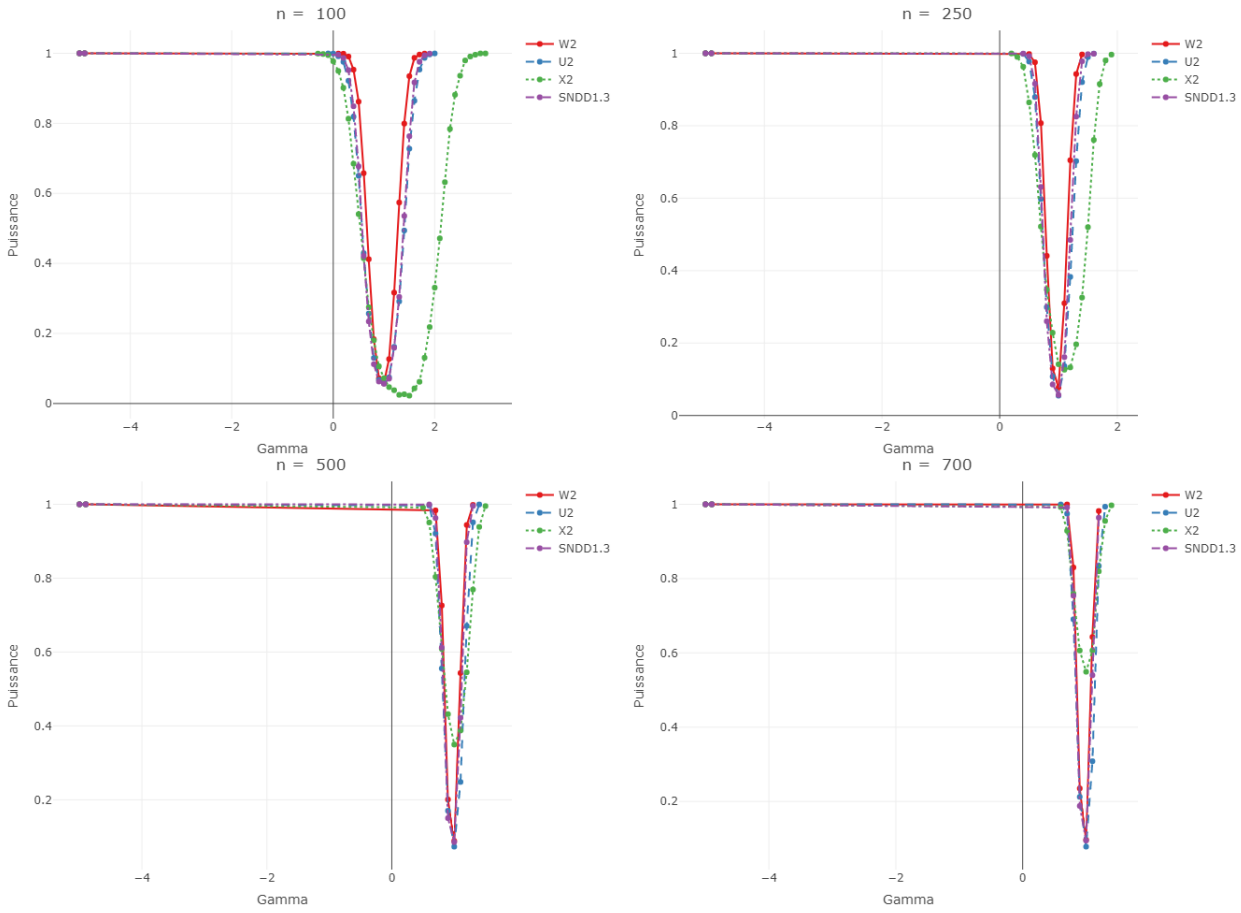

FIGURE A.2.9 – Indépendance Newcomb-Benford Généralisée Newcomb-Benford Généralisée ($\mathcal{GB}_1(\gamma) \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Rodriguez sachant Rodriguez

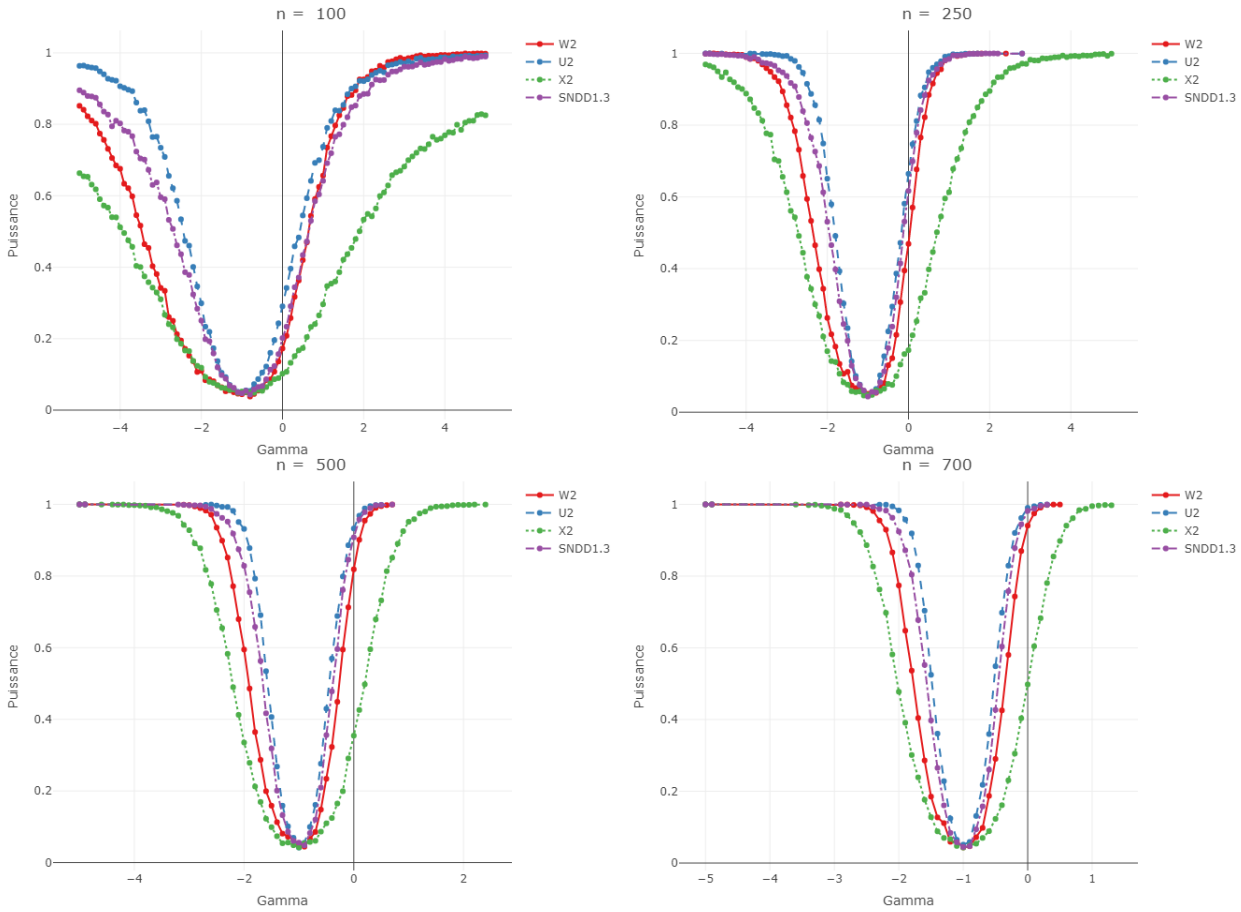

FIGURE A.2.10 – Rodriguez sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Benford

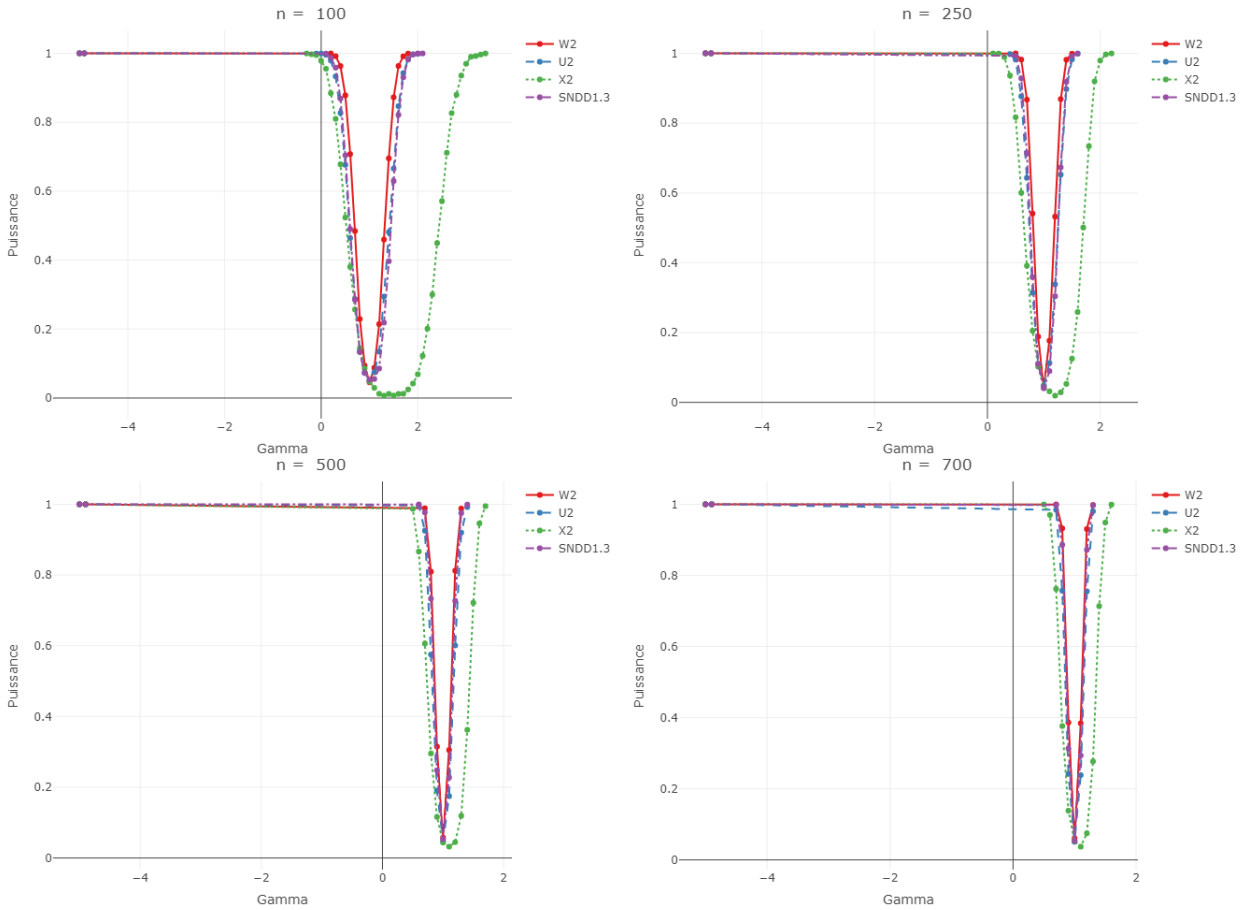

FIGURE A.2.11 – Newcomb-Benford Généralisée sachant Benford $\left(\mathcal{GB}_1(\gamma), \mathscr{B}_{(2|1)}\right)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Benford sachant Newcomb-Benford Généralisée

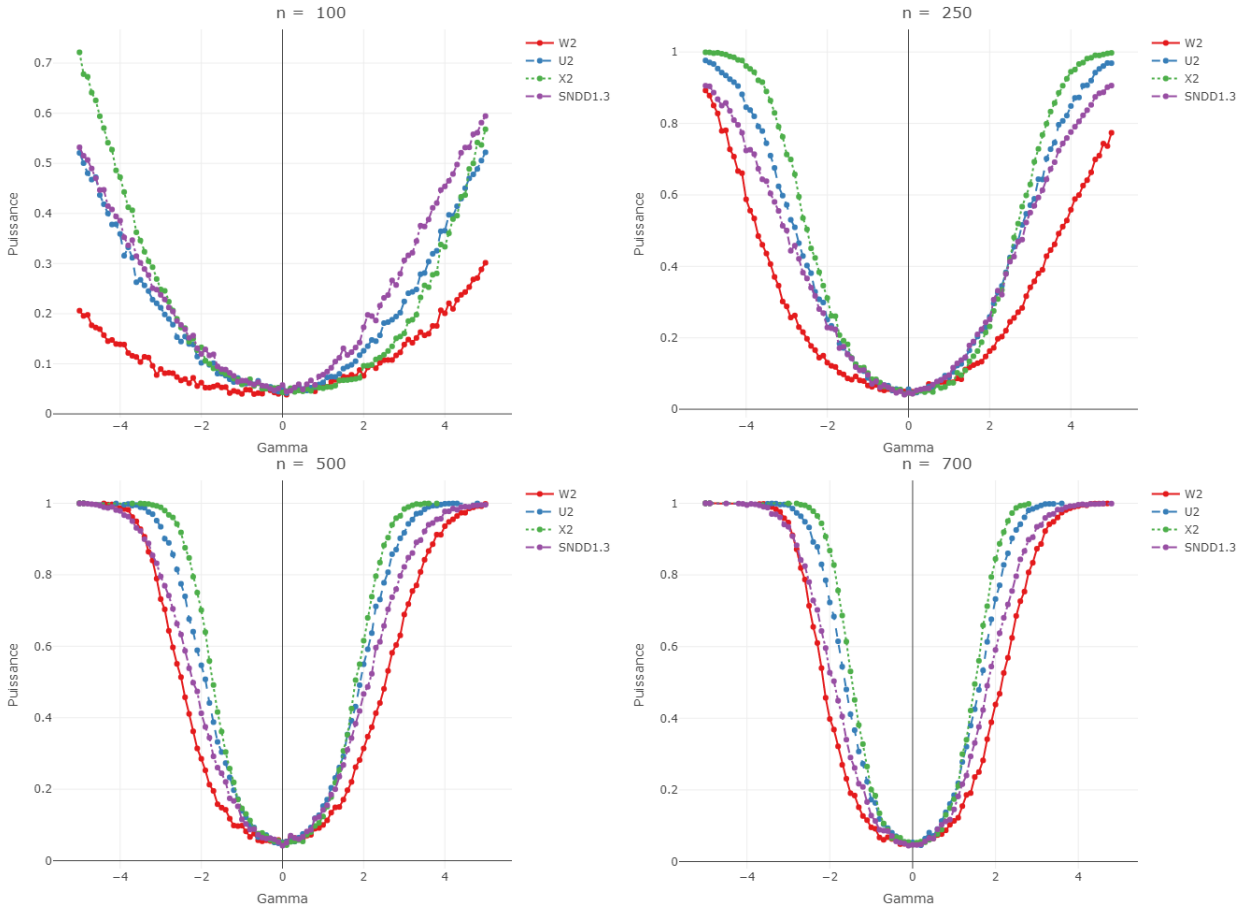

FIGURE A.2.12 – Benford sachant Newcomb-Benford Généralisée $\left(\mathscr{B}_1, \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée

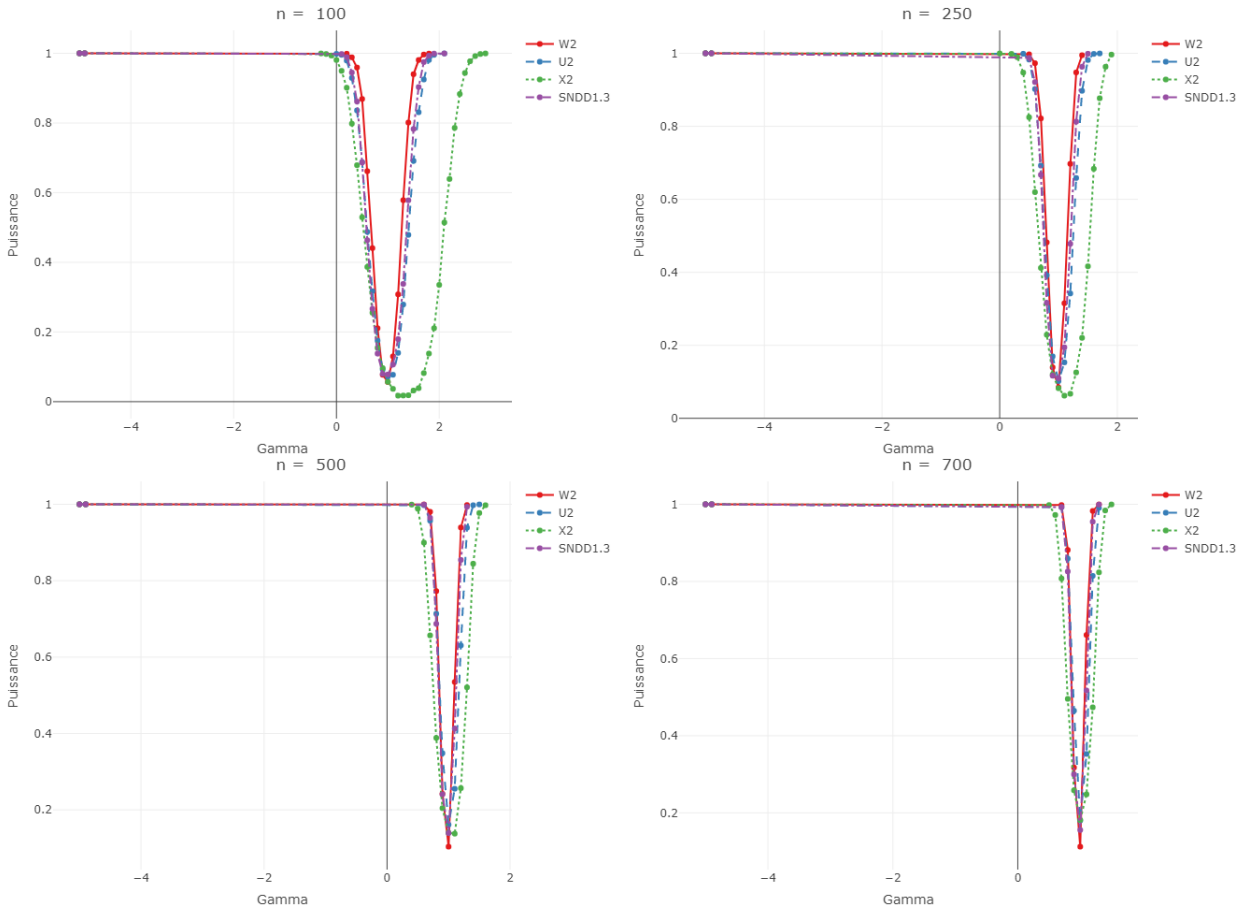

FIGURE A.2.13 – Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée $\left(\mathcal{GB}_1(\gamma), \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance en fonction du paramètre γ, des tests $\mathcal{S}_{NDD}$, U^2, χ^2, W^2 (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

A.3 Analyse des alternatives « Testing »

Nous représentons dans cette section les fréquences d'apparitions des chiffres significatifs et les fonctions de répartition des alternatives « Testing ».
Le lecteur remarquera que les conclusions tirées des alternatives « Training » dans la section 3.3.7 tiennent pour les alternatives « Testing » à savoir que pour les alternatives où le test de χ^2 est meilleur, les courbes respectives de leur fonction de répartition (à gauche $\gamma = 0$, à droite $\gamma = 1$) se positionnent à peu près au même niveau que la fonction de répartition de la loi de Newcomb-Benford et dans le cas où W^2 ou U^2 sont

plus puissants que le test du χ^2, les courbes des fonctions de répartition des alternatives concernées sont notablement différentes de la courbe de la fonction de répartition de la loi de Newcomb-Benford.

`Famille des mixtures: Benford Hill`

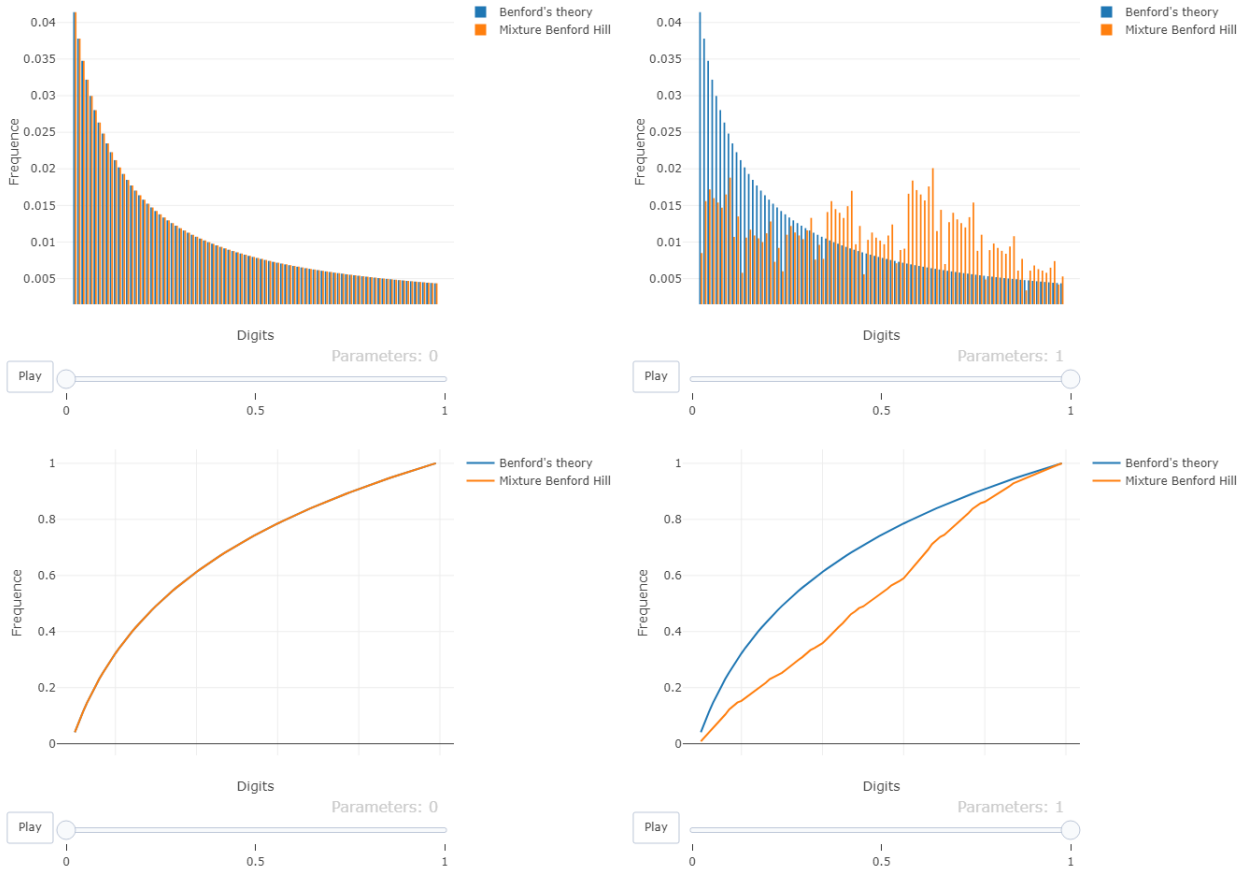

FIGURE A.3.1 – Mixture Benford Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{H}_{(1,2)}$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des mixtures: Benford Stigler

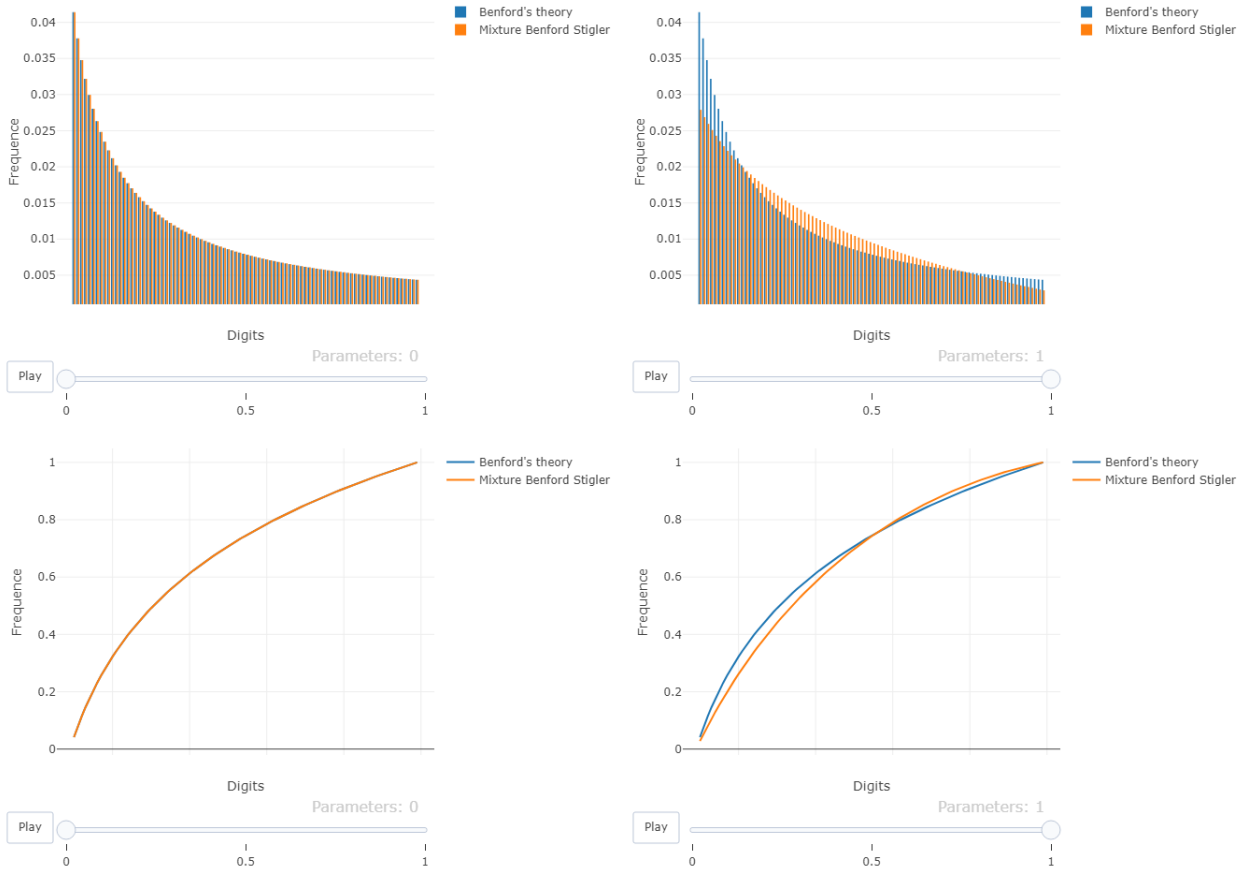

FIGURE A.3.2 – Mixture Benford Stigler $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{S}_{(1,2)}$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des mixtures: Benford Uniforme Stigler

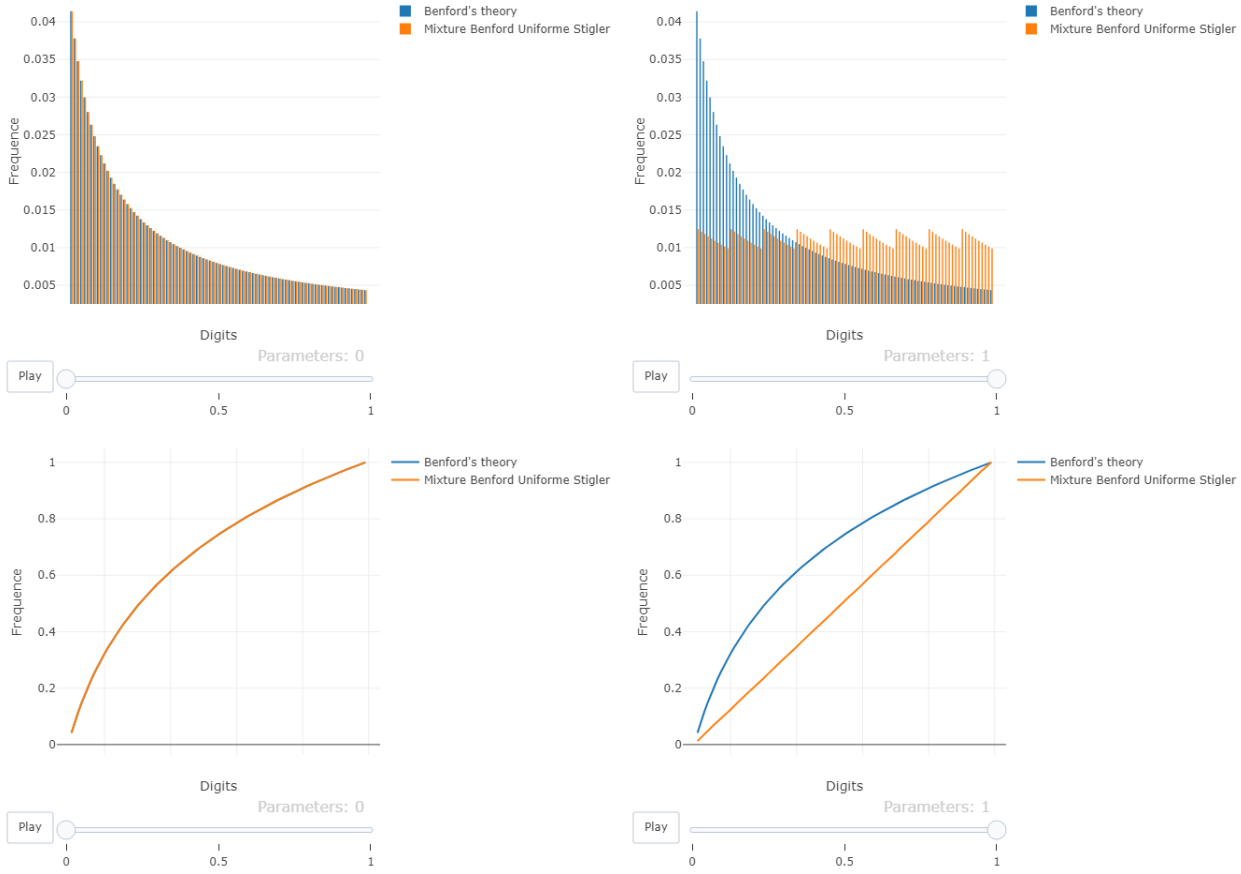

FIGURE A.3.3 – Mixture Benford Uniforme Stigler $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, U_{[1,9]} \otimes S_2$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille des mixtures: Benford Hill Uniforme`

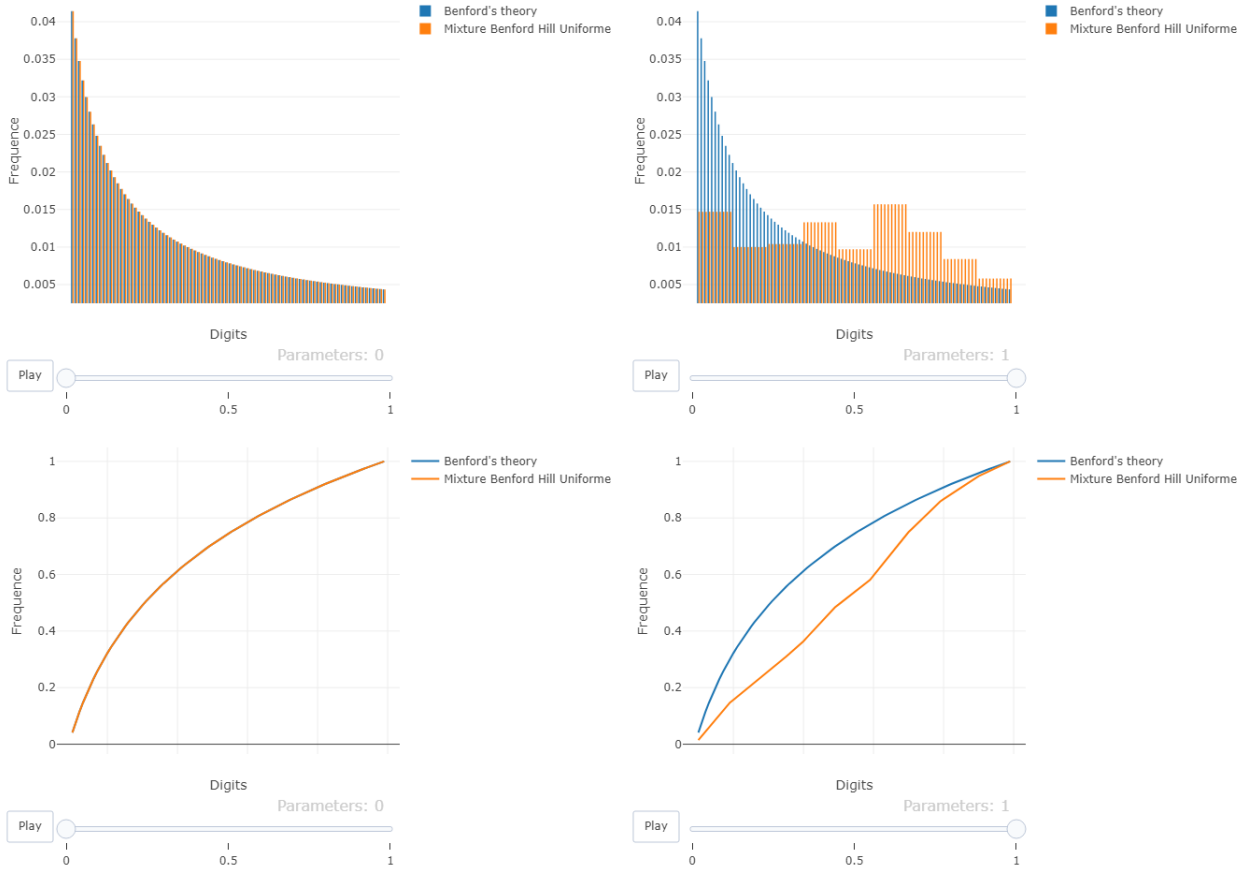

FIGURE A.3.4 – Mixture Benford Hill Uniforme $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, \mathcal{H}_1 \otimes U_{[0,9]}$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des mixtures: Benford Uniforme Hill

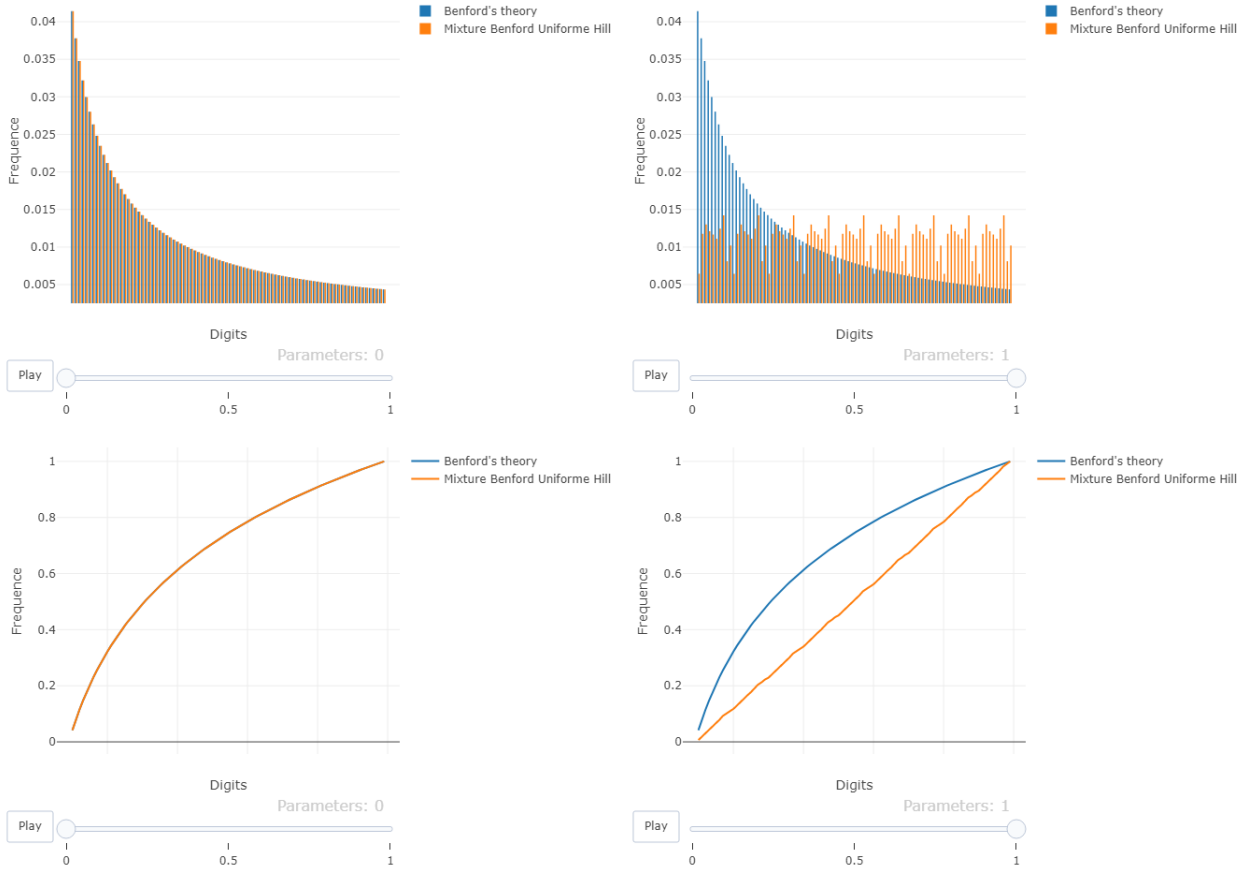

FIGURE A.3.5 – Mixture Benford Uniforme Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma(1-\gamma)\, U_{[1,9]} \otimes \mathcal{H}_2$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des Copules: Benford Stigler

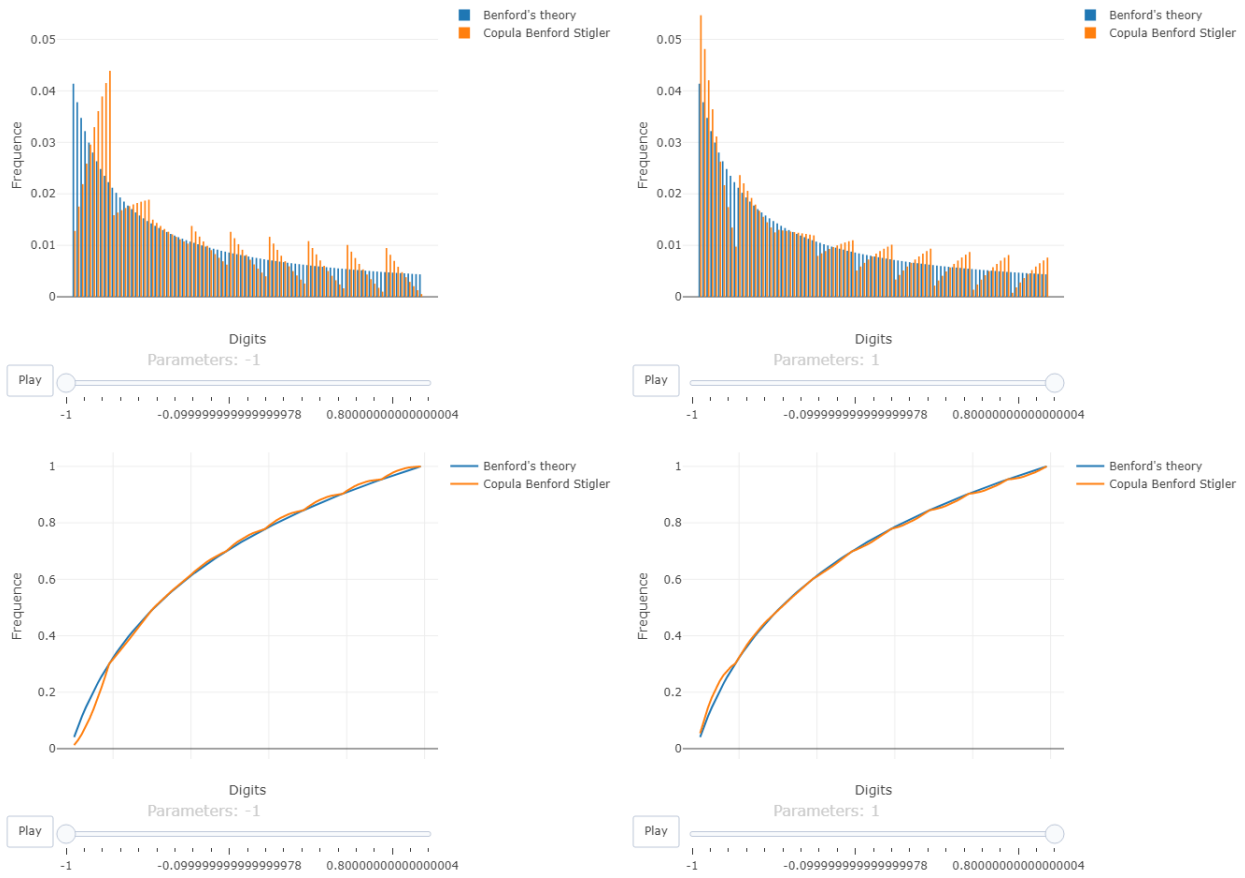

FIGURE A.3.6 – Copule Benford Stigler $C(\gamma, \mathscr{B}_1, S_2)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des Copules: Stigler Benford

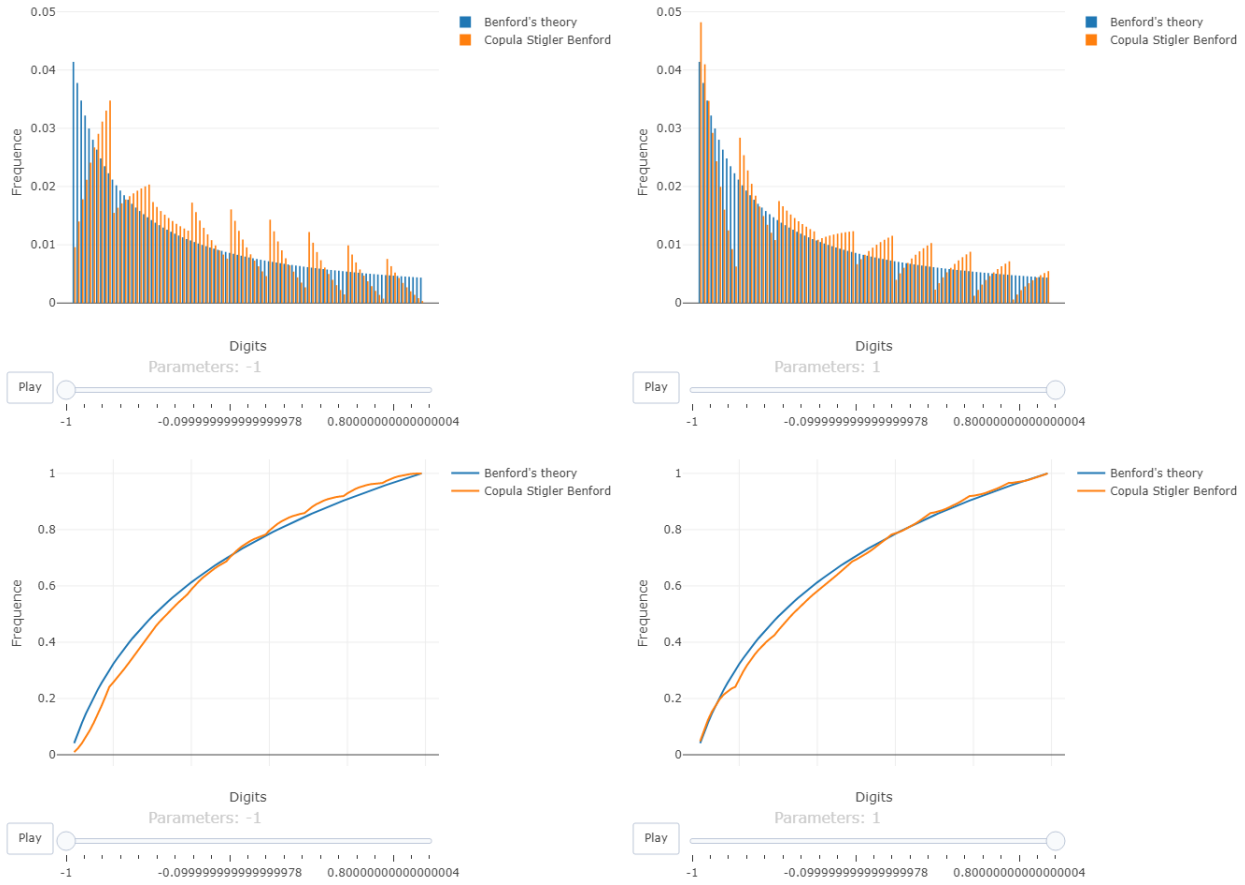

FIGURE A.3.7 – Copule Stigler Benford $C(\gamma, \mathcal{S}_1, \mathscr{B}_2)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille Indépendance: Benford Newcomb-Benford Généralisée

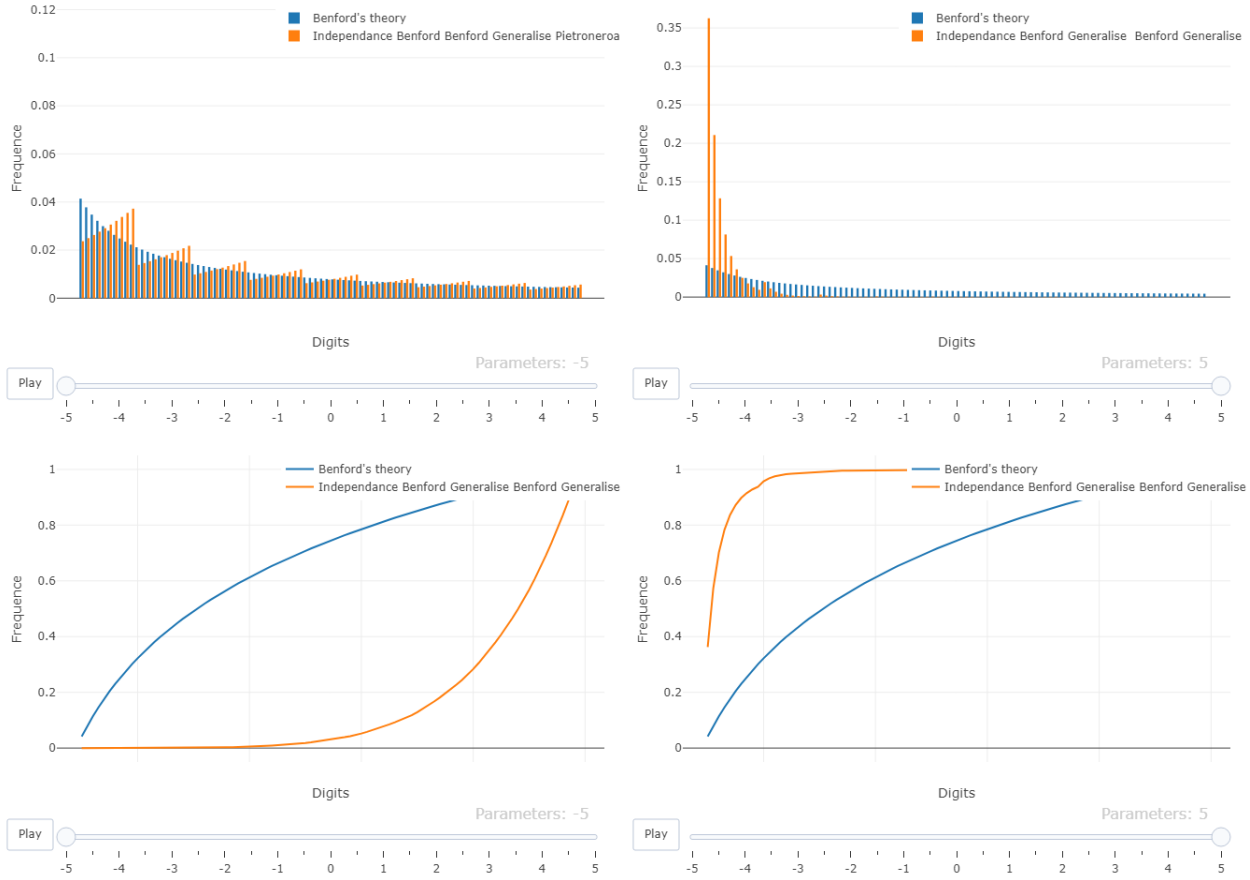

FIGURE A.3.8 – Indépendance Benford Newcomb-Benford Généralisée ($\mathcal{B}_1 \perp \mathcal{GB}_2(\gamma)$) :
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

`Famille Indépendance: Newcomb-Benford Généralisée Newcomb-Benford Généralisée`

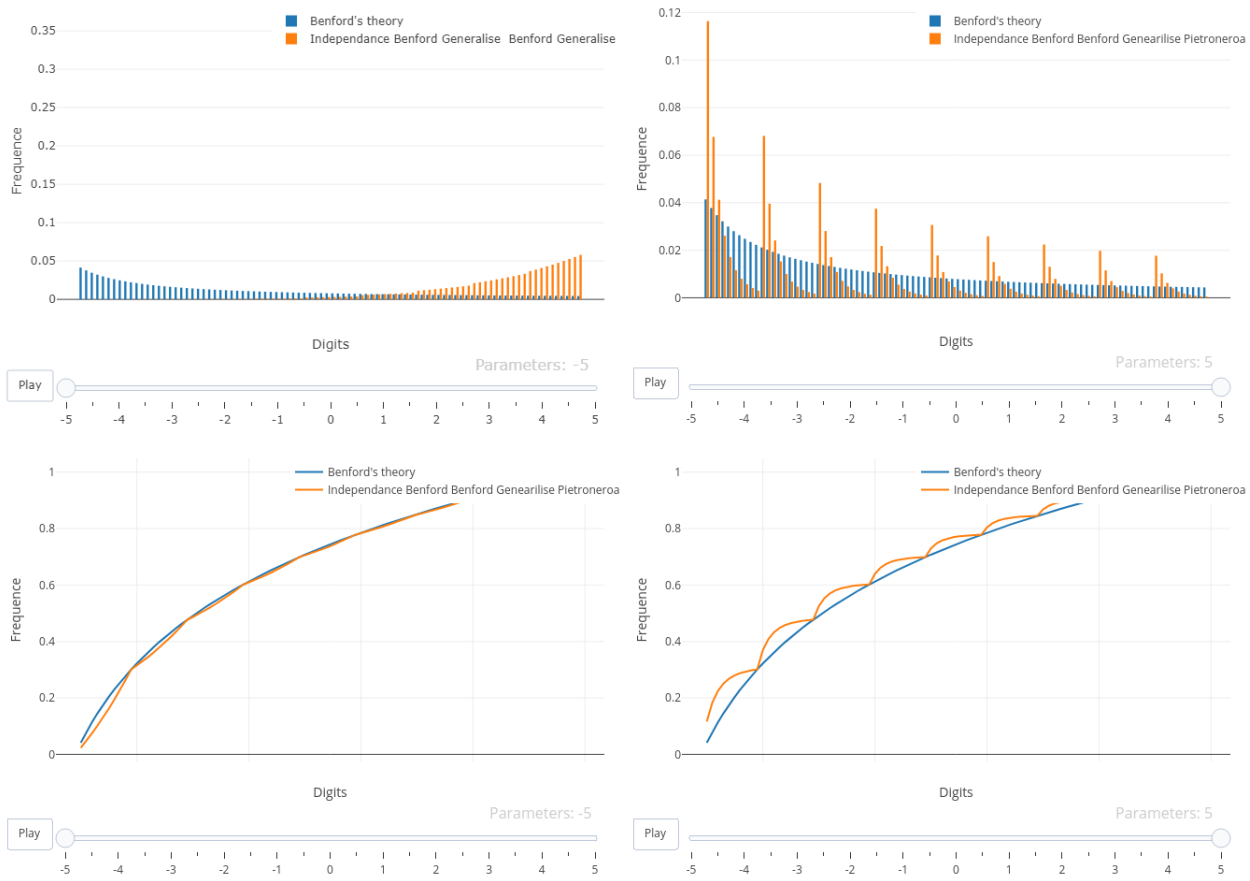

FIGURE A.3.9 – Indépendance Newcomb-Benford Généralisée Newcomb-Benford Généralisée ($\mathcal{GB}_1(\gamma) \perp \mathcal{GB}_2(\gamma)$) :
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des conditionnelles: Rodriguez sachant Rodriguez

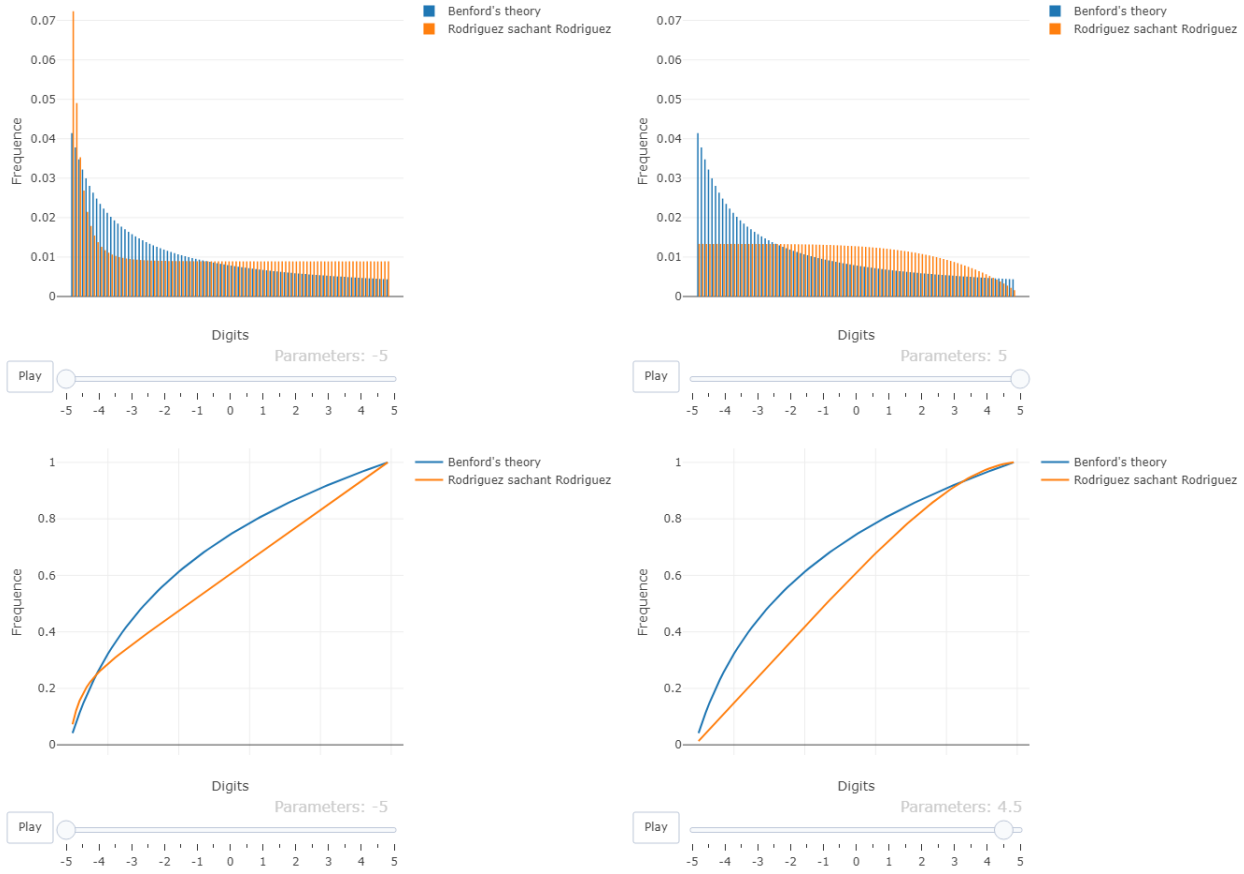

FIGURE A.3.10 – Rodriguez sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathcal{R}_{(2|1)}(\gamma)\right)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Benford

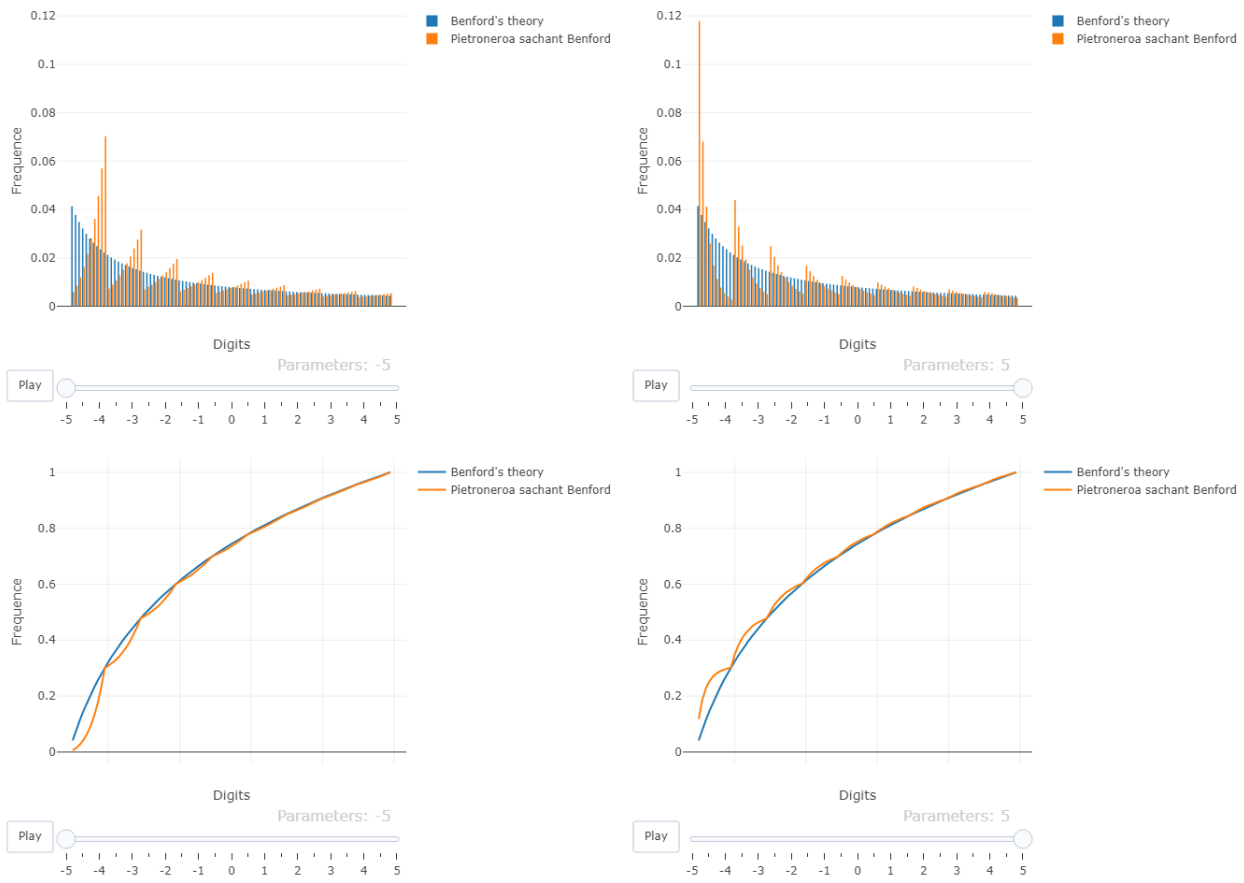

FIGURE A.3.11 – Newcomb-Benford Généralisée sachant Benford $\left(\mathcal{GB}_1(\gamma), \mathscr{B}_{(2|1)}\right)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des conditionnelles: Benford sachant Newcomb-Benford Généralisée

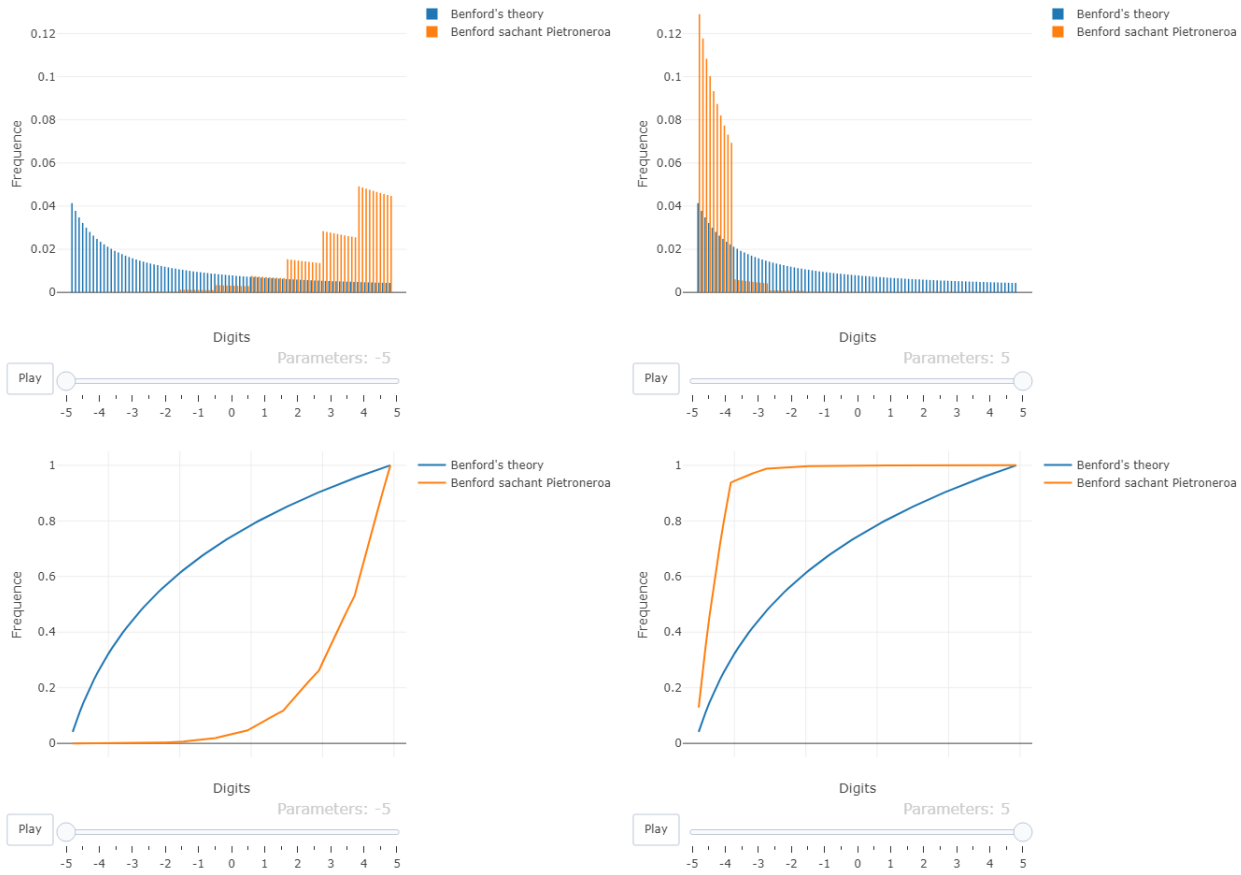

FIGURE A.3.12 – Benford sachant Newcomb-Benford Généralisée $\left(\mathscr{B}_1, \mathcal{GB}_{(2|1)}(\gamma)\right)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée

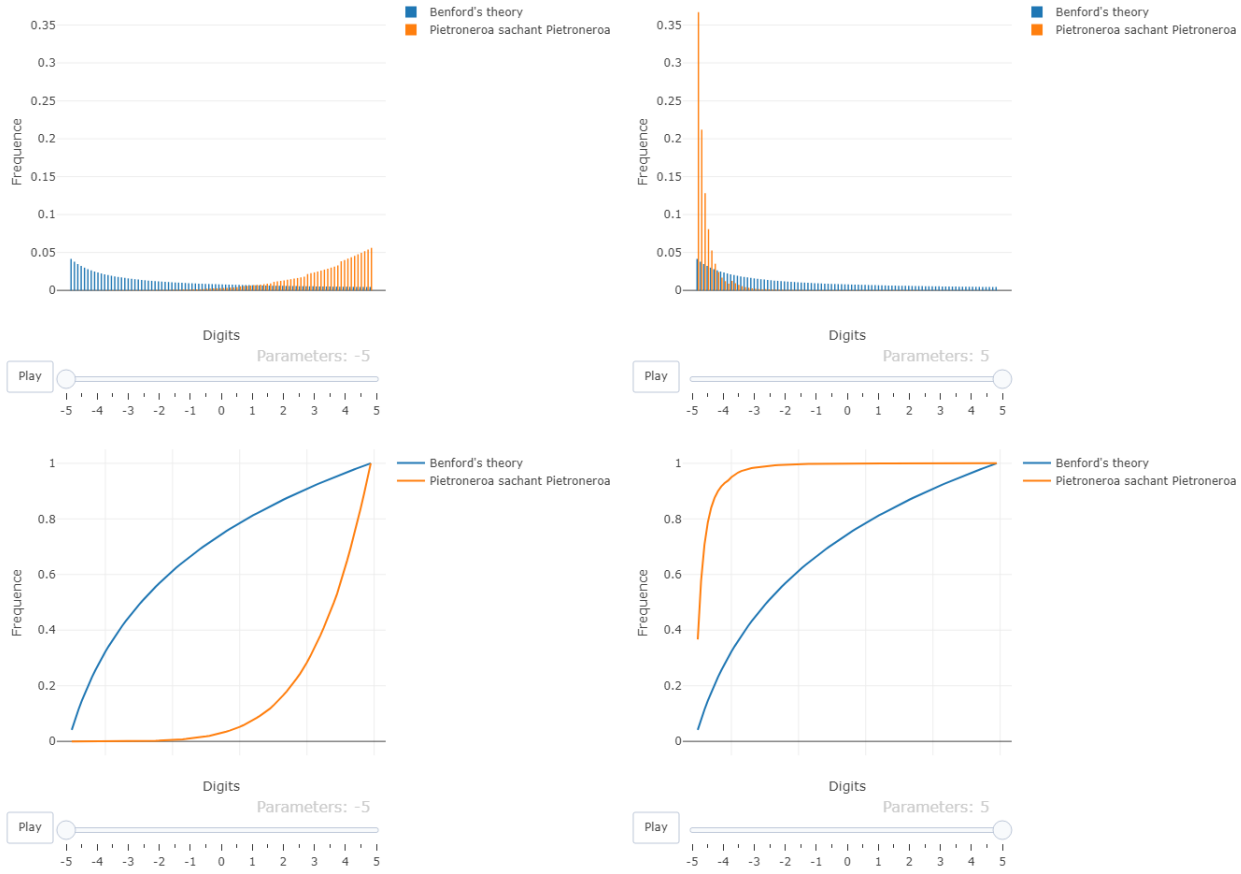

FIGURE A.3.13 – Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée $\left(\mathcal{GB}_1(\gamma), \mathcal{GB}_{(2|1)}(\gamma)\right)$:
Figures de hauts : Comparaison des fonctions de densités de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.
Figures de bas : Comparaison des fonctions de répartitions de $\mathscr{B}_{(1,2)}$ (en bleu) et de la loi alternative (en orange) pour $\gamma = 0$ à gauche et $\gamma = 1$ à droite.

A.4 Comparaison du test $\mathcal{S}_{NDD}$, $\mathcal{S}_{Cond,DD}$ et des tests classiques

Les graphiques des courbes de puissances sur les alternatives « Testing » dans le cadre de la comparaison des test $\mathcal{S}_{NDD}$ et $\mathcal{S}_{Cond,DD}$ développés dans le cadre de nos travaux à la section 3.3.10 sont présentés ci dessous.
Le lecteur remarquera que les conclusions tirées à la section 3.3.10 des alternatives « Training » tiennent pour les alternatives « Testing » à savoir que si W^2 ou U^2 sont plus puissants que le test du χ^2 alors$\mathcal{S}_{NDD}$ offre un bon compromis. Par contre si χ^2 est plus puissant que les tests W^2 ou U^2, $\mathcal{S}_{Cond,DD}$ offre un bon compromis.

Famille des mixtures: Benford Hill

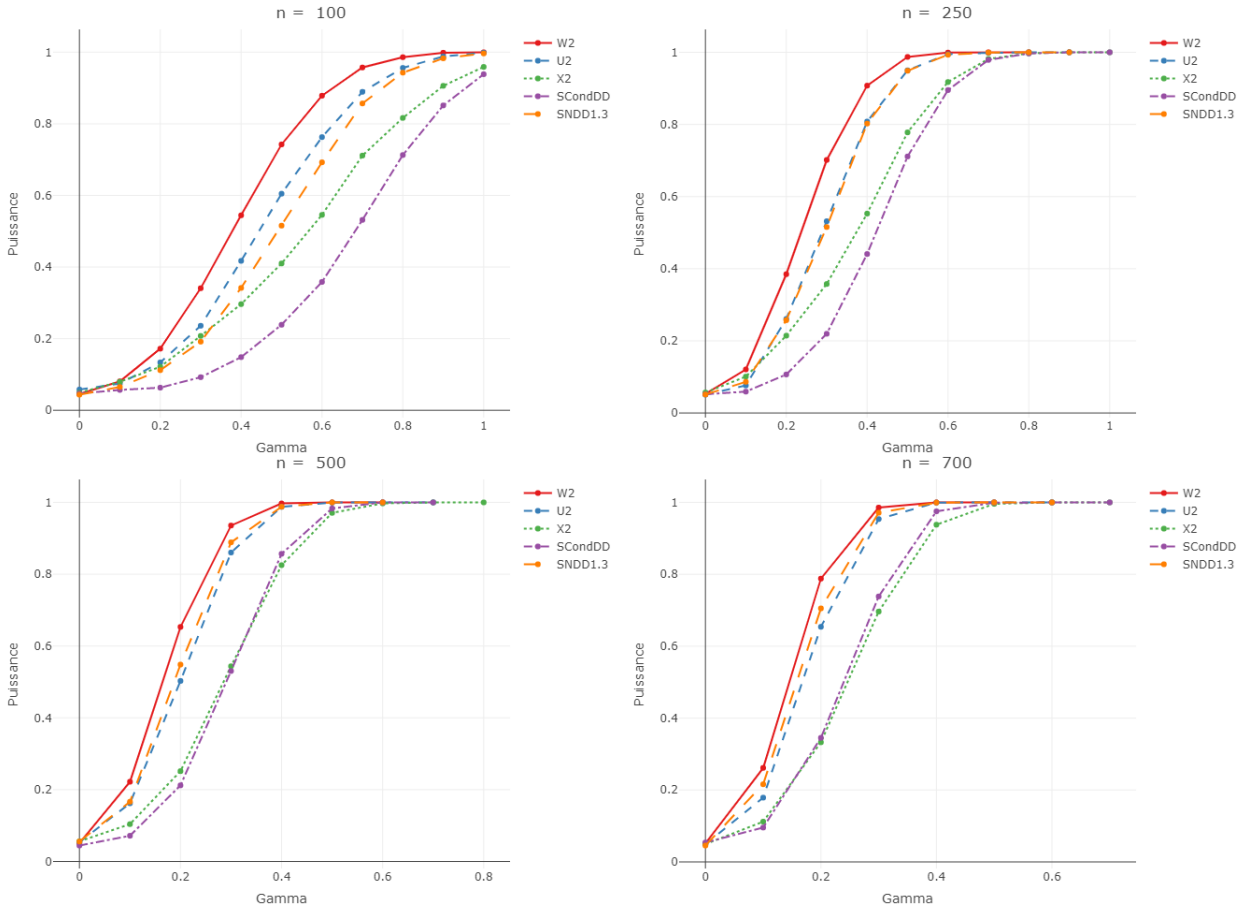

FIGURE A.4.1 – Mixture Benford Hill $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{H}_{(1,2)}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Stigler

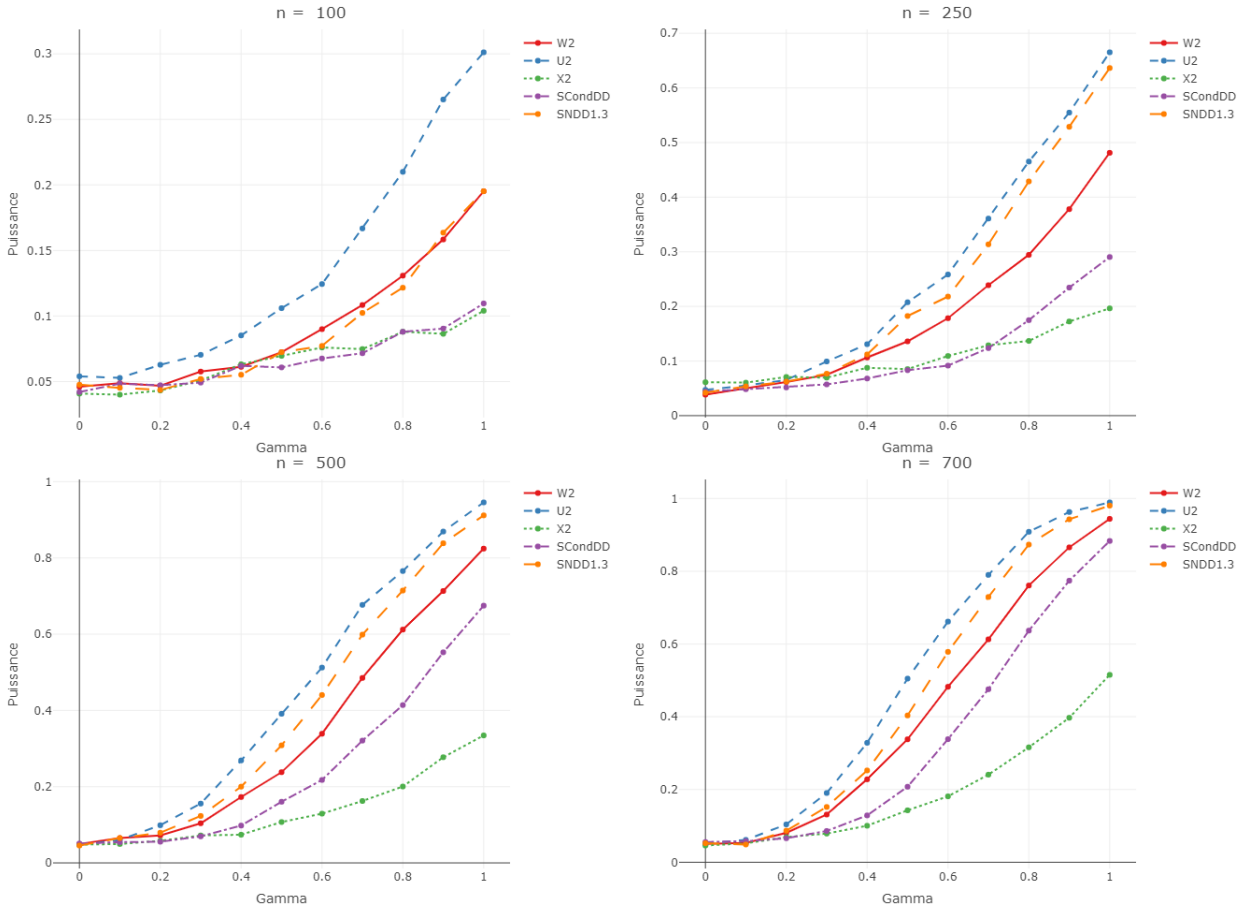

FIGURE A.4.2 – Mixture Benford Stigler $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\mathcal{S}_{(1,2)}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Uniforme Stigler

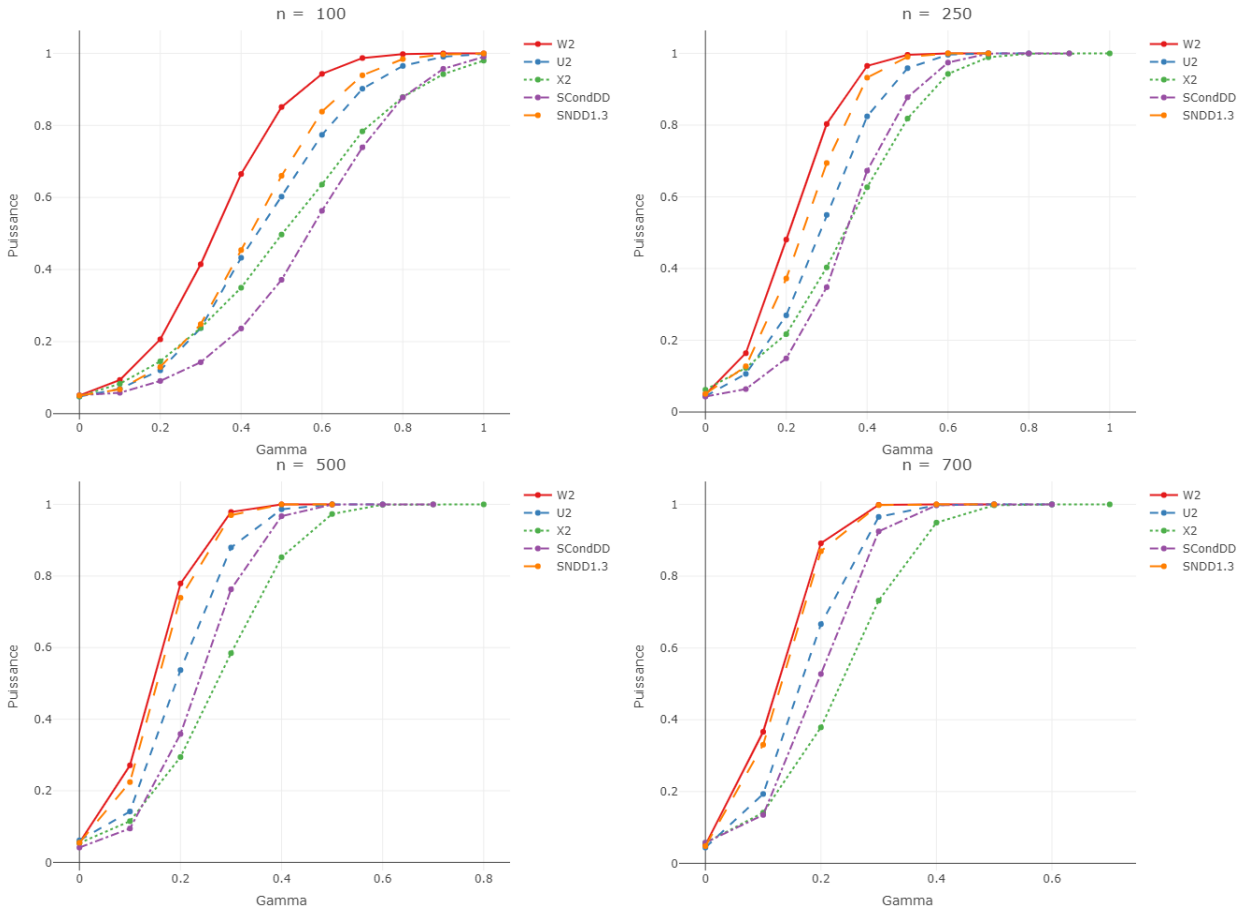

FIGURE A.4.3 – Mixture Benford Uniforme Stigler $(1-\gamma)\mathscr{B}_{(1,2)} + \gamma\, U_{[1,9]} \otimes S_2$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Hill Uniforme

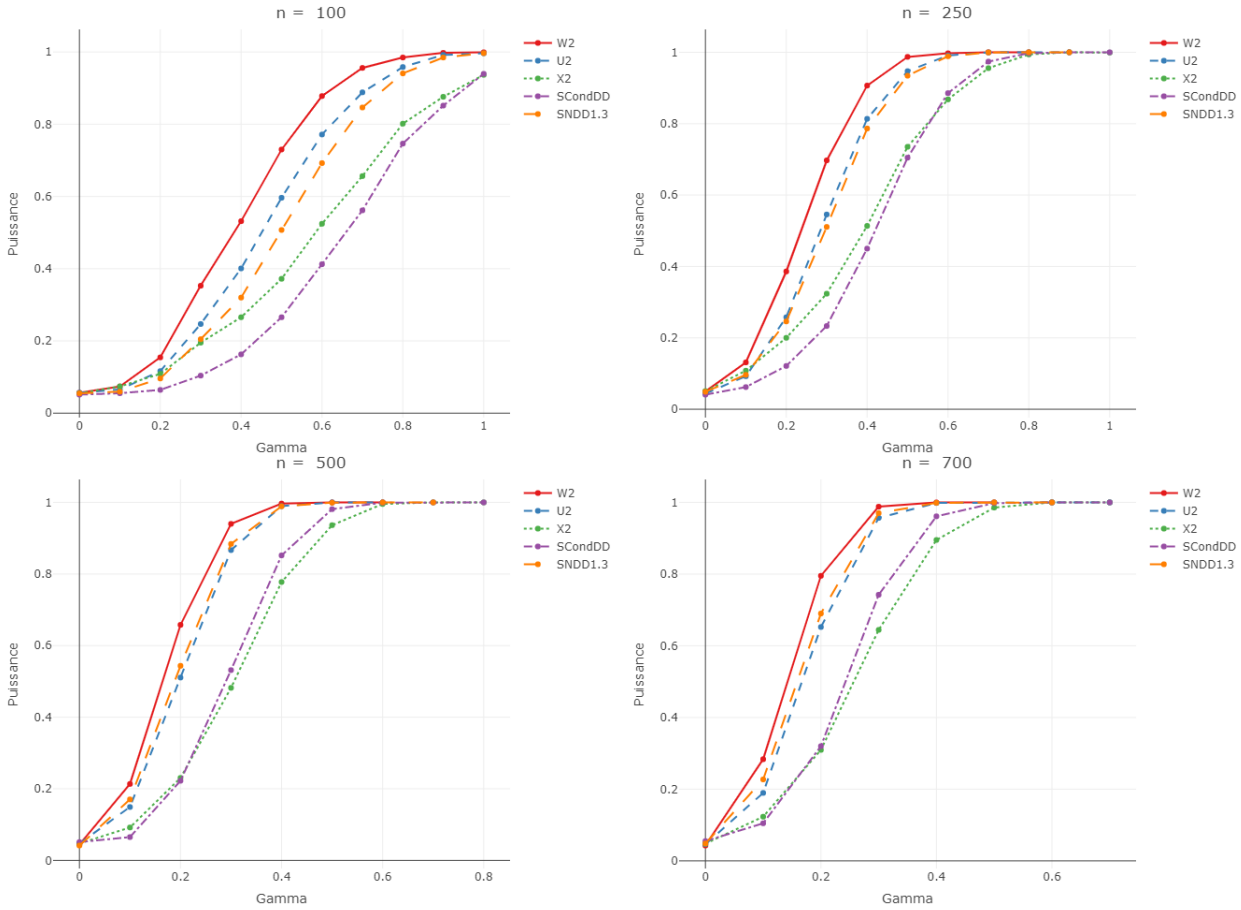

FIGURE A.4.4 – Mixture Benford Hill Uniforme $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma\,\mathcal{H}_1\otimes U_{[0,9]}$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des mixtures: Benford Uniforme Hill

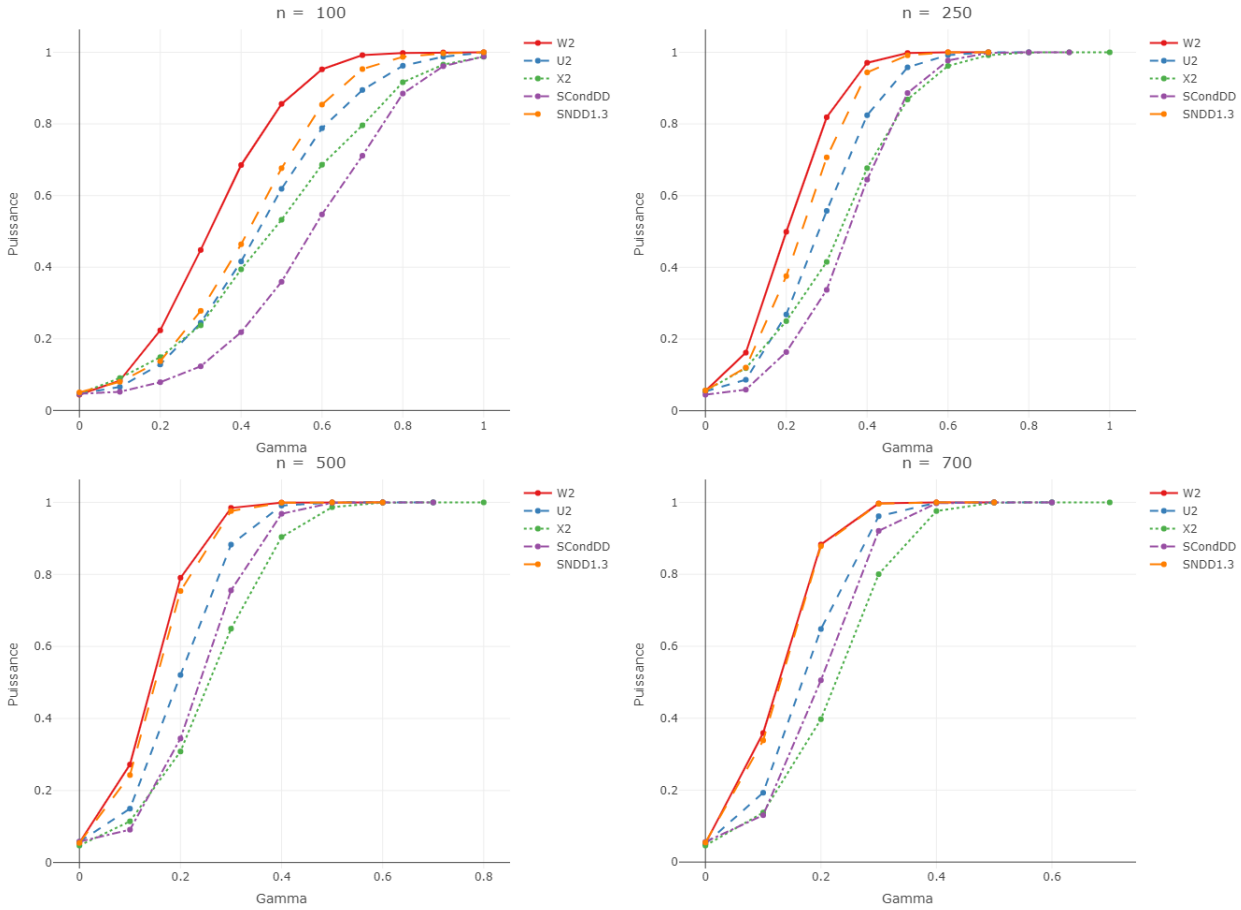

FIGURE A.4.5 – Mixture Benford Uniforme Hill $(1-\gamma)\mathscr{B}_{(1,2)}+\gamma(1-\gamma)U_{[1,9]}\otimes\mathcal{H}_2$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des Copules: Benford Stigler

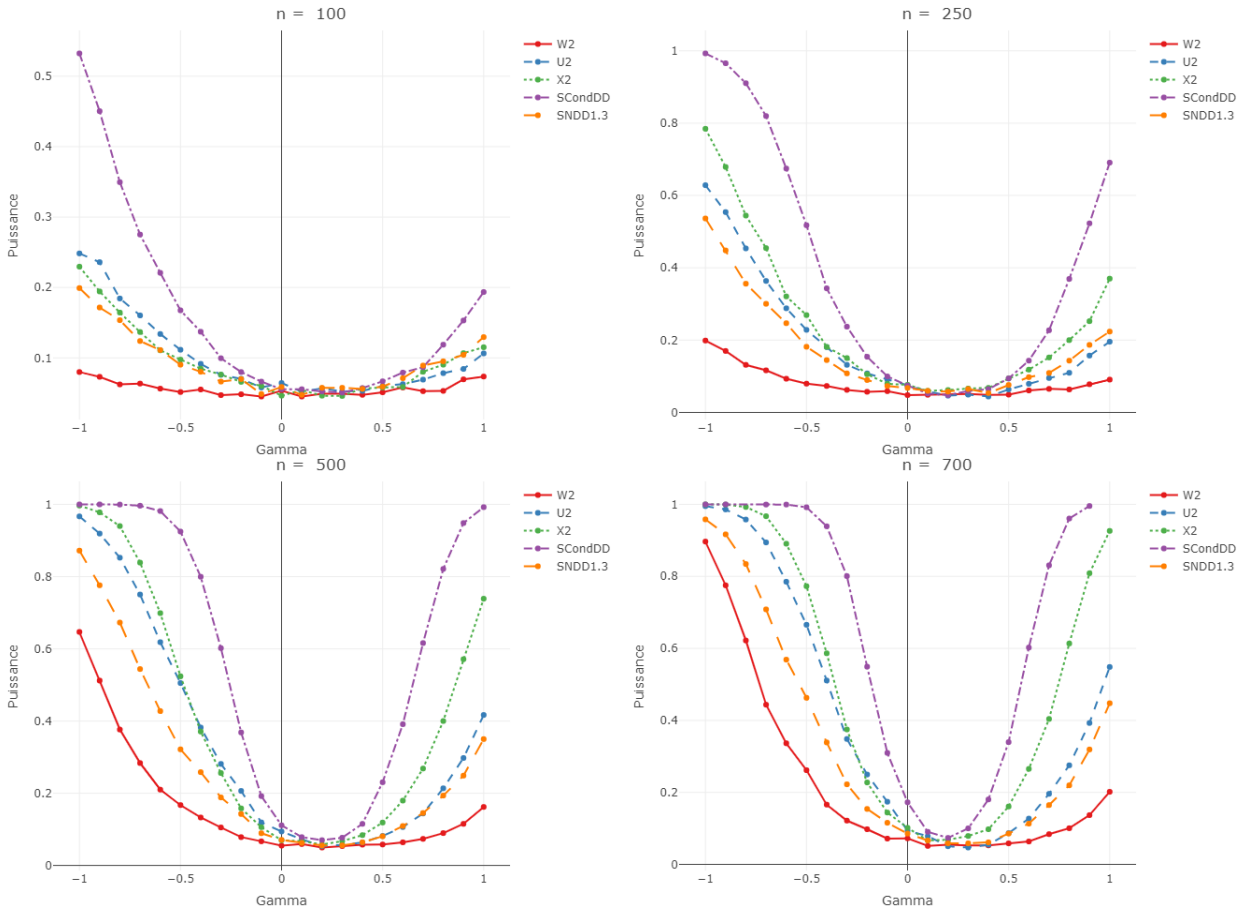

FIGURE A.4.6 – Copule Benford Stigler $C(\gamma, \mathscr{B}_1, S_2)$:Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des Copules: Stigler Benford

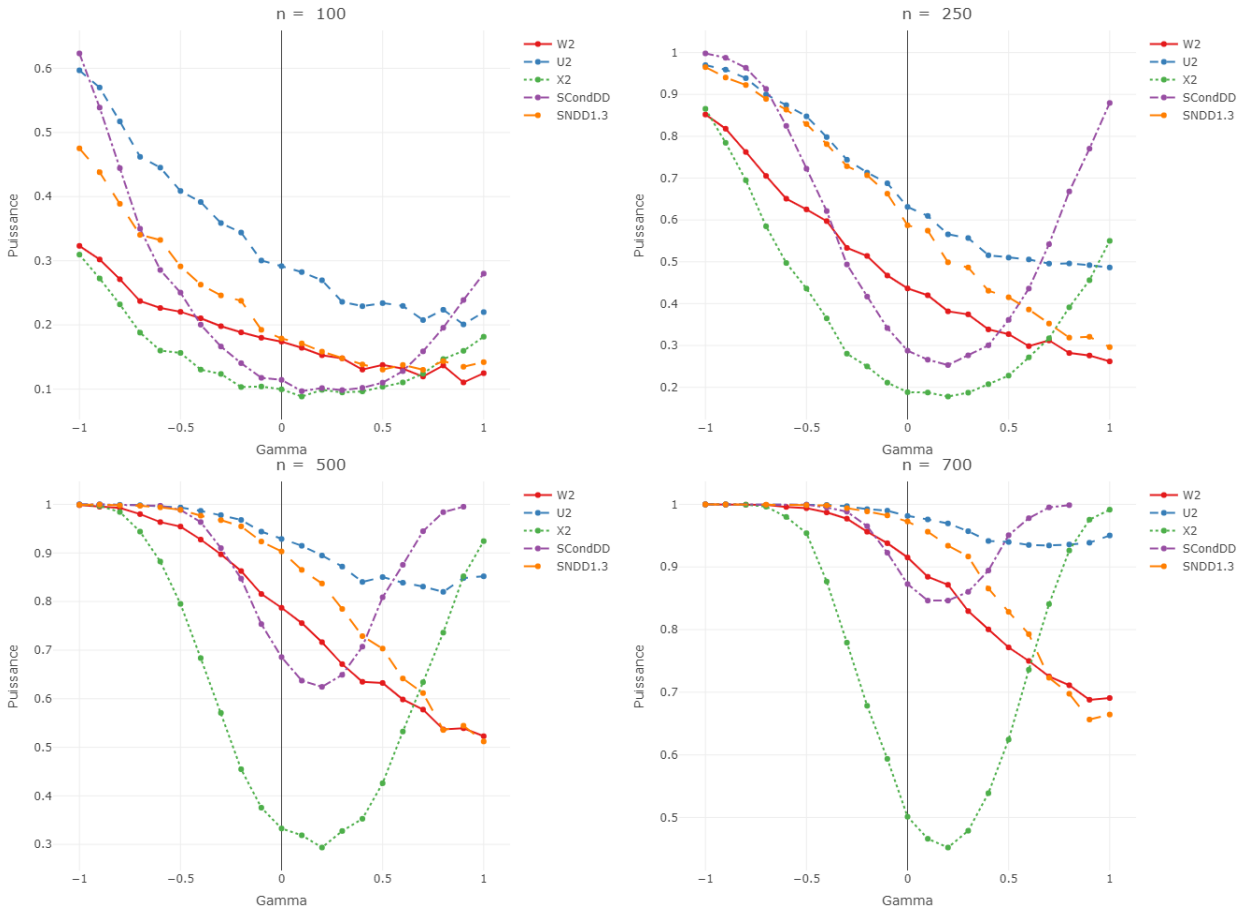

FIGURE A.4.7 – Copule Stigler Benford $C(\gamma, \mathcal{S}_1, \mathscr{B}_2)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Benford Newcomb-Benford Généralisée

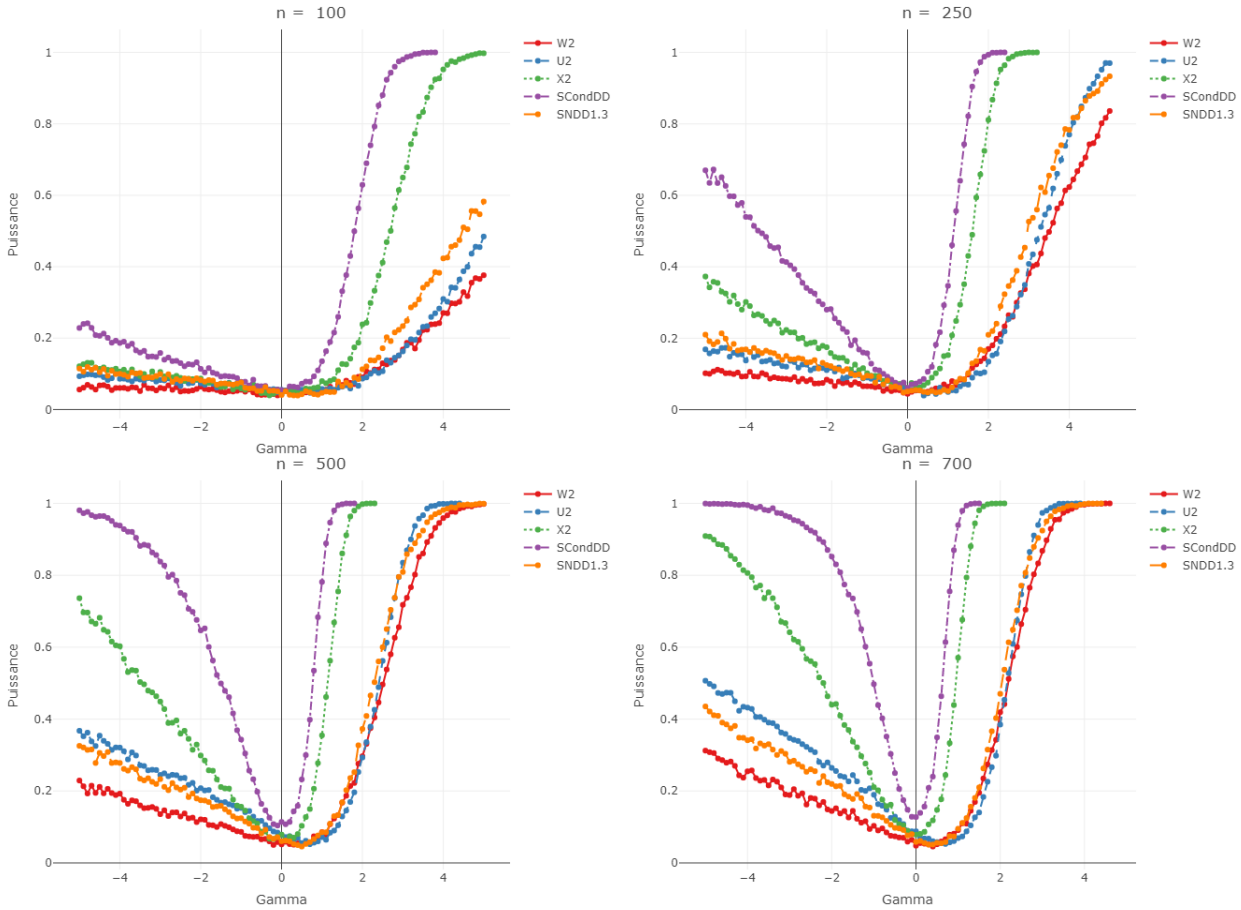

FIGURE A.4.8 – Indépendance Benford Newcomb-Benford Généralisée ($\mathcal{B}_1 \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille Indépendance: Newcomb-Benford Généralisée Newcomb-Benford Généralisée

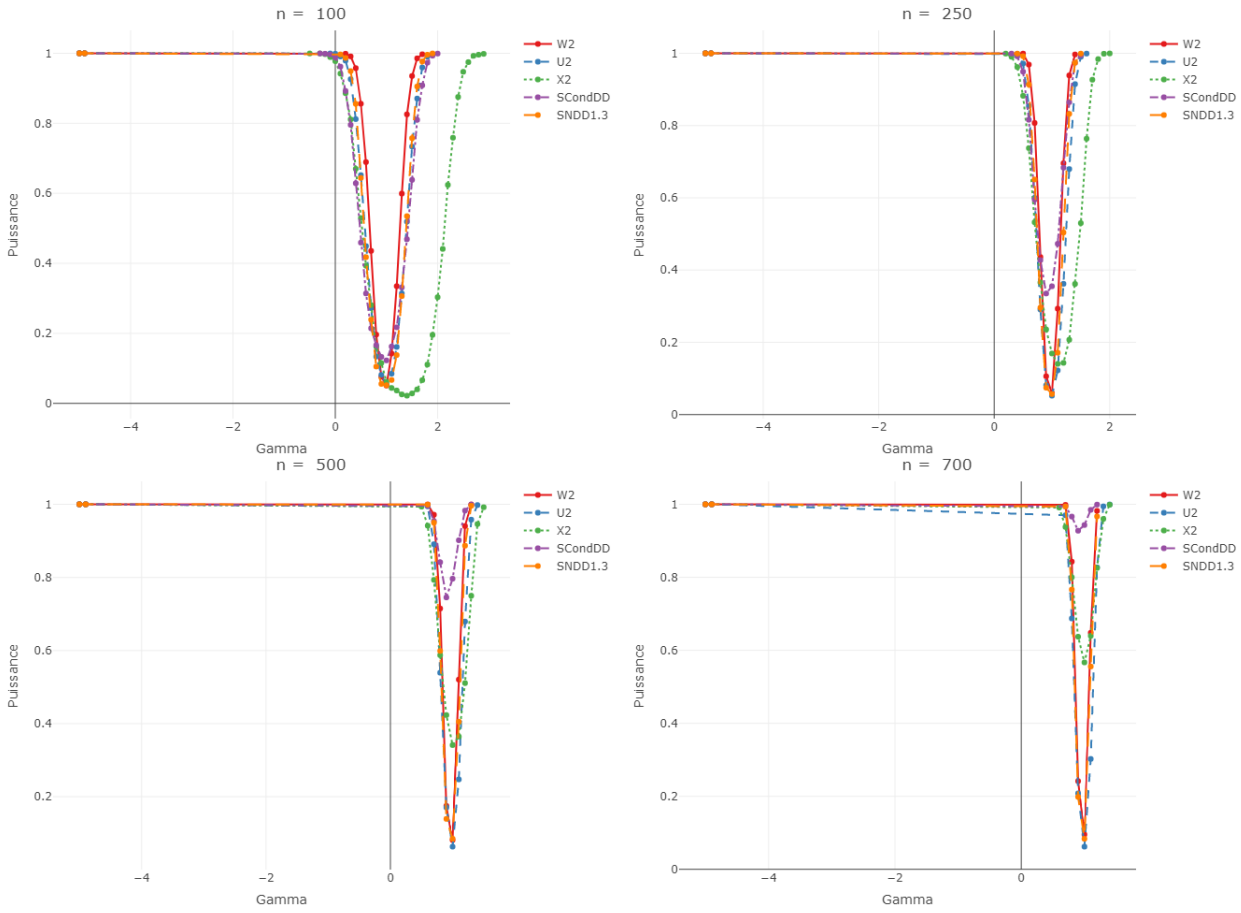

FIGURE A.4.9 – Indépendance Newcomb-Benford Généralisée Newcomb-Benford Généralisée ($\mathcal{GB}_1(\gamma) \perp \mathcal{GB}_2(\gamma)$) : Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Rodriguez sachant Rodriguez

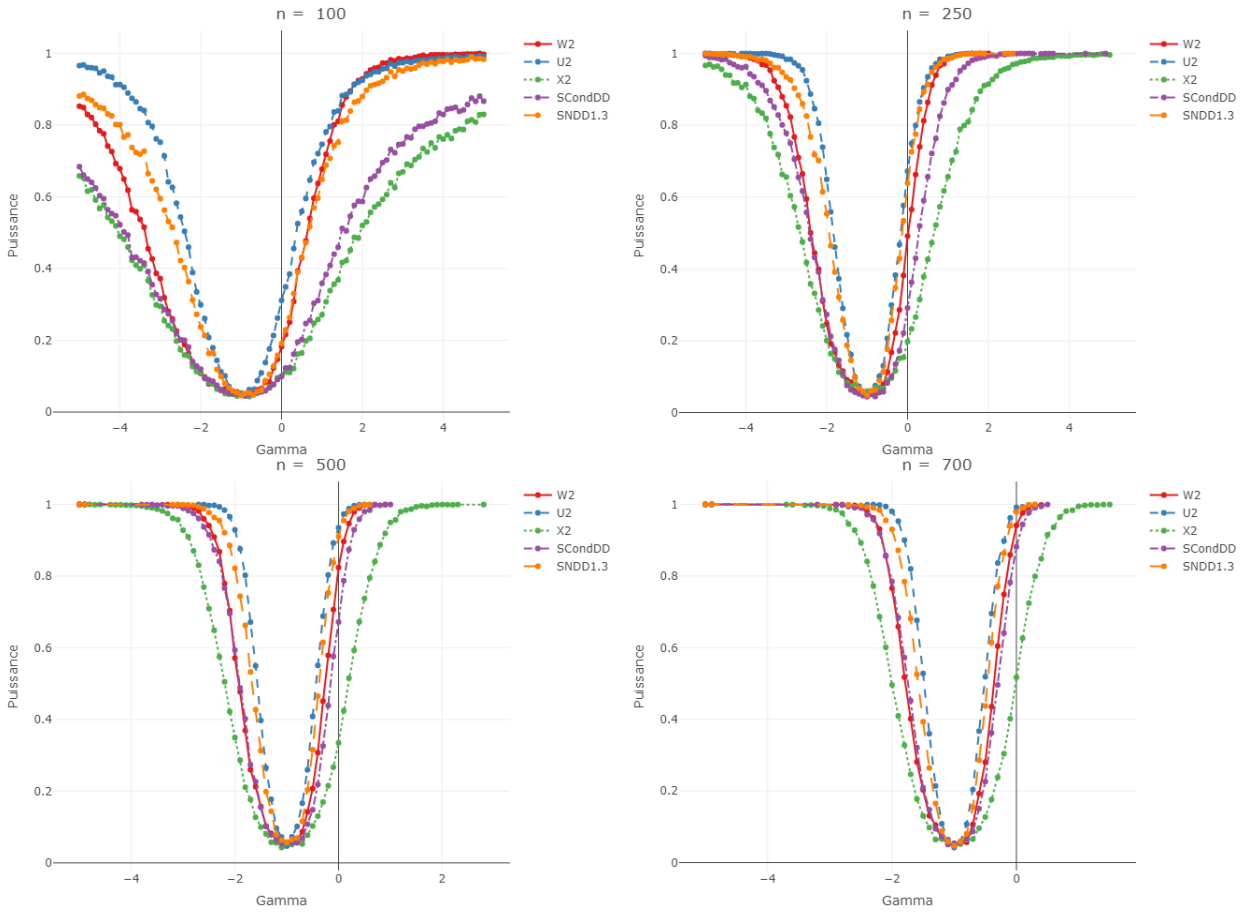

FIGURE A.4.10 – Rodriguez sachant Rodriguez $\left(\mathcal{R}_1(\gamma), \mathcal{R}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Benford

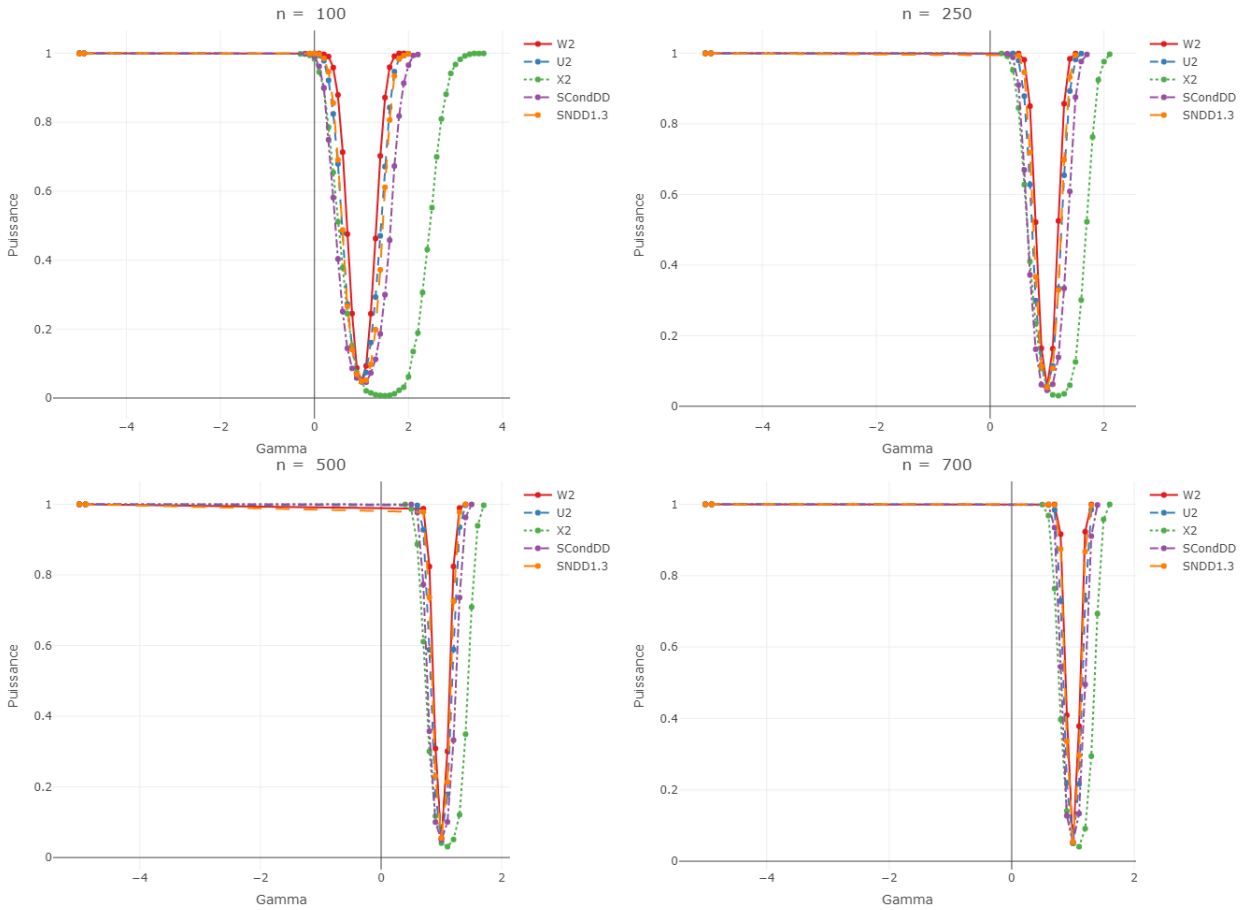

FIGURE A.4.11 – Newcomb-Benford Généralisée sachant Benford $\left(\mathcal{GB}_1(\gamma), \mathscr{B}_{(2|1)}\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Benford sachant Newcomb-Benford Généralisée

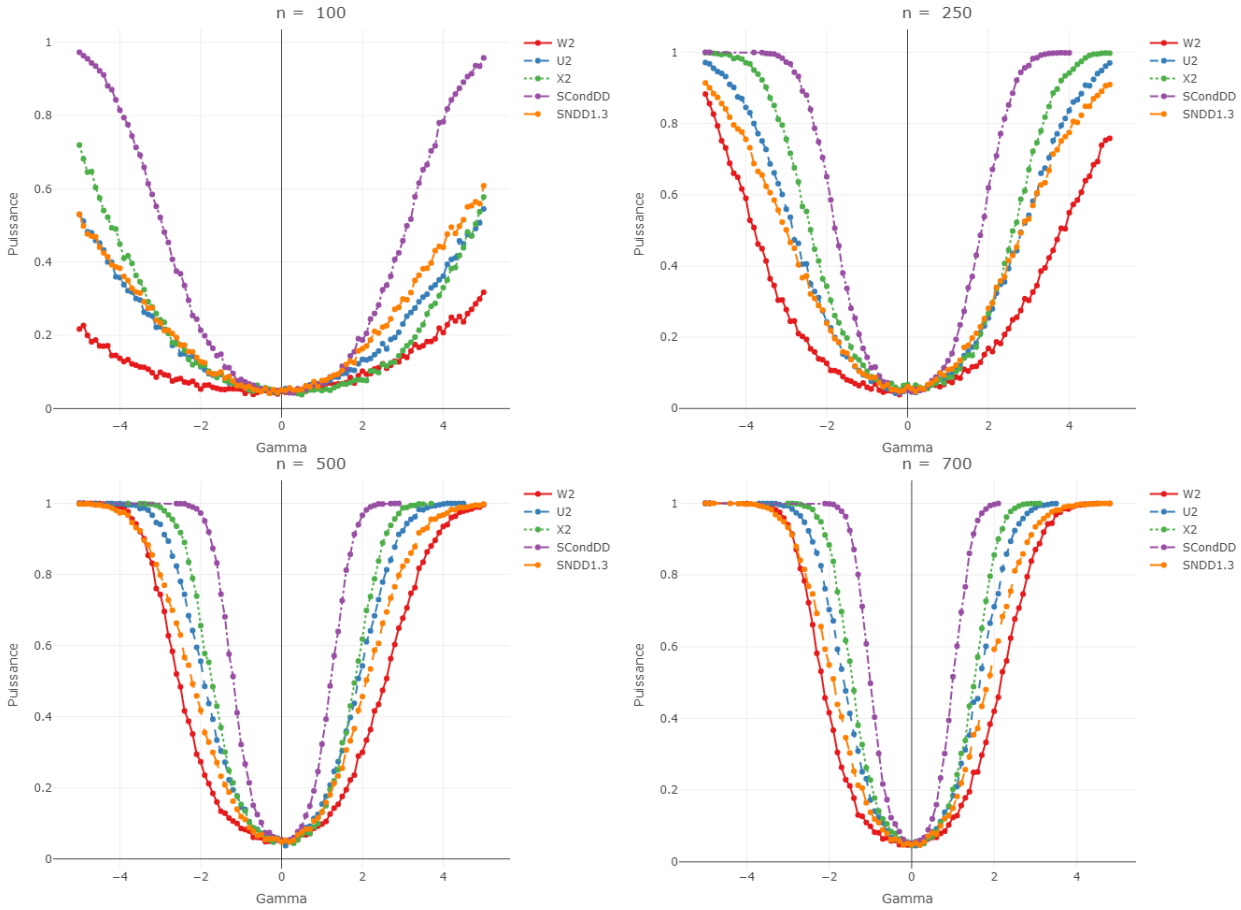

FIGURE A.4.12 – Benford sachant Newcomb-Benford Généralisée $\left(\mathscr{B}_1, \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.

Famille des conditionnelles: Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée

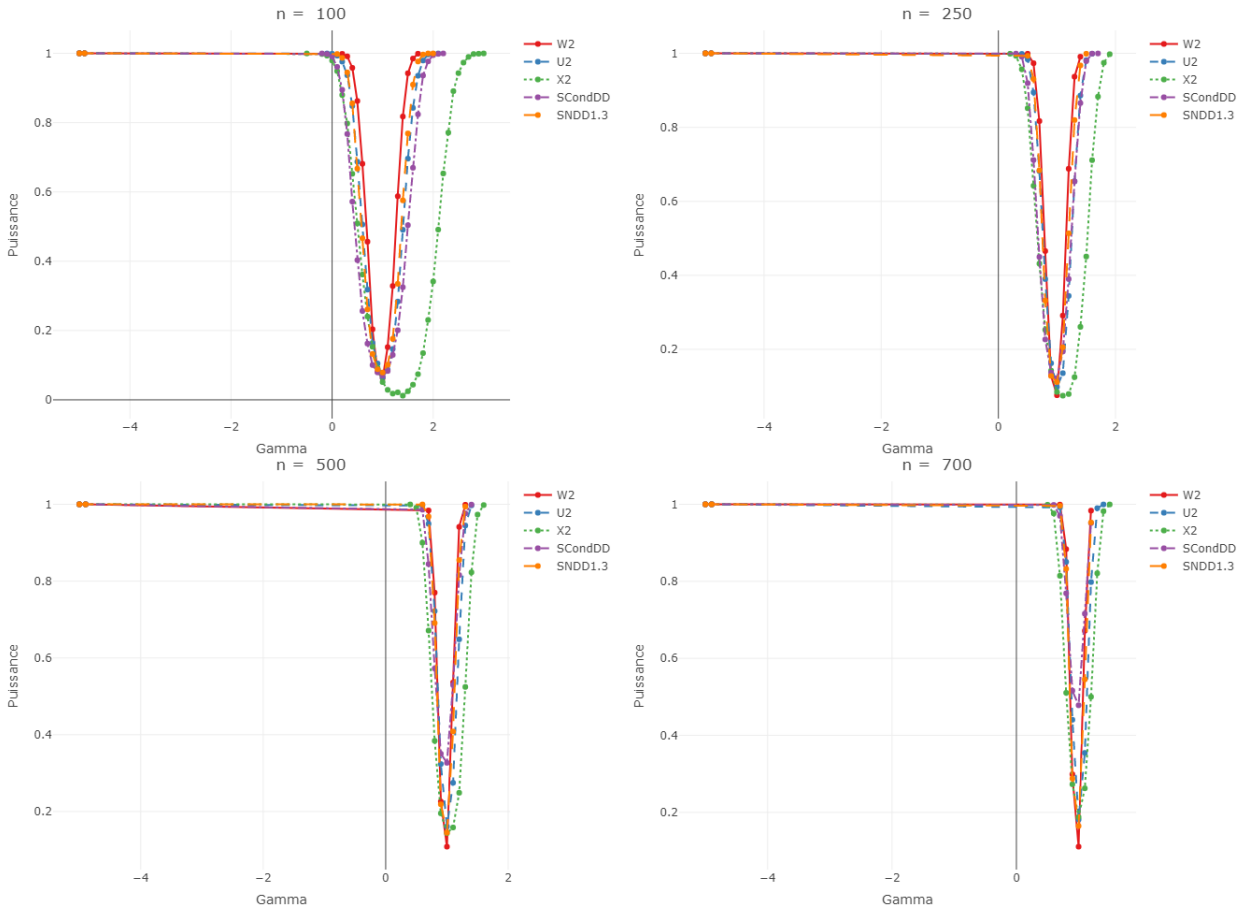

FIGURE A.4.13 – Newcomb-Benford Généralisée sachant Newcomb-Benford Généralisée $\left(\mathcal{GB}_1(\gamma), \mathcal{GB}_{(2|1)}(\gamma)\right)$: Courbes de puissance, en fonction du paramètre γ, de divers tests (basés sur 10000 réplications) au niveau 5% pour l'hypothèse nulle de la loi $\mathscr{B}_{(1,2)}$. Les tests représentés sont : W^2 (couleur rouge), χ^2(couleur verte) , U^2 (couleur bleue), $\mathcal{S}_{NDD}$ (couleur orange) et $\mathcal{S}_{Cond,DD}$ (couleur violette) dont les expressions se trouvent à la Section 3.3.3. Les quantiles de référence sont approximés par Monte-Carlo en utilisant 1000000 répétitions.